Lecture Notes
Nonlinear Optimization

Spring Semester 2023

Roland Herzog[*]

2023-07-14

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

These lecture notes are partly based on content from the books Nocedal, Wright, 2006; Ulbrich, Ulbrich, 2012 as well as on previous lecture notes by Roland Herzog and Gerd Wachsmuth (BTU Cottbus).

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 $f, g_i, h_j \colon \mathbb{R}^n \to \mathbb{R}$ are sufficiently smooth (C^2 functions),

- we have a finite number (possibly zero) of inequality and equality constraints, i. e., n_{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*

$$F := \left\{x \in \mathbb{R}^n \,\middle|\, g_i(x) \le 0 \textit{ for all } i = 1, \ldots, n_{\text{ineq}},\ h_j(x) = 0 \textit{ for all } j = 1, \ldots, n_{\text{eq}}\right\} \tag{1.2}$$

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 $\mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ with the following properties:

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

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

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

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

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

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

In particular, the Euclidean inner product $x^\mathsf{T} y$ corresponds to the identity matrix $M = \operatorname{Id}$, 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.

Notice that for vectors $x, y \in \mathbb{R}^n$, we have

$$\begin{aligned} a^\mathsf{T} b &= a^\mathsf{T} M^{-1} M\, b \\ &\le \|M^{-1} a\|_M \|b\|_M \quad \text{by the Cauchy-Schwarz inequality w.r.t. the } M\text{-inner product} \\ &= \|a\|_{M^{-1}} \|b\|_M. \end{aligned} \tag{2.3}$$

§ 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

$$\|A\|_{M_2 \leftarrow M_1} \coloneqq \max_{x \neq 0} \frac{\|A\,x\|_{M_2}}{\|x\|_{M_1}}. \tag{2.4}$$

We thus have

$$\|A\,x\|_{M_2} \le \|A\|_{M_2 \leftarrow M_1} \|x\|_{M_1} \quad \text{for all } x \in \mathbb{R}^n. \tag{2.5}$$

When M_1 and M_2 are both the Euclidean inner products, $\|A\|_{\operatorname{Id} \leftarrow \operatorname{Id}}$ 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.

There are matrix norm which are not operator norms. The most prominent one is induced by the inner product

$$A : B \coloneqq \operatorname{trace}(A^\mathsf{T} B) = \sum_{i=1}^m \sum_{j=1}^n a_{ij}\, b_{ij}. \tag{2.6}$$

The associated norm

$$\|A\|_F \coloneqq \Big(\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2 \Big)^{1/2} \tag{2.7}$$

is termed the **Frobenius norm** of A.

§ 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

$$A\,V = V\Lambda, \quad \text{i. e.,} \quad A = V \Lambda V^\mathsf{T} \tag{2.8}$$

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**

$$A\,v = \lambda\,v. \tag{2.9}$$

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

$$A\,v = \lambda\,M\,v \tag{2.10}$$

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**

$$A\,V = M\,V\Lambda, \quad \text{i. e.,} \quad A = M\,V\Lambda V^\intercal M, \tag{2.11}$$

where now V is orthogonal w.r.t. the M-inner product, i. e., $V^\intercal M\,V = \mathrm{Id}$ holds. This implies $V V^\intercal = M^{-1}$. We also refer to the solutions of (2.10) 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

$$\lambda_{\min}(A; M) \le \frac{x^\intercal A\,x}{x^\intercal M\,x} \le \lambda_{\max}(A; M) \quad \text{for all } x \neq 0. \tag{2.12}$$

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. Using (2.3) and (2.5), we also have

$$-\|A\|_{M^{-1}\leftarrow M} \le -\frac{\|x\|_M \|A\,x\|_{M^{-1}}}{\|x\|_M^2} \le \frac{x^\intercal A\,x}{\|x\|_M^2} \le \frac{\|x\|_M \|A\,x\|_{M^{-1}}}{\|x\|_M^2} \le \|A\|_{M^{-1}\leftarrow M}$$

and thus

$$\lambda_{\max}(H; M) \le \|H\|_{M^{-1}\leftarrow M} \quad \text{and} \quad -\lambda_{\min}(H; M) \le \|H\|_{M^{-1}\leftarrow M}. \tag{2.13}$$

Notice that the generalized eigenvalue problems (2.10) and

$$M\,v = \lambda\,M\,A^{-1}M\,v \tag{2.14a}$$

as well as

$$A\,M^{-1}A\,v = \lambda\,A\,v \tag{2.14b}$$

have the same eigenvalues and eigenvectors (provided ~~in case of (2.14a)~~ 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 (2.14):

$$\lambda_{\min}(A; M) \le \frac{x^\intercal M\,x}{x^\intercal M\,A^{-1}M\,x} \le \lambda_{\max}(A; M) \quad \text{for all } x \neq 0, \tag{2.15a}$$

$$\lambda_{\min}(A; M) \le \frac{x^\intercal A\,M^{-1}A\,x}{x^\intercal A\,x} \le \lambda_{\max}(A; M) \quad \text{for all } x \neq 0. \tag{2.15b}$$

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 $A = V\Lambda V^\intercal$ is a spectral decomposition of A with orthogonal V, then

$$A^{1/2} = V\Lambda^{1/2}V^\intercal \tag{2.16}$$

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

$$\kappa := \frac{\beta}{\alpha}. \tag{2.17}$$

Notice that a condition number always satisfies $\kappa \ge 1$. From the Rayleigh quotient estimate (2.12) (with $M = \mathrm{Id}$), we have

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

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

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

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

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

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

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*

$$1 \le \frac{(x^\mathsf{T} A\, x)\,(x^\mathsf{T} A^{-1} x)}{\|x\|^4} \le \frac{(\alpha+\beta)^2}{4\,\alpha\,\beta} \le \frac{\beta}{\alpha} \tag{2.18a}$$

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

$$1 \le \frac{(x^\mathsf{T} A\, x)\,(x^\mathsf{T} A^{-1} x)}{\|x\|^4} \le \frac{(\kappa+1)^2}{4\,\kappa} \le \kappa. \tag{2.18b}$$

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 (2.18).

[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 $\sigma_{\min}(A) = \lambda_{\min}(A) = \alpha$.

From here on, the proof follows Anderson, 1971, as reproduced in the Master's thesis Alpargu, 1996, Section 1.2.2. Let $\lambda_1, \ldots, \lambda_n > 0$ be the eigenvalues of A (in any order), and let $v_1, \ldots, v_n$ 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

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

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

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

This shows that essential upper bound in (2.18). 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 $A \in \mathbb{R}^{n \times n}$ and 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*

$$1 \le \frac{(x^\intercal A\, x)\,(x^\intercal M\, A^{-1} M\, x)}{\|x\|_M^4} \le \frac{(\alpha + \beta)^2}{4\,\alpha\,\beta} \le \frac{\beta}{\alpha} \tag{2.19a}$$

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

$$1 \le \frac{(x^\intercal A\, x)\,(x^\intercal A^{-1} x)}{\|x\|_M^4} \le \frac{(\kappa + 1)^2}{4\,\kappa} \le \kappa. \tag{2.19b}$$

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

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

  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]

- 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, 1971, 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*
$$\begin{aligned} &\textit{in case } k = 0: \quad \left|f(x^{(0)} + d) - f(x^{(0)}) - f'(x^{(0)})\, d\right| \le \varepsilon\, \|d\|, \\ &\textit{in case } k = 1: \quad \left|f(x^{(0)} + d) - f(x^{(0)}) - f'(x^{(0)})\, d - \frac{1}{2} d^\intercal f''(x^{(0)}) d\right| \le \varepsilon\, \|d\|^2. \end{aligned}$$
for all $\|d\| < \delta$.

[3] See for instance Cartan, 1971, Proposition 5.2.2

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

$$\begin{aligned}
&\textit{in case } k = 0: \quad f(x^{(0)} + d) = f(x^{(0)}) + f'(x^{(0)} + \xi\, d)\, d \quad (\textbf{\textit{mean value theorem}}),\\
&\textit{in case } k = 1: \quad f(x^{(0)} + d) = f(x^{(0)}) + f'(x^{(0)})\, d + \frac{1}{2} d^\top f''(x^{(0)} + \xi\, d)\, d.
\end{aligned}$$

§ 2.7 Convergence Rates

We denote (vector-valued) sequences $\mathbb{N} \to \mathbb{R}^n$ by $\big(x^{(k)}\big)$ and not (x_k) etc., in order to avoid a conflict of notation with the components of a vector $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$. 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*

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

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

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

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

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

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

[4] "Q" stands for "quotient".
[5] "R" stands for "root".

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

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

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

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

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

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

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

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

and $\big(\varepsilon^{(k)}\big)$ 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.

Definition 2.7 (Convex set).
A set $C \subseteq \mathbb{R}^n$ is termed ***convex*** *if $x, y \in C$ and $\alpha \in [0,1]$ imply $\alpha\, x + (1-\alpha)\, y \in C$.*

The condition in Definition 2.7 means that the entire line segment between x and y belongs to C.

Definition 2.8 (Convex function).
A function $f\colon \mathbb{R}^n \to \mathbb{R}$ is termed

(i) ***convex*** *in case*

$$f(\alpha\, x + (1-\alpha)\, y) \le \alpha f(x) + (1-\alpha) f(y) \tag{2.20}$$

holds for all $x, y \in \mathbb{R}^n$ and $\alpha \in [0,1]$.

(ii) ***strictly convex*** *in case*

$$f(\alpha\, x + (1-\alpha)\, y) < \alpha f(x) + (1-\alpha) f(y) \tag{2.21}$$

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*

$$f(\alpha\, x + (1-\alpha)\, y) + \frac{\mu}{2}\,\alpha\,(1-\alpha)\|x-y\|^2 \le \alpha f(x) + (1-\alpha) f(y) \tag{2.22}$$

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

$$f(x) - f(y) \ge f'(y)(x-y) \tag{2.23}$$

holds.

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

$$\big(f'(x) - f'(y)\big)(x-y) \ge 0 \tag{2.24}$$

holds. Equation (2.24) *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 $x \neq y$,*

$$f(x) - f(y) > f'(y)(x-y) \tag{2.25}$$

holds.

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

$$\big(f'(x) - f'(y)\big)(x-y) > 0. \tag{2.26}$$

Equation (2.26) *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 $x, y \in \mathbb{R}^n$,*

$$f(x) - f(y) \ge f'(y)(x-y) + \frac{\mu}{2}\,\|x-y\|^2 \tag{2.27}$$

holds.

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

$$\big(f'(x) - f'(y)\big)(x - y) \ge \mu \, \|x - y\|^2. \tag{2.28}$$

Equation (2.28) means that f' *is a* ***strongly monotone operator***.

Theorem 2.10 (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* $\lambda_{\min}(f''(x)) \ge \mu > 0$ *for all* $x \in \mathbb{R}^n$.

§ 2.9 Hyperplanes and Half Spaces

Suppose that $a \in \mathbb{R}^n$, $a \neq 0$ and $\beta \in \mathbb{R}$. Then the set

$$H(a, \beta) \coloneqq \{x \in \mathbb{R}^n \mid a^\intercal x = \beta\} \tag{2.29}$$

is termed a **hyperplane** in $\mathbb{R}^n$ with **normal vector** a.

A hyperplane separates $\mathbb{R}^n$ into two closed **half spaces**

$$\begin{aligned} H^-(a, \beta) &\coloneqq \{x \in \mathbb{R}^n \mid a^\intercal x \le \beta\} \quad &\textbf{negative half space}, \\ H^+(a, \beta) &\coloneqq \{x \in \mathbb{R}^n \mid a^\intercal x \ge \beta\} \quad &\textbf{positive half space}. \end{aligned} \tag{2.30}$$

§ 2.10 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 $x \in \mathbb{R}^n$,

$$B_\varepsilon^M(\overline{x}) \coloneqq \left\{x \in \mathbb{R}^n \,\middle|\, \|x - \overline{x}\|_M < \varepsilon\right\}$$

denotes the **open ε-ball** w.r.t. the M-norm about $\overline{x}$ (centered at $\overline{x}$). Similarly, the **closed ε-ball** is

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

A **neighborhood** of a point $\overline{x} \in \mathbb{R}$ is a set containing some open ball centered at $\overline{x}$. We often write $U(\overline{x})$ for such a neighborhood.

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

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, the first-order optimality condition is already sufficient for optimality, and we do not require a second-order condition.

§ 3 Optimality Conditions

We suppose 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. Suppose that $d \in \mathbb{R}^n$ is arbitrary. 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 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

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

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

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

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 (2.12) (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

§ 4 Minimization of Quadratic Functions

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

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

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

If we knew a spectral decomposition of $A = V \Lambda V^\intercal$ (which of course we usually don't), we could represent the objective as $\phi(x) = \frac{1}{2} x^\intercal V \, \Lambda \, V^\intercal x - b^\intercal V \, V^\intercal x + c$. After a substitution of variables $x = V^\intercal y$, this becomes $\widetilde{\phi}(y) = \frac{1}{2} y^\intercal \, \Lambda \, y - b^\intercal V \, y + c$. Consequently, in these coordinates, the problem decomposes into a sum of n independent quadratic minimization problems in the components y_i.

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:*

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

(*b*) *The objective ϕ is bounded below.*

(*c*) *$A x = b$ is solvable.*

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

(*ii*) *In case A is not positive semidefinite[4], the objective ϕ 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 $A x = b$. 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 $A = L L^\mathsf{T}$ with the lower triangular matrix L.[5] However, when the problem is high-dimensional (such as $n \ge 10\,000$), 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 $\big(x^{(k)}\big)$ 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 $A x = b$ (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

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

[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

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 $x \mapsto A\,x$. **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?

Our objective ϕ from (4.1) satisfies

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

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 ϕ), 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

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

□

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**.

[7] Sometimes the residual is defined in the literature with opposite sign. We do not write $r(x)$ to keep the notation concise. It will be clear from the context which vector x the residual is associated with.

[8] You can learn more about iterative solvers for more general linear systems (not related to optimization) in the class *Numerische lineare Algebra*.

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 $x^{(k+1)}$.

§ 4.1 Direction of Steepest Descent

We first need to clarify what **descent directions** and the **directions of steepest descent** of a function $f\colon \mathbb{R}^n \to \mathbb{R}$ at a point x are.

Definition 4.4 (Descent direction).
Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable at $x \in \mathbb{R}^n$. A vector $d \in \mathbb{R}^n$ is termed a ***descent direction*** *for f at x if*

$$f'(x)\,d < 0. \tag{4.3}$$

holds.

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:

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

We note that we could be considering the equivalent problem

$$\begin{aligned} \text{Minimize} \quad & f'(x)\,d \quad \text{where } d \in \mathbb{R}^n \\ \text{subject to} \quad & \|d\|_M \le 1. \end{aligned} \tag{4.5}$$

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

Problems (4.4), (4.5) 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]

$$f'(x)\,d = \nabla f(x)^\intercal d = \nabla f(x)^\intercal M^{-1} M\,d = \big(M^{-1}\nabla f(x)\big)^\intercal M\,d,$$

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 4.5 (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.

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

(*i*) *The vector*

$$\nabla_M f(x) := M^{-1} \nabla f(x) \tag{4.6}$$

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 the* ***directions 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

$$M\, d^* = -\nabla f(x). \tag{4.7}$$

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**.

§ 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 size. We have the following expression for the difference of function values before and after a step:

$$\begin{aligned} \phi(x + \alpha\, d) - \phi(x) &= \frac{1}{2}(x + \alpha\, d)^\mathsf{T} A\, (x + \alpha\, d) - b^\mathsf{T}(x + \alpha\, d) + c - \frac{1}{2} x^\mathsf{T} A\, x + b^\mathsf{T} x - c \\ &= \frac{1}{2}(d^\mathsf{T} A\, d)\, \alpha^2 + (A\, x - b)^\mathsf{T} d\, \alpha \\ &= \frac{1}{2}(d^\mathsf{T} A\, d)\, \alpha^2 + (r^\mathsf{T} d)\, \alpha. \end{aligned} \tag{4.8}$$

Note: This formula holds for arbitrary directions d and step sizes α.

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 α such that $\phi(x + \alpha\, d)$ is minimized. According to (4.8), 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

$$\alpha^* = -\frac{r^\mathsf{T} d}{d^\mathsf{T} A\, d}. \tag{4.9}$$

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

This "optimal" step size is also known as the **Cauchy step size**. For this choice, the difference of function values (4.8) before and after a step becomes

$$\begin{aligned}\phi(x+\alpha^* d)-\phi(x) &= \frac{1}{2}(d^\intercal A\, d)\,(\alpha^*)^2+(r^\intercal d)\,\alpha^* \\ &= \frac{1}{2}(d^\intercal A\, d)\left(\frac{r^\intercal d}{d^\intercal A\, d}\right)^2-(r^\intercal d)\,\frac{r^\intercal d}{d^\intercal A\, d} \\ &= -\frac{1}{2}\frac{(r^\intercal d)^2}{d^\intercal A\, d}. \end{aligned} \tag{4.10}$$

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

We can now state the steepest descent method w.r.t. the M-inner product and the Cauchy step size (4.9) 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.

