Lecture Notes
Nonlinear Optimization

Spring Semester 2023

Roland Herzog*

2023-04-17

*Interdisciplinary Center for Scientific Computing, Heidelberg University, 69120 Heidelberg, Germany (roland.herzog@iwr.uni-heidelberg.de, `https://scoop.iwr.uni-heidelberg.de/team/roland-herzog`).

These lecture notes are partly based on content from the books Nocedal, Wright, 2006; Ulbrich, Ulbrich, 2012.

Material for 14 weeks.

Please send comments to roland.herzog@iwr.uni-heidelberg.de.

Contents

Chapter 0 Introduction

§ 1 Elementary Notions

Mathematical optimization is about solving problems of the form

$$\left.\begin{array}{rlll} \text{Minimize} & f(x) & \text{where } x \in \Omega & (\textbf{objective function}) \\ \text{subject to} & g_i(x) \le 0 & \text{for } i = 1, \ldots, n_{\text{ineq}} & (\textbf{inequality constraints}) \\ \text{and} & h_j(x) = 0 & \text{for } j = 1, \ldots, n_{\text{eq}}. & (\textbf{equality constraints}) \end{array}\right\} \tag{1.1}$$

$\Omega \subseteq \mathbb{R}^n$ is the **basic set** and x is the **optimization variable** or simply the **variable** of the problem. We will assume that

- the functions $f, g_i, h_j : \mathbb{R}^n \to \mathbb{R}$ are sufficiently smooth (C^2 functions),
- we have a finite number (possibly zero) of inequality and equality constraints, i. e., n_{ineq} and n_{eq} are in $\mathbb{N}_0$.

We will assume $\Omega = \mathbb{R}^n$, i. e., we consider only **continuous optimization** problems and without implicit constraints.

Definition 1.1 (Elementary notions).

(*i*) *The set*

$$F := \left\{ x \in \mathbb{R}^n \,\middle|\, g_i(x) \le 0 \text{ for all } i = 1, \ldots, n_{\text{ineq}},\ h_j(x) = 0 \text{ for all } j = 1, \ldots, n_{\text{eq}} \right\}$$

associated with an optimization problem (1.1) is termed the ***feasible set****. Any $x \in F$ is termed a* ***feasible point****.*

(*ii*) *The inequality $g_i(x) \le 0$ is called* ***active*** *at a point x if $g_i(x) = 0$ holds. It is called* ***inactive*** *in case $g_i(x) < 0$. It is called* ***violated*** *if $g_i(x) > 0$ holds.*

(*iii*) *The value*

$$f^* := \inf \{ f(x) \,|\, x \in F \}$$

is termed the ***infimal value*** *of problem (1.1).*

(*iv*) *In case $F = \emptyset$, the problem (1.1) is said to be* ***infeasible****. In that case, we have $f^* = +\infty$. In case $f^* = -\infty$, the problem is said to be* ***unbounded****.*

(*v*) *A point* $x^* \in F$ *is a* ***global minimizer*** *or* ***globally optimal solution*** *of* (1.1) *if*

$$f(x^*) \le f(x) \text{ for all } x \in F$$

holds. Equivalently, $x^* \in F$ *is a global minimizer if* $f(x^*) = f^*$ *holds. In this case, the infimal value* f^* *is also referred to as the* ***global minimum*** *or* ***globally optimal value*** *of* (1.1).

(*vi*) *A global minimizer* x^* *is* ***strict*** *in case*

$$f(x^*) < f(x) \text{ for all } x \in F, \ x \neq x^*.$$

(*vii*) *A point* $x^* \in F$ *is a* ***local minimizer*** *or* ***locally optimal solution*** *of* (1.1) *if there exists a neighborhood* $U(x^*)$ *such that*

$$f(x^*) \le f(x) \text{ for all } x \in F \cap U(x^*)$$

holds. In this case, $f(x^*)$ *is also referred to as a* ***local minimum*** *or a* ***locally optimal value*** *of* (1.1).

(*viii*) *A local minimizer* x^* *is* ***strict*** *in case*

$$f(x^*) < f(x) \text{ for all } x \in F \cap U(x^*), \quad x \neq x^*.$$

(*ix*) *An optimization problem* (1.1) *is* ***solvable*** *if it has at least one global minimizer, i. e., if the optimal value is attained at some point. Otherwise, the problem is* ***unsolvable***.

Definition 1.2 (Classification of optimization problems)**.**

(*i*) *An optimization problem* (1.1) *is said to be* ***unconstrained*** *in case* $n_{\text{ineq}} = n_{\text{eq}} = 0$. *Otherwise, it is said to be* ***equality constrained*** and/or ***inequality constrained***.

(*ii*) *Inequality constraints of the simple kind*

$$\ell_i \le x_i \le u_i, \quad i = 1, \ldots, n$$

with bounds $\ell_i \in \mathbb{R} \cup \{-\infty\}$ *and* $u_i \in \mathbb{R} \cup \{\infty\}$ *are called* ***bound constraints***.

(*iii*) *When* f *is a quadratic polynomial and* g *and* h *are affine linear functions, then* (1.1) *is called a* ***quadratic optimization problem*** *or a* ***quadratic program*** *(**QP**).*

(*iv*) *In the general case, i. e., when* (1.1) *is not a quadratic program, we refer to* (1.1) *as a* ***nonlinear optimization problem*** *or* ***nonlinear program*** *(**NLP**).*

The emphasis in this class is on numerical techniques for unconstrained and constrained nonlinear programs. We will see that fast algorithms take into account the optimality conditions of the respective problem. Therefore we will also discuss optimality conditions.

