# Lecture Notes
# Nonlinear Optimization

Spring Semester 2023

Roland Herzog*

2023-04-21

*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

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

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

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

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

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

(*ii*) *The inequality $g_i(x) \leq 0$ 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*

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

*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*

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

*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*

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

*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*

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

*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:

$$(x, y) = (y, x) \qquad \text{(symmetry)} \tag{2.1a}$$
$$(\alpha_1 x_1 + \alpha_2 x_2, y) = \alpha_1 (x_1, y) + \alpha_2 (x_2, y) \qquad \text{(bilinearity part 1)} \tag{2.1b}$$
$$(x, \beta_1 y_1 + \beta_2 y_2) = \beta_1 (x, y_1) + \beta_2 (x, y_2) \qquad \text{(bilinearity part 2)} \tag{2.1c}$$
$$(x, x) \geq 0 \quad \text{and} \quad x \neq 0 \Rightarrow (x, x) > 0 \qquad \text{(positive definiteness)} \tag{2.1d}$$

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

$$(x, y)_M := x^\mathsf{T} M\, y,$$

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

$$\|x\|_M := \sqrt{x^\mathsf{T} M\, x}. \tag{2.2}$$

In particular, the Euclidean inner product $x^\mathsf{T} y$ 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

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

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

$$A\,V = V\Lambda, \quad \text{i. e.,} \quad A = V\Lambda V^\mathsf{T} \tag{2.4}$$

holds. The diagonal of $\Lambda$ contains the eigenvalues $\lambda_i$, 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**

$$A\,V = M\,V\Lambda, \quad \text{i. e.,} \quad A = M\,V\Lambda V^\mathsf{T} M, \tag{2.7}$$

where now $V$ is orthogonal w.r.t. the $M$ inner product, i. e., $V^\mathsf{T} M\,V = \mathrm{Id}$ holds. This implies $VV^\mathsf{T} = M^{-1}$. 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

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

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 same 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):

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

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

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 $A = V \Lambda V^{\mathsf{T}}$ is a spectral decomposition of $A$ with orthogonal $V$, then

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

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

## § 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 $\kappa \geq 1$. From the Rayleigh quotient estimate (2.8) (with $M = \mathrm{Id}$), we have

$$\frac{x^{\mathsf{T}} A x}{\|x\|^2} \leq \beta.$$

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

$$\frac{x^{\mathsf{T}} A^{-1} x}{\|x\|^2} \leq \frac{1}{\alpha}.$$

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

$$\frac{(x^{\mathsf{T}} A x)\,(x^{\mathsf{T}} A^{-1} x)}{\|x\|^4} \leq \frac{\beta}{\alpha}.$$

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*

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

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

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

*Proof.* The Cauchy-Schwarz inequality implies

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

By squaring this, we obtain

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

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

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

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" $\gamma_i^2$. 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 $0 < \alpha \le \beta$. □

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*

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

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

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

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 $e^{(i)} = (0, \dots, 0, 1, 0, \dots, 0)^\mathsf{T}$ 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

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

  The transposed vector (a column vector)

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix} = f'(x)^\mathsf{T} \in \mathbb{R}^n$$

  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 $F_1, \ldots, F_m$ are differentiable at $x$. In this case, the derivative $F'(x)$ is given by the **Jacobian** of $F$ at $x$, i. e., by

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

- $F$ 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

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

  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 $f\colon \mathbb{R}^n \to \mathbb{R}$ 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 $f\colon G \to \mathbb{R}$ $k$ times differentiable, and $(k+1)$ times differentable at $x_0 \in G$. Then for all $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\text{in case } k = 0: \quad \left|f(x_0 + d) - f(x_0) - f'(x_0)\, d\right| \leq \varepsilon \|d\|,$$
$$\text{in case } k = 1: \quad \left|f(x_0 + d) - f(x_0) - f'(x_0)\, d - \frac{1}{2} d^\mathsf{T} f''(x_0) d\right| \leq \varepsilon \|d\|^2.$$

*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 $f\colon G \to \mathbb{R}$ $(k+1)$ 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*

$$\text{in case } k = 0: \quad f(x_0 + d) = f(x_0) + f'(x_0 + \xi\, d)\, d \quad (\textbf{\textit{mean value theorem}}),$$
$$\text{in case } k = 1: \quad f(x_0 + d) = f(x_0) + f'(x_0)\, d + \frac{1}{2} d^\mathsf{T} f''(x_0 + \xi\, d)\, d.$$

## § 2.7 Convergence Rates

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

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 $\left(x^{(k)}\right) \subset \mathbb{R}^n$ is a sequence and $x^* \in \mathbb{R}^n$. Moreover, let $M$ be an inner product on $\mathbb{R}^n$.*

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

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

[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*

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

(*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*

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

**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*

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

*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*

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

*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*

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

*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*

$$f(\alpha\, x + (1-\alpha)\, y) \leq \alpha f(x) + (1-\alpha) f(y) \tag{2.15}$$

*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*) ***$\mu$-strongly convex*** *or* ***strongly convex*** *with parameter $\mu > 0$ in case*

$$f(\alpha\, x + (1-\alpha)\, y) + \frac{\mu}{2}\, \alpha\, (1-\alpha) \|x - y\|^2 \leq \alpha f(x) + (1-\alpha) f(y) \tag{2.17}$$

*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 $x, y \in \mathbb{R}^n$,*

$$f(x) - f(y) \geq f'(y)(x - y) \tag{2.18}$$

*holds.*

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

$$\big(f'(x) - f'(y)\big)(x - y) \geq 0 \tag{2.19}$$

*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 $x \neq y$,*

$$f(x) - f(y) > f'(y)(x - y) \tag{2.20}$$

*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 $f : \mathbb{R}^n \to \mathbb{R}$ 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 $S \subseteq \mathbb{R}^n$ by $\operatorname{int} S$ and its **closure** by $\operatorname{cl} S$.

Given $\varepsilon > 0$ and $x \in \mathbb{R}^n$,

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

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

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

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

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

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 $\gamma(t) := x^* + t\,d$ and again consider the objective along the curve, i. e., $\varphi := f \circ \gamma$, which has a local minimizer at $t = 0$. Since $\varphi$ is twice differentiable at $t = 0$, Theorem 2.3 implies the following: for all $\varepsilon > 0$ there exists $\delta > 0$ such that

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

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

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

for all $|t|$ sufficiently small, whence

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

Since $\varepsilon > 0$ was arbitrary, we conclude $\varphi''(0) = d^\mathsf{T} f''(x^*)\,d \geq 0$. 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*

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

*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

$$\left|f(x^* + d) - f(x^*) - f'(x^*)\,d - \frac{1}{2}d^\mathsf{T} f''(x^*)d\right| \leq \varepsilon\,\|d\|^2$$

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

$$-\varepsilon\,\|d\|^2 \leq f(x^* + d) - f(x^*) - \frac{1}{2}d^\mathsf{T} f''(x^*)d$$

[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

$$f(x^* + d) \geq f(x^*) + \frac{1}{2} d^\mathsf{T} f''(x^*)\, d - \varepsilon \|d\|^2$$

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

$$d^\mathsf{T} f''(x^*)\, d \geq \mu \|d\|^2 \quad \text{for all } d \in \mathbb{R}^n.$$

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

$$\begin{aligned} f(x^* + d) &\geq f(x^*) + \frac{1}{2} d^\mathsf{T} f''(x^*)\, d - \varepsilon \|d\|^2 \\ &\geq f(x^*) + \frac{\mu}{2} \|d\|^2 - \varepsilon \|d\|^2 \\ &= f(x^*) + \frac{\beta}{2} \|d\|^2 \end{aligned}$$

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

## § 4 Minimization of Quadratic Functions

In this section we consider the simplest reasonable class of unconstrained optimization problems, namely the minimization of quadratic polynomials:

$$\text{Minimize} \quad \phi(x) := \frac{1}{2} x^\mathsf{T} A\, x - b^\mathsf{T} x + c \quad \text{where } x \in \mathbb{R}^n. \tag{4.1}$$

The data of the problem is $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ und $c \in \mathbb{R}$. We can assume w.l.o.g. that $A$ is symmetric.
**Quiz 4.1:** Why?