Algorithm 4.6 (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 $A\,x = b$
Set $k := 0$
Set $r^{(0)} := A\,x^{(0)} - b$                         // evaluate the initial residual
Set $d^{(0)} := -M^{-1}r^{(0)}$                          // evaluate the initial negative $M$-gradient
Set $\delta^{(0)} := -(r^{(0)})^\intercal d^{(0)}$       // $\delta^{(0)} = \|\nabla_M\phi(x^{(0)})\|_M^2 = \|r^{(0)}\|_{M^{-1}}^2$
while stopping criterion not met do
    Set $q^{(k)} := A\,d^{(k)}$
    Set $\theta^{(k)} := (q^{(k)})^\intercal d^{(k)}$
    Set $\alpha^{(k)} := \delta^{(k)}/\theta^{(k)}$      // evaluate the Cauchy step size
    Set $x^{(k+1)} := x^{(k)} + \alpha^{(k)}d^{(k)}$     // update the iterate
    Set $r^{(k+1)} := r^{(k)} + \alpha^{(k)}q^{(k)}$     // update the residual
    Set $d^{(k+1)} := -M^{-1}r^{(k+1)}$                  // evaluate the negative $M$-gradient
    Set $\delta^{(k+1)} := -(r^{(k+1)})^\intercal d^{(k+1)}$   // $\delta^{(k+1)} = \|\nabla_M\phi(x^{(k+1)})\|_M^2 = \|r^{(k+1)}\|_{M^{-1}}^2$
    Set $k := k + 1$
end while
return $x^{(k)}$
```

The following can be said about Algorithm 4.6.

Remark 4.7 (on Algorithm 4.6)**.**

(*i*) *Algorithm 4.6 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 $A\,x = b$.*

(*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 4.6 is often termed a* ***preconditioned gradient descent method****. The case $M = \mathrm{Id}$ corresponds to the classical gradient descent method (without preconditioning).*

(*iv*) *We also 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 4.6 requires the storage of four vectors, which are iteratively overwritten: iterates $x^{(k)}$, residuals $r^{(k)}$, negative gradient directions $d^{(k)}$, and vectors $q^{(k)} = A\,d^{(k)}$.*

(*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 $r^{(k)} \coloneqq A\,x^{(k)} - b$, rather than update it.*

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

$$0 < \lambda_{\min}(A;M) \le \frac{1}{\alpha^{(k)}} = \frac{(d^{(k)})^\intercal A\,d^{(k)}}{(d^{(k)})^\intercal M\,d^{(k)}} \le \lambda_{\max}(A;M), \tag{4.11}$$

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 4.6 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:*

$$\phi(x^{(0)}) = c + \frac{1}{2}(r^{(0)} - b)^\intercal(x^{(0)}) \quad \textit{initialization} \tag{4.12a}$$

$$\phi(x^{(k+1)}) = \phi(x^{(k)}) - \frac{1}{2}\,\alpha^{(k)}\delta^{(k)} \quad \textit{update.} \tag{4.12b}$$

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 4.6. The function values at the iterates satisfy

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

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 $\alpha^{(k)} = -\frac{(r^{(k)})^\top d^{(k)}}{(d^{(k)})^\top A\, d^{(k)}}$ and obtain

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

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

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

The fraction is precisely the type of expression estimated by the generalized Kantorovich inequality (2.19). This yields

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

We have thus shown the following classical convergence result for Algorithm 4.6:

Theorem 4.8 (Convergence of Algorithm 4.6). *Suppose that $A \in \mathbb{R}^{n\times n}$ and 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 sizes 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*

$$
\phi(x^{(k+1)}) - \phi(x^*) \le \left(\frac{\kappa-1}{\kappa+1}\right)^2 \big(\phi(x^{(k)}) - \phi(x^*)\big) \tag{4.13a}
$$

$$
\|x^{(k+1)} - x^*\|_A \le \left(\frac{\kappa-1}{\kappa+1}\right) \|x^{(k)} - x^*\|_A \tag{4.13b}
$$

and consequently

$$
\phi(x^{(k)}) - \phi(x^*) \le \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \big(\phi(x^{(0)}) - \phi(x^*)\big) \tag{4.13c}
$$

$$
\|x^{(k)} - x^*\|_A \le \left(\frac{\kappa-1}{\kappa+1}\right)^{k} \|x^{(0)} - x^*\|_A. \tag{4.13d}
$$

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 (4.13c) and (4.13d) have been decreased relative to their initial values.

Corollary 4.9 (Maximal number of iterations required in Algorithm 4.6)**.** *Given positive numbers ε_1 and ε_2, it takes*

$$k \le \left\lceil \frac{\kappa}{4} \ln\left(\frac{1}{\varepsilon_1}\right) \right\rceil \textit{ iterations until } \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \le \varepsilon_1,$$
$$k \le \left\lceil \frac{\kappa}{2} \ln\left(\frac{1}{\varepsilon_2}\right) \right\rceil \textit{ iterations until } \left(\frac{\kappa-1}{\kappa+1}\right)^{k} \le \varepsilon_2.$$

Proof. (1) We first show that

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

holds for all $\kappa \ge 1$. At $\kappa = \frac{e+1}{e-1}$, we have

$$-\ln\left(\frac{\kappa-1}{\kappa+1}\right) = -\ln\left(\frac{1}{e}\right) = 1 > \frac{2}{\kappa} = 2\,\frac{e-1}{e+1} \approx 0.92.$$

We now show that

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

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

$$\frac{-2}{(\kappa-1)(\kappa+1)} < \frac{-2}{\kappa^2} < 0 \quad \text{for all } \kappa > 1.$$

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

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

for all $\kappa > 1$.

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

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

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

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

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

The second claim follows similarly. □

Remark 4.10 (on Theorem 4.8).

(i) (4.13b) *shows the Q-linear convergence of* $\big(x^{(k)}\big)$ *to the solution* x^* *in the* A*-norm.*

(ii) *The contraction factor is* $0 \le \frac{\kappa-1}{\kappa+1} < 1$, *i. e., the convergence estimate depends on the ratio* κ *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* $A\,x = b$ *directly.*

(iv) *A good preconditioner is a compromise between a moderate generalized condition number* κ *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 4.8 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 for some initial guesses than the estimate* (4.13c) *suggests.*

Figure 4.1 illustrates the convergence behavior of Algorithm 4.6 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 a generalized eigenvector of A w.r.t. M, 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?

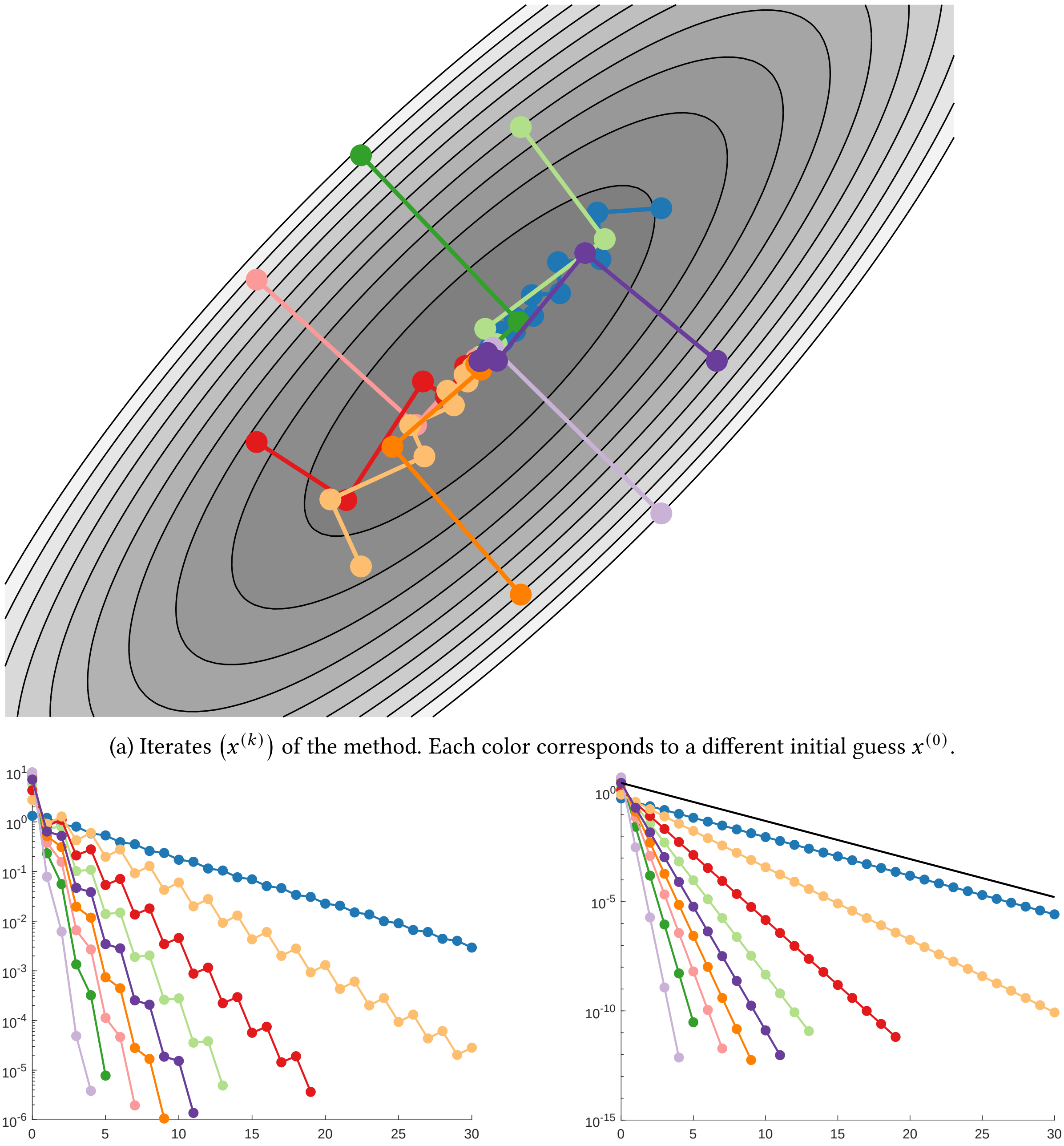
The zig-zagging behavior of the iterates $x^{(k)}$, as well as the non-monotone behavior of $\|r^{(k)}\|_{M^{-1}}$ have been analyzed in detail in the literature; see for instance Akaike, 1959; Forsythe, 1968; Nocedal, Sartenaer, Zhu, 2002. Essentially what happens is that, asymptotically, the error $x^{(k)} - x^*$ 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?



(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 $\sqrt{\delta^{(k)}} = \|\nabla_M \phi(x^{(k)})\|_M = \|r^{(k)}\|_{M^{-1}}$ does not necessarily converge monotonically.

(c) The objective values $\phi(x^{(k)}) - \phi(x^*)$ converge monotonically. The black line illustrates the bound (4.13c).

Figure 4.1: Illustration of the convergence behavior of Algorithm 4.6 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 $\kappa = 10$.

(*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 $\|r^{(k)}\|_{M^{-1}}$:

$$\|r^{(k)}\|_{M^{-1}} \le \varepsilon_{\text{rel}} \, \|r^{(0)}\|_{M^{-1}}, \qquad \text{i. e., } \delta^{(k)} \le \varepsilon_{\text{rel}}^2 \, \delta^{(0)}, \tag{4.14a}$$

$$\|r^{(k)}\|_{M^{-1}} \le \varepsilon_{\text{abs}}, \qquad \text{i. e., } \delta^{(k)} \le \varepsilon_{\text{abs}}^2, \tag{4.14b}$$

$$\|r^{(k)}\|_{M^{-1}} \le \varepsilon_{\text{rel}} \, \|r^{(0)}\|_{M^{-1}} + \varepsilon_{\text{abs}}, \qquad \text{i. e., } (\delta^{(k)})^{1/2} \le \varepsilon_{\text{rel}} \, (\delta^{(0)})^{1/2} + \varepsilon_{\text{abs}}, \tag{4.14c}$$

$$\|r^{(k)}\|_{M^{-1}} \le \max\big\{\varepsilon_{\text{rel}} \, \|r^{(0)}\|_{M^{-1}}, \ \varepsilon_{\text{abs}}\big\}, \qquad \text{i. e., } \delta^{(k)} \le \max\big\{\varepsilon_{\text{rel}}^2 \, \delta^{(0)}, \ \varepsilon_{\text{abs}}^2\big\}. \tag{4.14d}$$

Let us see which consequences either of the implementable stopping criteria (4.14) has on the other two quantities of interest:

Lemma 4.11 (Implications). *The criteria from* (4.14) *imply, respectively,*

$$\left.\begin{aligned} \|x^{(k)} - x^*\|_A &\le \sqrt{\kappa} \, \varepsilon_{\text{rel}} \, \|x^{(0)} - x^*\|_A \\ \|x^{(k)} - x^*\|_M &\le \kappa \, \varepsilon_{\text{rel}} \, \|x^{(0)} - x^*\|_M \end{aligned}\right\} \tag{4.15a}$$

$$\left.\begin{aligned} \|x^{(k)} - x^*\|_A &\le (1/\sqrt{\alpha}) \, \varepsilon_{\text{abs}} \\ \|x^{(k)} - x^*\|_M &\le (1/\alpha) \, \varepsilon_{\text{abs}} \end{aligned}\right\} \tag{4.15b}$$

$$\left.\begin{aligned} \|x^{(k)} - x^*\|_A &\le \sqrt{\kappa} \, \varepsilon_{\text{rel}} \, \|x^{(0)} - x^*\|_A + (1/\sqrt{\alpha}) \, \varepsilon_{\text{abs}} \\ \|x^{(k)} - x^*\|_M &\le \kappa \, \varepsilon_{\text{rel}} \, \|x^{(0)} - x^*\|_M + (1/\alpha) \, \varepsilon_{\text{abs}} \end{aligned}\right\} \tag{4.15c}$$

$$\left.\begin{aligned} \|x^{(k)} - x^*\|_A &\le \max\big\{\sqrt{\kappa} \, \varepsilon_{\text{rel}} \, \|x^{(0)} - x^*\|_A, \ (1/\sqrt{\alpha}) \, \varepsilon_{\text{abs}}\big\} \\ \|x^{(k)} - x^*\|_M &\le \max\big\{\kappa \, \varepsilon_{\text{rel}} \, \|x^{(0)} - x^*\|_M, \ (1/\alpha) \, \varepsilon_{\text{abs}}\big\} \end{aligned}\right\} \tag{4.15d}$$

Proof. homework problem 2.3 □

§ 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

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

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

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

Here we need to convert the last term into $d^\intercal M\,A^{-1} M\,d$, which is equal to $r^\intercal A^{-1} r$, 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 (2.15a):

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

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 4.12 (on the convergence of Algorithm 4.6 with constant step sizes).

(*i*) *We have shown that Algorithm 4.6, where Line 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 β 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 $\overline{\alpha} \in (0, \frac{2}{\beta})$.*

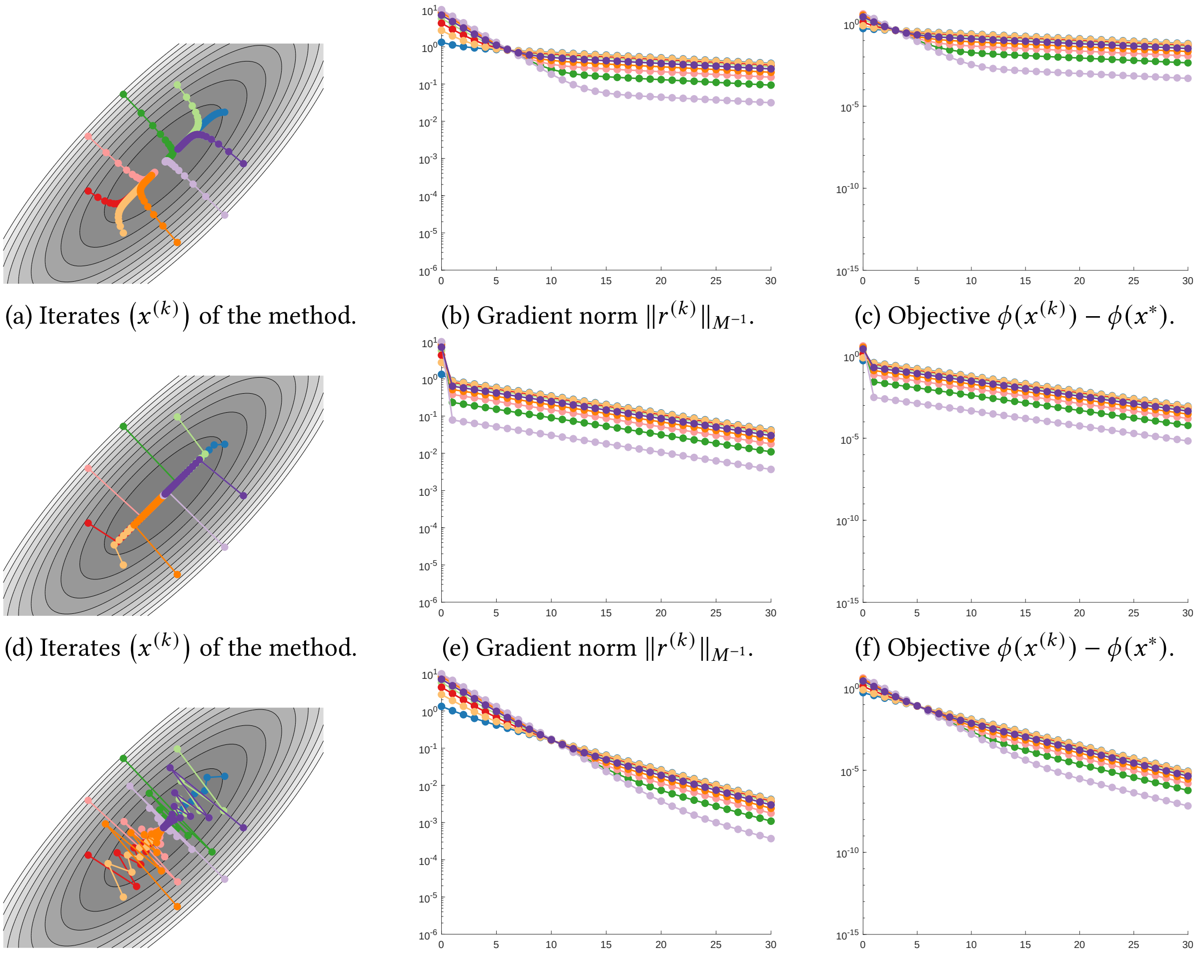
(*iii*) *The choice $\overline{\alpha} = \frac{1}{\beta}$ yields the optimal estimate. In this case, we obtain*

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

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 (4.13a). Consequently, there is no reason to prefer the gradient descent method with constant step sizes over the version with Cauchy step sizes.

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

Figure 4.2 illustrates the convergence behavior of Algorithm 4.6 with constant step sizes for a 2-dimensional example problem from a number of different initial guesses $x^{(0)}$.



(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 4.6 with various constant step sizes instead of the Cauchy step size. 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 sizes is $\overline{\alpha} \in (0, \frac{2}{\beta}) = (0, 0.2)$.

§ 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 $\alpha \coloneqq \lambda_{\min}(A;M)$ and $\beta \coloneqq \lambda_{\max}(A;M)$ are the extremal generalized eigenvalues of A w.r.t. M, and $\kappa \coloneqq \frac{\beta}{\alpha}$ is the generalized condition number. Suppose that $\kappa \ge 1.06$ and that

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

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

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

Then the gradient descent method Algorithm 4.6 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 4.9 for the Cauchy step size, the bound on the iteration numbers is proportional only to $\sqrt{\kappa}$, not to κ. The result can be modified so that 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 § 4.6.

§ 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

$$\begin{aligned} \dot{x}(t) &= -\nabla_M f(x(t)), \quad t \ge 0 \\ x(0) &= x^{(0)}. \end{aligned} \tag{4.16}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t)) = f'(x(t))\,\dot{x}(t) = -f'(x(t))\,M^{-1}\nabla f(x(t)) = -\|\nabla f(x(t))\|^2_{M^{-1}} = -\|\nabla_M f(x(t))\|^2_M, \tag{4.17}$$

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

When we discretize (4.16) 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,

$$x^{(k+1)} = x^{(k)} - \Delta t^{(k)} \, M^{-1} \nabla f(x^{(k)}). \tag{4.18}$$

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

§ 4.6 Conjugate Gradient Method

The typical inefficient zig-zaggging pattern of the directions $d^{(k)}$ is a consequence of the fact that gradient descent is a memoryless method. That is, we could restart the method at any iterate and it would produce the same iterates, whether restarted or not. This is where the **conjugate gradient method** (**CG method**, introduced in Hestenes, Stiefel, 1952) takes a different turn. It works with search directions $d^{(k)}$ which are pairwise A-orthogonal (also known as A-conjugate), and builds a memory of previously visited directions.

Definition 4.13 (Conjugate directions)**.** *Suppose that $A \in \mathbb{R}^{n \times n}$ is s. p. d. A set of non-zero vectors $\{d^{(0)}, \ldots, d^{(k)}\} \subset \mathbb{R}^n$ is termed **A-conjugate** if*

$$(d^{(i)})^\intercal A \, d^{(j)} = 0 \quad \textit{for } 0 \le i, j \le k, \quad i \neq j.$$

In other words, A-conjugate vectors are pairwise orthogonal w.r.t. the A-inner product. In particular, $\{d^{(0)}, \ldots, d^{(k)}\}$ is a linearly independent set. (**Quiz 4.4:** Can you prove that?)

The CG method is a member of the class of **conjugate direction methods**. We begin by describing the properties of a generic conjugate direction method first before we particularize to the CG method. A conjugate direction method chooses its search directions $d^{(0)}, d^{(1)}, \ldots$ so that they are A-conjugate, and the iterates satisfy

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}. \tag{4.19}$$

The step size $\alpha^{(k)}$ is the Cauchy step size, which minimizes the one-dimensional quadratic polynomial

$$\alpha \mapsto \phi(x^{(k)} + \alpha \, d^{(k)}).$$

That is, we have

$$\alpha^{(k)} := -\frac{(r^{(k)})^\intercal d^{(k)}}{(d^{(k)})^\intercal A \, d^{(k)}}, \tag{4.20}$$

compare (4.9). As in the gradient descent method, the residuals satisfy the recursion

$$r^{(k+1)} = r^{(k)} + \alpha^{(k)} A \, d^{(k)}. \tag{4.21}$$

Conjugate direction methods have the remarkable property that the sequence of one-dimensional minimizations in the A-conjugate directions $d^{(0)}, d^{(1)}, \ldots$ is equivalent to the minimization over the entire affine subspace $x^{(0)} + \operatorname{span}\{d^{(0)}, d^{(1)}, \ldots\}$. This is shown in the following result.

Lemma 4.14 (Properties of conjugate direction methods)**.** *Suppose that $A \in \mathbb{R}^{n \times n}$ is s. p. d. Given an initial guess $x^{(0)}$ and a set $\{d^{(0)}, d^{(1)}, \dots, d^{(k-1)}\}$, $k \ge 1$ of A-conjugate search directions, suppose that the iterates $x^{(0)}, \dots, x^{(k)}$ are generated according to (4.19) with Cauchy step size (4.20). Then the following holds.*

(*i*)
$$(r^{(k)})^\intercal d^{(i)} = 0 \quad \textit{for all } i = 0, 1, \dots, k-1. \tag{4.22}$$

(*ii*) *$x^{(k)}$ minimizes ϕ over the affine subspace $x^{(0)} + \operatorname{span}\{d^{(0)}, d^{(1)}, \dots, d^{(k-1)}\}$.*

Proof. We can show Statement (*i*) via induction over k. For $k = 1$,

$$\begin{aligned}
(r^{(1)})^\intercal d^{(0)} &= (A\,x^{(1)} - b)^\intercal d^{(0)} && \text{by definition of the residual} \\
&= (A\,x^{(0)} + \alpha^{(0)} A\,d^{(0)} - b)^\intercal d^{(0)} && \text{by (4.19)} \\
&= (r^{(0)})^\intercal d^{(0)} + \alpha^{(0)}\,(d^{(0)})^\intercal A\,d^{(0)} && \text{by definition of the residual} \\
&= 0 && \text{since } \alpha^{(0)} \text{ is the Cauchy step size (4.20).}
\end{aligned}$$

The induction step assumes $(r^{(k-1)})^\intercal d^{(i)} = 0$ for all $i = 0, 1, \dots, k-2$ and proceeds as follows.

$$\begin{aligned}
(r^{(k)})^\intercal d^{(k-1)} &= (r^{(k-1)} + \alpha^{(k-1)} A\,d^{(k-1)})^\intercal d^{(k-1)} && \text{by the residual recursion (4.21)} \\
&= 0 && \text{since } \alpha^{(k-1)} \text{ is the Cauchy step size (4.20).}
\end{aligned}$$

For the remaining search directions $d^{(i)}$, $i = 0, 1, \dots, k-2$ we have

$$\begin{aligned}
(r^{(k)})^\intercal d^{(i)} &= \big(r^{(k-1)} + \alpha^{(k-1)} A\,d^{(k-1)}\big)^\intercal d^{(i)} && \text{by the residual recursion (4.21)} \\
&= \underbrace{(r^{(k-1)})^\intercal d^{(i)}}_{=0 \text{ by assumption}} + \alpha^{(k-1)} \underbrace{(d^{(k-1)})^\intercal A\,d^{(i)}}_{=0 \text{ due to } A\text{-conjugacy}} \\
&= 0.
\end{aligned}$$

For Statement (*ii*) we consider the function $h\colon \mathbb{R}^k \to \mathbb{R}$

$$h(\sigma) \coloneqq \phi\left(x^{(0)} + \sum_{j=0}^{k-1} \sigma_j\, d^{(j)}\right).$$

h is strongly convex (**Quiz 4.5:** Why?), and the unique minimizer σ^* is characterized by

$$\frac{\partial h(\sigma^*)}{\partial \sigma_i} = \nabla\phi\Big(x^{(0)} + \sum_{j=0}^{k-1} \sigma_j^*\, d^{(j)}\Big)^\intercal d^{(i)} = 0, \quad i = 0, \dots, k-1. \tag{$*$}$$

However, we already know that it is the iterate

$$x^{(k)} = x^{(0)} + \sum_{j=0}^{k-1} \alpha^{(j)}\, d^{(j)} \in x^{(0)} + \operatorname{span}\{d^{(0)}, d^{(1)}, \dots, d^{(k-1)}\}$$

which satisfies (∗), since

$$\nabla\phi\Big(x^{(0)} + \sum_{j=0}^{k-1} \alpha^{(j)}\, d^{(j)}\Big)^\intercal d^{(i)} = \nabla\phi(x^{(k)})^\intercal d^{(i)} = (r^{(k)})^\intercal d^{(i)} = 0$$

holds for all $i = 0, \ldots, k-1$, as shown in Statement (i). □

Corollary 4.15 (Properties of conjugate direction methods). *Any iterative method* (4.19) *using A-conjugate directions $d^{(k)}$ and Cauchy step sizes* (4.20) *converges to the unique solution of* (4.1) *in at most n steps.*

Proof. The search directions $d^{(k)}$ are A-conjugate and thus linearly independent. Therefore,

$$\operatorname{span}\{d^{(0)}, d^{(1)}, \ldots, d^{(n-1)}\}$$

is all of $\mathbb{R}^n$, so that $x^{(n)}$ minimizes ϕ over all of $\mathbb{R}^n$ by Lemma 4.14. □

In practice, the statement of Corollary 4.15 is weakened by floating point error. Moreover, the result of Corollary 4.15 is not really relevant for high-dimensional problems since performing n iterations is prohibitively expensive. We will later see more practical converge estimates.

There are many possibilities to generate pairwise A-conjugate directions $d^{(k)}$, each of which leads to a different conjugate direction method. The **conjugate gradient method** (**CG method**) determines the current direction $d^{(k)}$ as a linear combination of the previous direction $d^{(k-1)}$ and the current steepest descent direction $-M^{-1}r^{(k)}$:[11]

$$\begin{aligned} d^{(0)} &:= -M^{-1}r^{(0)} && \text{for } k = 0, \\ d^{(k)} &:= -M^{-1}r^{(k)} + \beta^{(k)}\, d^{(k-1)} && \text{for } k \ge 1. \end{aligned} \tag{4.23}$$

The coefficient $\beta^{(k)}$ is determined in such a way that at least $d^{(k)}$ and $d^{(k-1)}$ are A-conjugate:

$$\beta^{(k)} := \frac{(r^{(k)})^\intercal M^{-1} A\, d^{(k-1)}}{(d^{(k-1)})^\intercal A\, d^{(k-1)}}. \tag{4.24}$$

Interestingly, the algorithm obtained in this way generates search directions which are fully A-conjugate, as shown in the following result.

Lemma 4.16 (Properties of the iterates in the CG algorithm, see Nocedal, Wright, 2006, Theorem 5.3). *Suppose that $x^{(0)} \in \mathbb{R}^n$ is given and that the search directions $\{d^{(0)}, d^{(1)}, \ldots, d^{(k)}\}$ and the subsequent iterates $x^{(1)}, \ldots, x^{(k)}$, $k \ge 1$, are generated according to* (4.19)–(4.20), (4.23)–(4.24)*, where $\alpha^{(k)} \neq 0$.*[12]

$$\begin{aligned} \operatorname{span}\{r^{(0)}, r^{(1)}, \ldots, r^{(k)}\} &= \phantom{M^{-1}}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\, r^{(0)}, \ldots, (A\,M^{-1})^k\, r^{(0)}\}, && (4.25) \\ \operatorname{span}\{d^{(0)}, d^{(1)}, \ldots, d^{(k)}\} &= M^{-1} \operatorname{span}\{r^{(0)}, (A\,M^{-1})\, r^{(0)}, \ldots, (A\,M^{-1})^k\, r^{(0)}\}, && (4.26) \\ (d^{(k)})^\intercal A\, d^{(i)} &= 0 \quad \textit{for all } i = 0, 1, \ldots, k-1, && (4.27) \\ (r^{(k)})^\intercal M^{-1} r^{(i)} &= 0 \quad \textit{for all } i = 0, 1, \ldots, k-1. && (4.28) \end{aligned}$$

[11] With $\beta^{(k)} = 0$, we obtain again the steepest descent method (Algorithm 4.6).

[12] $\alpha^{(k)} = 0$ would mean that $x^{(k)}$ is the unique solution x^*. Due to the form of the Cauchy step (4.20), this is clear for $k = 0$, as the nominator is $\|r^{(k)}\|_{M^{-1}}$. (4.22) shows that this is also true for $k > 0$.

The subspace

$$\mathcal{K}^{(k+1)}(A\,M^{-1}; r^{(0)}) \coloneqq \operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^k\,r^{(0)}\} \tag{4.29}$$

is termed the **Krylov subspace** (of order $k+1$) of the matrix $A\,M^{-1}$ with initial vector $r^{(0)}$. Therefore, the CG method is a representative of the class of **Krylov subspace methods**. The properties (4.25) and (4.26) imply that the method creates, simultaneously, an expanding sequence of M^{-1}-orthogonal basis vectors of the spaces $\mathcal{K}^{(k)}(A\,M^{-1}; r^{(0)})$, as well as an expanding sequence of A-orthogonal basis vectors of the spaces $M^{-1}\mathcal{K}^{(k)}(A\,M^{-1}; r^{(0)})$.

Proof. We first prove (4.25)–(4.27), by induction. For $k = 0$, statement (4.25) holds trivially. Statement (4.26) holds since the CG method starts with $d^{(0)} = -M^{-1}r^{(0)}$. Statement (4.27) is void for $k = 0$.

Suppose now that (4.25) and (4.26) have been shown up to some $k \ge 0$. We need to show that they also hold for $k+1$. By hypothesis,

$$\begin{aligned}
r^{(k)} &\in \phantom{A\,M^{-1}}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^k\,r^{(0)}\},\\
d^{(k)} &\in M^{-1}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^k\,r^{(0)}\},\\
\text{hence } A\,d^{(k)} &\in A\,M^{-1}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^k\,r^{(0)}\}\\
&= \phantom{A\,M^{-1}}\operatorname{span}\{(A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^{k+1}\,r^{(0)}\}.
\end{aligned}$$

Due to the residual recursion (4.21), we therefore have

$$\begin{aligned}
r^{(k+1)} &= r^{(k)} + \alpha^{(k)} A\,d^{(k)}\\
&\in \operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^k\,r^{(0)}\} + \operatorname{span}\{(A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^{k+1}\,r^{(0)}\}\\
&= \operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^{k+1}\,r^{(0)}\}.
\end{aligned} \tag{$*$}$$

Due to the induction hypothesis for (4.25), the same statement ($*$) holds when $k+1$ is replaced by a smaller index. Therefore, we have shown that

$$\operatorname{span}\{r^{(0)}, r^{(1)}, \ldots, r^{(k+1)}\} \subseteq \operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^{k+1}\,r^{(0)}\}$$

holds. Now for the reverse inequality. By the induction hypothesis for (4.26), we find

$$A\,M^{-1}(A\,M^{-1})^k r^{(0)} \in A\ \operatorname{span}\{d^{(0)}, d^{(1)}, \ldots, d^{(k)}\} = \operatorname{span}\{A\,d^{(0)}, A\,d^{(1)}, \ldots, A\,d^{(k)}\}.$$

By the residual recursion (4.21), specifically

$$A\,d^{(i)} = \frac{1}{\alpha^{(i)}}\big(r^{(i+1)} - r^{(i)}\big) \in \operatorname{span}\{r^{(i)}, r^{(i+1)}\}$$

for $i = 0, 1, \ldots, k$, it follows that

$$A\,M^{-1}(A\,M^{-1})^k r^{(0)} \in \operatorname{span}\{r^{(0)}, r^{(1)}, \ldots, r^{(k+1)}\}.$$

When combined with the induction hypothesis for (4.25), i. e.,

$$\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \ldots, (A\,M^{-1})^k\,r^{(0)}\} = \operatorname{span}\{r^{(0)}, r^{(1)}, \ldots, r^{(k)}\},$$

we find the desired reverse inequality

$$\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \dots, (A\,M^{-1})^{k+1}\,r^{(0)}\} \subseteq \operatorname{span}\{r^{(0)}, r^{(1)}, \dots, r^{(k+1)}\}.$$

Thus the induction step for (4.25) is complete.

To see (4.26),

$$\begin{aligned}
&\operatorname{span}\{d^{(0)}, \dots, d^{(k)}, d^{(k+1)}\} \\
&\qquad = \operatorname{span}\{d^{(0)}, \dots, d^{(k)}, M^{-1}r^{(k+1)}\} && \text{by (4.23)} \\
&\qquad = M^{-1}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \dots, (A\,M^{-1})^{k}\,r^{(0)}, r^{(k+1)}\} && \text{by (4.26)} \\
&\qquad = M^{-1}\operatorname{span}\{r^{(0)}, r^{(1)}, \dots, r^{(k)}, r^{(k+1)}\} && \text{by (4.25)} \\
&\qquad = M^{-1}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}, \dots, (A\,M^{-1})^{k}\,r^{(0)}, (A\,M^{-1})^{k+1}\,r^{(0)}\} && \text{by (4.25) for } k+1.
\end{aligned}$$

This concludes the induction step for (4.26).

Next we address the A-conjugacy of search directions, (4.27). By the induction hypothesis, the directions $d^{(0)}, \dots, d^{(k)}$ are pairwise A-conjugate. Consider

$$(d^{(k+1)})^\intercal A\,d^{(i)} = (-M^{-1}r^{(k+1)} + \beta^{(k+1)}\,d^{(k)})^\intercal A\,d^{(i)} \tag{$**$}$$

for $i = 0, \dots, k$. In case $i = k$, we have

$$(d^{(k+1)})^\intercal A\,d^{(k)} = 0$$

by construction of the search direction $d^{(k+1)}$, see (4.23) and (4.24). When $i \le k-1$, we argue as follows. From (4.26), we obtain

$$\begin{aligned}
M^{-1}A\,d^{(0)} &\in M^{-1}A\,M^{-1}\operatorname{span}\{r^{(0)}\} && \subseteq \operatorname{span}\{d^{(0)}, d^{(1)}\}, \\
M^{-1}A\,d^{(1)} &\in M^{-1}A\,M^{-1}\operatorname{span}\{r^{(0)}, (A\,M^{-1})\,r^{(0)}\} && \subseteq \operatorname{span}\{d^{(0)}, d^{(1)}, d^{(2)}\}, \\
\vdots \quad & \quad \vdots && \quad \vdots \\
M^{-1}A\,d^{(k-1)} &\in M^{-1}A\,M^{-1}\operatorname{span}\{r^{(0)}, \dots, (A\,M^{-1})^{k-1}\,r^{(0)}\} && \subseteq \operatorname{span}\{d^{(0)}, \dots, d^{(k)}\}.
\end{aligned}$$

We thus find that, for any $i \le k-1$, the term $(r^{(k+1)})^\intercal M^{-1}A\,d^{(i)}$ in ($**$) belongs to

$$(r^{(k+1)})^\intercal \operatorname{span}\{d^{(0)}, \dots, d^{(i+1)}\} = \operatorname{span}\{(r^{(k+1)})^\intercal d^{(0)}, \dots, (r^{(k+1)})^\intercal d^{(i+1)}\}.$$

By (4.22), however, $(r^{(k+1)})^\intercal d^{(j)} = 0$ for $j = 0, \dots, k$. Therefore, ($**$) reduces to

$$(d^{(k+1)})^\intercal A\,d^{(i)} = \beta^{(k+1)}\,(d^{(k)})^\intercal A\,d^{(i)}. \tag{$***$}$$

By the induction hypothesis, this is equal to zero, which concludes the induction step for (4.27).

Finally, we consider the M^{-1}-conjugacy of residuals, (4.28), for $k \ge 1$. We do not need an induction argument for this. We consider two cases for $(r^{(k)})^\intercal M^{-1}r^{(i)}$:

(1) In case $i = k-1$, we have

$$(r^{(k)})^\top M^{-1} r^{(k-1)} = \begin{cases} \overbrace{(r^{(k-1)} + \alpha^{(k-1)}\, A\, d^{(k-1)})^\top(-d^{(k-1)}}^{(\square)} + \beta^{(k-1)}\, d^{(k-2)}) & \text{for } k \ge 2 \\ \underbrace{(r^{(k-1)} + \alpha^{(k-1)}\, A\, d^{(k-1)})^\top(-d^{(k-1)})}_{(\square)} & \text{for } k = 1 \end{cases}$$

by the residual recursion (4.21) and the construction of search directions (4.23). Since the Cauchy step size satisfies $\alpha^{(k-1)} = -\frac{(d^{(k-1)})^\top r^{(k-1)}}{(d^{(k-1)})^\top A\, d^{(k-1)}}$, the term $(\square)$ is equal to zero for all $k \ge 1$. Let us consider the remaining terms when $k \ge 2$. We obtain

$$\begin{aligned} (r^{(k-1)})^\top d^{(k-2)} &= 0 \quad \text{due to (4.22)}, \\ (A\, d^{(k-1)})^\top (d^{(k-2)}) &= 0 \quad \text{owing to the } A\text{-conjugacy of search directions.} \end{aligned}$$

Therefore we conclude that $(r^{(k)})^\top M^{-1} r^{(k-1)} = 0$ holds for all $k \ge 1$.

(2) in case $i < k-1$, we have

$$(r^{(k)})^\top M^{-1} r^{(i)} = \begin{cases} (r^{(k-1)} + \alpha^{(k-1)}\, A\, d^{(k-1)})^\top(-d^{(i)} + \beta^{(i)}\, d^{(i-1)}) & \text{for } i \ge 1 \\ (r^{(k-1)} + \alpha^{(k-1)}\, A\, d^{(k-1)})^\top(-d^{(i)}) & \text{for } i = 0 \end{cases}$$

When expanding, we obtain terms of the types (note $i < k-1$)

$$\begin{aligned} (r^{(k-1)})^\top d^{(i)} &= 0 \quad \text{due to (4.22)}, \\ (A\, d^{(k-1)})^\top d^{(i)} &= 0 \quad \text{owing to the } A\text{-conjugacy of search directions}, \\ (r^{(k-1)})^\top d^{(i-1)} &= 0 \quad \text{due to (4.22)}, \\ (A\, d^{(k-1)})^\top d^{(i-1)} &= 0 \quad \text{owing to the } A\text{-conjugacy of search directions.} \end{aligned}$$

Therefore we conclude that $(r^{(k)})^\top M^{-1} r^{(i)} = 0$ holds for all $k \ge 1$ and $0 \le i < k-1$. □

Using the properties of the iterates shown above, the equations (4.20) for $\alpha^{(k)}$ as well as (4.24) for $\beta^{(k)}$ in the CG method can be equivalently formulated as follows:

$$\begin{aligned} \alpha^{(k)} &= -\frac{(r^{(k)})^\top d^{(k)}}{(d^{(k)})^\top A\, d^{(k)}} && \text{by the Cauchy step size formula (4.20)} \\ &= \frac{(r^{(k)})^\top M^{-1} r^{(k)}}{(d^{(k)})^\top A\, d^{(k)}} - \beta^{(k)} \frac{(r^{(k)})^\top d^{(k-1)}}{(d^{(k)})^\top A\, d^{(k)}} && \text{by the search direction recursion (4.23)} \\ &= \frac{(r^{(k)})^\top M^{-1} r^{(k)}}{(d^{(k)})^\top A\, d^{(k)}} && \text{by (4.22)} \end{aligned} \tag{4.20'}$$

and

$$\begin{aligned} \beta^{(k+1)} &= \frac{(r^{(k+1)})^\top M^{-1} A\, d^{(k)}}{(d^{(k)})^\top A\, d^{(k)}} && \text{by the orthogonalization coefficient (4.24)} \\ &= \frac{(r^{(k+1)})^\top M^{-1} (r^{(k+1)} - r^{(k)})}{(d^{(k)})^\top (r^{(k+1)} - r^{(k)})} && \text{by the residual recursion (4.21)} \\ &= \frac{(r^{(k+1)})^\top M^{-1} (r^{(k+1)} - r^{(k)})}{(-M^{-1} r^{(k)} + \beta^{(k)}\, d^{(k-1)})^\top (r^{(k+1)} - r^{(k)})} && \text{by the construction of search directions (4.23)} \\ &= \frac{(r^{(k+1)})^\top M^{-1} r^{(k+1)}}{(r^{(k)})^\top M^{-1} r^{(k)}} && \text{by (4.22) and (4.25).} \end{aligned} \tag{4.24'}$$

The relations (4.20') and (4.24') are also true for $k = 0$.

We have now obtained the common form of the CG method w.r.t. the M-inner product, commonly referred to as the **preconditioned conjugate gradient method**.

Algorithm 4.17 (Conjugate gradient method for (4.1) w.r.t. the M-inner product).