We will begin in Chapter 1 with algorithms for unconstrained optimization. Some of the content was already part of the class *Grundlagen der Optimierung* (Herzog, 2022), but we will revisit the material in more detail here. The theory for constrained problems is relatively involved and merits its own chapter (Chapter 2). We will subsequently discuss major algorithmic ideas for constrained problems in Chapter 3. Finally, we will review in Chapter 4 some computer-aided techniques to obtain derivatives of functions, which the algorithms under consideration generally require.

Throughout the class, we will emphasize the connections between optimization and numerical linear algebra.

§ 2 Notation and Background Material

In these lecture notes we use color codes for **definitions** and highlights. The natural numbers are $\mathbb{N} = \{1, 2, \ldots\}$, and we write $\mathbb{N}_0$ for $\mathbb{N} \cup \{0\}$. We denote open intervals by (a, b) and closed intervals by $[a, b]$. We usually use Latin capital letters for matrices, Latin lowercase letters for vectors and Greek or Latin lowercase letters for scalars. We use Id for the identity matrix. We distinguish the vector space $\mathbb{R}^n$ of column vectors from the vector space $\mathbb{R}_n$ of row vectors.

§ 2.1 Vector Norms

An **inner product** $(\cdot, \cdot)$ on $\mathbb{R}^n$ is a symmetric and positive definite bilinear form, i. e., a map $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with the following properties:

$$(x, y) = (y, x) \qquad \text{(symmetry)} \tag{2.1a}$$

$$(\alpha_1\, x_1 + \alpha_2\, x_2, y) = \alpha_1\, (x_1, y) + \alpha_2\, (x_2, y) \qquad \text{(bilinearity part 1)} \tag{2.1b}$$

$$(x, \beta_1\, y_1 + \beta_2\, y_2) = \beta_1\, (x, y_1) + \beta_2\, (x, y_2) \qquad \text{(bilinearity part 2)} \tag{2.1c}$$

$$(x, x) \ge 0 \quad \text{and} \quad x \neq 0 \Rightarrow (x, x) > 0 \qquad \text{(positive definiteness)} \tag{2.1d}$$

for all $x, x_1, x_2, y, y_1, y_2 \in \mathbb{R}^n$ and all $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{R}$.

Inner products on $\mathbb{R}^n$ are in one-to-one correspondence with symmetric and positive definite (s. p. d.) $n \times n$ matrices. That is, every s. p. d. matrix $M \in \mathbb{R}^{n \times n}$ induces an inner product

$$(x, y)_M \coloneqq x^\top M\, y,$$

and, on the other hand, every inner product $(\cdot, \cdot)$ on $\mathbb{R}^n$ is induced by an s. p. d. matrix M. For simplicity, we will refer to M itself as the inner product it induces, or use the term "M-inner product".

Every inner product $(\cdot, \cdot)_M$ induces a norm[1] by way of

$$\|x\|_M \coloneqq \sqrt{x^\top M\, x}. \tag{2.2}$$

In particular, the Euclidean inner product $x^\top y$ corresponds to the identity matrix $M = \mathrm{Id}$, and we denote the associated norm by $\|x\|$. We won't be writing $\langle x\,, y \rangle$ or $x \cdot y$ for the Euclidean inner product.

[1]We are only considering norms induced by inner products.

§ 2.2 Matrix Norms

A matrix $A \in \mathbb{R}^{m \times n}$ represents a linear map by way of $\mathbb{R}^n \ni x \mapsto A\,x \in \mathbb{R}^m$. When $\mathbb{R}^n$ is equipped with the M_1-inner product and $\mathbb{R}^m$ is equipped with the M_2-inner product, we define the **matrix norm** or **operator norm** of A as

$$\|A\|_{M_2 \leftarrow M_1} \coloneqq \max_{x \neq 0} \frac{\|A\,x\|_{M_2}}{\|x\|_{M_1}}. \tag{2.3}$$

When M_1 and M_2 are both the Euclidean inner products, $\|A\|_{\mathrm{Id} \leftarrow \mathrm{Id}}$ or simply $\|A\|$ is the largest singular value of A. In the general case, $\|A\|_{M_2 \leftarrow M_1}$ is the largest singular value of a suitably generalized singular value decomposition.

§ 2.3 Eigenvalues and Eigenvectors

Every symmetric matrix $A \in \mathbb{R}^{n \times n}$ possesses an orthogonal transformation to a diagonal matrix, known as **eigen decomposition** or **spectral decomposition**. That is, there exists an orthogonal matrix $V \in \mathbb{R}^{n \times n}$ and a diagonal matrix $\Lambda \in \mathbb{R}^{n \times n}$, such that

$$A = V \Lambda V^\intercal \tag{2.4}$$

holds. The diagonal of Λ contains the eigenvalues λ_i, and the columns v_i of V are the corresponding eigenvectors. This decomposition yields the complete solution to the **eigenvalue problem**

$$A\,v = \lambda\,v. \tag{2.5}$$

We also work with the **generalized eigenvalue problem**

$$A\,v = \lambda\,M\,v \tag{2.6}$$

for the particular case where A is still symmetric and the second matrix $M \in \mathbb{R}^{n \times n}$ is s. p. d.. There exists an analogous **generalized spectral decomposition**

$$A = V \Lambda V^\intercal, \tag{2.7}$$

where now V is orthogonal w.r.t. the M^{-1} inner product, i. e., $V^\intercal M^{-1} V = \mathrm{Id}$ holds. This implies $V V^\intercal = M$. We also refer to the solutions of (2.6) as the **eigenvalues/eigenvectors of A w.r.t. M** or **eigenvalues/eigenvectors of the pair $(A; M)$.**