Being able to solve (4.1) is an essential building block for subsequent tasks.

**Lemma 4.1** (Solvability and global solutions of (4.1)[3]). *Suppose that $A \in \mathbb{R}^{n \times n}$ is symmetric, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then the following holds:*

(*i*) *If $A$ is positive semidefinite, then the objective in (4.1) is convex. In this case, the following are equivalent:*

[3] compare Nocedal, Wright, 2006, Lemma 4.7

(*i*) *The problem* (4.1) *possess at least one (global) minimizer.*

(*ii*) *The objective* $\phi$ *is bounded below.*

(*iii*) $Ax = b$ *is solvable.*

*The global minimizers of* (4.1) *are precisely the solutions of the linear system* $Ax = b$.

(*ii*) *In case* $A$ *is not positive semidefinite*[4]*, the objective* $\phi$ *is not bounded below, thus problem* (4.1) *is unbounded.*

*Proof.* □

**Corollary 4.2** (Unique solvability of (4.1))**.** *Problem* (4.1) *possesses a unique (global) solution* $x^*$ *if and only if* $A$ *is s. p. d. In this case,* $x^* = A^{-1}b$*, and the optimal value is*

$$\phi(x^*) = c - \frac{1}{2}\|x^*\|_A^2 = c - \frac{1}{2}\|A^{-1}b\|_A^2 = c - \frac{1}{2}\|b\|_{A^{-1}}^2.$$

We will assume for the remainder of § 4 that $A$ is symmetric and positive definite (s. p. d.). Hence, the solution of (4.1) is equivalent to the solution of the linear system $Ax = b$. We denote that solution by $x^* = A^{-1}b$. Of course, we could be using a **direct solver**, such as Gaussian elimination, which computes an LU decomposition of $A$, or rather its s. p. d. variant without pivoting, which computes the Cholesky decomposition $A = LL^\mathsf{T}$ with the lower triangular matrix $L$.[5] However, when the problem is high-dimensional (such as $n \geq 10\,000$), then the generic $\sim n^3$ effort for solving the linear system becomes prohibitive. Even when $A$ is sparse, as is often the case for high-dimensional problems, and a direct solver which exploits this is used[6], this is no longer feasible for very high dimension $n$.

This is where **iterative solvers** for linear systems come into play. They do not solve the problem at once, but rather generate a sequence $\left(x^{(k)}\right)$ which converges to the solution. Beyond the ability to deal with very high-dimensional problems, iterative solvers have another advantage: Any iterate $x^{(k)}$ of the method can be viewed as an approximate solution of $Ax = b$ (or an approximate solution of (4.1)), and we can stop the iteration as soon as the desired tolerance is reached, when the time budget is used up, or when something unexpected happens, e. g., $A$ turns out not to be positive definite after all. Recall that direct solvers do not yield any usable approximate solutions of the system while they are running; they have to carry through to the end, and only then return a solution, which is exact up to the influence of floating-point error. Iterative solvers have the additional advantage that they do not require access to the matrix $A$ entry by entry. Rather they only require matrix-vector products, i. e., a function which evaluates $x \mapsto Ax$. **Quiz 4.2:** Can you think of an example where matrix-vector products are available, but you typically don't have access to the entries of the underlying matrix?

[4] The matrix $A$ possesses at least one negative eigenvalue.

[5] We assume you have seen these methods, e. g., in the class *Einführung in die Numerik*.

[6] such as a sparse Cholesky decomposition

Our objective $\phi$ from (4.1) satisfies

$$\phi(x) = \frac{1}{2}x^\mathsf{T}A\,x - b^\mathsf{T}x + c$$
$$\nabla\phi(x) = A\,x - b =: r.$$

We call $r = \nabla\phi(x)$ the **residual** of the linear system $A\,x = b$ at $x$.[7] Independently of any method we might be using to solve $A\,x = b$ (or minimize $\phi$), we have the following relation between the values of the objective, the **error** $x - x^*$ at a point $x$, and the residual at $x$:

**Lemma 4.3.** *We have*

$$\phi(x) - \phi(x^*) = \frac{1}{2}\|x - x^*\|_A^2 = \frac{1}{2}\|r\|_{A^{-1}}^2 = \frac{1}{2}\|\nabla\phi(x)\|_{A^{-1}}^2. \tag{4.2}$$

*Proof.* Direct calculation shows

$$\begin{aligned}
\phi(x) - \phi(x^*) &= \frac{1}{2}x^\mathsf{T}A\,x - b^\mathsf{T}x + c - \frac{1}{2}(x^*)^\mathsf{T}A\,x^* + b^\mathsf{T}x^* - c \\
&= \frac{1}{2}x^\mathsf{T}A\,x - (x^*)^\mathsf{T}A\,x - \frac{1}{2}(x^*)^\mathsf{T}A\,x^* + (x^*)^\mathsf{T}A\,x^* \quad \text{since } b = A\,x^* \\
&= \frac{1}{2}x^\mathsf{T}A\,x - (x^*)^\mathsf{T}A\,x + \frac{1}{2}(x^*)^\mathsf{T}A\,x^* \\
&= \frac{1}{2}\|x - x^*\|_A^2 \\
&= \frac{1}{2}(x - x^*)^\mathsf{T}r = \frac{1}{2}r^\mathsf{T}A^{-1}r \qquad \text{since } r = A\,(x - x^*) \\
&= \frac{1}{2}\|r\|_{A^{-1}}^2 \\
&= \frac{1}{2}\|\nabla\phi(x)\|_{A^{-1}}^2.
\end{aligned}$$

□

We will discuss in the remainder of this section two different iterative methods for the solution of (4.1), and equivalently the solution of the linear system $A\,x = b$, where $A$ is s. p. d.[8] These methods are the **gradient descent method** (also known as **steepest descent method**), and the **conjugate gradient method**.

We begin with the gradient descent method, which is based on the following simple
**Idea:** from the current iterate $x^{(k)}$, move a bit along the direction of steepest descent of the objective, and take the point reached as the next iterate $x^{(k+1)}$.

[7] Sometimes the residual is defined in the literature with opposite sign.
[8] You can learn more about iterative solvers for more general linear systems (not related to optimization) in the class *Numerische lineare Algebra*.

## § 4.1 Direction of Steepest Descent

We first need to clarify what the **direction of steepest descent** of a function $f\colon \mathbb{R}^n \to \mathbb{R}$ at a point $x$ is. By definition, the direction of steepest descent minimizes the directional derivative $f'(x)\,d$ over all vectors $d \in \mathbb{R}^n$ of constant length. What we mean by "length" is defined through the inner product $M$ in use:

$$\begin{aligned} &\text{Minimize} && f'(x)\,d \quad \text{where } d \in \mathbb{R}^n \\ &\text{subject to} && \|d\|_M = 1. \end{aligned} \tag{4.3}$$

We note that we could be considering the equivalent problem

$$\begin{aligned} &\text{Minimize} && f'(x)\,d \quad \text{where } d \in \mathbb{R}^n \\ &\text{subject to} && \|d\|_M \leq 1. \end{aligned} \tag{4.4}$$

The normalization to unit length is, by the way, arbitrary.