```
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 $A\,x = b$
Set $k := 0$
Set $r^{(0)} := A\,x^{(0)} - b$    // evaluate the initial residual
Set $d^{(0)} := -M^{-1} r^{(0)}$    // evaluate the initial negative $M$-gradient
Set $\delta^{(0)} := -(r^{(0)})^\intercal d^{(0)}$    // $\delta^{(0)} = \|\nabla_M \phi(x^{(0)})\|_M^2 = \|r^{(0)}\|_{M^{-1}}^2$
while stopping criterion not met do
    Set $q^{(k)} := A\,d^{(k)}$
    Set $\theta^{(k)} := (q^{(k)})^\intercal d^{(k)}$
    Set $\alpha^{(k)} := \delta^{(k)} / \theta^{(k)}$    // evaluate the Cauchy step size
    Set $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$    // update the iterate
    Set $r^{(k+1)} := r^{(k)} + \alpha^{(k)} q^{(k)}$    // update the residual
    Set $d^{(k+1)} := -M^{-1} r^{(k+1)}$    // evaluate the negative $M$-gradient
    Set $\delta^{(k+1)} := -(r^{(k+1)})^\intercal d^{(k+1)}$    // $\delta^{(k+1)} = \|\nabla_M \phi(x^{(k+1)})\|_M^2 = \|r^{(k+1)}\|_{M^{-1}}^2$
    Set $\beta^{(k+1)} := \delta^{(k+1)} / \delta^{(k)}$    // evaluate the $A$-orthogonalization coefficient
    Set $d^{(k+1)} := d^{(k+1)} + \beta^{(k+1)}\, d^{(k)}$    // make $d^{(k+1)}$ $A$-orthogonal w.r.t. $d^{(k)}$
    Set $k := k + 1$
end while
return $x^{(k)}$
```

Remark 4.18 (on Algorithm 4.17).

(*i*) *From Lemma 4.16 we know that the CG method generates pairwise A-orthogonal directions, although it only needs to orthogonalize any new direction $d^{(k+1)}$ against the most recent one, $d^{(k)}$. This phenomenon, known as **short-term recurrence**, is possible due to the symmetry of A.*

(*ii*) *The conjugate thus keeps a memory of previously visited directions, although this memory is mainly implicit. As shown in Algorithm 4.17, we can implement the method with a constant amount of storage.*

(*iii*) *The implementation of the CG method is very similar to the steepest descent method (Algorithm 4.6). The only (but significant!) difference lies in the fact that we A-orthogonalize the steepest descent direction against $d^{(k)}$ before we use it as the new search direction $d^{(k+1)}$. The initial search direction $d^{(0)}$ is the steepest descent direction for ϕ at $x^{(0)}$. Consequently, the iterate $x^{(1)}$ is the same for the conjugate gradient method and the steepest descent method with Cauchy step size (Algorithm 4.6).*

(*iv*) *The name* ***conjugate gradient method*** *is a bit of a misnomer, since it is not the gradients which are A-conjugate, but rather the search directions $d^{(k)}$.*

(*v*) *Remark 4.7 remains valid for the conjugate gradient method as well, with minor modifications. We need to store one additional vector since $d^{(k)}$ and $d^{(k+1)}$ are needed simultaneously.*

(*vi*) *The stopping criteria (4.14) and their consequences (4.15) continue to hold since they depend on the same computable quantity $\|r^{(k)}\|_{M^{-1}}$ as in the steepest descent method.*

Our next goal is to establish a convergence result for the conjugate gradient method, and to compare it to Theorem 4.8 for the steepest descent method with Cauchy step size. A major difference is that we will not obtain a result about the reduction of the error from iteration to iteration, but rather a result about the reduction of the error compared with its initial value.

Theorem 4.19 (Convergence of Algorithm 4.17, compare Theorem 4.8)**.** *Suppose that $A \in \mathbb{R}^{n\times n}$ and 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 conjugate gradient method 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*[13]

$$\phi(x^{(k)}) - \phi(x^*) \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \left(\phi(x^{(0)}) - \phi(x^*)\right) \tag{4.30a}$$

$$\|x^{(k)} - x^*\|_A \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{k} \|x^{(0)} - x^*\|_A, \tag{4.30b}$$

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

Proof. Since the search directions, by (4.26), span $M^{-1}\mathcal{K}^{(k)}(A\,M^{-1}; r^{(0)})$, we have

$$x^{(k)} - x^{(0)} \in M^{-1}\mathcal{K}^{(k)}(A\,M^{-1}; r^{(0)}).$$

In other words, we have

$$x^{(k)} - x^{(0)} = q^{(k-1)}(M^{-1}A)\,M^{-1}r^{(0)}$$

for some polynomial $q^{(k-1)}$ in the matrix $M^{-1}A$ of degree at most $k-1$. Abbreviating $e^{(k)} := x^{(k)} - x^*$ and using $A\,e^{(0)} = A\,x^{(0)} - A\,x^* = r^{(0)}$, we can manipulate this equation into

$$\begin{aligned} e^{(k)} &= e^{(0)} + q^{(k-1)}(M^{-1}A)\,M^{-1}r^{(0)} \\ &= e^{(0)} + q^{(k-1)}(M^{-1}A)\,M^{-1}A\,e^{(0)} \\ &= \left[\mathrm{Id} + q^{(k-1)}(M^{-1}A)\,M^{-1}A\right] e^{(0)} \\ &= p^{(k)}(M^{-1}A)\,e^{(0)}, \end{aligned}$$

where now $p^{(k)}$ is a polynomial of degree at most k satisfying $p^{(k)}(0) = 1$.

[13] compare (4.13c), (4.13d)

By construction, the conjugate gradient method minimizes $\|e^{(k)}\|_A$ in every iteration. We can now express this in terms of a minimization over the vector space Π_k of polynomials of degree $\le k$:

$$\|e^{(k)}\|_A = \min\Big\{\|p(M^{-1}A)\, e^{(0)}\|_A \,\Big|\, p \in \Pi_k,\ p(0) = 1\Big\}. \tag{4.31}$$

We expand the initial error $e^{(0)}$ in terms of the basis of eigenvectors of A w.r.t. M; see (2.10), (2.11). Suppose we denote the generalized eigenpairs by $(\lambda^{(j)}, v^{(j)})$, we can write

$$e^{(0)} = \sum_{j=1}^{n} \gamma^{(j)} v^{(j)}$$

with some coefficients $\gamma^{(j)}$ determined by $e^{(0)}$. We can thus manipulate the objective in the minimization problem above as follows:

$$\begin{aligned}\|p(M^{-1}A)\, e^{(0)}\|_A &= \Big\|p(M^{-1}A)\,\Big(\sum_{j=1}^{n} \gamma^{(j)} v^{(j)}\Big)\Big\|_A \\ &= \Big\|\sum_{j=1}^{n} \gamma^{(j)} p(M^{-1}A)\, v^{(j)}\Big\|_A\end{aligned}$$

In view of $A\, v^{(j)} = \lambda^{(j)} M\, v^{(j)}$ and thus $M^{-1}A\, v^{(j)} = \lambda^{(j)} v^{(j)}$, this is

$$= \Big\|\sum_{j=1}^{n} \gamma^{(j)} p(\lambda^{(j)})\, v^{(j)}\Big\|_A.$$

By pulling the maximal value of $|p(\lambda^{(j)})|$ out of the sum (**Quiz 4.6:** Can you fill in the details why this is possible?), we can estimate this quantity further:

$$\begin{aligned}&\le \max_{j=1,\ldots,n} |p(\lambda^{(j)})| \,\Big\|\sum_{j=1}^{n} \gamma^{(j)} v^{(j)}\Big\|_A \\ &= \max_{j=1,\ldots,n} |p(\lambda^{(j)})| \,\big\|e^{(0)}\big\|_A.\end{aligned}$$

Combining this with (4.31), we see

$$\begin{aligned}\|e^{(k)}\|_A &\le \min\Big\{\max_{j=1,\ldots,n} |p(\lambda^{(j)})|\, \|e^{(0)}\|_A \,\Big|\, p \in \Pi_k,\ p(0) = 1\Big\} \\ &= \min\Big\{\max_{j=1,\ldots,n} |p(\lambda^{(j)})| \,\Big|\, p \in \Pi_k,\ p(0) = 1\Big\}\, \|e^{(0)}\|_A\end{aligned}$$

and since the eigenvalues lie in the interval $[\alpha, \beta]$,

$$\|e^{(k)}\|_A \le \min\Big\{\max_{z \in [\alpha,\beta]} |p(z)| \,\Big|\, p \in \Pi_k,\ p(0) = 1\Big\}\, \|e^{(0)}\|_A. \tag{4.32}$$

We have thus estimated $\frac{\|e^{(k)}\|_A}{\|e^{(0)}\|_A}$ by the smallest maximal absolute value any polynomial $p \in \Pi_k$ with $p(0) = 1$ can attain on the interval $[\alpha, \beta]$ spanning all generalized eigenvalues of A w.r.t. M.

The question about the *optimal* polynomial in (4.32) can be answered by Chebyshev polynomials; we refer you to Elman, Silvester, Wathen, 2014, Theorem 2.4 if you want to know more details. It turns out that the optimal value

$$\min\Big\{ \max_{z\in[\alpha,\beta]} |p(z)| \,\Big|\, p \in \Pi_k,\ p(0) = 1\Big\}$$

depends only on $\kappa = \beta/\alpha$ and it is given by

$$\begin{aligned} &= 2 \left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} \right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^k \right]^{-1} \\ &\le 2 \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^k . \end{aligned}$$

From there, we finally obtain

$$\|e^{(k)}\|_A \le 2 \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^k \|e^{(0)}\|_A,$$

which is precisely (4.30b). Squaring both sides and dividing by 2, we also obtain (4.30a). □

Corollary 4.20 (Maximal number of iterations required in Algorithm 4.17, compare Corollary 4.9)**.** *Given positive numbers ε_1 and ε_2, it takes*

$$\begin{aligned} k &\le \left\lceil \frac{\sqrt{\kappa}}{4} \ln\left(\frac{2}{\varepsilon_1}\right) \right\rceil \text{ iterations until } 2 \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^{2k} \le \varepsilon_1, \\ k &\le \left\lceil \frac{\sqrt{\kappa}}{2} \ln\left(\frac{2}{\varepsilon_2}\right) \right\rceil \text{ iterations until } 2 \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^{k} \le \varepsilon_2. \end{aligned}$$

Proof. The proof is similar to Corollary 4.9 and it uses that

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

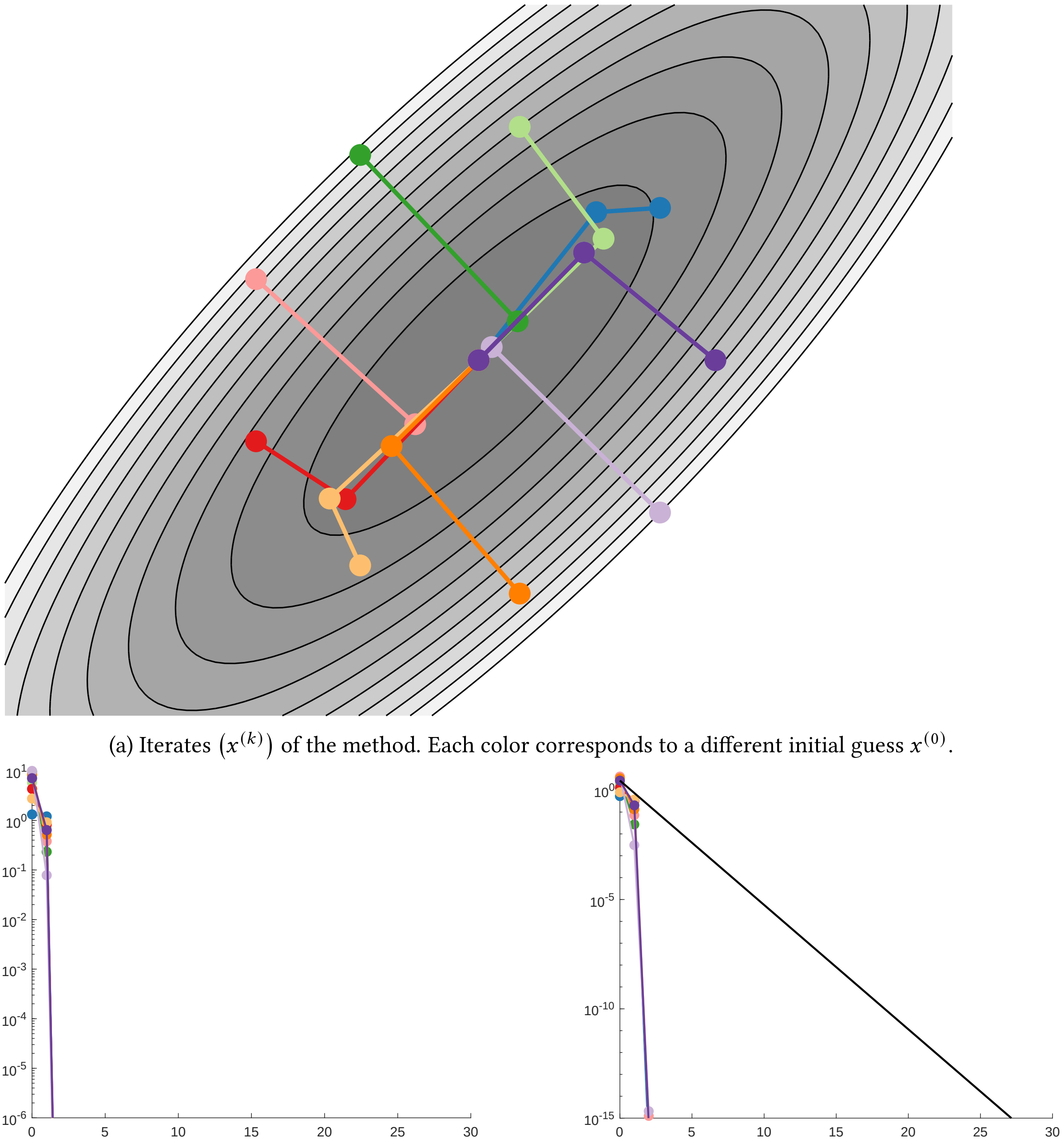
holds for all $\kappa \ge 1$. □

Remark 4.21 (on Theorem 4.19)**.**

(*i*) *The estimates* (4.30a) *and* (4.30b) *establish the R-linear convergence of the respective quantities to zero.*

(*ii*) *Compared to the estimates* (4.13c) *and* (4.13d) *for the gradient descent method, we obtain the reduction factor* $\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k$ *in place of* $\left(\frac{\kappa-1}{\kappa+1}\right)^k$*, which is generally much better.*

(*iii*) *The superiority of the CG method compared to the gradient descent method is also reflected in the estimates for the maximal iteration numbers to achieve a certain reduction in the quantities* $\phi(x^{(k)}) - \phi(x^*)$ *and* $\|x^{(k)} - x^*\|_A$*, respectively. The bounds for the maximal iteration numbers are proportional to* $\sqrt{\kappa}$ *for the CG method, not proportional to* κ*.*

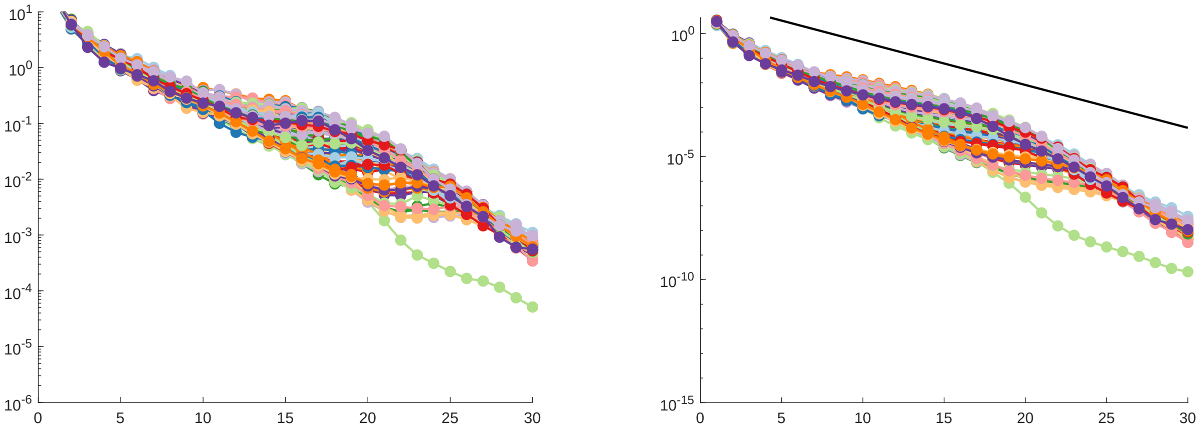


(a) Iterates $(x^{(k)})$ of the method. Each color corresponds to a different initial guess $x^{(0)}$.

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

(c) The objective values $\phi(x^{(k)}) - \phi(x^*)$ converge monotonically. The black line illustrates the bound (4.30a).

Figure 4.3: Illustration of the convergence behavior of Algorithm 4.17 from a number of initial guesses $x^{(0)}$. No preconditioning ($M = \text{Id}$) is used. The two eigenvalues of the matrix are $\alpha = 1$ and $\beta = 10$ so the condition number is $\kappa = 10$.



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

(b) The objective values $\phi(x^{(k)}) - \phi(x^*)$ converge monotonically. The black line illustrates the bound (4.30a).

Figure 4.4: Illustration of the convergence behavior of Algorithm 4.17 from a number of initial guesses $x^{(0)}$. No preconditioning ($M = \mathrm{Id}$) is used. Here A is a random matrix of dimension 100×100 with eigenvalues in the interval $[\alpha, \beta] = [1, 100]$ so that the condition number is $\kappa = 100$.

(*iv*) *As was the case for Theorem 4.8, the estimates of Theorem 4.19 are worst-case estimates since they do not depend on the initial guess $x^{(0)}$. In fact, as can be seen in Figure 4.3c and Figure 4.4b, the actual contraction factor for the objective values can be significantly smaller for some initial guesses than the estimate* (4.30a) *suggests.*

(*v*) *Other informative error bounds than* (4.30) *and* (4.30b) *and convergence results can be obtained by proceeding as in the proof of Theorem 4.19 and choosing other polynomials to bound the error with.*

The iterates of the conjugate gradient method have a further remarkable property, which we will exploit later on:

Lemma 4.22 (Growth of the distance from the initial guess[14]). *Consider the iterates $x^{(k)}$ of the conjugate gradient method (Algorithm 4.17). As long as $x^{(k)} \neq x^*$ holds, the sequence $\|x^{(k)} - x^{(0)}\|_M$ is strictly increasing.*

Note: The steepest descent method does not have this property.

Proof. Statement (*i*) in Lemma 4.14 implies that

$$(r^{(k)})^\intercal (x^{(k)} - x^{(0)}) = \sum_{i=0}^{k-1} \alpha_i \underbrace{(r^{(k)})^\intercal d^{(i)}}_{=0} = 0 \quad \text{for all } k \geq 0. \tag{$*$}$$

[14] In the literature, we find this result often only for the case $x^{(0)} = 0$, see for instance Nocedal, Wright, 2006, Theorem 7.3.

We now show by induction that $(x^{(k)} - x^{(0)})^\intercal M\, d^{(k)} > 0$ holds for $k \ge 1$. Initially, for $k = 1$, Statement (i) in Lemma 4.14 once again yields

$$\begin{aligned}(x^{(1)} - x^{(0)})^\intercal M\, d^{(1)} &= \alpha^{(0)}\, (d^{(0)})^\intercal M\, (\overbrace{-M^{-1} r^{(1)}}^{=0} + \beta^{(1)} d^{(0)}) \\ &= \underbrace{\alpha^{(0)}}_{>0}\, \underbrace{\beta^{(1)}}_{>0}\, \underbrace{(d^{(0)})^\intercal M\, d^{(0)}}_{>0} \\ &> 0.\end{aligned}$$

We now proceed with the step from index k to $k+1$:

$$\begin{aligned}(x^{(k+1)} - x^{(0)})^\intercal M\, d^{(k+1)} &= (x^{(k+1)} - x^{(0)})^\intercal M\, (-M^{-1} r^{(k+1)} + \beta^{(k+1)}\, d^{(k)}) \\ &= \beta^{(k+1)}\, (x^{(k+1)} - x^{(0)})^\intercal M\, d^{(k)} && \text{by } (*) \\ &= \beta^{(k+1)}\, (x^{(k)} + \alpha^{(k)} d^{(k)} - x^{(0)})^\intercal M\, d^{(k)} \\ &= \beta^{(k+1)}\, (x^{(k)} - x^{(0)})^\intercal M\, d^{(k)} + \alpha^{(k)} \beta^{(k+1)}\, (d^{(k)})^\intercal M\, d^{(k)} \\ &> 0. && (**)\end{aligned}$$

Due to the induction hypothesis as well as $\alpha^{(k)} > 0$, $\beta^{(k+1)} > 0$ and $(d^{(k)})^\intercal M\, d^{(k)} > 0$, the entire expression is positive.

The desired result now easily follows from

$$\begin{aligned}\|x^{(k+1)} - x^{(0)}\|_M^2 &= \|x^{(k)} + \alpha^{(k)} d^{(k)} - x^{(0)}\|_M^2 \\ &= \|x^{(k)} - x^{(0)}\|_M^2 + 2\, \underbrace{\alpha^{(k)}}_{>0} \underbrace{(x^{(k)} - x^{(0)})^\intercal M\, d^{(k)}}_{>0} + \underbrace{(\alpha^{(k)})^2\, \|d^{(k)}\|_M^2}_{>0}. \qquad (***)\end{aligned}$$

□

The relations $(**)$ and $(***)$ allow us to compute the informative quantities

$$\omega^{(k)} := \|x^{(k)} - x^{(0)}\|_M^2 \tag{4.33a}$$

$$\xi^{(k)} := (x^{(k)} - x^{(0)})^\intercal M\, d^{(k)} \tag{4.33b}$$

$$\gamma^{(k)} := \|d^{(k)}\|_M^2 \tag{4.33c}$$

on the side without any noticeable effort. This can be achieved by inserting, at the appropriate positions in Algorithm 4.17 (**Quiz 4.7:** Where?), the relations

$$\omega^{(0)} := 0, \qquad \omega^{(k+1)} := \omega^{(k)} + 2\, \alpha^{(k)} \xi^{(k)} + (\alpha^{(k)})^2\, \gamma^{(k)} \qquad \text{see } (***) \tag{4.34a}$$

$$\xi^{(0)} := 0, \qquad \xi^{(k+1)} := \beta^{(k+1)}\, (\xi^{(k)} + \alpha^{(k)} \gamma^{(k)}) \qquad \text{see } (**) \tag{4.34b}$$

$$\gamma^{(0)} := \delta^{(0)}, \qquad \gamma^{(k+1)} := \delta^{(k+1)} + (\beta^{(k+1)})^2\, \gamma^{(k)} \qquad \text{(confirm for yourself).} \tag{4.34c}$$

The remarkable fact about this is the possibility to keep track of (4.33) without requiring access to the matrix M, or even matrix-vector products with M. Notice that we usually do not have the latter since we only need matrix-vector products with M^{-1} in Algorithm 4.17.

End of Week 3

§ 5 Line Search Methods for Nonlinear Unconstrained Problems

We consider in this section a large class of methods to solve general, nonlinear unconstrained problems

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

The methods we consider are so-called **line search methods**. In every iteration, a line search method first determines a **search direction** and subsequently finds a **step size** (or **step length**) $\alpha^{(k)}$, that leads to the next iterate via

$$x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}.$$

Assumption 5.1. *Throughout § 5 we are assuming that $f\colon \mathbb{R}^n \to \mathbb{R}$ is a C^1 function.*

Most line search methods, in particular the ones we consider, require that $d^{(k)}$ is a **descent direction** for the objective f at the current iterate $x^{(k)}$, i. e., that

$$f'(x^{(k)})\, d^{(k)} < 0 \tag{5.1}$$

holds, see Definition 4.4. This implies that we have descent at least for sufficiently small positive step sizes $\alpha^{(k)}$,

$$f(x^{(k+1)}) = f(x^{(k)} + \alpha^{(k)} d^{(k)}) < f(x^{(k)})$$

and it motivates the term **descent method**.

Most methods[15] we are discussing in § 5 determine the search direction $d^{(k)}$ by considering a local **quadratic model** of the objective:

$$q^{(k)}(d) = f(x^{(k)}) + f'(x^{(k)})\, d + \frac{1}{2}\, d^\mathsf{T} H^{(k)} d. \tag{5.2}$$

This model uses the data $f(x^{(k)})$ and $f'(x^{(k)})$ at the iterate $x^{(k)}$ and it agrees with f regarding that data at $d = 0$:

$$\begin{aligned} q^{(k)}(0) &= f(x^{(k)}) \\ \text{and} \quad (q^{(k)})'(0) &= f'(x^{(k)}) \end{aligned}$$

The matrix $H^{(k)}$ is the Hessian of the model, briefly: the **model Hessian**. In case $H^{(k)} = f''(x^{(k)})$, the model $q^{(k)}$ is the second-order Taylor polynomial of f at $x^{(k)}$. However, in general, the model Hessian is chosen to be any symmetric and possibly positive definite matrix. In fact, different line search methods differ w.r.t. their choice of the model Hessians $H^{(k)}$, and thus with respect to the search directions they use.

The search direction $d^{(k)}$ is obtained by minimizing (possibly only to a certain accuracy) the quadratic polynomial $q^{(k)}$:

$$\text{Minimize} \quad q^{(k)}(d), \quad d \in \mathbb{R}^n. \tag{5.3}$$

As we know from Lemma 4.1, the following cases can occur:

[15] with the exception of nonlinear conjugate gradient methods in § 5.8

(*i*) When $H^{(k)}$ is s. p. d., then the unique solution of (5.3) is given by the unique solution of the linear system

$$H^{(k)} d^{(k)} = -\nabla f(x^{(k)}). \tag{5.4}$$

(*ii*) When $H^{(k)}$ is symmetric and only positive semidefinite, then (5.3) is either unbounded, or else has infinitely many minimizers. In any case, the minimizers of (5.3) are precisely the solutions of the linear system (5.4).[16]

(*iii*) When $H^{(k)}$ is symmetric but not positive semidefinite (i. e., at least one eigenvalue of $H^{(k)}$ is negative), then (5.3) is an unbounded problem. However, the linear system (5.4) may still be uniquely solvable, or solvable with multiple solutions, or not solvable. The solutions of the linear systems (if any) are either all saddle points[17] of $q^{(k)}$, or they are all global maximizers. (**Quiz 5.1:** Is this statement clear?)

To solve (5.3) and (5.4), respectively, we can employ the conjugate gradient (CG) method from § 4.6. However, it would be useful to enhance it so that it checks and reacts to the potential occurrence of non-positive eigenvalues in the model Hessian $H^{(k)}$. We will see more details on that later.

§ 5.1 A Generic Descent Method

We begin by considering the following model algorithm of a generic line-search descent method:

Algorithm 5.2 (Generic line-search descent method).