In view of the Courant-Fischer theorem for (generalized) eigenvalues of symmetric matrices, the **generalized Rayleigh quotient** of A w.r.t. M satisfies

$$\lambda_{\min}(A; M) \le \frac{x^\intercal A\,x}{x^\intercal M\,x} \le \lambda_{\max}(A; M) \quad \text{for all } x \neq 0. \tag{2.8}$$

The eigenvectors associated with the smallest and largest generalized eigenvalues $\lambda_{\min}(A; M)$ and $\lambda_{\max}(A; M)$ satisfy the first respectively the second inequality with equality.

Notice that the generalized eigenvalue problems (2.6) and

$$M\,v = \lambda\,M\,A^{-1}M\,v \tag{2.9a}$$

as well as

$$A\,M^{-1}A\,v = \lambda\,A\,v \tag{2.9b}$$

have the save eigenvalues and eigenvectors (provided in case of (2.9a) that A is not only symmetric but also invertible) since $M\,v = \lambda\,M\,A^{-1}M\,v \Leftrightarrow v = \lambda\,A^{-1}M\,v \Leftrightarrow A\,v = \lambda\,M\,v$ and $A\,M^{-1}A\,v = \lambda\,A\,v \Leftrightarrow M^{-1}A\,v = \lambda\,v \Leftrightarrow A\,v = \lambda\,M\,v$. Consequently, we obtain the following estimate for the generalized Rayleigh quotients associated with (2.9):

$$\lambda_{\min}(A;M) \le \frac{x^\top M\,x}{x^\top M\,A^{-1}M\,x} \le \lambda_{\max}(A;M) \quad \text{for all } x \neq 0, \tag{2.10a}$$

$$\lambda_{\min}(A;M) \le \frac{x^\top A\,M^{-1}A\,x}{x^\top A\,x} \le \lambda_{\max}(A;M) \quad \text{for all } x \neq 0. \tag{2.10b}$$

Every s. p. d. matrix $A \in \mathbb{R}^{n\times n}$ possesses a unique s. p. d. **matrix square root** $A^{1/2}$. When $A = V\Lambda V^\top$ is a spectral decomposition of A with orthogonal V, then

$$A^{1/2} = V\Lambda^{1/2}V^\top \tag{2.11}$$

holds. Herein, $\Lambda^{1/2}$ is the elementwise square root of the diagonal matrix Λ.

§ 2.4 Kantorovich Inequality

Suppose that A is an s. p. d. matrix. Let us denote the extremal eigenvalues by $\alpha := \lambda_{\min}(A)$ and $\beta := \lambda_{\max}(A)$. Moreover, since A is s. p. d., it follows that its **condition number**[2] is given by

$$\kappa := \frac{\beta}{\alpha}. \tag{2.12}$$

Notice that a condition number always satisfies $\kappa \ge 1$. From the Rayleigh quotient estimate (2.8) (with $M = \mathrm{Id}$), we have

$$\frac{x^\top A\,x}{\|x\|^2} \le \beta.$$

Moreover, since the eigenvalues of A^{-1} are the reciprocals of those of A, we have $\lambda_{\max}(A^{-1}) = 1/\lambda_{\min}(A) = 1/\alpha$ and thus

$$\frac{x^\top A^{-1}\,x}{\|x\|^2} \le \frac{1}{\alpha}.$$

These inequalities hold for all $x \in \mathbb{R}^n \setminus \{0\}$, and they imply

$$\frac{(x^\top A\,x)\,(x^\top A^{-1}\,x)}{\|x\|^4} \le \frac{\beta}{\alpha}.$$

This estimate, however, is not sharp in general. (**Quiz 2.1:** Can you explain why not?) The Kantorovich inequality improves this estimate.

[2] Generally, the condition of an invertible matrix A is $\kappa = \|A\|\,\|A^{-1}\|$. This is equal to $\sigma_{\max}(A)/\sigma_{\min}(A)$ with the extremal singular values $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$. Since A is symmetric, its singular values are just the absolute values of its eigenvalues, and since A is also positive definite, we have $\sigma_{\max}(A) = \lambda_{\max}(A) = \beta$ and $\sigma_{\min}(A) = \lambda_{\min}(A) = \alpha$.

Lemma 2.1 (Kantorovich inequality). *Suppose that $A \in \mathbb{R}^{n\times n}$ is s. p. d., $\alpha := \lambda_{\min}(A)$ and $\beta := \lambda_{\max}(A)$ are its extremal eigenvalues, and $\kappa = \beta/\alpha$ is its condition number. Then*

$$1 \le \frac{(x^\top A\,x)\,(x^\top A^{-1}x)}{\|x\|^4} \le \frac{(\alpha+\beta)^2}{4\,\alpha\,\beta} \le \frac{\beta}{\alpha} \tag{2.13a}$$

holds for all $x \in \mathbb{R}^n \setminus \{0\}$, or equivalently, in terms of the condition number $\kappa = \beta/\alpha$,

$$1 \le \frac{(x^\top A\,x)\,(x^\top A^{-1}x)}{\|x\|^4} \le \frac{(\kappa+1)^2}{4\,\kappa} \le \kappa. \tag{2.13b}$$

Proof. The Cauchy-Schwarz inequality implies

$$\|x\|^2 = x^\top x = x^\top A^{-1/2}\,A^{1/2}x \le \|A^{-1/2}x\|\,\|A^{1/2}x\|.$$

By squaring this, we obtain

$$\|x\|^4 \le \|A^{-1/2}x\|^2\,\|A^{1/2}x\|^2 = (x^\top A\,x)\,(x^\top A^{-1}x)$$

and thus the lower bound in (2.13).