Problems (4.3), (4.4) are constrained problems, but we can solve them without an elaborated theory. We rewrite the objective so that the directional derivative is expressed using the $M$-inner product[9]

$$f'(x)\,d = \nabla f(x)^{\mathsf{T}} d = \nabla f(x)^{\mathsf{T}} M^{-1} M\,d = \big(M^{-1}\nabla f(x)\big)^{\mathsf{T}} M\,d,$$

where we used the symmetry of $M$ (actually of $M^{-1}$) in the last step. The Cauchy-Schwarz inequality w.r.t. the $M$-inner product shows that this expression is minimal precisely when $d$ is antiparallel to $M^{-1}\nabla f(x)$.

We summarize these findings:

**Definition 4.4** ($M$-gradient, direction of steepest descent w.r.t. the $M$-inner product)**.**
*Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$ and that $f'(x) \neq 0$ holds.*

(*i*) *The vector*

$$\nabla_M f(x) := M^{-1}\nabla f(x) \tag{4.5}$$

*is termed the **gradient of $f$ at $x$ w.r.t. the $M$-inner product** or briefly: the **$M$-gradient**.*

(*ii*) *The vector $-\nabla_M f(x)$ and all of its positive multiples are termed a **direction of steepest descent of $f$ at $x$ w.r.t. the $M$-inner product**.*

We evaluate the negative $M$-gradient (direction of steepest descent) by solving the linear system

$$M\,d^* = -\nabla f(x). \tag{4.6}$$

When using the Euclidean inner product ($M = \mathrm{Id}$), we continue to write $\nabla f(x)$ instead of $\nabla_{\mathrm{Id}} f(x)$. Sometimes, the use of $\nabla_M f(x)$ instead of the Euclidean gradient direction $\nabla f(x)$ is referred to as **preconditioning**.

[9] In case this means something to you, we determine the Riesz representer of $f'(x)$ w.r.t. the $M$-inner product.

## § 4.2 Gradient Descent Method with Cauchy Step Sizes

The direction of steepest descent at $x$ used by the gradient method is thus[10]

$$d = -\nabla_M \phi(x) = -M^{-1} r.$$

Now that the choice of direction is clear, let us analyze the choice of the step length. We have the following expression for the difference of function values before and after a step:

$$\begin{aligned}\phi(x + \alpha\, d) - \phi(x) &= \frac{1}{2}(x + \alpha\, d)^\mathsf{T} A\,(x + \alpha\, d) - b^\mathsf{T}(x + \alpha\, d) + c - \frac{1}{2} x^\mathsf{T} A\, x + b^\mathsf{T} x - c \\ &= \frac{1}{2}(d^\mathsf{T} A\, d)\,\alpha^2 + (A\, x - b)^\mathsf{T} d\,\alpha \\ &= \frac{1}{2}(d^\mathsf{T} A\, d)\,\alpha^2 + (r^\mathsf{T} d)\,\alpha. \end{aligned} \tag{4.7}$$

This expression holds for arbitrary directions $d$ and step sizes $\alpha$.

When $d \neq 0$, then the one-dimensional quadratic polynomial $\alpha \mapsto \phi(x + \alpha\, d)$ is strongly convex. It is therefore an obvious idea to choose $\alpha$ such that $\phi(x + \alpha\, d)$ is minimized. According to (4.7), we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}\phi(x + \alpha\, d) &= (d^\mathsf{T} A\, d)\,\alpha + r^\mathsf{T} d, \\ \frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\phi(x + \alpha\, d) &= d^\mathsf{T} A\, d > 0. \end{aligned}$$

Due to the positivity of the second derivative, the second-order sufficient condition (Theorem 3.3) is satisfied when $\frac{\mathrm{d}}{\mathrm{d}\alpha}\phi(x + \alpha\, d) = 0$, which amounts to

$$\alpha^* = -\frac{r^\mathsf{T} d}{d^\mathsf{T} A\, d}. \tag{4.8}$$

This "optimal" step size is also known as the **Cauchy step size**. For this choice, the difference of function values (4.7) before and after a step becomes

$$\begin{aligned}\phi(x + \alpha^*\, d) - \phi(x) &= \frac{1}{2}(d^\mathsf{T} A\, d)\,(\alpha^*)^2 + (r^\mathsf{T} d)\,\alpha^* \\ &= \frac{1}{2}(d^\mathsf{T} A\, d)\left(\frac{r^\mathsf{T} d}{d^\mathsf{T} A\, d}\right)^2 - (r^\mathsf{T} d)\,\frac{r^\mathsf{T} d}{d^\mathsf{T} A\, d} \\ &= -\frac{1}{2}\frac{(r^\mathsf{T} d)^2}{d^\mathsf{T} A\, d}. \end{aligned} \tag{4.9}$$

**Note:** This formula holds for arbitrary directions $d \neq 0$ but it uses the Cauchy step size $\alpha^*$.

We can now state the steepest descent method w.r.t. the $M$-inner product and the Cauchy step size (4.8) for the iterative solution of the unconstrained quadratic minimization problem (4.1) with s. p. d. $A$. This method, with $M = \mathrm{Id}$, was already published by Cauchy, 1847.

[10] We avoid iteration indices for now in order to avoid cluttered notation.

**Algorithm 4.5** (Gradient descent method for (4.1) w.r.t. the $M$-inner product with Cauchy step size)**.**
***Input:*** *initial guess* $x^{(0)} \in \mathbb{R}^n$
***Input:*** *right-hand side* $b \in \mathbb{R}^n$
***Input:*** *s. p. d. matrix* $A$ *(or matrix-vector products with* $A$*)*
***Input:*** *s. p. d. matrix* $M$ *(or matrix-vector products with* $M^{-1}$*)*
***Output:*** *approximate solution of* (4.1)*, i. e., of* $Ax = b$

1: Set $k := 0$
2: Set $r^{(0)} := Ax^{(0)} - b$ // Evaluate the initial residual
3: Set $d^{(0)} := -M^{-1}r^{(0)}$ // Evaluate the initial negative $M$-gradient
4: Set $\delta^{(0)} := -(r^{(0)})^\intercal d^{(0)}$ // $\delta^{(0)} = \|\nabla_M \phi(x^{(0)})\|_M^2 = \|r^{(0)}\|_{M^{-1}}$
5: ***while*** *stopping criterion not met* ***do***
6: Set $q^{(k)} := Ad^{(k)}$
7: Set $\theta^{(k)} := (q^{(k)})^\intercal d^{(k)}$
8: Set $\alpha^{(k)} := \delta^{(k)}/\theta^{(k)}$ // Evaluate the Cauchy step size
9: Set $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$ // Update the iterate
10: Set $r^{(k+1)} := r^{(k)} + \alpha^{(k)} q^{(k)}$ // Update the residual
11: Set $d^{(k+1)} := -M^{-1}r^{(k+1)}$ // Evaluate the new negative $M$-gradient
12: Set $\delta^{(k+1)} := -(r^{(k+1)})^\intercal d^{(k+1)}$ // $\delta^{(k+1)} = \|\nabla_M \phi(x^{(k+1)})\|_M^2 = \|r^{(k+1)}\|_{M^{-1}}$
13: Set $k := k + 1$
14: ***end while***
15: ***return*** $x^{(k)}$

The following can be said about Algorithm 4.5.