```
Input: initial guess $x^{(0)} \in \mathbb{R}^n$
Input: routine to evaluate $f$ and $f'$ (or $\nabla f$)
Output: approximately stationary point of (UP)
Set $k := 0$
while stopping criterion not met do
    Determine a search direction $d^{(k)}$ such that $f'(x^{(k)})\, d^{(k)} < 0$      // descent direction
    Choose a step size $\alpha^{(k)} > 0$ such that $f(x^{(k)} + \alpha^{(k)} d^{(k)}) < f(x^{(k)})$      // obtain descent
    Set $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$      // take the step
    Set $k := k + 1$
end while
return $x^{(k)}$
```

In order to analyze the convergence properties of this generic algorithm and to determine further requirements for the descent directions and step sizes, we ignore the stopping criterion for now, so that Algorithm 5.2 produces infinite sequences of iterates $x^{(k)}$, search directions $d^{(k)}$ and step sizes $\alpha^{(k)}$. In practice, of course, we will use a stopping criterion to be discussed later.

[16] The solution set of the linear system (5.4) is either the empty set or an affine subspace of $\mathbb{R}^n$ whose dimension agrees with the dimension of $\ker H^{(k)}$.

[17] A stationary point x of f is called a **saddle point** of f if the Hessian $f''(x)$ is indefinite, i. e., has at least one positive and at least one negative eigenvalue.

We will see that, in general, we cannot expect the iterates $x^{(k)}$ to converge overall, but there may be convergent subsequences with different limit points (although this rarely occurs in practice). We recall that the limit points of convergent subsequences $\big(x^{(k^{(\ell)})}\big)$ are precisely the **accumulation points** of $\big(x^{(k)}\big)$.

We would like the accumulation points of the sequence of iterates$\{x^{(k)}\}$ to be "special" points. Therefore, it would be desirable to have the following property:

$$\text{When } x^* \text{ is an accumulation of } \big(x^{(k)}\big)\text{, then } f'(x^*) = 0\text{, i. e., } x^* \text{ is stationary.} \tag{5.5}$$

The relatively weak property (5.5) is often referred to as the **global convergence** of an algorithm. In particular, global convergence does not mean that one obtains a global minimizer. By contrast, it means that one obtains a convergence result (5.5) that is valid for arbitrary initial guesses $x^{(0)}$. Notice that (5.5) does not assert that an accumulation point even exists.[18] It turns out that, in general, we cannot expect more. Under additional assumptions on f, one may be able to show stronger results, for instance

$$\|\nabla f(x^{(k)})\| \text{ has an accumulation point at } 0. \tag{5.6a}$$

$$\text{The entire sequence } \|\nabla f(x^{(k)})\| \text{ converges to } 0. \tag{5.6b}$$

$$\text{Accumulation points of } \big(x^{(k)}\big) \text{ are stationary.} \tag{5.6c}$$

$$\text{The entire sequence } \big(x^{(k)}\big) \text{ converges to a stationary point.} \tag{5.6d}$$

$$\text{The entire sequence } \big(x^{(k)}\big) \text{ converges to a local miminizer.} \tag{5.6e}$$

We will now investigate the minimal requirements on the search directions $d^{(k)}$ and step sizes $\alpha^{(k)}$ in Algorithm 5.2 that ensure global convergence in the sense of (5.5). To this end, two properties are essential:

(1) The search directions $d^{(k)}$ are "good descent directions".

(2) The step sizes $\alpha^{(k)}$ are chosen so that the achievable descent along the search direction $d^{(k)}$ is "sufficiently exploited".

We use the user-defined M-inner product in the space of optimization variables and search directions $\mathbb{R}^n$. Since all norms in $\mathbb{R}^n$ are equivalent, all concepts and properties of algorithms in the remainder of § 5 are *qualitatively independent* of the choice of M. However, the choice of M is still important through its impact on the convergence properties and stopping criteria.

Requirements on the Descent Directions

Definition 5.3 (Admissible search directions). *Suppose that $x^{(k)}$ and $d^{(k)}$ the sequences of iterates and search (descent) directions generated by an algorithm of type Algorithm 5.2. The sequence $d^{(k)}$ of search*

[18] Indeed, an example such as $f(x) = x$ for $x \in \mathbb{R}$ shows that any algorithm with the global convergence property (5.5) couldn't produce an accumulation point, since f has no stationary point.

directions is termed ***admissible*** *in case*

$$\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \to 0 \quad \Rightarrow \quad f'(x^{(k)}) \to 0. \tag{5.7}$$

Note: The admissibility is a property that the sequence of search directions generated by a particular algorithm, applied to a particular problem (objective), started from a particular initial guess may or may not possess. One is, of course, interested in designing algorithms which generate admissible search directions for arbitrary objectives f and initial guesses $x^{(0)}$.

The expression $\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M}$ is the directional derivative of f at $x^{(k)}$ in the direction $d^{(k)}$ normalized. Therefore, we can interpret the condition (5.7) as follows: when the directional derivatives in the normalized search directions converge to zero, then it is due to the derivatives converging to zero and not due to the search directions becoming inefficient (which would be the case if they become essentially M-orthogonal to the steepest descent direction $-\nabla_M f$). This reflects our first goal (item (1) above) that the search directions are "good descent directions". (**Quiz 5.2:** Which search directions are admissible for functions $f : \mathbb{R} \to \mathbb{R}$?)

Condition (5.7) is purely qualitative. By contrast, the **angle condition**

$$\cos \measuredangle \big(\underbrace{-\nabla_M f(x^{(k)})}_{\text{steepest descent direction}}, \overbrace{d^{(k)}}^{\text{chosen search direction}}\big) = \frac{(-\nabla_M f(x^{(k)}), d^{(k)})_M}{\|\nabla_M f(x^{(k)})\|_M \, \|d^{(k)}\|_M} = \frac{-f'(x^{(k)})\, d^{(k)}}{\|f'(x^{(k)})\|_{M^{-1}} \|d^{(k)}\|_M} \ge \eta \tag{5.8}$$

with some $\eta \in (0, 1)$ is a stronger, quantitative condition, which is moreover easy to verify. It means that the angles (as measured in the M-inner product) between the chosen search directions $d^{(k)}$ and the directions of steepest descent $-\nabla_M f(x^{(k)})$ are uniformly bounded away from 90°.

Lemma 5.4 (Angle condition implies admissibility)**.** *Suppose that $x^{(k)}$ and $d^{(k)}$ are the sequences of iterates and search (descent) directions generated by an algorithm of type Algorithm 5.2. If the angle condition (5.8) holds with some $\eta \in (0, 1)$, then the sequence $d^{(k)}$ of search directions is admissible.*

Proof. We have

$$f'(x^{(k)})\, d^{(k)} = \big(\nabla f(x^{(k)}), d^{(k)}\big) = \big(\nabla_M f(x^{(k)}), d^{(k)}\big)_M.$$

The angle condition (5.8) implies

$$-\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \ge \eta \, \|\nabla_M f(x^{(k)})\|_M = \eta \, \|f'(x^{(k)})^\intercal\|_{M^{-1}} \ge 0.$$

When the left-hand term goes to zero, then $f'(x^{(k)})$ must go to zero as well. □

As we already mentioned, almost all of the algorithms we will discuss in detail determine their search directions from the solutions of linear systems

$$H^{(k)} d^{(k)} = -\nabla f(x^{(k)}) \tag{5.4}$$

with a symmetric and possibly positive definite matrix $H^{(k)}$, the model Hessian. In the s. p. d. case, in view of

$$f'(x^{(k)})\, d^{(k)} = -f'(x^{(k)})\big[(H^{(k)})^{-1}\,\nabla f(x^{(k)})\big] = -\nabla f(x^{(k)})^\intercal\,(H^{(k)})^{-1}\,\nabla f(x^{(k)}) < 0, \tag{5.9}$$

$d^{(k)}$ is a descent direction as long as $f'(x^{(k)}) \neq 0$ holds. However, when $H^{(k)}$ is not positive definite, then $d^{(k)}$ may fail to be a descent direction.

In the s. p. d. case, we can show that as long as the sequence of model Hessians remains "well behaved", the sequence of search directions satisfies the angle condition (5.8) and thus is admissible as well.

Lemma 5.5 (Bounded condition numbers imply the angle condition[19]). *Suppose that $x^{(k)}$ and $d^{(k)}$ are the sequences of iterates and search (descent) directions generated by an algorithm of type Algorithm 5.2. Suppose that the search directions are obtained from (5.4), where $H^{(k)} \in \mathbb{R}^{n\times n}$ is a sequence of s. p. d. model Hessians. Suppose, moreover, that the generalized condition numbers of $H^{(k)}$ w.r.t. M satisfy*

$$\kappa(H^{(k)}; M) \coloneqq \frac{\lambda_{\max}(H^{(k)}; M)}{\lambda_{\min}(H^{(k)}; M)} \le \overline{\kappa}.$$

Then the sequence of search directions $d^{(k)}$ satisfies the angle condition (5.8) with

$$\eta = \frac{2\sqrt{\overline{\kappa}}}{\overline{\kappa}+1} \ge \frac{1}{\sqrt{\overline{\kappa}}}.$$

Proof. We perform a couple of equivalent reformulations of the claim to obtain

$$\begin{aligned}
& -\nabla f(x^{(k)})^\intercal d^{(k)} \ge \frac{2\sqrt{\kappa}}{\kappa+1}\|\nabla_M f(x^{(k)})\|_M\,\|d^{(k)}\|_M \\
\Leftrightarrow\quad & (d^{(k)})^\intercal H^{(k)} d^{(k)} \ge \frac{2\sqrt{\kappa}}{\kappa+1}\|M^{-1}H^{(k)}d^{(k)}\|_M\,\|d^{(k)}\|_M \qquad \text{since } H^{(k)}d^{(k)} = -\nabla f(x^{(k)}) \\
\Leftrightarrow\quad & ((d^{(k)})^\intercal H^{(k)} d^{(k)})^2 \ge \frac{4\,\kappa}{(\kappa+1)^2}\|M^{-1}H^{(k)}d^{(k)}\|_M^2\,\|d^{(k)}\|_M^2 \\
\Leftrightarrow\quad & \frac{((d^{(k)})^\intercal H^{(k)} M^{-1} H^{(k)} d^{(k)})\,((d^{(k)})^\intercal M\, d^{(k)})}{((d^{(k)})^\intercal H^{(k)} d^{(k)})^2} \le \frac{(\kappa+1)^2}{4\,\kappa}.
\end{aligned}$$

The statement in the previous line, however, is true due to the generalized Kantorovich inequality (Corollary 2.2). □

We summarize our findings on search directions:

the model Hessians $H^{(k)}$ have bounded condition numbers
⇒ the angle condition (5.8) holds
⇒ the search directions are admissible (5.7).

[19] In the literature, one often finds this result only in the case $M = \mathrm{Id}$, and with the non-optimal bound $\eta = \frac{1}{\kappa}$; see for instance Ulbrich, Ulbrich, 2012, p.32 or Nocedal, Wright, 2006, eq.(3.19).

Requirements on the Step Sizes

We now address the step sizes $\alpha^{(k)}$. The following example shows that the mere requirement

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) < f(x^{(k)})$$

is not sufficient to obtain a reasonable convergence behavior.

Example 5.6 (Too small step sizes[20]). *Consider the objective $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$, initial guess $x^{(0)} = 1$, search directions $d^{(k)} = -1$ and the Euclidean inner product $M = 1$. With step sizes $\alpha^{(k)} = \left(\frac{1}{2}\right)^{k+2}$, we obtain the sequences of iterates according to*

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)}\,(-1) = x^{(0)} - \sum_{i=0}^{k} \left(\tfrac{1}{2}\right)^{i+2} = \tfrac{1}{2} + \left(\tfrac{1}{2}\right)^{k+2}.$$

This implies $x^{(k+1)} < x^{(k)}$ and $f(x^{(k+1)}) < f(x^{(k)})$. However, $x^{(k)} \to x^* = 1/2$, which is not a stationary point of f.

The step sizes in the previous example are too small and thus they violate our second goal (item (2) above) since they do not exploit the achievable descent sufficiently well. We therefore introduce the following qualitative condition on the step sizes.

Definition 5.7 (Admissible step sizes). *Suppose that $x^{(k)}$ and $d^{(k)}$ are the sequences of iterates and search (descent) directions generated by an algorithm of type Algorithm 5.2. The sequence $\alpha^{(k)}$ of step sizes is termed* ***admissible*** *in case*

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) \le f(x^{(k)}) \quad \text{for all } k \in \mathbb{N}_0, \tag{5.10a}$$

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) \to 0 \quad \Rightarrow \quad \frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \to 0. \tag{5.10b}$$

We can interpret (5.10b) as follows: when the progress in the objective values converges to zero, then it is due to the normalized directional derivatives converging to zero and not due to the step sizes becoming too small. In other words, admissible step sizes do make sufficient use of the descent available in the direction $d^{(k)}$.

Condition (5.10) is purely qualitative. By contrast, the condition that the step sizes be **efficient**, i. e., there exists $\theta > 0$ such that

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) \le f(x^{(k)}) - \theta \left(\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \right)^2 \tag{5.11}$$

for all $k \in \mathbb{N}_0$ is a stronger, quantitative condition, which is moreover easy to verify.

[20] from Alt, 2002, Beispiel 4.4.1

Lemma 5.8 (Efficiency implies admissibility). *Suppose that $x^{(k)}$ and $d^{(k)}$ are the sequences of iterates and search (descent) directions generated by an algorithm of type Algorithm 5.2. If the sequence of step sizes $\alpha^{(k)}$ is efficient, then it is also admissible.*

Proof. Suppose that $\alpha^{(k)}$ is efficient, i. e.,

$$0 \le \theta \left(\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \right)^2 \le f(x^{(k)}) - f(x^{(k)} + \alpha^{(k)} d^{(k)})$$

Therefore (5.10a) is clear. To show (5.10b), suppose

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) \longrightarrow 0.$$

Since θ is strictly positive, this implies

$$\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \longrightarrow 0,$$

which confirms (5.10b). □

Using the assumptions of admissible search directions and admissible step sizes, we will obtain a theorem (see Theorem 5.9 below) on the global convergence of Algorithm 5.2. However, in view of the expected convergence result (5.5), we will have to work with accumulation points (limits of subsequences) of the iterates. This means that we should refine the notion of admissible search directions (5.7), the notions of admissible step sizes (5.10) as well as efficient step sizes (5.11) to subsequences. We denote such subsequences here with $\left(x^{(k)}\right)_{k \in K}$, where $K \subseteq \mathbb{N}_0$ is an infinite subset of the index set $\mathbb{N}_0$. (**Quiz 5.3:** How does this notation relate to the notation for subsequences $\left(x^{(k^{(\ell)})}\right)$ introduced in § 2.7?)

In detail, the refined conditions on subsequences read as follows:

admissible search directions:

$$\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \xrightarrow{k \in K} 0 \quad \Rightarrow \quad f'(x^{(k)}) \xrightarrow{k \in K} 0, \tag{5.7'}$$

angle condition:

$$\frac{-f'(x^{(k)})\, d^{(k)}}{\|\nabla_M f(x^{(k)})\|_M \, \|d^{(k)}\|_M} \ge \eta \quad \text{for all } k \in K \tag{5.8'}$$

admissible step sizes:

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) \le f(x^{(k)}) \quad \text{for all } k \in \mathbb{N}_0, \tag{5.10a'}$$

$$f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) \xrightarrow{k \in \mathbb{N}_0} 0 \quad \Rightarrow \quad \frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \xrightarrow{k \in K} 0, \tag{5.10b'}$$

efficient step sizes:

$$f(x^{(k)}+\alpha^{(k)}d^{(k)}) \le f(x^{(k)}) - \theta\left(\frac{f'(x^{(k)})\,d^{(k)}}{\|d^{(k)}\|_M}\right)^2 \quad \text{for all } k\in K. \tag{5.11'}$$

The statements of Lemma 5.4 and Lemma 5.5 continue to hold when restricted to subsequences. For the analog of Lemma 5.8, we have to make (5.10a') an assumption rather than a conclusion.

We now show a global convergence theorem for the model Algorithm 5.2.

Theorem 5.9 (Global convergence of model Algorithm 5.2)**.** *Suppose that Algorithm 5.2 generates an infinite sequence of iterates* $x^{(k)}$*, search directions* $d^{(k)} \neq 0$ *and step sizes* $\alpha^{(k)}$*. Suppose that* x^* *is an accumulation point of* $x^{(k)}$ *and that* $\big(x^{(k)}\big)_{k\in K}$ *is a subsequence converging to* x^**. Finally, suppose that the subsequences* $\big(d^{(k)}\big)_{k\in K}$ *and* $\big(\alpha^{(k)}\big)_{k\in K}$ *of search directions and step sizes are both admissible. Then* $f'(x^*) = 0$.

Note: In other words, when a generic descent algorithm (Algorithm 5.2) produces admissible search directions and admissible step sizes, then any accumulation point of the iterates is stationary.

Quiz 5.4: What goes wrong in Example 5.6?

Proof. Due to the continuity of f, we have $f(x^{(k)}) \xrightarrow{k\in K} f(x^*)$. Moreover, by admissibility of the step sizes (5.10a'), the entire sequence $f(x^{(k)})$ is monotone decreasing. Therefore, the entire sequence in fact converges: $f(x^{(k)}) \to f(x^*)$. Consequently, we also have

$$f(x^{(k+1)}) - f(x^{(k)}) = f(x^{(k)}+\alpha^{(k)}d^{(k)}) - f(x^{(k)}) \to 0.$$

The admissibility of step sizes along the subsequence, (5.10b'), implies

$$\frac{f'(x^{(k)})\,d^{(k)}}{\|d^{(k)}\|} \xrightarrow{k\in K} 0.$$

Since the search directions along the subsequence are in turn admissible, (5.7'), we can conclude

$$f'(x^{(k)}) \xrightarrow{k\in K} 0.$$

On the other hand, since f is of class C^1, we also have

$$f'(x^{(k)}) \xrightarrow{k\in K} f'(x^*).$$

This shows $f'(x^*) = 0$. □

§ 5.2 Step Size Strategies

In this section we will see how efficient step sizes (5.11) or at least admissible step sizes (5.10) can be found in general.

Armijo Backtracking Line Search

The Armijo backtracking line search is the simplest step size strategy and it is sufficient in many situations. Suppose that $d^{(k)}$ is a descent direction for f at $x^{(k)}$. In order to obtain sufficient decrease, the **Armijo condition** requires that the step size α satisfy

$$f(x^{(k)} + \alpha\, d^{(k)}) \le f(x^{(k)}) + \sigma\,\alpha\, f'(x^{(k)})\, d^{(k)} \tag{5.12}$$

holds. Here $\sigma \in (0,1)$ is the given **Armijo parameter**. Using the auxiliary function (**line search function**)

$$\varphi(\alpha) \coloneqq f(x^{(k)} + \alpha\, d^{(k)})$$

to simplify notation, we can write the Armijo condition (5.12) equivalently in the form

$$\varphi(\alpha) \le \varphi(0) + \sigma\,\alpha\,\varphi'(0). \tag{5.12}$$

Step sizes $\alpha \ge 0$ which satify (5.12) are termed **Armijo step sizes**. Condition (5.12) requires that the step size α realizes at least the σ-fraction of the first-order descent suggested by the tangent of φ at $\alpha = 0$.

Notice that due to the chain rule, φ inherits the C^1 property of f, and we have

$$\varphi'(\alpha) = f'(x^{(k)} + \alpha\, d^{(k)})\, d^{(k)} \tag{5.13a}$$

$$\text{and, in particular,}\quad \varphi'(0) = f'(x^{(k)})\, d^{(k)}. \tag{5.13b}$$

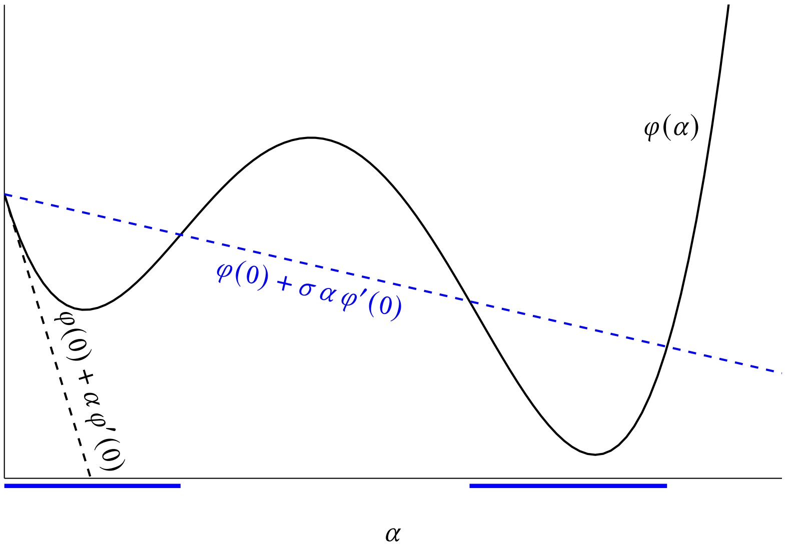


Figure 5.1: Illustration of step sizes $\alpha \ge 0$ satisfying the Armijo condition (5.12) (blue). As an example, the Armijo parameter is chosen as $\sigma = 0.07$.

We will now answer the question whether Armijo step sizes exist, and how to find them.

Lemma 5.10 (Existence of Armijo step sizes). *Suppose that d is a descent direction for f at x, and that the Armijo parameter satisfies $\sigma \in (0,1)$. Then there exists $\overline{\alpha} > 0$ such that (5.12) holds for all $\alpha \in [0, \overline{\alpha}]$.*

Proof. φ' is continuous at 0, which implies that there exists $\overline{\alpha} > 0$ such that

$$\varphi'(\alpha) < \sigma\,\varphi'(0) \quad \text{holds for all } \alpha \in [0, \overline{\alpha}].$$

From Taylor's theorem 2.4 we obtain that there exists $\xi \in [0, \alpha]$ such that

$$\begin{aligned} \varphi(\alpha) &= \varphi(0) + \alpha\,\varphi'(\xi) \\ &\le \varphi(0) + \sigma\,\alpha\,\varphi'(0). \end{aligned}$$

Therefore, the Armijo condition (5.12) holds for all $\alpha \in [0, \overline{\alpha}]$. □

We have seen that the Armijo condition is always satisfied in an interval starting at $\alpha = 0$. However, we need to select a step size which is not too small, as demonstrated by Example 5.6. This can be achieved by a **backtracking strategy**: run through a sequence of trial step sizes from large to small until the Armijo conditon (5.12) is satisfied for the first time.

Algorithm 5.11 (Armijo backtracking line search).

```
Input: initial trial step size $\alpha$
Input: routine to evaluate $\varphi$
Input: pre-computed function values $\varphi(0)$ and $\varphi'(0)$
Input: Armijo parameter $\sigma \in (0,1)$
Input: backtracking parameter $\beta \in (0,1)$
Output: step size $\alpha$ satisfying the Armijo condition (5.12)
Set $\ell := 0$
while Armijo condition (5.12) does not hold for $\alpha$ do
    Set $\alpha := \beta\,\alpha$                // new trial step size
    Set $\ell := \ell + 1$
end while
return $\alpha$
```

Remark 5.12 (on Algorithm 5.11).

(i) *In Algorithm 5.11, we did not number the trial step sizes $\alpha^{(0)}, \alpha^{(1)}, \ldots$ by an index in order to avoid confusion with the step size $\alpha^{(k)}$ which eventually gets used in the k-th iteration of the outer algorithm (Algorithm 5.2).*

(ii) *Every trial step size that fails to satisfy the Armijo condition "costs" one additional evaluation of φ, i. e., one additional evaluation of f.*

(iii) *The Armijo parameter is often chosen to be small, e. g., $\sigma = 10^{-2}$ or even $\sigma = 10^{-4}$. A typical value for the backtracking parameter is $\beta = 1/2$.*

(iv) *It follows from Lemma 5.10 that Algorithm 5.11 terminates successfully after finitely many iterations with a step size α that satisfies $\alpha \ge \overline{\alpha}\,\beta$. Here $\overline{\alpha} > 0$ is the upper bound of any interval $[0, \overline{\alpha}]$ containing only Armijo step sizes.*

(v) *In a practical implementation, one often adds further checks and stopping criteria to Algorithm 5.11. For instance, we need to safeguard against $\varphi'(0) \ge 0$ (d is not a descent direction) and against too many unsuccessful trial steps.*

Suitable values for the initial trial step size α in Algorithm 5.11 depend on how the search directions $d^{(k)}$ are generated in the outer method. We will see more on that when we discuss concrete instances of Algorithm 5.2. Since the backtracking strategy only shortens the initial trial step size, we need to ensure that the initial trial step size is sufficiently large in order to obtain admissible step sizes that exploit the achievable descent sufficiently well. This is what the following result is about.

Lemma 5.13 (Armijo backtracking line search produces admissible step sizes)**.** *Suppose that Algorithm 5.2 generates an infinite sequence of iterates $x^{(k)}$ and search (descent) directions $d^{(k)} \neq 0$. Suppose moreover that the step sizes $\alpha^{(k)}$ are obtained by the Armijo backtracking line search (Algorithm 5.11) with initial trial step size $\alpha^{(k,0)}$. Assume that $K \subseteq \mathbb{N}_0$ is an infinite index set such that the subsequence $\big(x^{(k)}\big)_{k\in K}$ is bounded. Finally, suppose that $\psi\colon [0,\infty \longrightarrow [0,\infty)$ is any monotone increasing function and that the initial trial step sizes satisfy*

$$\alpha^{(k,0)} \, \|d^{(k)}\|_M \ge \psi\Big(\frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M}\Big) \quad \text{for all } k \in K. \tag{5.14}$$

Then the step sizes $\big(\alpha^{(k)}\big)_{k\in K}$ are admissible.

Proof. We need to show (5.10a') and (5.10b'). The first condition is a direct consequence of the Armijo condition holding at $\alpha^{(k)} > 0$

$$f(x^{(k)} + \alpha^{(k)}\, d^{(k)}) \le f(x^{(k)}) + \sigma\, \alpha^{(k)} \underbrace{f'(x^{(k)})\, d^{(k)}}_{<0},$$

the fact that $d^{(k)}$ is a descent direction and that σ is positive. It remains to verify (5.10b').

By assumption, the sequence $\big(x^{(k)}\big)_{k\in K}$ is bounded. Therefore, it has a convergent subsequence with index set K'. By continuity of f, $\big(f(x^{(k)})\big)_{k\in K'}$ converges. Due to the Armijo condition (5.12), the sequence $f(x^{(k)})$ is monotone decreasing, so that in fact the entire sequence $f(x^{(k)})$ converges. From there and the Armijo condition (5.12) we conclude

$$f(x^{(k+1)}) - f(x^{(k)}) = f(x^{(k)} + \alpha^{(k)} d^{(k)}) - f(x^{(k)}) \le \sigma\, \alpha^{(k)} f'(x^{(k)})\, d^{(k)} < 0.$$

The left-hand side converges to 0, therefore we must have

$$\alpha^{(k)} f'(x^{(k)})\, d^{(k)} \longrightarrow 0. \tag{$*$}$$

In order to verify (5.10b'), we need to show

$$\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \xrightarrow{k\in K} 0.$$

In the remainder of the proof, we distinguish indices $k \in K$ according to the following cases:

$$\text{When } \alpha^{(k)} \|d^{(k)}\|_M \text{ is "large", then } \frac{\alpha^{(k)} f'(x^{(k)})\, d^{(k)}}{\alpha^{(k)} \|d^{(k)}\|_M} \text{ is small.}$$

$$\text{When } \alpha^{(k)} \|d^{(k)}\|_M \text{ is "small", then } \begin{cases} \text{we use the assumption (5.14)} & \text{in case } \alpha^{(k)} = \alpha^{(k,0)}. \\ \text{we use the Armijo condition (5.12)} & \text{in case } \alpha^{(k)} < \alpha^{(k,0)}. \end{cases}$$

By assumption, the sequence $\big(x^{(k)}\big)_{k\in K}$ is bounded, hence the continuous function f' is uniformly continuous "near the $\big(x^{(k)}\big)_{k\in K}$". More precisely, suppose that $R > 0$ is any fixed number, then f' is uniformly continuous on the compact set

$$A_R \coloneqq \operatorname{cl} \bigcup_{k\in K} B_R^M(x^{(k)}).$$

(**Quiz 5.5:** Why is this set compact?) Now suppose that $\varepsilon > 0$ is given. Then there exists $\overline{\delta} > 0$ such that

$$\|f'(y) - f'(z)\|_{M^{-1}} \le (1-\sigma)\, \varepsilon$$

holds for all $y, z \in A_R$ such that $\|y - z\|_M \le \overline{\delta}$. Possibly by making $\overline{\delta}$ smaller, we can assume $\overline{\delta} \le R$. Thus, in particular, we obtain

$$\Big\|f'(\underbrace{x^{(k)} + e}_{\in A_R}) - f'(\underbrace{x^{(k)}}_{\in A_R})\Big\|_{M^{-1}} \le (1-\sigma)\, \varepsilon \quad \text{for all } k \in K,\ \|e\|_M \le \overline{\delta}. \tag{$**$}$$

We now set

$$\delta \coloneqq \min\big\{\overline{\delta}\, \beta,\ \psi(\varepsilon)\big\} \in (0, \overline{\delta}).$$

Due to the convergence in ($*$), there exists an index $k_0 \in \mathbb{N}_0$ such that

$$\alpha^{(k)} \big|f'(x^{(k)})\, d^{(k)}\big| \le \varepsilon\, \delta \quad \text{holds for all } k \ge k_0. \tag{$***$}$$

From now on, let $k \in K$, $k \ge k_0$, be arbitrary. We are going to show that

$$0 \le \frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \le \varepsilon$$

holds, which proves (5.10b'). We distinguish the following cases, as anticipated above:

Case 1: $\alpha^{(k)} \|d^{(k)}\|_M \ge \delta$
In this case we immediately conclude

$$\begin{aligned} 0 &\le \frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} && \text{since } d^{(k)} \text{ is a descent direction} \\ &= \frac{-\alpha^{(k)} f'(x^{(k)})\, d^{(k)}}{\alpha^{(k)} \|d^{(k)}\|_M} \\ &\le \frac{\varepsilon\, \delta}{\delta} && \text{by } (***) \text{ and the assumption in case 1} \\ &= \varepsilon. \end{aligned}$$

Case 2: $\alpha^{(k)} \|d^{(k)}\|_M < \delta$ and $\alpha^{(k)} = \alpha^{(k,0)}$
We obtain

$$\begin{aligned} \psi\Big(\frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M}\Big) &\le \alpha_{k,0}\, \|d^{(k)}\|_M && \text{by assumption (5.14)} \\ &< \delta && \text{by the assumption in case 2} \\ &\le \psi(\varepsilon) && \text{by the choice of } \delta. \end{aligned}$$

Since ψ is monotone increasing, we conclude

$$0 \le \frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \le \varepsilon.$$

Case 3: $\alpha^{(k)} \|d^{(k)}\|_M < \delta$ and $\alpha^{(k)} < \alpha^{(k,0)}$
The assumption $\alpha^{(k)} < \alpha^{(k,0)}$ means that the initial trial step size (and possibly some of the subsequent trial step sizes) did not satisfy the Armijo condition. Since $\alpha^{(k)}$ was the first trial step size to satisfy the Armijo condition (5.12), the previous trial step size, $\beta^{-1}\alpha^{(k)}$, violated it:

$$\sigma\, \beta^{-1}\alpha^{(k)} f'(x^{(k)}) d^{(k)} < f\big(x^{(k)} + \beta^{-1}\alpha^{(k)} d^{(k)}\big) - f(x^{(k)}).$$

By Taylor's theorem 2.4, there exists $\xi^{(k)} \in (0,1)$ such that

$$\sigma\, \beta^{-1}\alpha^{(k)} f'(x^{(k)}) d^{(k)} < \beta^{-1}\alpha^{(k)}\, f'\big(x^{(k)} + \beta^{-1}\alpha^{(k)}\, \xi^{(k)} d^{(k)}\big)\, d^{(k)}$$

and thus

$$\begin{aligned} &\sigma\, f'(x^{(k)})\, d^{(k)} \\ &\quad < f'\big(x^{(k)} + \beta^{-1}\alpha^{(k)}\, \xi^{(k)} d^{(k)}\big)\, d^{(k)} \\ &\quad = f'(x^{(k)})\, d^{(k)} + \big[f'\big(x^{(k)} + \beta^{-1}\alpha^{(k)}\, \xi^{(k)} d^{(k)}\big) - f'(x^{(k)})\big] d^{(k)} \\ &\quad \le f'(x^{(k)})\, d^{(k)} + \|f'\big(x^{(k)} + \underbrace{\beta^{-1}\alpha^{(k)}\, \xi^{(k)} d^{(k)}}_{=:e^{(k)}}\big) - f'(x^{(k)})\|_{M^{-1}}\, \|d^{(k)}\|_M \quad \text{by (2.3)}. \end{aligned}$$

The vector $e^{(k)}$ satisfies

$$\begin{aligned} \|e^{(k)}\|_M &= \beta^{-1}\alpha^{(k)} \xi^{(k)} \|d^{(k)}\|_M \\ &< \beta^{-1}\delta && \text{by the assumption in case 3 and since } \xi^{(k)} \in (0,1) \\ &\le \overline{\delta} && \text{by the choice of } \delta. \end{aligned}$$

We may thus apply estimate (**) to the inequality above to obtain

$$\sigma\, f'(x^{(k)})\, d^{(k)} \le f'(x^{(k)})\, d^{(k)} + (1-\sigma)\, \varepsilon\, \|d^{(k)}\|_M.$$

Sorting terms and dividing by $\|d^{(k)}\|_M$ finally yields

$$0 \le \frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \le \varepsilon.$$

□

Remark 5.14 (Armijo backtracking line search produces efficient step sizes)**.** *When we choose $\psi(z) = c\,z$ with some $c > 0$, i. e., when we use initial trial step sizes satisfying*

$$\alpha_{k,0}\,\|d^{(k)}\|_M \ge c\,\frac{-f'(x^{(k)})\,d^{(k)}}{\|d^{(k)}\|_M}, \tag{5.15}$$

and if f' is Lipschitz continuous on the sublevel set $\mathcal{M}_f(x^{(0)}) \coloneqq \{x \in \mathbb{R}^n \,|\, f(x) \le f(x^{(0)})\}$, then one can show that Algorithm 5.11 produces not only admissible, but efficient step sizes.

To conclude the presentation of Armijo backtracking strategies, we consider a modification of Algorithm 5.11 which often produces trial step sizes more effectively than simple backtracking $\alpha \rightsquigarrow \beta\,\alpha$ in case the Armijo condition fails on the initial trial step size.

The modification is based on the fact that we have available the data of the line search function φ

$$\varphi(0), \quad \varphi'(0) < 0 \quad \text{and} \quad \varphi(\alpha)$$

for the current trial step size α. Using this data, we can fit a quadratic polynomial

$$p(\alpha) = a + b\,\alpha + c\,\alpha^2.$$

The conditions[21] $p(0) = \varphi(0)$, $p'(0) = \varphi'(0)$ and $p(\alpha) = \varphi(\alpha)$ uniquely define the coefficients

$$a = \varphi(0), \qquad b = \varphi'(0), \qquad c = \frac{1}{\alpha^2}\big(\varphi(\alpha) - \varphi(0) - \varphi'(0)\,\alpha\big). \tag{5.16}$$

Naturally, this quadratic model of φ will be used only when the Armijo condition (5.12) failed at the trial step size α, i. e., in case

$$\varphi(\alpha) - \varphi(0) - \varphi'(0)\,\alpha > \varphi(\alpha) - \varphi(0) - \sigma\,\varphi'(0)\,\alpha > 0$$

holds, which implies $c > 0$. This in turn means that the unique global minimizer $\alpha^* = -\frac{b}{2c}$ of p satisfies

$$\alpha^* = \frac{-\varphi'(0)\,\alpha^2}{2\,\big(\varphi(\alpha) - \varphi(0) - \varphi'(0)\,\alpha\big)} > 0.$$

We then choose α^* as the next trial step size α^+, but in order to avoid drastic changes or even an increase from α to α^+, we clip α^* to the interval $[\underline{\beta}\,\alpha, \overline{\beta}\,\alpha]$ according to

$$\alpha^+ \coloneqq \min\big\{\max\{\alpha^*, \underline{\beta}\,\alpha\}, \overline{\beta}\,\alpha\big\} = \begin{cases} \underline{\beta}\,\alpha, & \text{if } \alpha^* < \underline{\beta}\,\alpha, \\ \alpha^*, & \text{if } \underline{\beta}\,\alpha \le \alpha^* \le \overline{\beta}\,\alpha, \\ \overline{\beta}\,\alpha, & \text{if } \alpha^* > \overline{\beta}\,\alpha, \end{cases}$$

where $0 < \underline{\beta} < \overline{\beta} < 1$ are the clipping parameters.[22] This modified Armijo backtracking line search maintains the essential properties of the simple Armijo backtracking line search. In particular, the admissibility (and potentially efficiency) of the accepted step sizes (see Lemma 5.13 and Remark 5.14) continue to hold.

For completeness, we present the modified Armijo backtracking line search procedure in Algorithm 5.15.

[21] Fitting a polynomial using function values and derivatives is known as **Hermite interpolation**. Using function values only is known as **Lagrange interpolation**.

[22] Using $\underline{\beta} = \overline{\beta} = \beta$ we get back our previous simple backtracking strategy where $\alpha^+ = \beta\,\alpha$.

Algorithm 5.15 (Modified Armijo backtracking line search with interpolation)**.**