From here on, the proof follows Anderson, 1971, as reproduced in the Master's thesis Alpargu, 1996, Section 1.2.2. Let $\lambda_1,\dots,\lambda_n > 0$ be the eigenvalues of A (in any order), and let $v_1,\dots,v_n$ be an orthonormal set of associated eigenvectors. We represent $x \in \mathbb{R}^n \setminus \{0\}$ as $x = \sum_{i=1}^n \gamma_i\,v_i$. Suppose, w.l.o.g., that $\|x\|^2 = \sum_{i=1}^n \gamma_i^2 = 1$ holds. Inserting the representation of x yields

$$\frac{(x^\top A\,x)\,(x^\top A^{-1}x)}{\|x\|^4} = \underbrace{\Big[\sum_{i=1}^n \lambda_i\,\gamma_i^2\Big]}_{=\mathbb{E}(T)}\;\underbrace{\Big[\sum_{i=1}^n \frac{1}{\lambda_i}\,\gamma_i^2\Big]}_{=\mathbb{E}(1/T)}.$$

It is helpful to think about the two factors on the right-hand side as expected values of a "random variable" T and $1/T$, respectively. Here T takes the values $\lambda_i \in [\alpha,\beta]$ with "probability" γ_i^2. For any $0 < \alpha \le T \le \beta$, we can estimate

$$0 \le (\beta - T)\,(T-\alpha) = (\beta+\alpha-T)\,T - \alpha\,\beta,$$

and thus

$$\frac{1}{T} \le \frac{\alpha+\beta-T}{\alpha\,\beta}.$$

Taking the expected value, this implies

$$\begin{aligned}\mathbb{E}(T)\,\mathbb{E}(1/T) &\le \mathbb{E}(T)\,\frac{\alpha+\beta-\mathbb{E}(T)}{\alpha\,\beta}\\ &= \frac{(\alpha+\beta)^2}{4\,\alpha\,\beta} - \frac{1}{\alpha\,\beta}\Big[\mathbb{E}(T) - \frac{1}{2}(\alpha+\beta)\Big]^2\\ &\le \frac{(\alpha+\beta)^2}{4\,\alpha\,\beta}.\end{aligned}$$

This shows that essential upper bound in (2.13). The remaining inequality follows directly from $0 < \alpha \le \beta$. $\square$

Instead of the Euclidean norm, we can also use the norm induced by the M-inner product.

Corollary 2.2 (Generalized Kantorovich inequality). *Suppose that $A \in \mathbb{R}^{n\times n}$ are M are both s. p. d., $\alpha := \lambda_{\min}(A;M)$ and $\beta := \lambda_{\max}(A;M)$ are the extremal generalized eigenvalues of A w.r.t. M. Then*

$$1 \le \frac{(x^\intercal A\, x)\,(x^\intercal M\, A^{-1} M\, x)}{\|x\|_M^4} \le \frac{(\alpha+\beta)^2}{4\,\alpha\,\beta} \le \frac{\beta}{\alpha} \tag{2.14a}$$

holds for all $x \in \mathbb{R}^n \setminus \{0\}$, or equivalently, in terms of the ***generalized condition number*** *$\kappa = \beta/\alpha$,*

$$1 \le \frac{(x^\intercal A\, x)\,(x^\intercal A^{-1} x)}{\|x\|_M^4} \le \frac{(\kappa+1)^2}{4\,\kappa} \le \kappa. \tag{2.14b}$$

We do not give a proof of Corollary 2.2 here; see for instance Herzog, 2022, Folgerung 4.14.

§ 2.5 Functions and Derivatives

- Given a function $f\colon \mathbb{R}^n \to \mathbb{R}$ and $x \in \mathbb{R}^n$, the derivative of the partial function $t \mapsto f(x + t\, e^{(i)})$ at $t = 0$ is the i-th **partial derivative** of f at x, briefly: $\frac{\partial}{\partial x_i} f(x)$. Here $e^{(i)} = (0, \ldots, 0, 1, 0, \ldots, 0)^\intercal$ is one of the standard basis vectors of $\mathbb{R}^n$. In other words,

$$\frac{\partial}{\partial x_i} f(x) = \lim_{t\to 0} \frac{f(x + t\, e^{(i)}) - f(x)}{t}.$$

- More generally, the derivative of the function $t \mapsto f(x + t\, d)$ at $t = 0$ is the **(two-sided) directional derivative** of f at x in the direction $d \in \mathbb{R}^n$, briefly:

$$\frac{\partial}{\partial d} f(x) = \lim_{t\to 0} \frac{f(x + t\, d) - f(x)}{t}.$$

- The right-sided derivative of the function $t \mapsto f(x + t\, d)$ at $t = 0$ is the **(one-sided) directional derivative** of f at x in the direction $d \in \mathbb{R}^n$, briefly:

$$f'(x; d) = \lim_{t\searrow 0} \frac{f(x + t\, d) - f(x)}{t}.$$

- A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is **differentiable** at $x \in \mathbb{R}^n$ if there exists a row vector $v \in \mathbb{R}_n$ such that

$$\frac{f(x + d) - f(x) - v\, d}{\|d\|} \to 0 \quad \text{for } d \to 0.$$

  In this case, the vector v is the **(total) derivative** of f at x, and it is denoted by $f'(x)$.