**Remark 4.6** (on Algorithm 4.5)**.**

(*i*) *Algorithm 4.5 is an iterative solver for the unconstrained quadratic minimization problem* (4.1) *with s. p. d.* $A$*, and simultaneously an iterative solver for the linear system* $Ax = b$*.*

(*ii*) *We do not require access to the matrix* $A$ *entry by entry, matrix-vector products with* $A$ *are enough.*

(*iii*) *The user gets to choose the inner product* $M$*. This is known as* ***preconditioning****, and therefore Algorithm 4.5 is often termed a* ***preconditioned gradient descent method****. The case* $M = \mathrm{Id}$ *corresponds to the classical gradient descent method (without preconditioning).*

(*iv*) *We do not require access to the inner product matrix* $M$ *entry by entry, matrix-vector products with* $M^{-1}$ *(i. e., solutions of linear systems with* $M$*) are enough.*

(*v*) *Algorithm 4.5 requires the storage of four vectors, which are iteratively overwritten: iterates* $x^{(k)}$*, residuals* $r^{(k)}$*, negative gradient directions* $d^{(k)}$*, and vectors* $q^{(k)} = Ad^{(k)}$*.*

(*vi*) *Every iteration requires one matrix-vector product with* $A$ *and one application of the preconditioner, i. e., one matrix-vector product with* $M^{-1}$*.*

(*vii*) *In order to mitigate the accumulation of round-off error, it is advisable to evaluate the residual every, say, 50 iterations according to* $r^{(k)} := Ax^{(k)} - b$*, rather than update it.*

(*viii*) *The Cauchy step sizes satisfy*

$$0 < \lambda_{\min}(A;M) \le \frac{1}{\alpha^{(k)}} = \frac{(d^{(k)})^\mathsf{T} A\, d^{(k)}}{(d^{(k)})^\mathsf{T} M\, d^{(k)}} \le \lambda_{\max}(A;M), \tag{4.10}$$

*as long as $d^{(k)} \neq 0$ holds, i. e., as long as $x^{(k)} \neq x^*$. Consequently, the Cauchy step sizes generated can be used to obtain estimates on the eigenvalues of $A$ w.r.t. $M$.*

(*ix*) *When Algorithm 4.5 is provided with the value of $c$, the following recursion can be added to the algorithm to keep track of the value of the objective:*

$$\phi(x^{(0)}) = c + \frac{1}{2}(r^{(0)} - b)^\mathsf{T}(x^{(0)}) \quad \textit{initialization} \tag{4.11a}$$

$$\phi(x^{(k+1)}) = \phi(x^{(k)}) - \frac{1}{2}\,\alpha^{(k)}\,\delta^{(k)} \quad \textit{update.} \tag{4.11b}$$

*This does not incur noticeable computational overhead and does not require the storage of extra vectors. Alternatively, the value of $\phi(x^{(0)})$ can be provided.*

We now seek to estimate the speed of convergence of Algorithm 4.5. The function values at the iterates satisfy

$$\begin{aligned}
&\phi(x^{(k+1)}) - \phi(x^*) \\
&\quad = \frac{1}{2}\|r^{(k+1)}\|^2_{A^{-1}} && \text{by (4.2)} \\
&\quad = \frac{1}{2}\|r^{(k)} + \alpha^{(k)} A\, d^{(k)}\|^2_{A^{-1}} \\
&\quad = \frac{1}{2}\|r^{(k)}\|^2_{A^{-1}} + \alpha^{(k)}\,(r^{(k)})^\mathsf{T} d^{(k)} + \frac{1}{2}\big[\alpha^{(k)}\big]^2\,(d^{(k)})^\mathsf{T} A\, d^{(k)}.
\end{aligned}$$

This formula so far holds for any choice of step size $\alpha^{(k)}$ and any choice of direction $d^{(k)}$. We now insert the Cauchy step size $\alpha^{(k)} = -\frac{(r^{(k)})^\mathsf{T} d^{(k)}}{(d^{(k)})^\mathsf{T} A\, d^{(k)}}$ and obtain

$$\begin{aligned}
&= \frac{1}{2}\|r^{(k)}\|^2_{A^{-1}} - \frac{\big[(r^{(k)})^\mathsf{T} d^{(k)}\big]^2}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} + \frac{1}{2}\frac{\big[(r^{(k)})^\mathsf{T} d^{(k)}\big]^2}{(d^{(k)})^\mathsf{T} A\, d^{(k)}} \\
&= \left(1 - \frac{\big[(r^{(k)})^\mathsf{T} d^{(k)}\big]^2}{\big[(d^{(k)})^\mathsf{T} A\, d^{(k)}\big]\,\big[(r^{(k)})^\mathsf{T} A^{-1} r^{(k)}\big]}\right)\big(\phi(x^{(k)}) - \phi(x^*)\big) && \text{by (4.2).}
\end{aligned}$$

The directions $d^{(k)}$ are still arbitrary. Inserting the relationship $d^{(k)} = -M^{-1} r^{(k)} = -\nabla_M \phi(x^{(k)})$ characteristic for gradient descent, in the form $r^{(k)} = -M\, d^{(k)}$, we obtain

$$= \left(1 - \frac{\big[(d^{(k)})^\mathsf{T} M\, d^{(k)}\big]^2}{\big[(d^{(k)})^\mathsf{T} A\, d^{(k)}\big]\,\big[(d^{(k)})^\mathsf{T} M A^{-1} M\, d^{(k)}\big]}\right)\big(\phi(x^{(k)}) - \phi(x^*)\big).$$

The fraction is precisely the type of expression estimated by the generalized Kantorovich inequality (2.14). This yields

$$\begin{aligned}
&\phi(x^{(k+1)}) - \phi(x^*) \\
&\quad \leq \left(1 - \frac{4\,\alpha\,\beta}{(\alpha+\beta)^2}\right)\left(\phi(x^{(k)}) - \phi(x^*)\right) \\
&\quad = \left(\frac{\beta-\alpha}{\beta+\alpha}\right)^2\left(\phi(x^{(k)}) - \phi(x^*)\right) \\
&\quad = \left(\frac{\kappa-1}{\kappa+1}\right)^2\left(\phi(x^{(k)}) - \phi(x^*)\right) \qquad \text{since } \kappa = \beta/\alpha.
\end{aligned}$$

We have thus shown the following classical convergence result for Algorithm 4.5:

**Theorem 4.7** (Convergence of Algorithm 4.5)**.** *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 for any choice of the initial guess $x^{(0)}$, the gradient descent method with Cauchy step converges to the unique solution $x^* = A^{-1}b$ of (4.1). In terms of the generalized condition number $\kappa = \beta/\alpha$, we have the estimates*

$$\phi(x^{(k+1)}) - \phi(x^*) \leq \left(\frac{\kappa-1}{\kappa+1}\right)^2\left(\phi(x^{(k)}) - \phi(x^*)\right) \tag{4.12a}$$

$$\|x^{(k+1)} - x^*\|_A \leq \left(\frac{\kappa-1}{\kappa+1}\right)\|x^{(k)} - x^*\|_A \tag{4.12b}$$

*and consequently*

$$\phi(x^{(k)}) - \phi(x^*) \leq \left(\frac{\kappa-1}{\kappa+1}\right)^{2k}\left(\phi(x^{(0)}) - \phi(x^*)\right) \tag{4.12c}$$

$$\|x^{(k)} - x^*\|_A \leq \left(\frac{\kappa-1}{\kappa+1}\right)^{k}\|x^{(0)} - x^*\|_A. \tag{4.12d}$$

*Moreover, the objective values $\phi(x^{(k)})$ and thus the norm of the error $\|x^{(k)} - x^*\|_A$ are monotonically decreasing.*

As an immediate consequence of this theorem, we can estimate the maximal number of iterations required until the left-hand terms in (4.12c) and (4.12d) have been decreased relative to their initial values.