```
Input: initial trial step size $\alpha$
Input: routine to evaluate $\varphi$
Input: pre-computed function values $\varphi(0)$ and $\varphi'(0)$
Input: Armijo parameter $\sigma \in (0,1)$
Input: backtracking parameters $0 < \underline{\beta} < \overline{\beta} < 1$
Output: step size $\alpha$ satisfying the Armijo condition (5.12)
Set $\ell := 0$
while Armijo condition (5.12) does not hold for $\alpha$ do
    Set $\alpha^* := \dfrac{-\varphi'(0)\,\alpha^2}{2\left(\varphi(\alpha) - \varphi(0) - \varphi'(0)\,\alpha\right)}$    // minimizer of quadratic polynomial
    Set $\alpha := \min\big\{\max\{\alpha^*, \underline{\beta}\,\alpha\}, \overline{\beta}\,\alpha\big\}$    // clip it and use as new trial step size
    Setze $\ell := \ell + 1$
end while
return $\alpha$
```

Wolfe-Powell Line Search

Recall from Lemma 5.10 that the Armijo condition

$$f(x^{(k)} + \alpha\, d^{(k)}) \le f(x^{(k)}) + \sigma\, \alpha\, f'(x^{(k)})\, d^{(k)} \quad \text{or} \quad \varphi(\alpha) \le \varphi(0) + \sigma\, \alpha\, \varphi'(0) \tag{5.12}$$

always holds in some interval $[0, \overline{\alpha}]$. Therefore, we combined the Armijo condition with backtracking, where we generate trial step sizes from large to small, in order to avoid overly small step sizes.

Alternatively, we could require, in addition to (5.12), the **curvature condition**

$$f'(x^{(k)} + \alpha\, d^{(k)})\, d^{(k)} \ge \tau\, f'(x^{(k)})\, d^{(k)} \quad \text{or} \quad \varphi'(\alpha) \ge \tau\, \varphi'(0) \tag{5.17}$$

or even the **strong curvature condition**

$$|f'(x^{(k)} + \alpha\, d^{(k)})\, d^{(k)}| \le -\tau\, f'(x^{(k)})\, d^{(k)} \quad \text{or} \quad |\varphi'(\alpha)| \le -\tau\, \varphi'(0) \tag{5.18}$$

to hold, where $\tau \in (\sigma, 1)$ is the **curvature parameter**. The curvature condition (5.17) demands that the derivative of φ at α is not too negative, namely that it is larger (has less descent) than at $\alpha = 0$. However, it would be fine for φ to increase near α; see Figure 5.2. This curvature condition already avoids too small step sizes α near 0.

The strong curvature condition (5.18) demands that, in addition, the derivative of φ at α it not too positive either. The condition can be interpreted as the requirement that α be an approximately stationary point of φ. **Note:** When α is a local minimizer of φ, then (5.18) holds even for $\tau = 0$.

The Armijo condition (5.12) and the curvature condition (5.17) together are referred to as the **Wolfe-Powell conditions**. The Armijo condition (5.12) and the strong curvature condition (5.18) together are referred to as the **strong Wolfe-Powell conditions**. Consequently, step sizes $\alpha \ge 0$ which satisfy the above conditions are referred to as **Wolfe-Powell step sizes** and **strong Wolfe-Powell step sizes**, respectively.

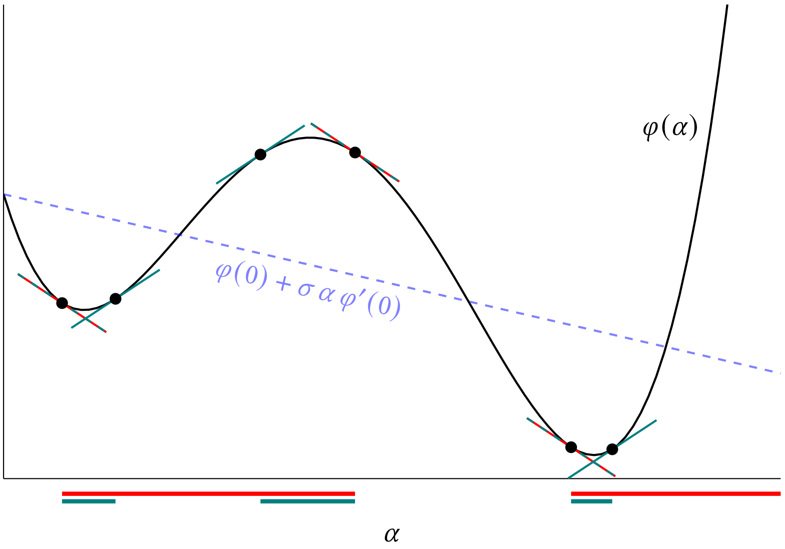


Figure 5.2: Illustration of step sizes $\alpha \ge 0$ satisfying the curvature condition (5.17) (red) and the strong curvature condition (5.18) (teal). As an example, the curvature parameter is chosen as $\tau = 0.2$.

A simple example such as $\varphi(\alpha) = -\alpha$ shows that the curvature condition may not be satisfiable without further assumptions on f. The following result gives a sufficient condition for strong Wolfe-Powell step sizes to exist.

Lemma 5.16 (Existence of (strong) Wolfe-Powell step sizes)**.** *Suppose that d is a descent direction for f at x and that the Armijo and curvature parameters satisfy $0 < \sigma < \tau < 1$. Suppose, moreover, that f is bounded below on the ray $\{x + \alpha\, d \,|\, \alpha \ge 0\}$. Then there exists a step size $\alpha_2 > 0$ such that the strong Wolfe-Powell conditions* (5.12) *and* (5.18) *(and thus also the regular Wolfe-Powell conditions* (5.12) *and* (5.17)*) hold in a neighborhood of α_2.*

Proof. We abbreviate as usual $\varphi(\alpha) \coloneqq f(x + \alpha\, d)$. Since by assumption, φ is bounded below on $\mathbb{R}_{\ge 0}$, φ intersects the Armijo line

$$\alpha \mapsto \varphi(0) + \sigma \underbrace{\varphi'(0)}_{<0} \alpha,$$

which is unbounded below, in at least one positive point. Suppose that α_1 is the smallest positive point of intersection (**Quiz 5.6:** Why does α_1 exist?). Then we have

$$\varphi(\alpha_1) = \varphi(0) + \sigma\, \varphi'(0)\, \alpha_1.$$

In view of $\varphi'(0) < 0$, the Armijo condition (5.12) holds for all $\alpha \in [0, \alpha_1]$, i. e., the Armijo line lies below φ on this interval. From the mean value theorem 2.4, we infer the existence of $\alpha_2 \in (0, \alpha_1)$ such that

$$\varphi'(\alpha_2) = \frac{\varphi(\alpha_1) - \varphi(0)}{\alpha_1} = \sigma\, \varphi'(0).$$

And thus we obtain the strong curvature condition (5.18) at α_2:

$$|\varphi'(\alpha_2)| = -\sigma\, \varphi'(0) < -\tau\, \varphi'(0).$$

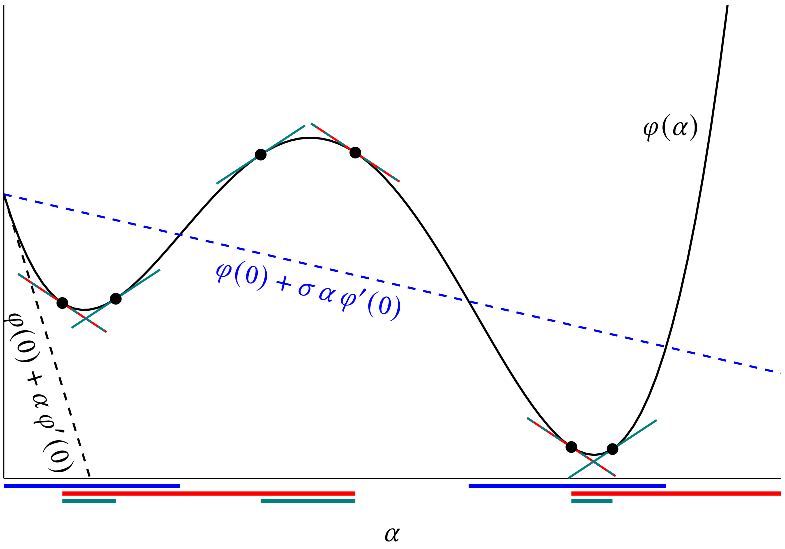


Figure 5.3: Illustration of step sizes $\alpha \ge 0$ satisfying both the Armijo condition (5.12) (blue), the curvature condition (5.17) (red) and the strong curvature condition (5.18) (teal). As an example, the Armijo parameter is chosen as $\sigma = 0.07$ and the curvature parameter is chosen as $\tau = 0.2$.

Due to the continuity of φ', the strong curvature condition (5.18) and thus also the regular curvature condition (5.17) continue to hold for all α in a neighborhood of α_2. □

We now address an algorithm to find a Wolfe-Powell step size. To simplify notation, we introduce the auxiliary function

$$\psi(\alpha) \coloneqq \varphi(\alpha) - \varphi(0) - \sigma\,\varphi'(0)\,\alpha,$$

which measures the signed gap between the function values of the one dimensional line search function φ and its σ-relaxed linearization at the origin, so that we can write

$$\begin{aligned}
&\text{the Armijo condition (5.12)} &&\iff && \psi(\alpha) \le 0, && \text{(5.12')}\\
&\text{the curvature condition (5.17)} &&\iff && -(\tau-\sigma)\,|\varphi'(0)| \le \psi'(\alpha), && \text{(5.17')}\\
&\text{the strong curvature condition (5.18)} &&\iff && -\underbrace{(\tau-\sigma)}_{>0}\,|\varphi'(0)| \le \psi'(\alpha) \le (\tau+\sigma)\,|\varphi'(0)|. && \text{(5.18')}
\end{aligned}$$

We restrict the discussion to the regular Wolfe-Powell condition, i. e., (5.12) and (5.17). See for instance Geiger, Kanzow, 1999, Kapitel 6.3 for the strong Wolfe-Powell condition.

Lemma 5.17 (Inclusion of Wolfe-Powell step sizes, see Geiger, Kanzow, 1999, Lemma 6.1). *Suppose that $0 \le a < b$ are chosen such the conditions*

$$\psi(a) \le 0 \quad \textit{and} \quad \psi'(a) < 0 \tag{5.19a}$$

$$\textit{as well as} \quad \psi(b) \ge 0 \tag{5.19b}$$

hold; see Figure 5.4. Then there exists $\alpha^* \in (a, b)$ such that

$$\psi(\alpha^*) < 0 \quad \textit{and} \quad \psi'(\alpha^*) = 0$$

holds. In particular, the Wolfe-Powell conditions (5.12') and (5.17') hold in a neighborhood of α^*.

Proof. Let us denote by α^* a global minimizer of

$$\text{Minimize } \psi(\alpha) \text{ on the compact interval } [a, b].$$

The assumptions on a and b imply that α^* belongs to the open interval (a, b). Consequently, α^* is also a local minimizer of the unconstrained problem “Minimize $\psi(\alpha)$ where $\alpha \in \mathbb{R}$”, and thus we have $\psi'(\alpha^*) = 0$. From $\psi(a) \le 0$ and $\psi'(a) < 0$ we infer $\psi(\alpha^*) < 0$. Since both (5.12') and (5.17') hold with strict inequalities at α^*, continuity implies that they hold in a neighborhood of α^*. □

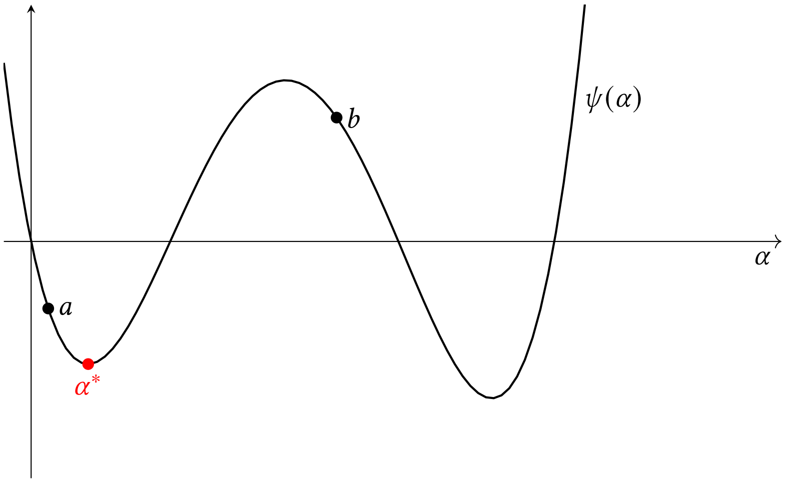


Figure 5.4: Illustration of the condition (5.19) and the statement of Lemma 5.17.

Note: The condition (5.19a) is readily seen to hold at $a = 0$. This motivates the strategy to first find a right boundary b so that (5.19b) holds as well, and then to approximate α^* by nesting intervals.

Algorithm 5.18 (Wolfe-Powell line search)**.**

```
Input: initial trial step size $\alpha$
Input: routine to evaluate $\varphi$ and $\varphi'$
Input: pre-computed function values $\varphi(0)$ and $\varphi'(0)$
Input: Armijo and curvature parameters $0 < \sigma < \tau < 1$
Input: expansion parameter $\gamma > 1$
Input: nesting parameters $\underline{\gamma}, \overline{\gamma} \in (0, 1/2]$
Output: step size $\alpha$ satisfying the Wolfe-Powell conditions (5.12) and (5.17)
Set $a := 0$ and $b := \alpha$
Set $\ell := 0$
while $\varphi(b) < \varphi(0) + \sigma\,\varphi'(0)\,b$ and $\varphi'(b) < \tau\,\varphi'(0)$ do    // phase 1 repeatedly expands $[0, b]$ until (5.19) holds
```

```
4:      Set $b := \gamma\, b$                                     // expand the right boundary $b$
5:      Set $\ell := \ell + 1$
6:  end while                                                      // now we have (5.19)
7:  Set $\alpha := b$
8:  while Armijo condition (5.12) or curvature condition (5.17) is violated at $\alpha$ do   // phase 2 repeatedly shrinks $[a, b]$ until (5.12) and (5.17) hold
9:      Choose $\alpha \in [a + \underline{\gamma}\,(b - a), b - \overline{\gamma}\,(b - a)]$   // for instance, choose the midpoint
10:     if $\varphi(\alpha) \ge \varphi(0) + \sigma\, \varphi'(0)\, \alpha$ then     // Armijo condition is violated at $\alpha$
11:         Set $b := \alpha$                                     // reduce the right boundary $b$
12:     else
13:         Set $a := \alpha$                                     // increase the left boundary $a$
14:     end if
15:     Set $\ell := \ell + 1$
16: end while
17: return $\alpha$
```

Remark 5.19 (on Algorithm 5.18, compare Remark 5.12)**.**

(*i*) *The Armijo parameter is often chosen to be small, e. g., $\sigma = 10^{-2}$ or even $\sigma = 10^{-4}$. Depending on the characteristics of the outer method (which determines the search directions), the curvature parameter $\tau > \sigma$ should be chosen "small" as well, e. g., $\tau = 0.1$, or otherwise "large", e. g., $\tau = 0.9$.*

(*ii*) *Each iteration of phase 1 "costs" one additional evaluation of φ and φ', i. e., one additional evaluation of f and f', or rather the directional derivative of f in the direction of the current search direction; compare (5.13). Each iteration of phase 2 "costs" one additional evaluation of φ.*

(*iii*) *Using Lemma 5.17, it is not difficult to see that Algorithm 5.18 terminates after finitely many steps under the conditions of Lemma 5.16:*

- *The while loop beginning at Line 3 terminates, since for b sufficiently large, the Armijo condition (5.12) is violated. For such b, we have $\psi(b) > 0$, i. e., (5.19b) holds.*

- *At the first iteration of the while loop beginning at Line 8, the conditions (5.19) of Lemma 5.17 are satisfied. Consequently, they continue to hold also in all subsequent iterations.*

- *The length of the intervals $[a, b]$ in phase 2 goes to zero if infinitely many iterations of the while loop beginning at Line 8 were performed. However, as shown in Lemma 5.17, there is an open set of points which satisfy both the Armijo condition (5.12) and the curvature condition (5.17) inside any of the intervals $[a, b]$ considered in phase 2. Therefore, phase 2 must terminate.*

(*iv*) *The step size accepted by Algorithm 5.18 may be larger or smaller than the initial trial step size provided by the user.*

(*v*) *As was already noted for the Armijo backtracking line search (Algorithm 5.11) in Remark 5.12, in a practical implementation, one often adds further checks and stopping criteria to Algorithm 5.11.*

For instance, we need to safeguard against $\varphi'(0) \ge 0$ (d is not a descent direction) and against too many unsuccessful trial steps.

(*vi*) *An algorithm for the strong Wolfe-Powell line search can be found in Geiger, Kanzow, 1999, Kapitel 6.3.*

The admissibility of step sizes generated by the Wolfe-Powell line search algorithm is shown in the following result. Clearly, this result also applies to step sizes satisfying the strong Wolfe-Powell conditions.

Lemma 5.20 (Wolfe-Powell line search produces admissible step sizes)**.** *Suppose that Algorithm 5.2 generates an infinite sequence of iterates $x^{(k)}$ and search (descent) directions $d^{(k)} \neq 0$. Suppose moreover that the step sizes $\alpha^{(k)}$ are chosen so that they satisfy the Wolfe-Powell conditions (5.12) and (5.17) (for instance by Algorithm 5.18).*[23] *Assume that $K \subseteq \mathbb{N}_0$ is an infinite index set such that the subsequence $\big(x^{(k)}\big)_{k\in K}$ is bounded. Then the step sizes $\big(\alpha^{(k)}\big)_{k\in K}$ are admissible.*

Proof. As in the proof of Lemma 5.13 we obtain the result

$$-\alpha^{(k)} f'(x^{(k)})\, d^{(k)} \to 0. \tag{$*$}$$

It remains to show

$$\frac{f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|} \xrightarrow{k\in K} 0.$$

To this end, let $\varepsilon > 0$.

Just like in the proof of Lemma 5.13, we can argue that the boundedness of $\big(x^{(k)}\big)_{k\in K}$ entails that the continuous function f' is uniformly continuous "near the $\big(x^{(k)}\big)_{k\in K}$". More precisely, there exists $\delta > 0$ such that

$$\big\|f'(x^{(k)} + e) - f'(x^{(k)})\big\|_{M^{-1}} \le (1-\tau)\,\varepsilon \quad \text{for all } k \in K,\ \|e\|_M \le \delta.$$

Because of ($*$), there exists an index $k_0 \in \mathbb{N}$ such that

$$\alpha^{(k)} |f'(x^{(k)})\, d^{(k)}| \le \varepsilon\,\delta \quad \text{for all } k \ge k_0. \tag{$**$}$$

From now on, let $k \in K$, $k \ge k_0$, be arbitrary. Similarly as in the proof of Lemma 5.13, we consider the following cases:

Case 1: $\alpha^{(k)} \|d^{(k)}\|_M \ge \delta$

[23] Notice that, in contrast to condition (5.14) in Lemma 5.13, there is no lower bound on the initial trial step size necessary to be observed.

Precisely as in the proof of Lemma 5.13, we obtain

$$\begin{aligned} 0 &\le \frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} && \text{since } d^{(k)} \text{ is a descent direction} \\ &= \frac{-\alpha^{(k)} f'(x^{(k)})\, d^{(k)}}{\alpha^{(k)}\|d^{(k)}\|_M} \\ &\le \frac{\varepsilon\,\delta}{\delta} && \text{by } (**) \text{ and the assumption in case 1} \\ &= \varepsilon. \end{aligned}$$

Case 2: $\alpha^{(k)}\|d^{(k)}\|_M < \delta$
In this we argue with the satisfaction of the curvature conditon (5.17) for $\alpha^{(k)}$:

$$\tau\, f'(x^{(k)})\, d^{(k)} \le f'(x^{(k)} + \alpha^{(k)} d^{(k)})\, d^{(k)}.$$

The addition of $|f'(x^{(k)})\, d^{(k)}| = -f'(x^{(k)})\, d^{(k)}$ on both sides yields

$$\begin{aligned} (1-\tau)\,|f'(x^{(k)})\, d^{(k)}| &\le f'(x^{(k)} + \alpha^{(k)} d^{(k)})\, d^{(k)} - f'(x^{(k)})\, d^{(k)} \\ &\le \left| f'\left(x^{(k)} + \alpha^{(k)} d^{(k)}\right) d^{(k)} - f'(x^{(k)})\, d^{(k)} \right| \\ &\le \left\| f'\left(x^{(k)} + \alpha^{(k)} d^{(k)}\right) - f'(x^{(k)}) \right\|_{M^{-1}} \|d^{(k)}\|_M. \end{aligned}$$

Invoking now the uniform continuity, we obtain

$$(1-\tau)\,|f'(x^{(k)})\, d^{(k)}| \le (1-\tau)\,\varepsilon\,\|d^{(k)}\|_M,$$

and hence

$$0 \le \frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \le \varepsilon.$$

□

Analogously as with the Armijo backtracking line search (Remark 5.14), one can also show the efficiency of step sizes when f' is Lipschitz continuous on the sublevel set $\mathcal{M}_f(x^{(0)}) \coloneqq \{x \in \mathbb{R}^n \mid f(x) \le f(x^{(0)})\}$. The proof is part of homework problem 4.2.

In concluding, we also remark that Line 9 in phase 2 of Algorithm 5.18 leaves some freedom in the choice of the next trial step size α. The available data $\varphi(a)$, $\varphi'(a)$, $\varphi(b)$ and $\varphi'(b)$ lends itself to a cubic Hermite interpolation, using the model

$$p(\alpha) = a + b\,\alpha + c\,\alpha^2 + d\,\alpha^3.$$

Provided that a unique local minimizuer α^* of p exists, we can calculate it explicitly and subsequently clip it to the interval $[a, b]$:

$$\alpha \coloneqq \max\{a, \min\{b, \alpha^*\}\}.$$

One needs to pay attention to the fact that not all of the data $\varphi'(a)$ and $\varphi'(b)$ is necessarily available in the current iteration of Algorithm 5.18. In this case one may proceed with a quadratic polynomial as in the modified Armijo backtracking line search method.

Remark 5.21 (Scaling invariance of the Armijo and curvature conditions)**.** *The Armijo and curvature conditions* (5.12), (5.17) *and* (5.18) *are invariant w.r.t. affine scaling in the domain and codomain spaces. Suppose that we consider, besides the objective f, another objective g related via*

$$f(x) \quad \rightsquigarrow \quad g(x) \coloneqq \gamma\, f(A\,x+b)+\delta,$$

where $A \in \mathbb{R}^{n\times n}$ is non-singular, $b \in \mathbb{R}^n$, $\gamma > 0$ and $\delta \in \mathbb{R}$.

Then the following holds: a step size α that satisfies any of the conditions (5.12), (5.17) *or* (5.18) *for g at x with search direction d, satisfies the same conditions for f at $A\,x+b$ with the search direction $A\,d$. Since the scaling of an optimization problem is often arbitrary, this is a desirable property.*

The proof is part of homework problem 4.3.

§ 5.3 Gradient Descent Method

In the remainder of § 5 we consider different concrete realizations of the generic descent method Algorithm 5.2. The methods differ w.r.t. the way the search directions $d^{(k)}$ are generated and w.r.t. the choice of the line search method (Armijo or Wolfe-Powell) to determine the step sizes $\alpha^{(k)}$. As was already mentioned, the methods discussed here obtain the search direction at an iterate $x^{(k)}$ by minimizing a quadratic model of the objective

$$q^{(k)}(d) = f(x^{(k)}) + f'(x^{(k)})\,d + \frac{1}{2}\,d^\intercal H^{(k)} d. \tag{5.2}$$

When the model Hessian $H^{(k)}$ is s. p. d., this is equivalent to the solution of the linear system

$$H^{(k)} d^{(k)} = -\nabla f(x^{(k)}). \tag{5.4}$$

The **gradient descent method** (also known as **steepest descent method**) for our generic unconstrained linear problem

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

generates its search directions in the same way we already know from § 4.2, when f was a quadratic polynomial. That is, we use

$$M\,d^{(k)} = -\nabla f(x^{(k)}) \quad \text{or} \quad d^{(k)} = -M^{-1}\nabla f(x^{(k)}) = -\nabla_M f(x^{(k)}). \tag{5.20}$$

This corresponds to using a constant model Hessian $H^{(k)} \equiv M$ in the model (5.2):

$$q^{(k)}(d) = f(x^{(k)}) + f'(x^{(k)})\,d + \frac{1}{2}\,d^\intercal M\,d.$$

The choice of the inner product M is due to the user. As was already mentioned in Remark 4.7, one refers to the case $M = \text{Id}$ as the classical **gradient descent method** without preconditioning. Otherwise one speaks of a **preconditioned gradient descent method** with **preconditioner** M.

The particular choice of $d^{(k)}$ in the gradient descent method clearly implies the angle condition (5.8) with the maximal possible value, $\eta = 1$. In particular, the search direction $d^{(k)}$ is a descent direction for f at $x^{(k)}$, as long as $f'(x^{(k)}) \neq 0$ holds.

A simple strategy is sufficient to determine admissible step sizes (5.10). One typically employs the Armijo backtracking line search (Algorithm 5.11) or the version with interpolation (Algorithm 5.15).

The efficiency condition (5.15) requires that the initial trial step size satisfy

$$\begin{aligned}
\alpha^{(k,0)} &\ge c \, \frac{-f'(x^{(k)}) \, d^{(k)}}{\|d^{(k)}\|_M^2} \\
&= c \, \frac{-(\nabla_M f(x^{(k)}), d^{(k)})_M}{\|d^{(k)}\|_M^2} \\
&= c \, \frac{\|d^{(k)}\|_M^2}{\|d^{(k)}\|_M^2} && \text{since } d^{(k)} = -\nabla_M f(x^{(k)}) \\
&= c
\end{aligned}$$

with some constant $c > 0$. This simply suggests to impose a lower bound on the initial trial step sizes in gradient descent methods. We will re-label c as $\underline{\alpha}$ in Algorithm 5.22 below.

In addition to observing this bound, it is useful to construct initial trial step sizes using information from past iterations. Assuming that the descent achievable in the current step is equal (to first order) to the descent in the previous step (when the accepted step size was $\alpha^{(k-1)}$), we obtain the following proposal for an initial trial step size $\alpha^{(k,0)}$ at iteration $k \ge 1$:

$$\begin{aligned}
& \alpha^{(k,0)} \, f'(x^{(k)}) \, d^{(k)} = \alpha^{(k-1)} f'(x^{(k-1)}) \, d^{(k-1)} \\
\Rightarrow \quad & \alpha^{(k,0)} = \alpha^{(k-1)} \, \frac{f'(x^{(k-1)}) \, d^{(k-1)}}{f'(x^{(k)}) \, d^{(k)}}.
\end{aligned}$$

Plugging in the descent directions used in the gradient descent method, this becomes

$$\alpha^{(k,0)} = \alpha^{(k-1)} \, \frac{\|\nabla_M f(x^{(k-1)})\|_M^2}{\|\nabla_M f(x^{(k)})\|_M^2} = \alpha^{(k-1)} \, \frac{\|d^{(k-1)}\|_M^2}{\|d^{(k)}\|_M^2}.$$

Alternatively, we could use the actual descent achieved in the previous step instead of its linearization, which would result in

$$\alpha^{(k,0)} = \frac{f(x^{(k-1)}) - f(x^{(k)})}{\|\nabla_M f(x^{(k)})\|_M^2} = \frac{f(x^{(k-1)}) - f(x^{(k)})}{\|d^{(k)}\|_M^2}.$$

We state the full gradient descent method in Algorithm 5.22, using the above considerations for the initial trial step size. As was the case for our methods in § 4 addressing the minimization of quadratic polynomials, we refer to the value of the derivative of f at an iterate $x^{(k)}$ as the **residual** $r^{(k)}$.

The global convergence of Algorithm 5.22, in the sense that every accumulation point of the sequence of iterates $x^{(k)}$ is a stationary point, follows directly from the global convergence theorem 5.9.

Algorithm 5.22 (Gradient descent method for (**UP**) w.r.t. the M-inner product and Armijo backtracking line search).