- When $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$, then

  $$f'(x) = \left(\tfrac{\partial f(x)}{\partial x_1}, \quad \dots, \quad \tfrac{\partial f(x)}{\partial x_n}\right) \in \mathbb{R}_n.$$

  The transposed vector (a column vector)

  $$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix} = f'(x)^\intercal \in \mathbb{R}^n$$

  is the **gradient** (w.r.t. the Euclidean inner product) of f at x.

- When $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$, then

  $$f'(x; d) = \frac{\partial}{\partial d} f(x) = f'(x)\, d$$

  holds for all $d \in \mathbb{R}^n$. That is, the one-sided and two-sided directional derivatives of f at x agree, and they can be evaluated by applying the derivative $f'(x)$ to the direction d.

- A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is **continuously partially differentiable** or briefly: $C^1(\mathbb{R}^n, \mathbb{R})$, if all partial derivatives $\frac{\partial f(x)}{\partial x_i}$, as functions of x, are continuous. C^1-functions are differentiable, and the derivative f' is continuous.

- A vector-valued function $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is **differentiable** at $x \in \mathbb{R}^n$ if all component function $F_1, \dots, F_m$ are differentiable at x. In this case, the derivative $F'(x)$ is given by the **Jacobian** of F at x, i. e., by

  $$\begin{pmatrix} \dfrac{\partial F_1(x)}{\partial x_1} & \dots & \dfrac{\partial F_1(x)}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial F_m(x)}{\partial x_1} & \dots & \dfrac{\partial F_m(x)}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{m \times n}.$$

- F is **continuously partially differentiable** if all entries of the Jacobian are continuous as functions of x. C^1-functions are differentiable, and the derivative F' is continuous.

- A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is **twice differentiable** at $x \in \mathbb{R}^n$ if f is differentiable in a neighborhood of x and the derivative $x \mapsto f'(x) \in \mathbb{R}^n$ is differentiable at x. In this case, the second derivative $f''(x)$ is given by the **Hessian** of f at x, i. e., by the matrix of second-order partial derivatives

  $$\left(\frac{\partial^2 f(x)}{\partial x_i\, \partial x_j}\right)_{i,j=1}^n = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \dots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \dots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{pmatrix}.$$

  When f is twice differentiable at x, then the Hessian is symmetric by Schwarz' theorem.[3]

[3] See for instance Cartan, 1967, Proposition 5.2.2

- A function $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ is **twice continuously partially differentiable** or briefly: $C^2(\mathbb{R}^n, \mathbb{R})$, if all entries of the Hessian are continuous as functions of x. C^2-functions are twice differentiable.

§ 2.6 Taylor's Theorem

We are going to state Taylor's theorem in two variants:

Theorem 2.3 (Taylor, see Cartan, 1967, Theorem 5.6.3)**.** *Suppose that $G \subseteq \mathbb{R}^n$ open, $k \in \mathbb{N}_0$ and $f\colon G \longrightarrow \mathbb{R}$ k times differentiable, and $(k+1)$ times differentable at $x_0 \in G$. Then for all $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\text{in case } k = 0: \quad \left| f(x_0 + d) - f(x_0) - f'(x_0)\, d \right| \le \varepsilon \, \|d\|,$$
$$\text{in case } k = 1: \quad \left| f(x_0 + d) - f(x_0) - f'(x_0)\, d - \frac{1}{2} d^\intercal f''(x_0) d \right| \le \varepsilon \, \|d\|^2.$$

for all $\|d\| < \delta$.

Theorem 2.4 (Taylor, see Geiger, Kanzow, 1999, Satz A.2 or Heuser, 2002, Satz 168.1)**.**
Suppose that $G \subseteq \mathbb{R}^n$ is open, $k \in \mathbb{N}_0$ and $f\colon G \longrightarrow \mathbb{R}$ $(k+1)$ times continuously partially differentiable, briefly a $C^{k+1}(G, \mathbb{R})$ function. Suppose that x_0 and $x_0 + d$ and the entire line segment between them lie in G. Then there exists $\xi \in (0,1)$ such that

$$\text{in case } k = 0: \quad f(x_0 + d) = f(x_0) + f'(x_0 + \xi\, d)\, d \quad (\textbf{\textit{mean value theorem}}),$$
$$\text{in case } k = 1: \quad f(x_0 + d) = f(x_0) + f'(x_0)\, d + \frac{1}{2} d^\intercal f''(x_0 + \xi\, d)\, d.$$

§ 2.7 Convergence Rates

We denote (vector-valued) sequences $\mathbb{N} \longrightarrow \mathbb{R}^n$ by $\big(x^{(k)}\big)$ and not (x_k) etc., in order to avoid a conflict of notation with the components of a vector $x = (x_1, \ldots, x_n)^\intercal \in \mathbb{R}^n$. The **subsequence** of $\big(x^{(k)}\big)$ obtained by the strictly increasing sequence $\mathbb{N} \ni \ell \mapsto k^{(\ell)} \in \mathbb{N}$ is denoted by $\big(x^{(k^{(\ell)})}\big)$.