**Corollary 4.8** (Maximal number of iterations required in Algorithm 4.5)**.** *Given positive numbers $\varepsilon_1$ and $\varepsilon_2$, it takes*

$$k \leq \left\lceil \frac{\kappa}{4}\ln\left(\frac{1}{\varepsilon_1}\right)\right\rceil \text{ iterations until } \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \leq \varepsilon_1,$$

$$k \leq \left\lceil \frac{\kappa}{2}\ln\left(\frac{1}{\varepsilon_2}\right)\right\rceil \text{ iterations until } \left(\frac{\kappa-1}{\kappa+1}\right)^{k} \leq \varepsilon_2.$$

*Proof.* (1) We first show that

$$-\ln\left(\frac{\kappa-1}{\kappa+1}\right) \geq \frac{2}{\kappa} > 0$$

holds for all $\kappa > 1$. At the limit $\kappa \to \infty$, both expressions converge to 0. We now show that

$$\frac{\mathrm{d}}{\mathrm{d}\kappa}\left[-\ln\left(\frac{\kappa-1}{\kappa+1}\right)\right] \geq \frac{\mathrm{d}}{\mathrm{d}\kappa}\frac{2}{\kappa}$$

holds for all $\kappa > 1$, which proves the claim. The derivative on the left is $\frac{-2}{\kappa-1}$, while the derivative on the right is $\frac{-2}{\kappa^2}$. In view of $0 < \kappa^2 < \kappa - 1$ for all $\kappa > 1$,

(2) Taking the reciprocal of the inequality shown above, we obtain

$$0 < \frac{-1}{\ln\left(\frac{\kappa-1}{\kappa+1}\right)} \leq \frac{\kappa}{2} \tag{$*$}$$

for all $\kappa > 1$.

(3) Given $\kappa > 1$, we easily infer that $\left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \leq \varepsilon_1$ holds if and only if

$$k \geq \frac{1}{2}\frac{-\ln\varepsilon_1}{-\ln\left(\frac{\kappa-1}{\kappa+1}\right)} = \frac{1}{2}\frac{-1}{\ln\left(\frac{\kappa-1}{\kappa+1}\right)}\ln\left(\frac{1}{\varepsilon_1}\right). \tag{$**$}$$

In view of the inequality ($*$) shown above, we obtain that

$$k \geq \left\lceil \frac{\kappa}{4}\ln\left(\frac{1}{\varepsilon_1}\right)\right\rceil \geq \frac{\kappa}{4}\ln\left(\frac{1}{\varepsilon_1}\right)$$

implies ($**$), which proves the first claim.

The second claim follows similarly.

□

**Remark 4.9** (on Theorem 4.7)**.**

(*i*) ($**$) *shows the Q-linear convergence of* $\left(x^{(k)}\right)$ *to the solution* $x^*$ *in the A-norm.*

(*ii*) *The contraction factor is* $0 \leq \frac{\kappa-1}{\kappa+1} < 1$*, i. e., the convergence estimate depends on the ratio* $\kappa$ *between the largest and the smallest generalized eigenvalue of A w.r.t. M. It is the purpose of the preconditioner/inner product M to keep this ratio small.*

(*iii*) *In the extreme case* $\kappa = 1$ *we obtain convergence in one step. This happens precisely when M is a multiple of A. However, we need a solve a linear system with M in every iteration. If we were able to do that, we might as well solve* $Ax = b$ *directly.*

(*iv*) *A good preconditioner is a compromise between a moderate generalized condition number* $\kappa$ *and the effort in applying* $M^{-1}$*. Finding a good preconditioner generally requires knowledge about the problem at hand.*

(v) *It is natural to measure convergence of the method in the A-norm of the error because, due to (4.2), that is the quantity being minimized.*

(vi) *The estimates of Theorem 4.7 are worst-case estimates since they do not depend on the initial guess $x^{(0)}$. In fact, as can be seen in Figure 4.1c, the actual contraction factor for the objective values can be significantly smaller than the estimate (∗) for some initial guesses.*

Figure 4.1 illustrates the convergence behavior of Algorithm 4.5 for a 2-dimensional example problem from a number of different initial guesses $x^{(0)}$. We observe the typical "zig-zagging" behavior of the iterates as they converge to the solution. This happens for any initial guess, except when $x^{(0)} - x^*$ happens to be an eigenvector of $M^{-1}A$, in which case convergence occurs in one step due to $x^{(1)} = x^*$. (Such a case is not shown in Figure 4.1). **Quiz 4.3:** Suppose $A$, $b$ and $M$ are given and you consider a random distribution of initial values $x^{(0)}$ in $\mathbb{R}^n$, which has a probability density. What is the probability of hitting an initial value such that convergence happens in one step?

The zig-zagging behavior of the iterates $x^{(k)}$, as well as the non-monotone behavior of $\|r^{(k)}\|_{M^{-1}}$ have been analyzed in detail in the literature; see for instance Akaike, 1959; Forsythe, 1968; Nocedal, Sartenaer, Zhu, 2002. Essentially what happens is that, asymptotically, the error $x^{(k)} - x^*$ alternates between elements of the eigenspaces belonging to the smallest and the largest eigenvalues of $A$ w.r.t. $M$. This is ultimately a consequence of the fact that gradient descent is a memoryless method.

It has also been shown that a necessary condition in order for the norm of the gradient $\|r^{(k)}\|_{M^{-1}}$ to converge non-monotonically is that the condition number satisfy $\kappa > 3 + 2\sqrt{2} \approx 5.83$.

It remains to discuss stopping criteria. Several quantities may be of interest in this respect:

(i) Are we happy with a point $x^{(k)}$ which is almost stationary, i. e., where $\|r^{(k)}\|_{M^{-1}}$ is small?

(ii) Are we happy with a point $x^{(k)}$ whose objective value is near the optimal value, i. e., where $\phi(x^{(k)}) - \phi(x^*)$ is small, or equivalently, where $\|x^{(k)} - x^*\|_A$ is small?

(iii) Are we happy with a point $x^{(k)}$ whose distance from the minimizer is small in the preconditioner-induced norm $M$, i. e., where $\|x^{(k)} - x^*\|_M$ is small?

The only of these three quantities which we can evaluate without knowing $x^*$ or $\phi(x^*)$ is $\delta^{(k)} = \|r^{(k)}\|^2_{M^{-1}}$. Therefore, many implementations use one of the following combinations of a relative and an absolute criterion based on $\|r^{(k)}\|_{M^{-1}}$:

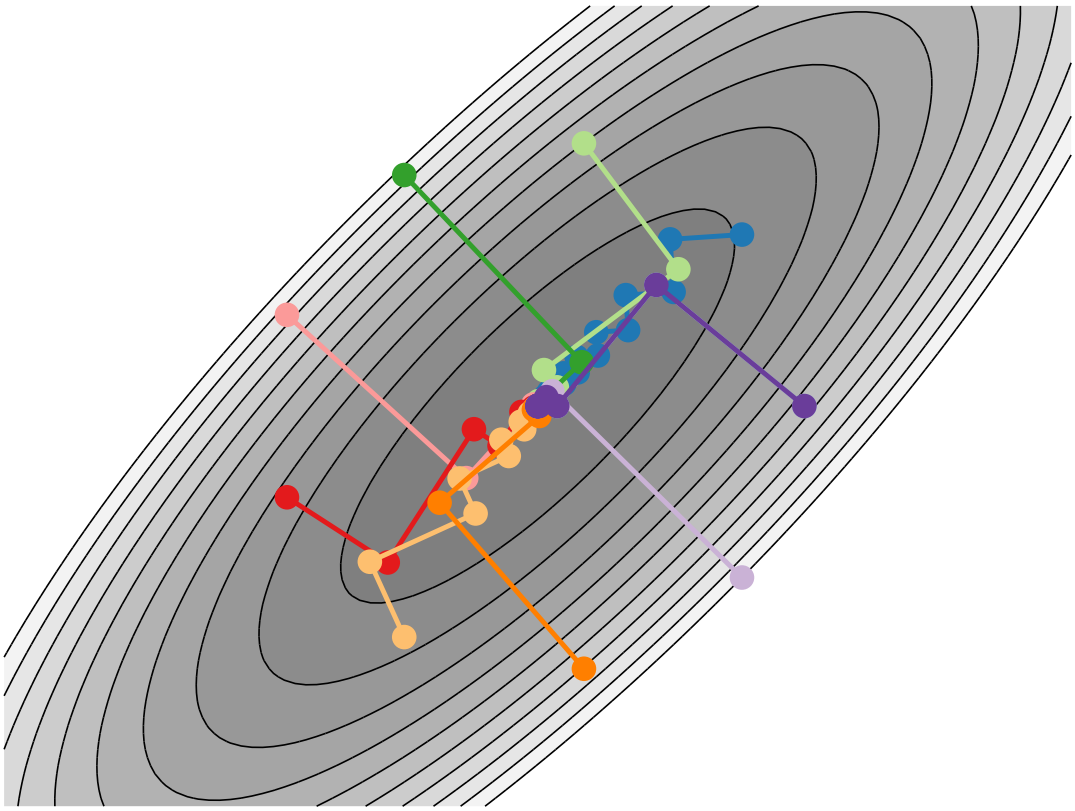
$$\|r_k\|_{M^{-1}} \le \varepsilon_{\text{rel}} \|r_0\|_{M^{-1}}, \qquad \text{i. e., } \delta_k \le \varepsilon^2_{\text{rel}} \delta_0, \tag{4.13a}$$

$$\|r_k\|_{M^{-1}} \le \varepsilon_{\text{abs}}, \qquad \text{i. e., } \delta_k \le \varepsilon^2_{\text{abs}}, \tag{4.13b}$$

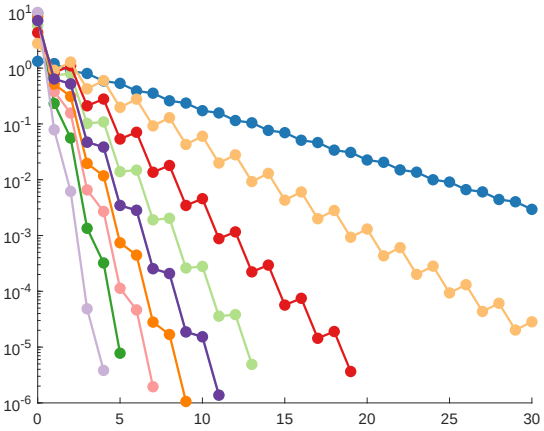
$$\|r_k\|_{M^{-1}} \le \varepsilon_{\text{rel}} \|r_0\|_{M^{-1}} + \varepsilon_{\text{abs}}, \qquad \text{i. e., } \delta_k^{1/2} \le \varepsilon_{\text{rel}} \delta_0^{1/2} + \varepsilon_{\text{abs}}, \tag{4.13c}$$

$$\|r_k\|_{M^{-1}} \le \max\{\varepsilon_{\text{rel}} \|r_0\|_{M^{-1}},\ \varepsilon_{\text{abs}}\}, \qquad \text{i. e., } \delta_k \le \max\{\varepsilon^2_{\text{rel}} \delta_0,\ \varepsilon^2_{\text{abs}}\}. \tag{4.13d}$$

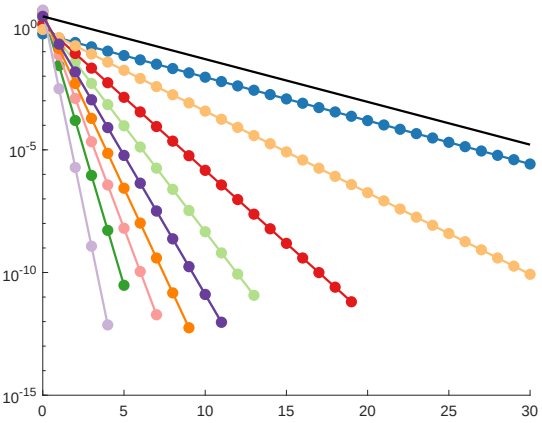
Let us see which consequences either of the implementable stopping criteria (4.13) has on the other two quantities of interest:



(a) Iterates $\left(x^{(k)}\right)$ of the method. Each color corresponds to a different initial guess $x^{(0)}$.




(b) The norm of the gradient $\sqrt{\delta^{(k)}} = \|\nabla_M \phi(x^{(k)})\|_{M^{-1}} = \|r^{(k)}\|_{M^{-1}}$ does not necessarily converge monotonically.




(c) The objective values $\phi(x^{(k)}) - \phi(x^*)$ converge monotonically. The black line illustrates the bound (∗).

Figure 4.1: Illustration of the convergence behavior of Algorithm 4.5 from a number of initial guesses $x^{(0)}$. No preconditioning ($M = \mathrm{Id}$) is used. The two eigenvalues of the matrix are $\alpha = 1$ and $\beta = 10$ so the condition number is $\kappa = 10$.

**Lemma 4.10** (Implications). *The criteria from* (4.13) *imply, respectively,*

$$\left.\begin{aligned} \|x_k - x^*\|_A &\le \sqrt{\kappa}\,\varepsilon_{\text{rel}}\,\|x_0 - x^*\|_A \\ \|x_k - x^*\|_M &\le \sqrt{\kappa}\,\varepsilon_{\text{rel}}\,\|x_0 - x^*\|_M \end{aligned}\right\} \tag{4.14a}$$

$$\left.\begin{aligned} \|x_k - x^*\|_A &\le (1/\sqrt{\alpha})\,\varepsilon_{\text{abs}} \\ \|x_k - x^*\|_M &\le \sqrt{\beta}\,\varepsilon_{\text{abs}} \end{aligned}\right\} \tag{4.14b}$$

$$\left.\begin{aligned} \|x_k - x^*\|_A &\le \sqrt{\kappa}\,\varepsilon_{\text{rel}}\,\|x_0 - x^*\|_A + (1/\sqrt{\alpha})\,\varepsilon_{\text{abs}} \\ \|x_k - x^*\|_M &\le \sqrt{\kappa}\,\varepsilon_{\text{rel}}\,\|x_0 - x^*\|_M + \sqrt{\beta}\,\varepsilon_{\text{abs}} \end{aligned}\right\} \tag{4.14c}$$

$$\left.\begin{aligned} \|x_k - x^*\|_A &\le \max\left\{\sqrt{\kappa}\,\varepsilon_{\text{rel}}\,\|x_0 - x^*\|_A,\ (1/\sqrt{\alpha})\,\varepsilon_{\text{abs}}\right\} \\ \|x_k - x^*\|_M &\le \max\left\{\sqrt{\kappa}\,\varepsilon_{\text{rel}}\,\|x_0 - x^*\|_M,\ \sqrt{\beta}\,\varepsilon_{\text{abs}}\right\} \end{aligned}\right\} \tag{4.14d}$$