```
Input: initial guess $x^{(0)} \in \mathbb{R}^n$
Input: routine to evaluate $f$ and $f'$ (or $\nabla f$)
Input: s. p. d. matrix $M$ (or matrix-vector products with $M^{-1}$)
Input: Armijo parameter $\sigma \in (0,1)$          // to be passed through to the Armijo backtracking line search
Input: backtracking parameter $\beta \in (0,1)$ // to be passed through to the Armijo backtracking line search
Input: lower bound $\underline{\alpha} > 0$ for the initial trial step sizes
Output: approximately stationary point of (UP)
 1: Set $k := 0$
 2: Set $f^{(0)} := f(x^{(0)})$                                   // evaluate the initial objective value
 3: Set $r^{(0)} := f'(x^{(0)})^\intercal = \nabla f(x^{(0)})$                        // evaluate the initial residual
 4: Set $d^{(0)} := -M^{-1} r^{(0)}$
 5: Set $\delta^{(0)} := -(r^{(0)})^\intercal d^{(0)}$                  // $\delta^{(0)} = \|\nabla_M f(x^{(0)})\|_M^2 = \|d^{(0)}\|_M^2$
 6: while stopping criterion not met do
 7:     if $k = 0$ then
 8:         Set $\alpha^{(k,0)} := \underline{\alpha}$             // no information from previous iteration available
 9:     else
10:         Set $\alpha^{(k,0)} := \max\{\underline{\alpha}, \frac{f^{(k-1)} - f^{(k)}}{\delta^{(k)}}\}$
11:     end if
12:     Determine a step size $\alpha^{(k)} > 0$ from an Armijo backtracking line search procedure (Algorithm 5.11),
        applied to $\varphi(\alpha) := f(x^{(k)} + \alpha\, d^{(k)})$, with initial trial step size $\alpha^{(k,0)}$, Armijo parameter $\sigma$ and
        backtracking parameter $\beta$      // $\varphi(0) = f^{(k)}$ and $\varphi'(0) = (r^{(k)})^\intercal d^{(k)} = -\delta^{(k)}$ are already known
13:     Set $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$
14:     Set $f^{(k+1)} := f(x^{(k+1)})$        // can be returned by the Armijo backtracking line search routine
15:     Set $r^{(k+1)} := f'(x^{(k+1)})^\intercal = \nabla f(x^{(k+1)})$
16:     Set $d^{(k+1)} := -M^{-1} r^{(k+1)}$
17:     Set $\delta^{(k+1)} := -(r^{(k+1)})^\intercal d^{(k+1)}$          // $\delta^{(k+1)} = \|\nabla_M f(x^{(k+1)})\|_M^2 = \|d^{(k+1)}\|_M^2$
18:     Set $k := k + 1$
19: end while
20: return $x^{(k)}$
```

In Line 12, we could also invoke the modified Armijo backtracking method (Algorithm 5.15), with the backtracking parameter β replaced by the pair of parameters $0 < \underline{\beta} < \overline{\beta} < 1$.

As a stopping criterion, we can choose again any of the conditions from (4.14), i. e., stop on the relative or absolute magnitude of the derivative or gradient

$$\|r^{(k)}\|_{M^{-1}} = \|f'(x^{(k)})\|_{M^{-1}} = \|\nabla_M f(x^{(k)})\|_M = \|d^{(k)}\|_M = (\delta^{(k)})^{1/2}.$$

These quantities are already available in the algorithm. A limited interpretation in the sense of Lemma 4.11 is also possible. In case the sequence $x^{(k)}$ converges to a local minimizer that satisfies the second-order sufficient optimality condition (Theorem 3.3), then we have: for all $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\|x^{(k)} - x^*\|_M \le \delta \quad \text{and} \quad \|f'(x^{(k)})\|_{M^{-1}} \le \varepsilon_{\text{abs}} \quad \Rightarrow \quad \|x^{(k)} - x^*\|_M \le \underbrace{\left(\frac{1}{\alpha} + \varepsilon\right)}_{\approx 1/\alpha} \varepsilon_{\text{abs}},$$

where $\alpha = \lambda_{\min}(f''(x^*); M)$ is the smallest eigenvalue of the Hessian at the solution w.r.t. M. In other words, when we are sufficiently close to a local minimizer satisfying the second-order sufficient optimality condition, then the norm of the derivative (or the gradient) is — up to the factor $1/\alpha$ — a useful measure of the distance to the solution.

Other often used stopping criteria are

$$\|x^{(k)} - x^{(k-1)}\|_M \le \varepsilon^x_{\text{abs}} + \varepsilon^x_{\text{rel}} \, \|x^{(k)} - x^{(0)}\|_M,$$
$$|f(x^{(k)}) - f(x^{(k-1)})| \le \varepsilon^f_{\text{abs}} + \varepsilon^f_{\text{rel}} \, |f(x^{(k)}) - f(x^{(0)})|.$$

These are triggered by slow progress in the iterates or the objective values, respectively. One typically sets $\varepsilon^f_{\text{rel}} = (\varepsilon^x_{\text{rel}})^2$.

It is remarkable that it is possible to monitor the quantities $\|x^{(k)} - x^{(k-1)}\|_M$ and $\|x^{(k)} - x^{(0)}\|_M$, although the matrix M (or matrix-vector products with M) may not be available. Matrix-vector products with M^{-1} are sufficient. The following quantities are useful for this purpose and can be recursively updated, compare (4.33):

$$\omega^{(k)} := \|x^{(k)} - x^{(0)}\|_M^2 \tag{5.21a}$$
$$\xi^{(k)} := (x^{(k)} - x^{(0)})^\intercal M \, d^{(k)} = -(x^{(k)} - x^{(0)})^\intercal r^{(k)} \tag{5.21b}$$
$$\delta^{(k)} := \|d^{(k)}\|_M^2 \tag{5.21c}$$

The details are left as an exercise.

End of Week 4

§ 5.4 Newton's Method

Newton's method is known as a method to solve a (nonlinear) equation $F(x) = 0$, where $F \colon \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function. For optimization purposes, we apply it to the first-order necessary optimality condition, i. e., we have $F(x) = \nabla f(x) = 0$, and thus f is assumed to be of class C^2.

The idea of Newton's method to find a zero (root) of F is as follows. Suppose $x^{(0)}$ is an initial guess. We replace F by its linear Taylor model at $x^{(0)}$ and determine the zero of this model instead. This results in

$$F(x^{(0)}) + F'(x^{(0)})(x - x^{(0)}) = 0 \quad \Leftrightarrow \quad x = x^{(0)} - F'(x^{(0)})^{-1} F(x^{(0)}),$$

provided that the Jacobian $F'(x^{(0)})$ is non-singular. This zero of the linear model is used as the next iterate $x^{(1)}$, etc. This procedure is known as the **(local) Newton's method**.

Algorithm 5.23 (Local Newton's method for $F(x) = 0$)**.**

```
Input: initial guess $x^{(0)} \in \mathbb{R}^n$
Input: routine to evaluate $F$ and $F'$
Output: approximate zero of $F$
Set $k := 0$
```

```
while stopping criterion not met do
    Determine the Newton direction by solving
        $F'(x^{(k)})\, d^{(k)} = -F(x^{(k)})$
    Set $x^{(k+1)} := x^{(k)} + d^{(k)}$
    Set $k := k+1$
end while
return $x^{(k)}$
```

Auxiliary Results

We recall some auxiliary results, which you may know from *Grundlagen der Optimierung* (Herzog, 2022) or other classes. As usual, we equip $\mathbb{R}^n$ with the M-inner product. Recall from § 2.2 that the operator norm of a matrix $K \in \mathbb{R}^{n \times n}$ that represents a map $K\colon \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$\|K\|_{M \leftarrow M} := \max_{x \neq 0} \frac{\|K\,x\|_M}{\|x\|_M}.$$

Although in finite dimensions all norms are equivalent, the above norm is not always the most appropriate choice: some matrices $A \in \mathbb{R}^{n \times n}$ actually represent maps $A\colon \mathbb{R}^n \to (\mathbb{R}^n)^*$, where $(\mathbb{R}^n)^*$ is the dual space of $\mathbb{R}^n$. The appropriate inner product in the dual space is the M^{-1}-inner product, leading to

$$\|A\|_{M^{-1} \leftarrow M} := \max_{x \neq 0} \frac{\|A\,x\|_{M^{-1}}}{\|x\|_M}.$$

Consequently, we would use

$$\|A^{-1}\|_{M \leftarrow M^{-1}} := \max_{r \neq 0} \frac{\|A^{-1}\,r\|_M}{\|r\|_{M^{-1}}}$$

for the inverse of A. We also need the case $B\colon (\mathbb{R}^n)^* \to \mathbb{R}^n$.

Lemma 5.24 (Banach' lemma)**.**

(*i*) *Suppose that* $K \in \mathbb{R}^{n \times n}$ *is a matrix* $\|K\|_{M \leftarrow M} < 1$. *Then* $\operatorname{Id} - K$ *is non-singular, and we have the following estimate on the norm of its inverse:*

$$\|(\operatorname{Id} - K)^{-1}\|_{M \leftarrow M} \le \frac{1}{1 - \|K\|_{M \leftarrow M}}.$$

(*ii*) *Suppose that* $A, B \in \mathbb{R}^{n \times n}$ *are such that* $\|\operatorname{Id} - BA\|_{M \leftarrow M} < 1$. *Then* A *and* B *are both non-singular, and we have*

$$\|B^{-1}\|_{M^{-1} \leftarrow M} \le \frac{\|A\|_{M^{-1} \leftarrow M}}{1 - \|\operatorname{Id} - BA\|_{M \leftarrow M}} \quad \textit{und} \quad \|A^{-1}\|_{M \leftarrow M^{-1}} \le \frac{\|B\|_{M \leftarrow M^{-1}}}{1 - \|\operatorname{Id} - BA\|_{M \leftarrow M}}.$$

Note: Statement (*i*) states that “small” perturbations of the identity matrix are still invertible. Statement (*ii*) states that $\operatorname{Id} - BA$ “small”, i. e., $B \approx A^{-1}$, entails that A and B are both necessarily invertible.

Proof. Statement (i): For $x \in \mathbb{R}^n$, we have

$$\begin{aligned}
\|(\mathrm{Id} - K)\,x\|_M &= \|x - Kx\|_M \\
&\ge \|x\|_M - \|Kx\|_M && \text{by the triangle inequality} \\
&\ge \underbrace{\left(1 - \|K\|_{M \leftarrow M}\right)}_{>0} \|x\|_M && \text{since } \|Kx\|_M \le \|K\|_{M \leftarrow M}\,\|x\|_M.
\end{aligned}$$

This implies $(\mathrm{Id} - K)\,x \neq 0$ for $x \neq 0$, ie, $\mathrm{Id} - K$ is injective and thus non-singular.

Now let $y \in \mathbb{R}^n$ be arbitrary and $x := (\mathrm{Id} - K)^{-1} y$. Then the above estimate shows

$$\begin{aligned}
& \|y\|_M \ge (1 - \|K\|_{M \leftarrow M}) \|(\mathrm{Id} - K)^{-1} y\|_M \\
\Rightarrow \quad & \|(\mathrm{Id} - K)^{-1}\|_{M \leftarrow M} = \max_{y \neq 0} \frac{\|(\mathrm{Id} - K)^{-1} y\|_M}{\|y\|_M} \le \frac{1}{1 - \|K\|_{M \leftarrow M}}.
\end{aligned}$$

Statement (ii): We set $K := \mathrm{Id} - BA$, whence $\|K\|_{M \leftarrow M} < 1$ holds. Due to Statement (i), we find that $\mathrm{Id} - K = \mathrm{Id} - (\mathrm{Id} - BA) = BA$ is non-singular, i. e., A and B are both non-singular. Moreover,

$$\begin{aligned}
& (\mathrm{Id} - K)^{-1} = (BA)^{-1} = A^{-1} B^{-1} \\
\Rightarrow \quad & B^{-1} = A\,(\mathrm{Id} - K)^{-1} \\
\Rightarrow \quad & \|B^{-1}\|_{M^{-1} \leftarrow M} \le \|A\|_{M^{-1} \leftarrow M} \|(\mathrm{Id} - K)^{-1}\|_{M \leftarrow M} \\
& \le \frac{\|A\|_{M^{-1} \leftarrow M}}{1 - \|K\|_{M \leftarrow M}} \quad \text{by Statement } (i) \\
& = \frac{\|A\|_{M^{-1} \leftarrow M}}{1 - \|\mathrm{Id} - BA\|_{M \leftarrow M}}.
\end{aligned}$$

The remaining inequality follows similarly. □

Lemma 5.25 (Implications of the invertibility of the Jacobian)**.** *Suppose that $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function and that $x^* \in \mathbb{R}^n$ is arbitrary with non-singular Jacobian $F'(x^*)$.*

(i) *Then there exists a neighborhood $B_\delta^M(x^*)$ and a constant $c > 0$ such that $F'(x)$ is invertible for all $x \in B_\delta^M(x^*)$. Moreover,*

$$\|F'(x)^{-1}\|_{M \leftarrow M^{-1}} \le c \quad \text{holds for all } x \in B_\delta^M(x^*). \tag{5.22}$$

(ii) *Suppose now in addition that $F(x^*) = 0$ holds. Then there exist a neighborhood $B_\delta^M(x^*)$ and a constant $\beta > 0$ such that*

$$\|x - x^*\|_M \le \beta\,\|F(x)\|_{M^{-1}} \quad \text{for all } x \in B_\delta^M(x^*). \tag{5.23}$$

β can be chosen as $2\,\|F'(x^*)^{-1}\|_{M \leftarrow M^1}$.

Note: Statement (i) is an instance of the fact from functional analysis that the set of boundedly invertible linear operators between two Banach spaces is open. Statement (ii) allows us to estimate the norm of the error $\|x - x^*\|_M$ from the norm of the residual $\|F(x)\|_{M^{-1}}$.

Proof. Statement (*i*): Since F' is continuous at x^*, there exists $\delta > 0$ such that

$$\|F'(x^*) - F'(x)\|_{M^{-1} \leftarrow M} \le \varepsilon := \frac{1}{2\,\|F'(x^*)^{-1}\|_{M \leftarrow M^{-1}}}$$

holds for all $x \in B_\delta^M(x^*)$. Consequently,

$$\begin{aligned} \|\mathrm{Id} - F'(x^*)^{-1} F'(x)\|_{M \leftarrow M} &= \|F'(x^*)^{-1}\big(F'(x^*) - F'(x)\big)\|_{M \leftarrow M} \\ &\le \|F'(x^*)^{-1}\|_{M \leftarrow M^{-1}} \|F'(x^*) - F'(x)\|_{M^{-1} \leftarrow M} \\ &\le \frac{1}{2} < 1. \end{aligned}$$

By Statement (*ii*) of Lemma 5.24 [with $A = F'(x)$ and $B = F'(x^*)^{-1}$], we can conclude that $F'(x)$ is non-singular for all $x \in B_\delta^M(x^*)$ with

$$\|F'(x)^{-1}\|_{M \leftarrow M^{-1}} \le \frac{\|F'(x^*)^{-1}\|_{M \leftarrow M^{-1}}}{1 - \|\mathrm{Id} - F'(x^*)^{-1} F'(x)\|_{M \leftarrow M}} \le 2\,\|F'(x^*)^{-1}\|_{M \leftarrow M^{-1}} =: c.$$

Statement (*ii*): Since F is differentiable in x^*, there exists — for the same $\varepsilon > 0$ as above — a $\delta > 0$ such that

$$\|F(x) - F(x^*) - F'(x^*)(x - x^*)\|_{M^{-1}} \le \varepsilon\,\|x - x^*\|_M \quad \text{for all } x \in B_\delta^M(x^*).$$

Therefore, for all $x \in B_\delta^M(x^*)$,

$$\begin{aligned} &\|F(x)\|_{M^{-1}} \\ &\quad \ge \|F'(x^*)(x - x^*)\|_{M^{-1}} - \|F(x) - \overbrace{F(x^*)}^{=0} - F'(x^*)(x - x^*)\|_{M^{-1}} \quad \text{by the triangle inequality.} \end{aligned}$$

In view of $\|x - x^*\|_M = \|F'(x^*)^{-1} F'(x^*)(x - x^*)\|_M \le \|F'(x^*)^{-1}\|_{M \leftarrow M^{-1}} \|F'(x^*)(x - x^*)\|_{M^{-1}}$, we can estimate this by

$$\begin{aligned} &\|F(x)\|_{M^{-1}} \\ &\quad \ge \frac{1}{\|F'(x^*)^{-1}\|_{M \leftarrow M^{-1}}} \|x - x^*\|_M - \varepsilon\,\|x - x^*\|_M \\ &\quad = \varepsilon\,\|x - x^*\|_M \quad \text{by the definition of } \varepsilon, \end{aligned}$$

and the claim follows with $\beta = \varepsilon^{-1}$. □

Lemma 5.26 (Auxiliary estimate). *Suppose that $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function and $x^* \in \mathbb{R}^n$. For all $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\|F(x) - F(x^*) - F'(x)(x - x^*)\|_{M^{-1}} < \varepsilon\,\|x - x^*\|_M$$

holds for all $x \in B_\delta^M(x^*)$.[24]

[24] Briefly, we can also denote this result as $\|F(x) - F(x^*) - F'(x)(x - x^*)\|_M \in o\left(\|x - x^*\|_M\right)$.

Proof. Take $\varepsilon > 0$. The triangle inequality implies

$$\begin{aligned}&\|F(x) - F(x^*) - F'(x)(x - x^*)\|_{M^{-1}} \\ &\qquad \le \|F(x) - F(x^*) - F'(x^*)(x - x^*)\|_{M^{-1}} + \|F'(x^*) - F'(x)\|_{M^{-1}\leftarrow M}\|x - x^*\|_M.\end{aligned}$$

Since by assumption, F is differentiable in x^*, there exists $\delta_1 > 0$ uch that

$$\|F(x) - F(x^*) - F'(x^*)(x - x^*)\|_{M^{-1}} < \frac{\varepsilon}{2}\,\|x - x^*\|_M$$

holds for all $x \in B^M_{\delta_1}(x^*)$. On the other hand, F' is continuous in x^*, which implies the existence of $\delta_2 > 0$ such that

$$\|F'(x^*) - F'(x)\|_{M^{-1}} < \frac{\varepsilon}{2}$$

holds for all $x \in B^M_{\delta_2}(x^*)$. The conclusion follows with $\delta \coloneqq \min\{\delta_1, \delta_2\}$. □

Local Newton's Method for $F(x) = 0$

We are now in a position to prove a convergence theorem for local Newton's method.

Theorem 5.27 (Convergence of local Newton's method). *Suppose that $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function and that $x^* \in \mathbb{R}^n$ is a point where $F(x^*) = 0$ and $F'(x^*)$ is non-singular. Then there exists a neighborhood $B^M_\delta(x^*)$ such that*

(*i*) *x^* is the unique zero of F in $B^M_\delta(x^*)$.*

(*ii*) *For any initial guess $x^{(0)} \in B^M_\delta(x^*)$, the local Newton's method is well-defined, and it generates a sequence $x^{(k)}$ which converges to x^*.*

(*iii*) *$(x^{(k)})$ converges to x^* Q-superlinearly w.r.t. the M-norm.*

(*iv*) *If F' is Lipschitz continuous in $B^M_\delta(x^*)$, then this convergence is even Q-quadratic.*

Proof. Statement (*i*): By Statement (*ii*) of Lemma 5.25, there exists $\delta_0 > 0$ such that x^* is the only zero of F in $B^M_\delta(x^*)$.

Statement (*ii*): By Statement (*i*) of Lemma 5.25, there exist $\delta_1 > 0$ and $c > 0$ such that $F'(x)$ is non-singular for all $x \in B^M_{\delta_1}(x^*)$ and

$$\|F'(x)^{-1}\|_{M\leftarrow M^{-1}} \le c \coloneqq 2\,\|F(x^*)^{-1}\|_{M\leftarrow M^{-1}}. \tag{$*$}$$

By Lemma 5.26, given $\varepsilon = 1/(2c)$, there exists $\delta_2 > 0$ such that

$$\|F(x) - F(x^*) - F'(x)(x - x^*)\|_{M^{-1}} \le \frac{1}{2c}\|x - x^*\|_M$$

holds for all $x \in B^M_{\delta_2}(x^*)$. Now set $\delta := \min\{\delta_0, \delta_1, \delta_2\}$ and choose $x^{(0)} \in B^M_\delta(x^*)$ arbitrarily. Then the next iterate $x^{(1)} := x^{(0)} - F'(x^{(0)})^{-1}F(x^{(0)})$ is well-defined, and we have

$$\begin{aligned}\|x^{(1)} - x^*\|_M &= \|x^{(0)} - x^* - F'(x^{(0)})^{-1}F(x^{(0)})\|_M \\ &= \|F'(x^{(0)})^{-1}\big[F'(x^{(0)})(x^{(0)} - x^*) - F(x^{(0)}) + \overbrace{F(x^*)}^{=0}\big]\|_M \\ &\le \|F'(x^{(0)})^{-1}\|_{M\leftarrow M^{-1}}\|F(x^{(0)}) - F(x^*) - F'(x^{(0)})(x^{(0)} - x^*)\|_{M^{-1}} \\ &\le c\,\frac{1}{2c}\|x^{(0)} - x^*\|_M \\ &= \frac{1}{2}\|x^{(0)} - x^*\|_M,\end{aligned}$$

and thus $x^{(1)}$ again belongs to $B^M_\delta(x^*)$. By induction, $x^{(k)}$ is well-defined, it belongs to $B^M_\delta(x^*)$, and $x^{(k)} \longrightarrow x^*$ Q-linearly w.r.t. the M-norm.

Statement (iii): We begin by setting up an equation for the error:

$$\begin{aligned}x^{(k+1)} - x^* &= x^{(k)} - x^* - F'(x^{(k)})^{-1}\big(F(x^{(k)}) - F(x^*)\big) \\ &= F'(x^{(k)})^{-1}\big[F'(x^{(k)})(x^{(k)} - x^*) - \big(F(x^{(k)}) - F(x^*)\big)\big] \\ &= F'(x^{(k)})^{-1}\Big[F'(x^{(k)})(x^{(k)} - x^*) - \int_0^1 F'(x^{(k)} + t\,(x^* - x^{(k)}))(x^{(k)} - x^*)\,\mathrm{d}t\Big] \\ &= F'(x^{(k)})^{-1}\Big[\int_0^1 F'(x^{(k)}) - F'(x^{(k)} + t\,(x^* - x^{(k)}))\,\mathrm{d}t\Big](x^{(k)} - x^*).\end{aligned}$$

This gives us the following fundamental estimate:

$$\begin{aligned}&\|x^{(k+1)} - x^*\|_M \\ &\quad\le \big\|F'(x^{(k)})^{-1}\big\|_{M\leftarrow M^{-1}} \underbrace{\int_0^1 \big\|\overbrace{F'(x^{(k)}) - F'(x^{(k)} + t\,(x^* - x^{(k)}))}^{=:D^{(k)}(t)}\big\|_{M^{-1}\leftarrow M}\,\mathrm{d}t}_{=:I^{(k)}}\,\|x^{(k)} - x^*\|_M. \qquad (**)\end{aligned}$$

Due to $x^{(k)} \longrightarrow x^*$, we infer that $x^{(k)} + t\,(x^* - x^{(k)}) \longrightarrow x^*$ uniformly for $t \in [0,1]$. Moreover, F' is continuous, and thus for any $\varepsilon > 0$, there exists an index $k_0 \in \mathbb{N}$ such that

$$\|D^{(k)}(t)\|_{M^{-1}\leftarrow M} \le \varepsilon \quad \text{for all } k \ge k_0 \text{ and all } t \in [0,1].$$

This implies

$$0 \le I^{(k)} = \int_0^1 \|D^{(k)}(t)\|_{M^{-1}\leftarrow M}\,\mathrm{d}t \le \varepsilon \quad \text{for all } k \ge k_0.$$

This in turn gives $I^{(k)} \longrightarrow 0$. But now $(*)$ and $(**)$ give us

$$\|x^{(k+1)} - x^*\|_M \le c\,I^{(k)}\,\|x^{(k)} - x^*\|_M \le c\,\varepsilon\,\|x^{(k)} - x^*\|_M$$

for all $k \ge k_0$, which is the Q-superlinear convergence.

Statement (iv): Since $x^{(k)}$ and $x^{(k)} + t\,(x^* - x^{(k)})$ belong to $B^M_\delta(x^*)$ for all $t \in [0,1]$, we can estimate the integral in a better way, using the stronger assumptions:

$$I^{(k)} = \int_0^1 \big\|F'(x^{(k)}) - F'(x^{(k)} + t\,(x^* - x^{(k)}))\big\|_{M^{-1}\leftarrow M}\,\mathrm{d}t \le \int_0^1 L\,t\,\|x^* - x^{(k)}\|_M\,\mathrm{d}t = \frac{L}{2}\|x^{(k)} - x^*\|_M.$$

From ($**$) we now obtain

$$\|x^{(k+1)} - x^*\|_M \le c\,\frac{L}{2}\|x^{(k)} - x^*\|_M^2.$$

□

Remark 5.28 (on local Newton's method (Algorithm 5.23)).

(i) *Local Newton's method (Algorithm 5.23) may break down since $F'(x^{(k)})$ is not guaranteed to be invertible, in case the initial guess $x^{(0)}$ lies outside the unknown neighborhood of local convergence $B_\delta^M(x^*)$.*

(ii) *The **simplified Newton's method**, which uses the fixed matrix $F'(x^{(0)})$ (assumed to be invertible) instead of $F'(x^{(k)})$, still converges Q-linearly w.r.t. the M-norm.*

Local Newton's Method in Optimization

Newton's method in optimization can be motivated in one of two ways:

(i) The first-order necessary optimality condition for (**UP**) reads

$$\nabla f(x) = 0,$$

see Theorem 3.1. When we employ Newton's method to solve this (generally nonlinear) equation $F(x) = \nabla f(x)$ with Jacobian $F'(x) = f''(x)$, we obtain the iteration

$$x^{(k+1)} = x^{(k)} - f''(x^{(k)})^{-1}\nabla f(x^{(k)}). \tag{5.24}$$

(ii) At the current iterate $x^{(k)}$, we replace (**UP**) by the minimization of the quadratic model

$$q^{(k)}(d) = f(x^{(k)}) + f'(x^{(k)})\,d + \frac{1}{2}\,d^\intercal H^{(k)} d \tag{5.2}$$

where the model Hessian is the symmetric matrix $H^{(k)} = f''(x^{(k)})$. That is, (5.2) becomes the second-order Taylor polynomial. If $H^{(k)}$ is positive definite, then the unique solution of (5.2) is characterized by the linear system

$$f''(x^{(k)})\,d^{(k)} = -\nabla f(x^{(k)}).$$

When one uses the fixed step size $\alpha^{(k)} = 1$ and sets

$$x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)} = x^{(k)} + d^{(k)},$$

we obtain again the iteration (5.24).

Remark 5.29 (on local Newton's method for (**UP**)).

(*i*) *Theorem 5.27 proves the local Q-superlinear (or local Q-quadratic) of local Newton's method towards a stationary point x^* of f, provided that $f''(x^*)$ is non-singular. The point x^* may be a local minimizer, a local maximizer, or a saddle point of f, unless we make an assumption or have knowledge about the definiteness of $f''(x^*)$.*

(*ii*) *If $f''(x^{(k)})$ is s. p. d., then the* ***Newton direction*** *$d^{(k)}$ obtained from the* ***Newton system***

$$f''(x^{(k)})\, d^{(k)} = -\nabla f(x^{(k)}) \tag{5.25}$$

is a descent direction for f at $x^{(k)}$, as long as $f'(x^{(k)}) \neq 0$; compare (5.9):

$$f'(x^{(k)})\, d^{(k)} = -\nabla f(x^{(k)})^\intercal f''(x^{(k)})^{-1} \nabla f(x^{(k)}) < 0.$$

Due to the fixed step size $\alpha^{(k)} = 1$ (instead of line search), descent from iterate to iterate, i. e., $f(x^{(k+1)}) < f(x^{(k)})$, is not guaranteed when $x^{(k)}$ is still "far" from the local minimizer x^*.

(*iii*) *Local Newton's method is invariant w.r.t. affine scaling (see homework problem 5.1). This is in contrast to the steepest descent method.*

(*iv*) *$f''(x)$ is a bilinear form accepting two directions and returning a number. Consequently, when we specify only a single direction, the resulting object becomes a linear form. It is thus appropriate to view the Hessian $f''(x)$ as a map $\mathbb{R}^n \longrightarrow (\mathbb{R}^n)^*$ and to use the associated operator norm.*

A Globalized Newton's Method in Optimization

We now seek to globalize the local Newton's method. In order to be able to apply the global convergence theorem 5.9, we require the search directions and the step sizes to be admissible. We will realize these requirements via a (generalized) angle condition and an Armijo backtracking line search. In addition, we pay attention not to disturb the local Q-superlinear convergence.

Algorithm 5.30 (Globalized Newton method for (**UP**)).

Input: *initial guess $x^{(0)} \in \mathbb{R}^n$*
Input: *routine to evaluate f and f' (or ∇f)*
Input: *routine to evaluate f'' (or matrix-vector products with f'')*
Input: *s. p. d. matrix M (or matrix-vector products with M^{-1})*
Input: *globalization parameters $\eta \in (0,1)$, $\rho > 0$ and exponent $p > 0$*
Input: *Armijo parameter $\sigma \in (0, 1/2)$* *// to be passed through to the Armijo backtracking line search*
Input: *backtracking parameter $\beta \in (0,1)$* *// to be passed through to the Armijo backtracking line search*
Output: *approximately stationary point of* (**UP**)