We introduce various convergence rates for sequences in order to characterize the speed of convergence, e. g., of iterates in an algorithm.

Definition 2.5 (Q-convergence rates[4])**.**
Suppose that $\big(x^{(k)}\big) \subset \mathbb{R}^n$ is a sequence and $x^* \in \mathbb{R}^n$. Moreover, let M be an inner product on $\mathbb{R}^n$.

(i) $\big(x^{(k)}\big)$ converges to x^* (at least) ***Q-linearly*** *w.r.t. the M-norm if there exists $c \in (0,1)$ such that*

$$\|x^{(k+1)} - x^*\|_M \le c\, \|x^{(k)} - x^*\|_M \quad \text{for all } k \in \mathbb{N} \text{ sufficiently large.}$$

[4] "Q" stands for "quotient".

(ii) $(x^{(k)})$ *converges to* x^* *(at least)* ***Q-superlinearly*** *w.r.t. the* M*-norm if there exists a null sequence* $(\varepsilon^{(k)})$ *such that*

$$\|x^{(k+1)} - x^*\|_M \le \varepsilon^{(k)} \, \|x^{(k)} - x^*\|_M \quad \text{for all } k \in \mathbb{N}.$$

(iii) *Suppose that* $x^{(k)} \to x^*$. $(x^{(k)})$ *converges to* x^* *(at least)* ***Q-quadratically*** *w.r.t. the* M*-norm if there exists* $C > 0$ *such that*

$$\|x^{(k+1)} - x^*\|_M \le C \, \|x^{(k)} - x^*\|_M^2 \quad \text{for all } k \in \mathbb{N}.$$

Note: Q-superlinear and Q-quadratic convergence of a sequence are independent of the norm (inner product) M. However, the property of Q-linear convergence can be lost when changing the norm.

Definition 2.6 (R-convergence rates[5]**).**
Suppose that $(x^{(k)}) \subset \mathbb{R}^n$ *is a sequence and* $x^* \in \mathbb{R}^n$. *Moreover, let* M *be an inner product on* $\mathbb{R}^n$.

(i) $(x^{(k)})$ *converges to* x^* *(at least)* ***R-linearly*** *w.r.t. the* M*-norm if there exists a null sequence* $(\varepsilon^{(k)})$ *such that*

$$\|x^{(k)} - x^*\|_M \le \varepsilon^{(k)} \quad \text{for all } k \in \mathbb{N},$$

and $(\varepsilon^{(k)})$ *converges to zero Q-linearly w.r.t.* $|\cdot|$.

(ii) $(x^{(k)})$ *converges to* x^* *(at least)* ***R-superlinearly*** *w.r.t. the* M*-norm if there exists a null sequence* $(\varepsilon^{(k)})$ *such that*

$$\|x^{(k)} - x^*\|_M \le \varepsilon^{(k)} \quad \text{for all } k \in \mathbb{N},$$

and $(\varepsilon^{(k)})$ *converges to zero Q-superlinearly w.r.t.* $|\cdot|$.

(iii) $(x^{(k)})$ *converges to* x^* *(at least)* ***R-quadratically*** *w.r.t. the* M*-norm if there exists a null sequence* $(\varepsilon^{(k)})$ *such that*

$$\|x^{(k)} - x^*\|_M \le \varepsilon^{(k)} \quad \text{for all } k \in \mathbb{N},$$

and $(\varepsilon^{(k)})$ *converges to zero Q-quadratically w.r.t.* $|\cdot|$.

Note: The R-convergence modes are slightly weaker than the respective Q-convergence rates. Q-convergence considers the decrease in the distance to the limit $\|x^{(k)} - x^*\|_M$ in every step of the sequence. By contrast, R-convergence considers the decrease overall.

§ 2.8 Convexity

Convexity plays a very important role in optimization in general. In this class, however, we will rely on it only scarcely. We briefly recall here some elements of convexity. You may study Herzog, 2022, § 13 if you wish to have more background information.

[5] "R" stands for "root".

Definition 2.7 (Convex function).
A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is termed

(i) ***convex*** *in case*
$$f(\alpha\, x + (1-\alpha)\, y) \le \alpha f(x) + (1-\alpha) f(y) \tag{2.15}$$
holds for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0,1]$.

(ii) ***strictly convex*** *in case*
$$f(\alpha\, x + (1-\alpha)\, y) < \alpha f(x) + (1-\alpha) f(y) \tag{2.16}$$
holds for all $x, y \in \mathbb{R}^n$ and $\alpha \in (0,1)$.

(iii) ***μ-strongly convex*** *or* ***strongly convex*** *with parameter $\mu > 0$ in case*
$$f(\alpha\, x + (1-\alpha)\, y) + \frac{\mu}{2}\, \alpha\, (1-\alpha) \|x - y\|^2 \le \alpha f(x) + (1-\alpha) f(y) \tag{2.17}$$
holds for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0,1]$.

(iv) ***concave*** *(concave) or* ***strictly concave*** *or* ***constrly concave*** *if $-f$ is convex or strictly convex or strongly convex, respectively.*

Theorem 2.8 (Characterization of convexity via first-order derivatives).
Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable.

(a) *The following are equivalent:*

  (i) *f is convex.*

  (ii) *For all $x, y \in \mathbb{R}^n$,*
$$f(x) - f(y) \ge f'(y)(x-y) \tag{2.18}$$
  holds.