*Proof.* The proof is part of homework problem 1.3. □

## § 4.3 Gradient Descent Method with Constant Step Sizes

We can show that the gradient descent method continues to converge Q-linearly when, in place of the Cauchy step sizes, we choose constant step sizes $\alpha^{(k)} \equiv \overline{\alpha}$ within a certain range. We obtain as above

$$\begin{aligned} &\phi(x^{(k+1)}) - \phi(x^*) \\ &\quad = \frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 + \overline{\alpha}\,(r^{(k)})^{\mathsf{T}} d^{(k)} + \frac{1}{2}\overline{\alpha}^2 (d^{(k)})^{\mathsf{T}} A\, d^{(k)}. \end{aligned}$$

We leave $\overline{\alpha}$ open for now and insert the gradient descent relation $r^{(k)} = -M\,d^{(k)}$ to obtain

$$\begin{aligned} &= \frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 - \overline{\alpha}\,(d^{(k)})^{\mathsf{T}} M\, d^{(k)} + \frac{1}{2}\overline{\alpha}^2 (d^{(k)})^{\mathsf{T}} A\, d^{(k)} \\ &\le \frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 - \overline{\alpha}\,(d^{(k)})^{\mathsf{T}} M\, d^{(k)} + \frac{1}{2}\overline{\alpha}^2 \beta\,(d^{(k)})^{\mathsf{T}} M\, d^{(k)} \quad \text{since } d^{\mathsf{T}} A\, d \le \beta\, d^{\mathsf{T}} M\, d \\ &= \frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 + \overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right)(d^{(k)})^{\mathsf{T}} M\, d^{(k)}. \end{aligned}$$

Here we need to convert the last term into $d^{\mathsf{T}} M A^{-1} M\, d$, which is equal to $r^{\mathsf{T}} A^{-1} r$, so that it can be combined with the first term. We require that the coefficient $\overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right)$ is negative to obtain convergence. Consequently, we use the first estimate in (2.10a):

$$\begin{aligned} &\le \frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 + \overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right)\alpha\,(d^{(k)})^{\mathsf{T}} M A^{-1} M\, d^{(k)} \quad \text{provided that } \overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right) < 0 \\ &= \left[1 + 2\,\overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right)\alpha\right]\frac{1}{2}\|r^{(k)}\|_{A^{-1}}^2 \\ &= \left[1 + 2\,\overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right)\alpha\right]\left(\phi(x^{(k)}) - \phi(x^*)\right). \end{aligned}$$

The condition that $\overline{\alpha}\left(\frac{1}{2}\overline{\alpha}\,\beta - 1\right)$ is negative amounts to $\overline{\alpha} \in (0, \frac{2}{\beta})$. It is precisely the midpoint $\overline{\alpha} = 1/\beta$ of this interval which minimizes this term and yields the optimal estimate, and the expression in $[\cdots]$ becomes $\frac{\kappa-1}{\kappa}$ in this case.

**Remark 4.11** (on the convergence of Algorithm 4.5 with constant step sizes)**.**

(*i*) *We have shown that Algorithm 4.5, where Step 8 is replaced by* $\alpha^{(k)} := \overline{\alpha}$*, still converges, provided that* $\overline{\alpha} \in (0, \frac{2}{\beta})$.

(*ii*) *From a practical perspective, we therefore need to know at least an upper bound for the largest eigenvalue* $\beta$ *of the generalized eigenvalue problem* $A\,x = \lambda\, M\, x$*. When we have* $\beta \le \beta_{\text{estimate}}$ *and choose* $\overline{\alpha} \in (0, \frac{2}{\beta_{\text{estimate}}})$*, we also have* $\overline{\alpha} \in (0, \frac{2}{\beta})$.

(*iii*) *The choice* $\overline{\alpha} = \frac{1}{\beta}$ *yields the optimal estimate. In this case, we obtain*

$$\phi(x^{(k+1)}) - \phi(x^*) \le \left(\frac{\kappa - 1}{\kappa}\right) \left(\phi(x^{(k)}) - \phi(x^*)\right).$$

*Since for all* $\kappa \ge 1$*, we have* $\left(\frac{\kappa-1}{\kappa+1}\right)^2 \le \frac{\kappa-1}{\kappa}$*, the contraction factor in the bound we obtained with constant step sizes is worse than the one for the Cauchy step sizes; see* (∗).

(*iv*) *The Kantorovich inequality was not needed in the proof.*

Figure 4.2 illustrates the convergence behavior of Algorithm 4.5 with constant step sizes for a 2-dimensional example problem from a number of different initial guesses $x^{(0)}$.

## § 4.4 Gradient Descent Method with Other Step Size Rules

Step size rules other than the Cauchy step sizes and constant step sizes have been proposed and analyzed in the literature with the goal of breaking the non-efficient zig-zaggging pattern; among them Barzilai, Borwein, 1988; De Asmundis, di Serafino, Riccio, et al., 2013; De Asmundis, di Serafino, Hager, et al., 2014; Gonzaga, Schneider, 2015. We do not go into the details here but mention one remarkable result from Gonzaga, 2016, Theorem 1. Suppose that $\alpha := \lambda_{\min}(A; M)$ and $\beta := \lambda_{\max}(A; M)$ are the extremal generalized eigenvalues of $A$ w.r.t. $M$, and $\kappa := \frac{\beta}{\alpha}$ is the generalized condition number. Suppose that $\kappa \ge 1.06$ and that

$$k := \left\lceil \sqrt{\kappa} \ln\left(\frac{2}{\varepsilon_1}\right) \right\rceil.$$

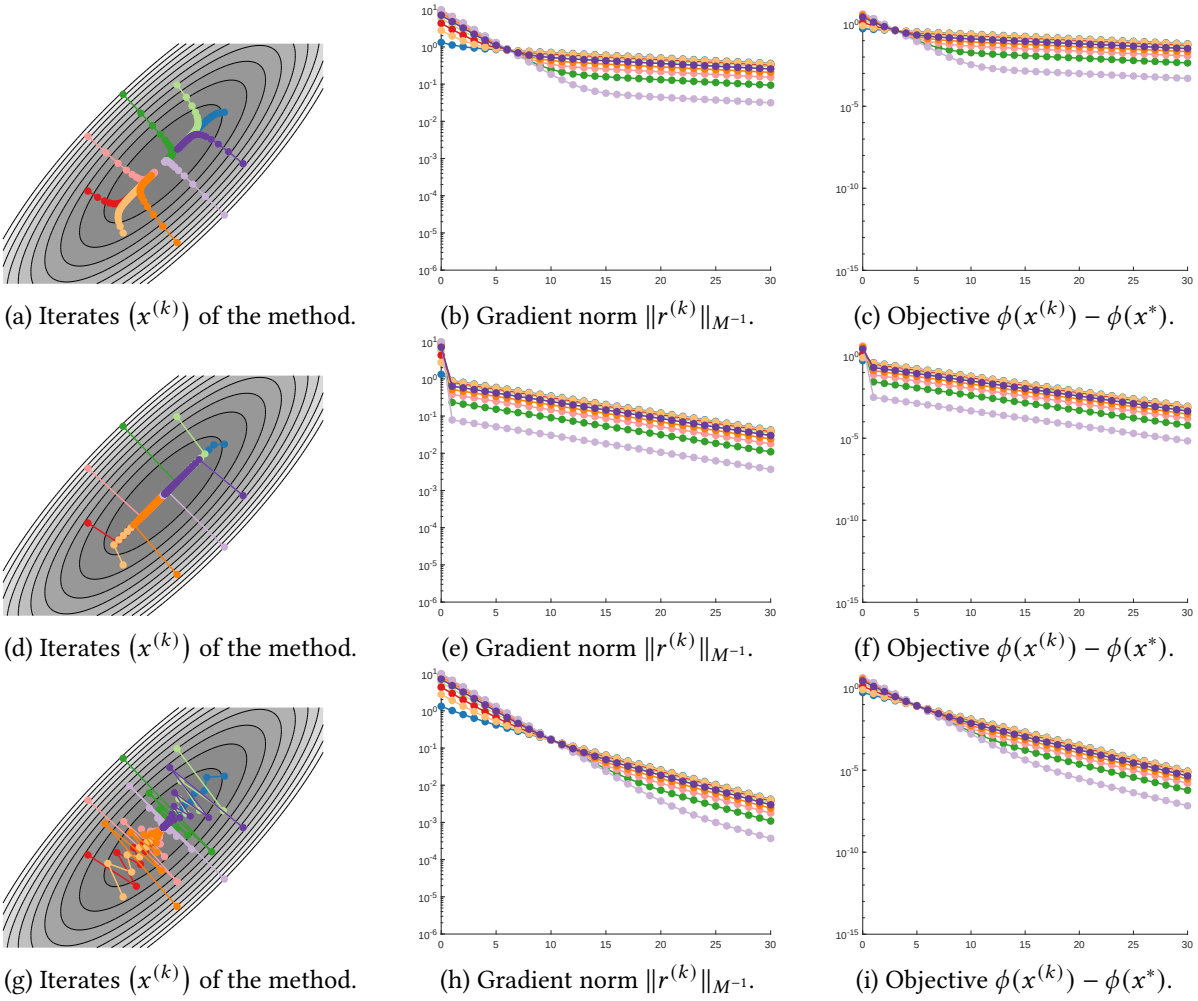
holds. Consider the set of mutually distinct, precomputed step lengths