```
Set $k := 0$
Set $f^{(0)} := f(x^{(0)})$                                   // evaluate the initial objective value
Set $r^{(0)} := f'(x^{(0)})^\intercal = \nabla f(x^{(0)})$          // evaluate the initial residual
Set $d_G^{(0)} := -M^{-1} r^{(0)}$                              // evaluate the negative $M$-gradient
Set $\delta^{(0)} := -(r^{(0)})^\intercal d_G^{(0)}$    // $\delta^{(0)} = \|\nabla_M f(x^{(0)})\|_M^2 = \|d_G^{(0)}\|_M^2$
while stopping criterion not met do
```

```
    Attempt to solve the Newton system
        $$f''(x^{(k)})\, d_N^{(k)} = -r^{(k)} \tag{5.26}$$
    if the Newton system is not solvable or not uniquely solvable then
        Set $d^{(k)} := d_G^{(k)}$                // use the steepest descent direction as fallback
    else                                          // Newton direction $d_N^{(k)}$ available
        Evaluate the generalized angle condition for the Newton direction
        $$f'(x^{(k)})\, d_N^{(k)} \le -\min\{\eta,\ \rho\, \|d_G^{(k)}\|_M^p\}\, \|d_G^{(k)}\|_M\, \|d_N^{(k)}\|_M \tag{5.27}$$
        if true then
            Set $d^{(k)} := d_N^{(k)}$            // use the Newton direction
        else
            Set $d^{(k)} := d_G^{(k)}$            // use the steepest descent direction as fallback
        end if
    end if
    Determine a step size $\alpha^{(k)} > 0$ from an Armijo backtracking line search procedure (Algorithm 5.11),
  applied to $\varphi(\alpha) := f(x^{(k)} + \alpha\, d^{(k)})$, with initial trial step size $\alpha^{(k,0)} = 1$, Armijo parameter $\sigma$ and
  backtracking parameter $\beta$  // $\varphi(0) = f^{(k)}$ and $\varphi'(0) = (r^{(k)})^\intercal d^{(k)} = -\delta^{(k)}$ in case of $d^{(k)} = d_G^{(k)}$, or
  $\varphi'(0) = f'(x^{(k)})\, d_N^{(k)}$ in case of $d^{(k)} = d_N^{(k)}$, are already known
    Set $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$
    Set $f^{(k+1)} := f(x^{(k+1)})$               // can be returned by the Armijo backtracking line search routine
    Set $r^{(k+1)} := f'(x^{(k+1)})^\intercal = \nabla f(x^{(k+1)})$
    Set $d_G^{(k+1)} := -M^{-1} r^{(k+1)}$        // evaluate the negative $M$-gradient
    Set $\delta^{(k+1)} := -(r^{(k+1)})^\intercal d_G^{(k+1)}$   // $\delta^{(k+1)} = \|\nabla_M f(x^{(k+1)})\|_M^2 = \|d_G^{(k+1)}\|_M^2$
    Set $k := k + 1$
end while
return $x^{(k)}$
```

So the basic idea of Algorithm 5.30 is to use the negative M-gradient $d_G^{(k)}$ in case the Newton direction $d_N^{(k)}$ is either not available, or in case it is not a good descent direction. To decide the latter, we verify its angle with the direction of steepest descent. We know that the steepest descent direction $d = d_G^{(k)}$ satisfies the angle condition (5.8), i. e.,

$$f'(x^{(k)})\, d \le -\eta\, \|\nabla_M f(x^{(k)})\|_M\, \|d\|_M = -\eta\, \|d_G^{(k)}\|_M\, \|d\|_M$$

with the maximal possible value, $\eta = 1$. In (5.27), we require qualitatively the same condition for the Newton direction, but with some smaller value $\eta \in (0, 1)$. Moreover, as the norm of the gradient $\|d_G^{(k)}\|_M$ becomes smaller, i. e., as we get closer to being stationary, we wish to encourage the Newton direction to be used in order to enable fast local convergence. In that phase, it is no longer necessary and, in fact, disadvantageous, to limit the angle between the Newton direction and the steepest descent direction. To be concrete, we use the term $\rho\, \|d_G^{(k)}\|_M^p$ to determine whether we are in that phase. This explains the condition

$$f'(x^{(k)})\, d \le -\min\{\eta,\ \rho\, \|d_G^{(k)}\|_M^p\}\, \|d_G^{(k)}\|_M\, \|d\|_M$$

that we employ the check the descent quality of the Newton direction $d_N^{(k)}$ in (5.27). A range of similar conditions achieving the same goal is also conceivable; see for instance Geiger, Kanzow, 1999, Kapitel 9.2 or Ulbrich, Ulbrich, 2012, p.49.

Remark 5.31 (on globalized Newton's method (Algorithm 5.30)).

(*i*) *The parameters ρ and p are often chosen relatively small, e. g.,*

$$\rho = 10^{-6} \quad \textit{and} \quad p = 10^{-1}.$$

(*ii*) *As in our previous algorithms, we may have available the preconditioner only in the form of matrix-vector products with M^{-1}. In order to evaluate* (5.27)*, however, we need to be able to compute $\|d_N^{(k)}\|_M$ as well, which appears to be unavailable.*

There is, however, an elegant way out. If we solve the Newton system (5.24) *using the CG method (Algorithm 4.17) with preconditioner M and initial guess 0, we have available by* (4.33)–(4.34) *the M-norm of the iterates and thus also the M-norm of the solution $d_N^{(k)}$.*

Moreover, the CG method can be easily modified to accommodate the situation that the Newton system is not solvable, or not uniquely solvable. This is the case when a direction of non-positive curvature is encountered during the CG iterations, i. e., when the quantity θ in Algorithm 4.17 becomes ≤ 0. We describe these modifications below (Algorithm 5.41) in the context of inexact Newton methods, where we also take advantage of the fact that it may not be necessary to solve (5.24) *exactly.*

(*iii*) *The approach to globalization taken in Algorithm 5.30 is to reject the Newton direction if it does not exist or does not offer a sufficiently negative directional derivative, and to replace it by the steepest descent direction. There are other approaches that modify the Newton direction so that it always exists and offers sufficient descent. One can, for instance, add a multiple of the identity matrix (or rather a multiple of the preconditioner) to $f''(x^{(k)})$ when the latter is found not to be "sufficiently positive definite". The modified Newton system then reads*

$$\left[f''(x^{(k)}) + \tau M\right] d^{(k)} = -\nabla f(x^{(k)})$$

with some $\tau > 0$; see for instance Geiger, Kanzow, 1999, p.93 and Nocedal, Wright, 2006, p.51.

We now proceed to show the global convergence of Algorithm 5.30.

Theorem 5.32 (Convergence of globalized Newton's method)**.** *Suppose that f is of class C^2. Suppose that x^* is an accumulation point of $x^{(k)}$ and that $\left(x^{(k)}\right)_{k\in K}$ is a subsequence converging to x^*. Then the search directions $\left(d^{(k)}\right)_{k\in K}$ and step sizes $\left(\alpha^{(k)}\right)_{k\in K}$ are admissible. Consequently, we have $f'(x^*) = 0$.*

Proof. We verify the prerequisites of the global convergence theorem 5.9, which then implies $f'(x^*) = 0$. To this end, we set

$$\begin{aligned} K_N &:= \{k \in K : d^{(k)} = d_N^{(k)}\} && \text{(index set of Newton steps)} \\ K_G &:= K \setminus K_N && \text{(index set of gradient steps).} \end{aligned}$$

Step (1) Wir first show the admissibility of the search directions. That is, we have to show that

$$\frac{f'(x^{(k)})\,d^{(k)}}{\|d^{(k)}\|_M} \xrightarrow{k\in K} 0 \quad \text{implies} \quad f'(x^{(k)}) \xrightarrow{k\in K} 0. \tag{5.7'}$$

For indices $k \in K_G$ we have $d^{(k)} = -M^{-1}\nabla f(x^{(k)})$ and thus

$$-\frac{f'(x^{(k)})\,d^{(k)}}{\|d^{(k)}\|_M} = \frac{\|\nabla_M f(x^{(k)})\|_M^2}{\|\nabla_M f(x^{(k)})\|_M} = \|\nabla_M f(x^{(k)})\|_M.$$

The left-hand side of (5.7') thus implies $\|\nabla_M f(x^{(k)})\|_M \xrightarrow{k\in K_G} 0$, which is equivalent to $f'(x^{(k)}) \xrightarrow{k\in K_G} 0$.

For the complementary indices $k \in K_N$, the generalized angle condition (5.27) reads

$$-\frac{f'(x^{(k)})\,d^{(k)}}{\|d^{(k)}\|_M} \ge \min\{\eta,\ \rho\,\|d_G^{(k)}\|_M^p\}\,\|d_G^{(k)}\|_M \ \ge 0.$$

The left-hand side of (5.7') thus implies $\|d_G^{(k)}\|_M = \|\nabla_M f(x^{(k)})\|_M \xrightarrow{K_N} 0$, which is the same as $f'(x^{(k)}) \xrightarrow{k\in K_N} 0$.

Step (2) The convergence of $\big(x^{(k)}\big)_{k\in K}$ and the C^2-property of the objective imply that the subsequence of Hessians $f''(x^{(k)})$ converges as well, and consequently the subsequence $f''(x^{(k)})$ is bounded (in any norm we might impose on the space of n-by-n matrices), so that we have $\|f''(x^{(k)})\|_{M^{-1}\leftarrow M} \le C$ for $k \in K$. For the Newton steps, we recall $f''(x)\,d = -\nabla f(x)$, which we can also write as $-M^{-1}f''(x)\,d = \nabla_M f(x)$. By the definition of matrix norms, see (2.5), we find

$$\|d^{(k)}\|_M \ge \frac{1}{\|f''(x^{(k)})\|_{M^{-1}\leftarrow M}}\|\nabla_M f(x^{(k)})\|_M \ge \frac{1}{C}\|\nabla_M f(x^{(k)})\|_M \quad \text{for } k \in K_N,$$

and clearly

$$\|d^{(k)}\|_M = 1\,\|\nabla_M f(x^{(k)})\|_M \qquad \text{for } k \in K_G,$$

so overall we have

$$\|d^{(k)}\|_M \ge \min\Big\{\frac{1}{C},\ 1\Big\}\,\|\nabla_M f(x^{(k)})\|_M \ge \min\Big\{\frac{1}{C},\ 1\Big\}\,\frac{-f'(x^{(k)})\,d^{(k)}}{\|d^{(k)}\|_M} \tag{5.28}$$

for all $k \in K$. In view of the initial Armijo trial step size being $\alpha^{(k,0)} = 1$, we satisfy condition (5.14) of Lemma 5.13 with $\psi(t) = \min\{t, t/C\}$, which in turn implies the admissibility of the step sizes along the subsequence.

□

Next we show that, under appropriate assumptions, Algorithm 5.30 eventually becomes identical to the local Newton's method, which means that

$$d^{(k)} = d_N^{(k)} \quad \text{and} \quad \alpha^{(k)} = 1 \tag{5.29}$$

holds for all k sufficiently large. Consequently, the local convergence theorem 5.27 applies, which yields the fast (at least Q-superlinear) convergence of the *entire sequence* of iterates, as soon as it is sufficiently close to a local minimizer satisfying second-order sufficient optimality condition.

Theorem 5.33 (Transition to fast local convergence in Algorithm 5.30, see Ulbrich, Ulbrich, 2012, Satz 10.14)**.** *Suppose that f is of class C^2. Suppose that x^* is an accumulation point of $x^{(k)}$ and that $(x^{(k)})_{k \in K}$ is a subsequence converging to x^*. Assume, moreover, that the Hessian $f''(x^*)$ is s. p. d. Then the following holds:*

(i) *$f'(x^*) = 0$ holds, i. e., x^* satisfies the second-order sufficient optimality condition.*

(ii) *The entire sequence $x^{(k)}$ converges to x^*.*

(iii) *There exists an index $k_0 \in \mathbb{N}_0$ such that (5.29) holds for all $k \ge k_0$. Consequently, $x^{(k)}$ converges to x^* Q-superlinearly w.r.t. the M-norm.*

(iv) *If f'' is Lipschitz continuous in a neighorbood of x^*, then the convergence is Q-quadratic.*

Proof. We do not provide the proof but refer the interested reader to Ulbrich, Ulbrich, 2012, Satz 10.14 for the time being. □

§ 5.5 Newton-Like Methods

From the point of convergence analysis, the globalized Newton's method (Algorithm 5.30) is superior to the steepest descent method (Algorithm 5.22) since it offers a Q-superlinear convergence phase. However, Newton's method has a number of drawbacks as well:

(1) The Hessian $f''(x)$ may be expensive to evaluate, and it is needed in addition to the first-order derivative $f'(x)$ of the objective.

(2) The solution of the Newton systems

$$f''(x^{(k)})\, d^{(k)} = -\nabla f(x^{(k)}) \tag{5.25}$$

is often more expensive compared to the evaluation of the gradient direction

$$M\, d^{(k)} = -\nabla f(d^{(k)}).$$

After all, M is constant and can be factorized using the Cholesky decomposition when the number of optimization variables is moderate.

We will address both issues simultaneously. To this end, we consider methods which allow us to

(1) replace the Hessian $f''(x^{(k)})$ by a (s. p. d.) model Hessian $H^{(k)}$ and

(2) solve the linear system

$$H^{(k)} d^{(k)} = -\nabla f(x^{(k)}) \tag{5.30}$$

iteratively, and possibly only inexactly.

The latter means that effectively we are solving a linear system

$$H^{(k)} d^{(k)} = -\nabla f(x^{(k)}) + \zeta^{(k)} \tag{5.31}$$

with an implicitly defined residual $\zeta^{(k)}$. To this end, we will typically specify a tolerance of the form $\|\zeta^{(k)}\|_{M^{-1}} \le \varepsilon^{(k)}$.

As a starting point, we consider a generic local Newton-like method with no line search.

Algorithm 5.34 (Generic Newton-like method for (**UP**)).

```
Input: initial guess $x^{(0)} \in \mathbb{R}^n$
Input: routine to evaluate $f$ and $f'$ (or $\nabla f$)
Input: initial symmetric model Hessian $H^{(0)} \in \mathbb{R}^{n \times n}$ (possibly s. p. d.)
Input: routine to determine the symmetric model Hessians $H^{(k)}$ (possibly s. p. d.)
Input: s. p. d. matrix $M$ (or matrix-vector products with $M^{-1}$)
Output: approximately stationary point of (UP)
Set $k := 0$
while stopping criterion not met do
    Determine a search direction $d^{(k)}$ by (inexactly) solving $H^{(k)} d^{(k)} = -\nabla f(x^{(k)})$
                                        // $H^{(k)} d^{(k)} = -\nabla f(x^{(k)}) + \zeta^{(k)}$ with some residual $\zeta^{(k)}$
    Set $x^{(k+1)} := x^{(k)} + d^{(k)}$
    Determine the next model Hessian $H^{(k+1)}$
    Set $k := k + 1$
end while
return $x^{(k)}$
```

The following questions arise:

(1) What are the requirements for $H^{(k)}$ and $\zeta^{(k)}$ in order to obtain fast ("Newton-like", i. e., Q-superlinear) convergence?

(2) What practical approaches exist to choose the matrices $H^{(k)}$ and to impose a bound for residual norm $\zeta^{(k)}$, with an eye to reducing the numerical effort?

As we did for Newton's method (§ 5.4), we begin by considering an analog of Algorithm 5.34 to find a zero of a C^1 function $F\colon \mathbb{R}^n \to \mathbb{R}^n$. In place of the exact Jacobians $F'(x^{(k)})$, we use model Jacobians $H^{(k)}$, which are supposed to be non-singular but not necessarily symmetric or positive definite.

Algorithm 5.35 (Generic Newton-like method for $F(x) = 0$).

```
Input: initial guess $x^{(0)} \in \mathbb{R}^n$
Input: routine to evaluate $F$
Input: routine to determine the non-singular model Jacobians $H^{(k)}$
Input: s. p. d. matrix $M$ (or matrix-vector products with $M^{-1}$)
Output: approximate zero of $F$
```