  (iii) *For all $x, y \in \mathbb{R}^n$,*
$$\big(f'(x) - f'(y)\big)(x-y) \ge 0 \tag{2.19}$$
  holds. Equation (2.19) means that f' is a ***monotone operator****.*

(b) *The following are equivalent:*

  (i) *f ist strictly convex.*

  (ii) *For all $x, y \in \mathbb{R}^n$ such that $x \ne y$,*
$$f(x) - f(y) > f'(y)(x-y) \tag{2.20}$$
  holds.

(*iii*) *For all* $x, y \in \mathbb{R}^n$ *such that* $x \neq y$,

$$\big(f'(x) - f'(y)\big)(x - y) > 0. \tag{2.21}$$

Equation (2.21) means that f' *is a* ***strictly monotone operator****.*

(c) The following are equivalent:

(*i*) f *ist strongly convex.*

(*ii*) *There exists* $\mu > 0$ *such that for all* $x, y \in \mathbb{R}^n$,

$$f(x) - f(y) \geq f'(y)(x - y) + \frac{\mu}{2}\,\|x - y\|^2 \tag{2.22}$$

holds.

(*iii*) *There exists* $\mu > 0$ *such that for all* $x, y \in \mathbb{R}^n$,

$$\big(f'(x) - f'(y)\big)(x - y) \geq \mu\,\|x - y\|^2. \tag{2.23}$$

Equation (2.23) means that f' *is a* ***strongly monotone operator****.*

Theorem 2.9 (Characterization of convexity via second-order derivatives)**.**
Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}$ *is twice differentiable.*

(a) The following are equivalent:

(*i*) f *ist convex.*

(*ii*) f'' *is everywhere positive semidefinite (has only non-negative eigenvalues).*

(b) When f'' *is everywhere positive definite, then* f *is strictly convex.*

(c) The following are equivalent:

(*i*) f *is strongly convex with parameter* $\mu > 0$.

(*ii*) *The smallest eigenvalue of* $f''(x)$ *satisfies* $\lambda_{\min}(f''(x)) \geq \mu > 0$ *for all* $x \in \mathbb{R}^n$.

§ 2.9 Miscellanea

We denote the **interior** of a set $M \subseteq \mathbb{R}^n$ by $\operatorname{int} M$ and its **closure** by $\operatorname{cl} M$. Given $\varepsilon > 0$ and $x \in \mathbb{R}^n$,

$$B_\varepsilon^M(x_0) \coloneqq \big\{x \in \mathbb{R}^n \,\big|\, \|x - x_0\|_M < \varepsilon\big\}$$

denotes the **open** ε**-ball** w.r.t. the M-norm about x_0. Similarly, the **closed** ε**-ball** is

$$\operatorname{cl} B_\varepsilon^M(x_0) \coloneqq \big\{x \in \mathbb{R}^n \,\big|\, \|x - x_0\|_M \leq \varepsilon\big\}.$$

The **ceiling function** $\lceil x \rceil$ returns the smallest integer $\geq x$.

Chapter 1 Numerical Techniques for Unconstrained Optimization Problems

We discuss in this chapter numerical methods for the unconstrained version of (1.1), i. e.,

$$\text{Minimize} \quad f(x) \quad \text{where } x \in \mathbb{R}^n. \tag{UP}$$

The reason for discussing the unconstrained problem first is that we can introduce the essential algorithmic techniques without the difficulties of any constraints present.

Up front, we mention that we can only hope to find *local* minimizers. Determining *global* minimizers is generally much harder and only possible under additional assumptions on the objective, and generally only in relatively small dimensions $n \in \mathbb{N}$. A notable case of an additional assumption is that of a *convex* objective f. In this case, every local minimizer is already a global minimizer. Morever, first-order optimality conditions are already sufficient for optimality, and we do not require second-order conditions.

§ 3 Optimality Conditions

We assume you have seen the following first- and second-order optimality conditions, so we only briefly recall them; see Herzog, 2022, § 3 for more details.

Theorem 3.1 (First-order necessary optimality condition)**.**
Suppose that x^* is a local minimizer of (UP) and that f is differentiable at x^*. Then $f'(x^*) = 0$.

Proof. Suppose that $d \in \mathbb{R}^n$ is arbitrary. We consider the curve $\gamma\colon (-\delta, \delta) \to \mathbb{R}^n$, $\gamma(t) \coloneqq x^* + t\,d$. For sufficiently small $\delta > 0$, this curve runs within the neighborhood of local optimality of x^*. This implies that $f \circ \gamma$ has a local minimizer at $t = 0$.

From this local optimality, we infer that the difference quotient satisfies

$$\frac{f(\gamma(t)) - f(\gamma(0))}{t} = \frac{f(x^* + t\,d) - f(x^*)}{t} \begin{cases} \ge 0 & \text{for } t > 0, \\ \le 0 & \text{for } t < 0. \end{cases}$$

On the other hand, this difference quotient converges to $f'(x^*)\,d$ as $t \to 0$. Consequently, we must have $f'(x^*)\,d = 0$. Since $d \in \mathbb{R}^n$ was arbitrary, this means $f'(x^*) = 0$. □

A point $x \in \mathbb{R}^n$ with the property $f'(x) = 0$ is termed a **stationary point** of f.