$$\left\{ \alpha^{(k)} := \frac{1}{\omega^{(k)}} \,\middle|\, \omega^{(k)} := \frac{\beta - \alpha}{2} \cos\left(\frac{1 + 2j}{2k}\,\pi\right) + \frac{\beta + \alpha}{2},\ j = 0, 1, \ldots, k-1 \right\}.$$

Then the gradient descent method Algorithm 4.5 with step sizes $\alpha^{(k)}$, applied in any order, requires at most

$$k \text{ iterations until } \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \le \varepsilon_1.$$

The interesting fact is that, compared to the estimate of Corollary 4.8 for the Cauchy step size, the bound on the iteration numbers is proportional only to $\sqrt{\kappa}$, not to $\kappa$. The result can be modified so that



(a) Iterates $\left(x^{(k)}\right)$ of the method. (b) Gradient norm $\|r^{(k)}\|_{M^{-1}}$. (c) Objective $\phi(x^{(k)}) - \phi(x^*)$.

(d) Iterates $\left(x^{(k)}\right)$ of the method. (e) Gradient norm $\|r^{(k)}\|_{M^{-1}}$. (f) Objective $\phi(x^{(k)}) - \phi(x^*)$.

(g) Iterates $\left(x^{(k)}\right)$ of the method. (h) Gradient norm $\|r^{(k)}\|_{M^{-1}}$. (i) Objective $\phi(x^{(k)}) - \phi(x^*)$.

Figure 4.2: Illustration of the convergence behavior of Algorithm 4.5 with various constant step sizes instead of the Cauchy step length. The step sizes, from top to bottom, are $\overline{\alpha} \in \{0.03, 0.10, 0.17\}$. No preconditioning ($M = \mathrm{Id}$) is used. The two eigenvalues of the matrix are $\alpha = 1$ and $\beta = 10$ so the admissible range of constant step lengths is $\overline{\alpha} \in (0, \frac{2}{\beta}) = (0, 0.2)$.

it is not required to know the extremal eigenvalues exactly, but knowledge of an interval containing them is sufficient.

We are going to obtain a similar complexity result for the conjgate gradient method in **??**.

## § 4.5 Gradient Descent Method as Discretized Gradient Flow

We conclude the discussion of the gradient descent method by interpreting it in another way. Consider the differential equation

$$\begin{aligned} \dot{x}(t) &= -\nabla_M f(x(t)), \quad t \geq 0 \\ x(0) &= x_0. \end{aligned} \tag{4.15}$$

This is known as the **gradient flow** associated with $f$. Its stationary points are precisely the stationary points of $f$. Due to

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t)) = f'(x(t))\,\dot{x}(t) = -f'(x(t))\,M^{-1}\nabla f(x(t)) = -\|\nabla f(x(t))\|^2_{M^{-1}}, \tag{4.16}$$

the value of $f$ is decreasing along the path $x(t)$.

When we discretize (4.15) by the explicit (forward) Euler method with time step size $\Delta t^{(k)}$, we obtain

$$\frac{x^{(k+1)} - x^{(k)}}{\Delta t^{(k)}} = -M^{-1}\nabla f(x^{(k)}),$$

or equivalently,

$$x^{(k+1)} = x^{(k)} - \Delta t^{(k)}\, M^{-1}\nabla f(x^{(k)}). \tag{4.17}$$

This is precisely a step of the gradient descent method with step size $\Delta t^{(k)}$. Therefore, we can interpret the gradient descent method as a discretization of the continuous gradient flow equation.

End of Week 2

# Chapter 2 Theory for Constrained Optimization Problems

# Chapter 3 Numerical Techniques for Constrained Optimization Problems

# Chapter 4 Differentiation Techniques

# Bibliography


Akaike, H. (1959). "On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method". *Annals of the Institute of Statistical Mathematics* 11, pp. 1–16. DOI: 10.1007/bf01831719.

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.

Barzilai, J.; J. M. Borwein (1988). "Two-point step size gradient methods". *IMA Journal of Numerical Analysis* 8.1, pp. 141–148. DOI: 10.1093/imanum/8.1.141.

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

Cauchy, A.-L. (1847). "Méthode générale pour la résolution des systemes d'équations simultanées". *Comptes Rendus de l'Académie des Sciences Paris* 25, pp. 536–538.

De Asmundis, R.; D. di Serafino; F. Riccio; G. Toraldo (2013). "On spectral properties of steepest descent methods". *IMA Journal of Numerical Analysis* 33.4, pp. 1416–1435. DOI: 10.1093/imanum/drs056.

De Asmundis, R.; D. di Serafino; W. W. Hager; G. Toraldo; H. Zhang (2014). "An efficient gradient method using the Yuan steplength". *Computational Optimization and Applications* 59.3, pp. 541–563. DOI: 10.1007/s10589-014-9669-5.

Forsythe, G. E. (1968). "On the asymptotic directions of the $s$-dimensional optimum gradient method". *Numerische Mathematik* 11, pp. 57–76. DOI: 10.1007/BF02165472.

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

Gonzaga, C. C. (2016). "On the worst case performance of the steepest descent algorithm for quadratic functions". *Mathematical Programming Series A* 160, pp. 307–320. DOI: 10.1007/s10107-016-0984-8.

Gonzaga, C. C.; R. M. Schneider (2015). "On the steepest descent algorithm for quadratic functions". *Computational Optimization and Applications* 63.2, pp. 523–542. DOI: 10.1007/s10589-015-9775-z.

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.; A. Sartenaer; C. Zhu (2002). "On the behavior of the gradient norm in the steepest descent method". *Computational Optimization and Applications. An International Journal* 22.1, pp. 5–35. DOI: 10.1023/A:1014897230089.

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.