Theorem 3.2 (Second-order necessary optimality condition).
Suppose that x^* is a local minimizer of (**UP**) *and that f is twice differentiable at x^*. Then the Hessian $f''(x^*)$ is positive semidefinite.*[1]

Proof. Es sei $d \in \mathbb{R}^n$ beliebig. Wie in Theorem 3.1 we define $\gamma(t) \coloneqq x^* + t\,d$ and again consider the objective along the curve, i. e., $\varphi \coloneqq f \circ \gamma$, which has a local minimizer at $t = 0$. Since φ is twice differentiable at $t = 0$, Theorem 2.3 implies the following: for all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\left|\varphi(t) - \varphi(0) - \varphi'(0)\,t - \frac{1}{2}\varphi''(0)\,t^2\right| \le \varepsilon\,t^2$$

holds for all $|t| < \delta$. In view of Theorem 3.1, $\varphi'(0) = 0$, and the local optimality implies $\varphi(0) \le \varphi(t)$ for all $|t|$ sufficiently small. We thus obtain

$$-\frac{1}{2}\varphi''(0)\,t^2 \le \varphi(t) - \varphi(0) - \frac{1}{2}\varphi''(0)\,t^2 \le \varepsilon\,t^2$$

for all $|t|$ sufficiently small, whence

$$\frac{1}{2}\varphi''(0) \ge -\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we conclude $\varphi''(0) = d^\intercal f''(x^*)\,d \ge 0$. And since $d \in \mathbb{R}^n$ was arbitrary, we have shown $f''(x^*)$ to be positive semidefinite. □

Theorem 3.3 (Second-order sufficient optimality condition).
Suppose that f is twice differentiable at x^* and

(*i*) $f'(x^*) = 0$ *and*

(*ii*) $f''(x^*)$ *is positive definite*[2]*, with minimal eigenvalue $\mu > 0$.*

Then for every $\beta \in (0, \mu)$, there exists a neighborhood $U(x^*)$ of x^* such that

$$f(x) \ge f(x^*) + \frac{\beta}{2}\|x - x^*\|^2 \quad \textit{for all } x \in U(x^*). \tag{3.1}$$

In particular, x^* is a strict local minimizer of f.

Proof. Here we use Theorem 2.3 directly for f (not along a curve). For every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\left|f(x^* + d) - f(x^*) - f'(x^*)\,d - \frac{1}{2}d^\intercal f''(x^*)d\right| \le \varepsilon\,\|d\|^2$$

holds for all $\|d\| < \delta$. According to the assumptions, $f'(x^*) = 0$ holds. Therefore,

$$-\varepsilon\,\|d\|^2 \le f(x^* + d) - f(x^*) - \frac{1}{2}d^\intercal f''(x^*)d$$

[1] Due to the symmetry of $f''(x^*)$ this is equivalent to all eigenvalues of $f''(x^*)$ being non-negative.
[2] Due to the symmetry of $f''(x^*)$ this is equivalent to all eigenvalues of $f''(x^*)$ being positive.

holds for all $\|d\| < \delta$. This implies

$$f(x^* + d) \ge f(x^*) + \frac{1}{2} d^\intercal f''(x^*)\, d - \varepsilon \|d\|^2$$

for all $\|d\| < \delta$.

From (2.8) (with $M = \mathrm{Id}$), the values of the Rayleigh quotient associated with the symmetric matrix $f''(x^*)$ are bounded above and below by the extremal eigenvalues of $f''(x^*)$. In particular, we have

$$d^\intercal f''(x^*)\, d \ge \mu \|d\|^2 \quad \text{for all } d \in \mathbb{R}^n.$$

We can now finalize the proof: for $\beta \in (0, \mu)$, choose $\varepsilon := (\mu - \beta)/2 > 0$ and an appropriate value of $\delta > 0$. Then we have

$$\begin{aligned} f(x^* + d) &\ge f(x^*) + \frac{1}{2} d^\intercal f''(x^*)\, d - \varepsilon \|d\|^2 \\ &\ge f(x^*) + \frac{\mu}{2} \|d\|^2 - \varepsilon \|d\|^2 \\ &= f(x^*) + \frac{\beta}{2} \|d\|^2 \end{aligned}$$

for all $\|d\| < \delta$. □

Property (3.1) means that f has at least **quadratic growth** near x^*. Equivalently, f is locally strongly convex with parameter $\beta \in (0, \mu)$.

End of Week 1

Chapter 2 Theory for Constrained Optimization Problems

Chapter 3 Numerical Techniques for Constrained Optimization Problems

Chapter 4 Differentiation Techniques

Bibliography

Alpargu, G. (1996). “The Kantorovich Inequality, with Some Extensions and with Some Statistical Applications”. MA thesis. Department of Mathematics and Statistics, McGill University, Montreal, Canada.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. John Wiley & Sons, Inc., New York-London-Sydney. DOI: 10.1002/9781118186428.

Cartan, H. (1967). *Calcul Différentiel*. Paris: Hermann, p. 178.

Geiger, C.; C. Kanzow (1999). *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. New York: Springer. DOI: 10.1007/978-3-642-58582-1.

Herzog, R. (2022). *Grundlagen der Optimierung*. Lecture notes. URL: https://tinyurl.com/scoop-gdo.

Heuser, H. (2002). *Lehrbuch der Analysis. Teil 2*. 12th ed. Stuttgart: B.G.Teubner. DOI: 10.1007/978-3-322-96826-5.

Nocedal, J.; S. J. Wright (2006). *Numerical Optimization*. 2nd ed. New York: Springer. DOI: 10.1007/978-0-387-40065-5.

Ulbrich, M.; S. Ulbrich (2012). *Nichtlineare Optimierung*. New York: Springer. DOI: 10.1007/978-3-0346-0654-7.