```
Set $k := 0$
while stopping criterion not met do
    Determine a search direction $d^{(k)}$ by (inexactly) solving $H^{(k)} d^{(k)} = -F(x^{(k)})$
                                        // $H^{(k)} d^{(k)} = -F(x^{(k)}) + \zeta^{(k)}$ with some residual $\zeta^{(k)}$
    Set $x^{(k+1)} := x^{(k)} + d^{(k)}$
    Set $k := k + 1$
end while
return $x^{(k)}$
```

In a nutshell, the sequence generated by Algorithm 5.35 is governed by

$$\begin{aligned} H^{(k)} d^{(k)} &= -F(x^{(k)}) + \zeta^{(k)} \\ x^{(k+1)} &= x^{(k)} + d^{(k)}. \end{aligned} \tag{5.32}$$

The following lemma shows that the fast local convergence of any sequence $x^{(k)}$ converging to a zero of F is related to the question how well the elements of that sequence satisfy the true Newton systems $F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)}) = 0$.

Lemma 5.36 (Characterization of fast local convergence)**.** *Suppose that $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function and that $x^{(k)}$ is any sequence in $\mathbb{R}^n$ converging to x^* with non-singular Jacobian $F'(x^*)$. Then the following are equivalent:*

(i) *$x^{(k)}$ converges Q-superlinearly w.r.t. the M-norm, and we have $F(x^*) = 0$.*

(ii) *For any $\varepsilon > 0$ there exists an index $k_0 \in \mathbb{N}_0$ such that*[25]

$$\|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \le \varepsilon\, \|x^{(k+1)} - x^{(k)}\|_M \quad \text{for all } k \ge k_0. \tag{5.33a}$$

(iii) *For any $\varepsilon > 0$ there exists an index $k_0 \in \mathbb{N}_0$ such that*[26]

$$\|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \le \varepsilon\, \|x^{(k)} - x^*\|_M \quad \text{for all } k \ge k_0. \tag{5.33b}$$

(iv) *For any $\varepsilon > 0$ there exists an index $k_0 \in \mathbb{N}_0$ such that*[27]

$$\|F(x^{(k)}) + F'(x^*)\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \le \varepsilon\, \|x^{(k+1)} - x^{(k)}\|_M \quad \text{for all } k \ge k_0. \tag{5.33c}$$

Proof. We begin with some preliminary estimates. Since F is of class C^1 and $F'(x^*)$ is non-singular, there exists a neighborhood $B_\delta^M(x^*)$ and constants $c, C > 0$ such that $\|F'(x)\|_{M^{-1}\leftarrow M}$ and $\|F'(x)^{-1}\|_{M\leftarrow M^{-1}}$ hold for all $x \in B_\delta^M(x^*)$; compare Lemma 5.25. The mean value theorem 2.4 gives us

$$F(x^{(k+1)}) - F(x^*) = F'\big(x^* + \xi^{(k)}\,(x^{(k+1)} - x^*)\big)\,(x^{(k+1)} - x^*)$$

[25] briefly: $\|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \in o(\|x^{(k+1)} - x^{(k)}\|_M)$

[26] briefly: $\|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \in o(\|x^{(k)} - x^*\|_M)$

[27] briefly: $\|F(x^{(k)}) + F'(x^*)\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \in o(\|x^{(k+1)} - x^{(k)}\|_M)$

with some $\xi^{(k)} \in (0,1)$. We thus conclude

$$c\,\|x^{(k+1)} - x^*\|_M \le \|F(x^{(k+1)}) - F(x^*)\|_{M^{-1}} \le C\,\|x^{(k+1)} - x^*\|_M. \tag{$*$}$$

for sufficiently large $k \in \mathbb{N}_0$.

Another application of the mean value theorem 2.4 yields

$$\begin{aligned}
&F(x^{(k+1)}) - F(x^{(k)}) \\
&\quad = F'\big(x^{(k)} + \widehat{\xi}^{(k)}\,(x^{(k+1)} - x^{(k)})\big)(x^{(k+1)} - x^{(k)}) \quad \text{where } \widehat{\xi}^{(k)} \in (0,1) \\
&\quad = F'(x^{(k)})(x^{(k+1)} - x^{(k)}) + \big[F'\big(x^{(k)} + \widehat{\xi}^{(k)}\,(x^{(k+1)} - x^{(k)})\big) - F'(x^{(k)})\big](x^{(k+1)} - x^{(k)})
\end{aligned}$$

and thus

$$\begin{aligned}
&\|F(x^{(k+1)}) - F(x^{(k)}) - F'(x^{(k)})(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\
&\qquad \le \|F'\big(x^{(k)} + \widehat{\xi}_k\,(x^{(k+1)} - x^{(k)})\big) - F'(x^{(k)})\|_{M^{-1} \leftarrow M}\,\|x^{(k+1)} - x^{(k)}\|_M.
\end{aligned}$$

As in the proof of Lemma 5.20 we can now exploit the uniform continuity of F' "near the $\big(x^{(k)}\big)$". This entails that, for any $\varepsilon > 0$, there exists an index $k_0 \in \mathbb{N}_0$ such that

$$\|F(x^{(k+1)}) - F(x^{(k)}) - F'(x^{(k)})(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \le \varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M \tag{$**$}$$

holds for all $k \ge k_0$.

Statement (i) $\Rightarrow$ Statement (ii) and Statement (iii): The triangle inequality and the Q-superlinear convergence imply that, for sufficiently large k, we have

$$\|x^{(k)} - x^*\|_M \le \|x^{(k)} - x^{(k+1)}\|_M + \|x^{(k+1)} - x^*\|_M \le \|x^{(k)} - x^{(k+1)}\|_M + \frac{1}{2}\,\|x^{(k)} - x^*\|_M,$$

and thus

$$\|x^{(k)} - x^*\|_M \le 2\,\|x^{(k+1)} - x^{(k)}\|_M. \tag{$***$}$$

On the other hand, the triangle inequality and the Q-superlinear convergence also imply that

$$\|x^{(k+1)} - x^{(k)}\|_M \le \|x^{(k+1)} - x^*\|_M + \|x^* - x^{(k)}\|_M \le 1\,\|x^{(k)} - x^*\|_M + \|x^* - x^{(k)}\|_M,$$

holds for sufficiently large k, whence

$$\|x^{(k+1)} - x^{(k)}\|_M \le 2\,\|x^{(k)} - x^*\|_M. \tag{$****$}$$

In other words, the quantities $\|x^{(k+1)} - x^{(k)}\|_M$ and $\|x^{(k)} - x^*\|_M$ "control each other" for k sufficiently large.

Let $\varepsilon > 0$ be arbitrary. We can estimate

$$\begin{aligned}
&\|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\
&\quad \le \|F(x^{(k+1)}) - F(x^{(k)}) - F'(x^{(k)})(x^{(k+1)} - x^{(k)})\|_{M^{-1}} + \|F(x^{(k+1)}) - \underbrace{F(x^*)}_{=0}\|_{M^{-1}} \\
&\quad \le \varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M + \|F(x^{(k+1)}) - F(x^*)\|_{M^{-1}} \quad \text{by } (**)
\end{aligned}$$

for k sufficiently large. We need to address the second term in the previous inequality:

$$\begin{aligned}\|F(x^{(k+1)}) - F(x^*)\|_{M^{-1}} &\le C\,\|x^{(k+1)} - x^*\|_M && \text{by } (*) \\ &\le C\,\|x^{(k)} - x^*\|_M && \text{by the Q-superlinear convergence.}\end{aligned}$$

for k sufficiently large. Plugging tihs estimate into the previous inequality is Statement (ii).

Moreover, as we demonstrated in $(****)$, $\|x^{(k)} - x^*\|_M$ and $\|x^{(k+1)} - x^{(k)}\|_M$ are different by at most a constant factor, we also have proved Statement (iii).

Statement (ii) or Statement (iii) $\Rightarrow$ Statement (i): We estimate

$$\begin{aligned}&\|F(x^{(k+1)})\|_{M^{-1}} \\ &\quad\le \|F(x^{(k+1)}) - F(x^{(k)}) - F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} + \|F(x^{(k)}) + F'(x^{(k)})(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\ &\quad\le \varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M + \|F(x^{(k)}) + F'(x^{(k)})(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \quad \text{by } (**).\end{aligned}$$

By Statement (ii) or Statement (iii), the second term can be bounded by $\varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M$ or $\varepsilon\,\|x^{(k)} - x^*\|_M$, respectively. In any case, we have

$$\|F(x^{(k+1)})\|_{M^{-1}} \le \varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M + 2\,\varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M = 3\,\varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M$$

for sufficiently large k. In view of $x^{(k)} \to x^*$, we find $F(x^{(k+1)}) \to 0$ and thus $F(x^*) = 0$.

It remains to show the Q-superlinear convergence of $x^{(k)}$. For any $\varepsilon \in (0, c)$, we have

$$\begin{aligned}c\,\|x^{(k+1)} - x^*\|_M &\le \|F(x^{(k+1)})\|_{M^{-1}} && \text{by } (*) \\ &\le 3\,\varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M \\ &\le 6\,\varepsilon\,\|x^{(k+1)} - x^*\|_M && \text{by } (****),\end{aligned}$$

and thus

$$\|x^{(k+1)} - x^*\|_M \le \frac{6\,\varepsilon}{c}\,\|x^{(k)} - x^*\|_M$$

for sufficiently large k. This shows the Q-superlinear convergence of $x^{(k)}$ to x^* w.r.t. the M-norm.

Statement (ii) $\Leftrightarrow$ Statement (iv): The difference of the terms inside the norms on the left hand sides in (5.33a) and (5.33c) is $[F'(x^{(k)}) - F'(x^*)]\,(x^{(k+1)} - x^{(k)})$. Its norm can be estimated as follows:

$$\begin{aligned}&\|[F'(x^{(k)}) - F'(x^*)]\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\ &\quad\le \|F'(x^{(k)}) - F'(x^*)\|_{M^{-1}\leftarrow M}\,\|x^{(k+1)} - x^{(k)}\|_M \\ &\quad\le \varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M \quad \text{by the continuity of } F'\end{aligned}$$

for sufficiently large k. The triangle inequality, either in the form

$$\begin{aligned}&\|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\ &\quad\le \|F(x^{(k)}) + F'(x^*)\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} + \|[F'(x^{(k)}) - F'(x^*)]\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\ &\quad\le \|F(x^{(k)}) + F'(x^*)\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} + \varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M\end{aligned}$$

or in the form

$$\begin{aligned}
&\|F(x^{(k)}) + F'(x^*)\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\
&\quad \le \|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} + \|[F'(x^{(k)}) - F'(x^*)]\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} \\
&\quad \le \|F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)})\|_{M^{-1}} + \varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M,
\end{aligned}$$

each for sufficiently large k, now shows the equivalence of Statement (ii) and Statement (iv). □

We now apply this lemma to Algorithm 5.34, where the sequence of iterates $x^{(k)}$ is generated via (5.32). The residual these iterates leave in the true Newton systems can be expressed as

$$\begin{aligned}
&F(x^{(k)}) + F'(x^{(k)})\,(x^{(k+1)} - x^{(k)}) \\
&\quad = F(x^{(k)}) + F'(x^{(k)})\,d^{(k)} \\
&\quad = F(x^{(k)}) + F'(x^{(k)})\,d^{(k)} \overbrace{- F(x^{(k)}) - H^{(k)}d^{(k)} + \zeta^{(k)}}^{=0} \\
&\quad = \big[F'(x^{(k)}) - H^{(k)}\big]\,d^{(k)} + \zeta^{(k)}.
\end{aligned}$$

We thus obtain from Lemma 5.36 the following corollary.

Corollary 5.37 (Characterization of fast local convergence)**.** *Suppose that $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function and that $x^{(k)}$ is a sequence generated by (5.32) that converges to x^* with non-singular Jacobian $F'(x^*)$. Then the following are equivalent:*

(i) *$x^{(k)}$ converges Q-superlinearly w.r.t. the M-norm, and we have $F(x^*) = 0$.*

(ii) *For any $\varepsilon > 0$ there exists an index $k_0 \in \mathbb{N}_0$ such that*

$$\big\|\big[F'(x^{(k)}) - H^{(k)}\big]\,d^{(k)} + \zeta^{(k)}\big\|_{M^{-1}} \le \varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M \quad \text{for all } k \ge k_0. \tag{5.34a}$$

(iii) *For any $\varepsilon > 0$ there exists an index $k_0 \in \mathbb{N}_0$ such that*

$$\big\|\big[F'(x^{(k)}) - H^{(k)}\big]\,d^{(k)} + \zeta^{(k)}\big\|_{M^{-1}} \le \varepsilon\,\|x^{(k)} - x^*\|_M \quad \text{for all } k \ge k_0. \tag{5.34b}$$

(iv) *For any $\varepsilon > 0$ there exists an index $k_0 \in \mathbb{N}_0$ such that*

$$\big\|\big[F'(x^*) - H^{(k)}\big]\,d^{(k)} + \zeta^{(k)}\big\|_{M^{-1}} \le \varepsilon\,\|x^{(k+1)} - x^{(k)}\|_M \quad \text{for all } k \ge k_0. \tag{5.34c}$$

This set of equivalent of conditions that are necessary and sufficient for the local Q-superlinear convergence are known as **Dennis-Moré conditions**, introduced in Dennis, Moré, 1974. They exhibit that two requisites are sufficient to ensure fast convergence:

(1) The residual in the linear system, $\|\zeta^{(k)}\|_{M^{-1}}$, goes to zero faster than $\|x^{(k+1)} - x^{(k)}\|_M$.

(2) The difference between the Jacobian $F'(x^{(k)})$ and the model Jacobian $H^{(k)}$, evaluated in the direction of $d^{(k)}$, goes to zero faster than $\|x^{(k+1)} - x^{(k)}\|_M$. **Note:** It is not necessary for $H^{(k)}$ to approximate the Jacobian $F'(x^{(k)})$ in its entirety!

We will discuss in the following two classes of methods that are specializations of Algorithm 5.34. The first class of methods are inexact Newton methods (§ 5.6), which use $H^{(k)} = F'(x^{(k)})$. The second class of methods are quasi-Newton algorithms (§ 5.7), which feature $\zeta^{(k)} = 0$.

§ 5.6 Inexact Newton Methods

Inexact Newton methods use the true Jacobian $H^{(k)} = F'(x^{(k)})$ in the linear systems (5.31), but they solve them only inexactly, leaving a residual $\zeta^{(k)}$:

$$F'(x^{(k)})\, d^{(k)} = -F(x^{(k)}) + \zeta^{(k)} \tag{5.35}$$

We measure the norm of the residual in the linear system (5.35) relative to the norm of the outer residual $F(x^{(k)})$ associated with the current iterate $x^{(k)}$. We require

$$\|\zeta^{(k)}\|_{M^{-1}} = \|F'(x^{(k)})\, d^{(k)} + F(x^{(k)})\|_{M^{-1}} \le \eta^{(k)}\, \|F(x^{(k)})\|_{M^{-1}} \tag{5.36}$$

with some $\eta^{(k)} \in (0,1)$. The sequence $\big(\eta^{(k)}\big)$ is known as a **forcing sequence**.

Note that $F(x^{(k)})$ is the residual associated with the zero vector, and hence

$$\frac{\|\text{residual associated with } d^{(k)}\|_{M^{-1}}}{\|\text{residual associated with } 0\|_{M^{-1}}} = \frac{\|\zeta^{(k)}\|_{M^{-1}}}{\|F(x^{(k)})\|_{M^{-1}}} \le \eta^{(k)}. \tag{5.37}$$

Thus we can interpret the forcing sequence as the relative reduction of the residual required in the linear system $F'(x^{(k)})\, d^{(k)} = -F(x^{(k)})$, compared to a zero initial guess. It is evident that we should demand $\eta^{(k)} < 1$. Otherwise, $d^{(k)} = 0$ would constitute a sufficiently accurate solution.

We refer to Algorithm 5.34 as an **inexact local Newton's method** in case $H^{(k)} = F'(x^{(k)})$. For completeness, we state the algorithm as

Algorithm 5.38 (local inexact Newton's method for $F(x) = 0$).

```
Input: initial guess $x^{(0)} \in \mathbb{R}^n$
Input: routine to evaluate $F$ and $F'$
Input: spd Matrix $M$ (oder Matrix-Vektor-Produkte mit $M^{-1}$)
Input: routine to determine the forcing sequence $\eta^{(k)}$
Output: approximate zero of $F$
Set $k := 0$
while stopping criterion not met do
    Determine a search direction $d^{(k)}$ by (inexactly) solving $F'(x^{(k)})\, d^{(k)} = -F(x^{(k)})$ so that the
    residual $\zeta^{(k)} := F'(x^{(k)})\, d^{(k)} + F(x^{(k)})$ satisfies the condition
        $\|\zeta^{(k)}\|_{M^{-1}} \le \eta^{(k)}\, \|F(x^{(k)})\|_{M^{-1}}$        (5.36)
    Set $x^{(k+1)} := x^{(k)} + d^{(k)}$
    Set $k := k + 1$
end while
```

```
7: return $x^{(k)}$
```

Note: With $\eta^{(k)} \equiv 0$, we obtain again the exact local Newton's method.

We can now specify a local convergence theorem for Algorithm 5.38.

Theorem 5.39 (Convergence of Algorithm 5.38; compare Theorem 5.27)**.** *Suppose that $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is a C^1 function and that $x^* \in \mathbb{R}^n$ is a point where $F(x^*) = 0$ and $F'(x^*)$ is non-singular. Suppose that $x^{(k)}$ is a sequence generated by Algorithm 5.38, where the elements of the forcing sequence satisfy $\eta^{(k)} \le \overline{\eta} < 1$ for all $k \in \mathbb{N}_0$. Then there exists a neighborhood $B_\delta^M(x^*)$ such that*

(*i*) *x^* is the unique zero of F in $B_\delta^M(x^*)$.*

(*ii*) *For any initial guess $x^{(0)} \in B_\delta^M(x^*)$, the local inexact Newton's method is well-defined, and it generates a sequence $x^{(k)}$ which converges to x^*.*

(*iii*) *$\left(x^{(k)}\right)$ converges to x^* Q-linearly* w.r.t. *the M-norm.*

(*iv*) *If, in addition, $\eta^{(k)} \searrow 0$ holds, then the convergence is Q-superlinear.*

(*v*) *If F' is Lipschitz continuous in $B_\delta^M(x^*)$, and if, in addition to $\eta^{(k)} \searrow 0$, we require $\eta^{(k)} \le C\,\|F(x^{(k)})\|_{M^{-1}}$ with some constant $C > 0$, then this convergence is even Q-quadratic.*

Proof. We only give a sketch of the proof. Statement (*i*) can be shown as Theorem 5.27. A guide to proving Statement (*ii*) and Statement (*iii*) can be found in Geiger, Kanzow, 1999, Satz 10.3. For Statement (*iv*), we use the characterization of Q-superlinear convergence by Corollary 5.37. We have

$$\begin{aligned}\|\underbrace{\left(F'(x^{(k)}) - H^{(k)}\right)}_{=0} d^{(k)} + \zeta^{(k)}\|_{M^{-1}} &= \|\zeta^{(k)}\|_{M^{-1}} \\ &\le \eta^{(k)}\,\|F(x^{(k)})\|_{M^{-1}} \quad \text{by (5.36).}\end{aligned}$$

As in the proof of Lemma 5.36, see $(*)$, we have $\|F(x^{(k)})\|_{M^{-1}} \le C\,\|x^{(k)} - x^*\|_M$ for sufficiently large indices k and thus

$$\|\left(F'(x^{(k)}) - H^{(k)}\right) d^{(k)} + \zeta^{(k)}\|_{M^{-1}} \le \eta^{(k)}\,C\,\|x^{(k)} - x^*\|_M.$$

Since $\eta^{(k)} \searrow 0$, we satisfy (5.34b).

Statement (*v*) follows similarly as in Theorem 5.27. □

A possible rule for the choice of the forcing sequence $\eta^{(k)}$ that guarantees the local Q-superlinear convergence is

$$\eta^{(k)} := \min\left\{\overline{\eta},\ \|F(x^{(k)})\|_{M^{-1}}^{\theta}\right\} \tag{5.38}$$

with some $\overline{\eta} < 1$ and $\theta \in (0,1]$, for instance $\overline{\eta} = 1/2$ and $\theta = 0.5$.[28]

[28] More precisely, we even obtain the Q-superlinear convergence with rate $1+\theta$ with this choice.

In the remainder of § 5.6 we consider a practical approach to the inexact solution of the Newton systems, while simultaneously globalizing the inexact local Newton's method (Algorithm 5.38). Since in this class we are discussing globalization for Newton-like methods only in the context of optimization (and not for general root-finding), we switch back to the optimization context now. That is, we have $F(x) = \nabla f(x)$ and $F'(x) = f''(x)$.

We need to take into account the following:

(1) The Newton system $f''(x^{(k)})\, d = -\nabla f(x^{(k)})$ is to be solved *iteratively*[29]. In this way, we can take advantage of the fact that an inexact solution is sufficient and we can stop once the residual norm for the linear system falls below the threshold dictated by the forcing sequence; see (5.36). We refer to the inexact Newton direction as $d_N^{(k)}$.

(2) The inexact Newton direction $d_N^{(k)}$ is required to be, at the very least, a descent direction for the objective f at the current outer iterate $x^{(k)}$.

(3) As we did in the globalized exact Newton's method (Algorithm 5.30), we need to verify whether the inexact Newton direction $d_N^{(k)}$ offers sufficient descent. If not, then we fall back to taking a step in the steepest descent direction.

It turns out that we can reach the first and the second goal simultaneously by a clever use of the conjugate gradient method (Algorithm 4.17), applied to the symmetric linear system $A\, d = b$, where

$$A = f''(x^{(k)}) \quad \text{and} \quad b = -\nabla f(x^{(k)}).$$

As a stopping criterion, we employ the relative criterion (4.14a) with $\varepsilon_{\text{rel}} = \eta^{(k)}$, and the zero vector serves as initial guess. In case the CG algorithm finishes "without an incidence", then — due to (5.37) — the solution returned is an inexact solution of the Newton system with sufficiently small residual norm in the sense of (5.36).

Remark 5.40 (inner and outer iterations)**.** *In what follows we will sometimes use the terms inner iterations and outer iterations. The* ***outer iterations*** *of those of the outer optimization method, which is the inexact Newton's method in this subsection. The quantities used in the outer iterations are the iterates $x^{(k)}$, search directions $d^{(k)}$, step sizes $\alpha^{(k)}$, etc.*

On the other hand, every search direction $d^{(k)}$ will now be found in an iterative way, which refer to as ***inner iterations****. In order to help avoid confusion, we will denote the inner iteration index by ℓ. Also, the iterates of the inner solver for the linear system $A\, d = b$ will be termed $d^{(\ell)}$ instead of $x^{(\ell)}$. The search directions in the inner solver will be $p^{(\ell)}$ instead of $d^{(\ell)}$. The residuals in the inner solver will be $\zeta^{(\ell)}$ instead of $r^{(\ell)}$.*

What could be the incidences that might occur in the CG algorithm in the present context? On the one hand, we might reach the maximum number of iterations before reaching the relative tolerance. On

[29] rather than using a direct solver such as Gaussian elimination

the other hand, the symmetric matrix A might not be positive definite. This means that the function

$$\phi(z) := \frac{1}{2} z^\intercal A\, z - b^\intercal z$$

has a least one direction $p \in \mathbb{R}^n$, $p \neq 0$, of non-positive curvature; i. e., $p^\intercal A\, p \le 0$ holds. A lack of positive definiteness does not mean that a search direction of non-positive curvature will actually be encountered during the inner iterations. On the one hand, the required tolerance may be reached beforehand. But even for exact solutions ($\varepsilon_{\text{rel}} = 0$), not all right hand sides b actually invoke directions of non-positive curvature.

In any case, if a direction $p^{(\ell)}$ with $\theta^{(\ell)} := (p^{(\ell)})^\intercal A\, p^{(\ell)} \le 0$ is encountered, a reaction is required since otherwise,

- in case $\theta^{(\ell)} = 0$, a division by zero would occur in Line 8 of the CG algorithm (Algorithm 4.17),
- in case $\theta^{(\ell)} < 0$, the CG algorithm could be continued; however, we might lose the property that the iterates $d^{(\ell)}$ are descent directions for f at $x^{(k)}$. This can be confirmed by examples. As long as all search directions $p^{(\ell)}$ are directions of positive curvature ($\theta^{(\ell)} > 0$), the descent property remains intact; see Lemma 5.42.

For the reasons above, it is customary to employ a variant of the CG method known as **truncated conjugate gradient method** (**truncated CG method**) as inner solver in a globalized inexact Newton method. Starting from a zero initial guess, iterate until either the relative stopping criterion (4.14a) is verified, or a search direction of non-positive curvature is encountered. In that case, the most recent iterate $d^{(\ell)}$ is returned as inexact solution.

For completeness, we state below the truncated CG algorithm. Notice that we chose the specific stopping criterion (5.37) instead of a general criterion.

Algorithm 5.41 (Truncated conjugate gradient method for symmetric systems $A\,d = b$ w.r.t. the M-inner product; compare Algorithm 4.17).

```
Input: right-hand side $b \in \mathbb{R}^n$
Input: symmetric matrix $A$ (or matrix-vector products with $A$)
Input: s. p. d. matrix $M$ (or matrix-vector products with $M^{-1}$)
Input: relative residual $\varepsilon_{\text{rel}}$
Output: approximate solution of $A\,d = b$
1:  Set $\ell := 0$
2:  Set $d^{(0)} := 0$                                  // zero initial guess
3:  Set $\zeta^{(0)} := -b$                             // evaluate the initial residual
4:  Set $p^{(0)} := -M^{-1}\zeta^{(0)}$
5:  Set $\delta^{(0)} := -(\zeta^{(0)})^\intercal p^{(0)}$          // $\delta^{(0)} = \|\zeta^{(0)}\|^2_{M^{-1}}$
6:  while $\delta^{(\ell)} \ge \varepsilon_{\text{rel}}^2 \delta^{(0)}$ do          // check stopping criterion (5.37)
7:      Set $q^{(\ell)} := A\,p^{(\ell)}$
8:      Set $\theta^{(\ell)} := (q^{(\ell)})^\intercal p^{(\ell)}$
9:      if $\theta^{(\ell)} > 0$ then
10:         Set $\alpha^{(\ell)} := \delta^{(\ell)} / \theta^{(\ell)}$
```

```
        Set $d^{(\ell+1)} := d^{(\ell)} + \alpha^{(\ell)} p^{(\ell)}$
        Set $\zeta^{(\ell+1)} := \zeta^{(\ell)} + \alpha^{(\ell)} q^{(\ell)}$
        Set $p^{(\ell+1)} := -M^{-1}\zeta^{(\ell+1)}$
        Set $\delta^{(\ell+1)} := -(\zeta^{(\ell+1)})^\top p^{(\ell+1)}$        // $\delta^{(\ell+1)} = \|\zeta^{(\ell+1)}\|^2_{M^{-1}}$
        Set $\beta^{(\ell+1)} := \delta^{(\ell+1)} / \delta^{(\ell)}$
        Set $p^{(\ell+1)} := p^{(\ell+1)} + \beta^{(\ell+1)} p^{(\ell)}$
        Set $\ell := \ell + 1$
    else
        Abort the while loop
    end if
end while
return $d^{(\ell)}$
```

It still remains to be proved that the approximate solution $d^{(\ell)}$ that the truncated CG method generates and that is to be used as inexact Newton direction $d_N^{(k)}$, is indeed a descent direction for the objective f at the current outer iterate $x^{(k)}$. This means that we need to show $f'(x^{(k)})\, d^{(\ell)} < 0$ or equivalently, $b^\top d^{(\ell)} > 0$.

Lemma 5.42 (The truncated CG method generates descent directions)**.** *Suppose that $b \neq 0$ and that $d^{(0)}, \ldots, d^{(\ell)}$ have been generated by Algorithm 5.41 for some $\ell \geq 1$. Then the following holds.*

(i) $b^\top M^{-1}\zeta^{(j)} = 0$ *for* $j = 1, \ldots, \ell$.

(ii) $b^\top p^{(j)} = \|\zeta^{(j)}\|^2_{M^{-1}}$ *for* $j = 0, \ldots, \ell$.

(iii) $b^\top d^{(\ell)} = \sum_{j=0}^{\ell-1} \alpha^{(j)}\, \|\zeta^{(j)}\|^2_{M^{-1}}$ *is positive and strictly monotonically increasing in* ℓ.

Proof. Statement *(i)*: Since we use the zero vector as initial guess, we have $\zeta^{(0)} = A\,0 - b = -b$ for the initial residual. Therefore,

$$b^\top M^{-1}\zeta^{(j)} = -(\zeta^{(0)})^\top M^{-1}\zeta^{(j)} = 0 \quad \text{for } j \geq 1$$

according to (4.28).

Statement *(ii)*: The initial search direction is $p^{(0)} = -M^{-1}\zeta^{(0)}$, and hence we have

$$b^\top p^{(0)} = (\zeta^{(0)})^\top M^{-1}\zeta^{(0)} = \|\zeta^{(0)}\|^2_{M^{-1}}.$$

By induction, we find for $j \geq 0$:

$$\begin{aligned} b^\top p^{(j+1)} &= b^\top\big(-M^{-1}\zeta^{(j+1)} + \beta^{(j+1)}\, p^{(j)}\big) \\ &= 0 + \beta^{(j+1)}\, b^\top p^{(j)} && \text{by Statement } (i) \\ &= \frac{\|\zeta^{(j+1)}\|^2_{M^{-1}}}{\|\zeta^{(j)}\|^2_{M^{-1}}}\, b^\top p^{(j)} && \text{by (4.24')} \\ &= \|\zeta^{(j+1)}\|^2_{M^{-1}} && \text{by the induction hypothesis.} \end{aligned}$$

Statement *(iii)*: Since Algorithm 5.41 generated the iterates $d^{(0)}, \dots, d^{(\ell)}$, the numbers $\theta^{(0)}, \dots, \theta^{(\ell-1)}$ are all strictly positive. Consequently, $\alpha^{(j)} = \delta^{(j)}/\theta^{(j)} > 0$ is also positive for $j = 0, \dots, \ell - 1$. We consider the expression

$$b^\mathsf{T} d^{(\ell)} = b^\mathsf{T} \sum_{j=0}^{\ell-1} \alpha^{(j)} \, p^{(j)} = \sum_{j=0}^{\ell-1} \alpha^{(j)} \, \|\zeta^{(j)}\|_{M^{-1}}^2$$

with the last equality due to Statement *(ii)*. The residuals $\zeta^{(0)}, \dots, \zeta^{(\ell-1)}$ are all $\neq 0$, otherwise the stopping criterion in Algorithm 5.41 would have been triggered. Therefore, the above expression is strictly increasing w.r.t. ℓ. □

Remark 5.43 (on Algorithm 5.41).

(i) *The first search direction is $p^{(0)} = M^{-1} b$, which is equal to the steepest descent direction $-M^{-1} \nabla f(x^{(k)})$ in the optimization context. When $p^{(0)}$ is a direction of positive curvature (if $\theta^{(0)} > 0$), then $d^{(1)}$ is the same as though we had applied the steepest descent method with Cauchy step size (Algorithm 4.6).*

(ii) *By contrast, when $p^{(0)}$ is a direction of non-positive curvature, Algorithm 5.41 stops and returns $d^{(0)} = 0$. This is, of course, not a useful descent direction for the outer, inexact Newton method, as will be detected by a quality test for the inexact Newton direction, and a fallback to a gradient step will be the consequence.*

(iii) *The strictly increasing monotonicity of $b^\mathsf{T} d^{(\ell)} = -f'(x^{(k)}) \, d^{(\ell)}$ w.r.t. the iteration counter ℓ means that the descent properties of the iterates $d^{(\ell)}$ progressively improve, as long as the search directions $p^{(\ell)}$ remain directions of positive curvature for A. Therefore, it is reasonable to continue performing CG iterations until either the desired tolerance is reached, or a direction of non-positive curvature is encountered. This is the strategy Algorithm 5.41 is following.*

As we already mentioned, the globalization of the inexact Newton method can be done along the same lines as in Algorithm 5.30. This leads to the following algorithm.

Algorithm 5.44 (Globalized inexact Newton method for (**UP**); compare Algorithm 5.30).

```
Input: initial guess $x^{(0)} \in \mathbb{R}^n$
Input: routine to evaluate $f$ and $f'$ (or $\nabla f$)
Input: routine to evaluate $f''$ (or matrix-vector products with $f''$)
Input: s. p. d. matrix $M$ (or matrix-vector products with $M^{-1}$)
Input: routine to determine the forcing sequence $\eta^{(k)}$
Input: globalization parameters $\eta \in (0,1)$, $\rho > 0$ and exponent $p > 0$
Input: Armijo parameter $\sigma \in (0, 1/2)$      // to be passed through to the Armijo backtracking line search
Input: backtracking parameter $\beta \in (0,1)$ // to be passed through to the Armijo backtracking line search
Output: approximately stationary point of (UP)
Set $k := 0$
Set $f^{(0)} := f(x^{(0)})$                          // evaluate the initial objective value
Set $r^{(0)} := f'(x^{(0)})^\mathsf{T} = \nabla f(x^{(0)})$             // evaluate the initial residual
Set $d_G^{(0)} := -M^{-1} r^{(0)}$                     // evaluate the negative M-gradient
Set $\delta^{(0)} := -(r^{(0)})^\intercal d_G^{(0)}$    // $\delta^{(0)} = \|\nabla_M f(x^{(0)})\|_M^2 = \|d_G^{(0)}\|_M^2$
while stopping criterion not met do
    Determine the inexact Newton direction $d_N^{(k)}$ using Algorithm 5.41 with $A = f''(x^{(k)})$, $b = -r^{(k)}$, preconditioner $M$ and relative residual $\varepsilon_{\text{rel}} = \eta^{(k)}$
    Evaluate the generalized angle condition for the Newton direction
        $$f'(x^{(k)})\, d_N^{(k)} \le -\min\{\eta,\ \rho \, \|d_G^{(k)}\|_M^p\} \, \|d_G^{(k)}\|_M \, \|d_N^{(k)}\|_M \qquad (5.27)$$
    if true then
        Set $d^{(k)} := d_N^{(k)}$    // use the inexact Newton direction
    else
        Set $d^{(k)} := d_G^{(k)}$    // use the steepest descent direction as fallback
    end if
    Determine a step size $\alpha^{(k)} > 0$ from an Armijo backtracking line search procedure (Algorithm 5.11), applied to $\varphi(\alpha) := f(x^{(k)} + \alpha\, d^{(k)})$, with initial trial step size $\alpha^{(k,0)} = 1$, Armijo parameter $\sigma$ and backtracking parameter $\beta$    // $\varphi(0) = f^{(k)}$ and $\varphi'(0) = (r^{(k)})^\intercal d^{(k)} = -\delta^{(k)}$ in case of $d^{(k)} = d_G^{(k)}$, or $\varphi'(0) = f'(x^{(k)})\, d_N^{(k)}$ in case of $d^{(k)} = d_N^{(k)}$, are already known
    Set $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$
    Set $f^{(k+1)} := f(x^{(k+1)})$    // can be returned by the Armijo backtracking line search routine
    Set $r^{(k+1)} := f'(x^{(k+1)})^\intercal = \nabla f(x^{(k+1)})$
    Set $d_G^{(k+1)} := -M^{-1} r^{(k+1)}$    // evaluate the negative $M$-gradient
    Set $\delta^{(k+1)} := -(r^{(k+1)})^\intercal d_G^{(k+1)}$    // $\delta^{(k+1)} = \|\nabla_M f(x^{(k+1)})\|_M^2 = \|d_G^{(k+1)}\|_M^2$
    Set $k := k + 1$
end while
return $x^{(k)}$
```

Remark 5.45 (on Algorithm 5.44).

(*i*) *See Remark 5.31 on choosing the globalization parameters ρ and p.*

(*ii*) *The quantity $\|d_N^{(k)}\|_M$ required to evaluate the generalized angle condition (5.27) can be returned at negligible additional cost by the truncated CG algorithm (Algorithm 5.41), as described in (4.33)–(4.34).*

The global convergence of Algorithm 5.44 can be verified very similarly as in Theorem 5.32. In fact, **Step** (1) in the proof (admissibility of search directions) remains exactly the same since the generalized angle condition and the fallback to steepest descent directions remains the same as in Algorithm 5.30. In **Step** (2) (admissibility of step sizes), we need to take into account the fact that the inexact Newton direction satisfies the Newton system only with a residual. We end up replacing the estimate (5.28) by

$$\|d^{(k)}\|_M \ge \min\Big\{\frac{1-\eta^{(k)}}{C},\ 1\Big\} \, \|\nabla_M f(x^{(k)})\|_M \ge \min\Big\{\frac{1-\eta^{(k)}}{C},\ 1\Big\} \, \frac{-f'(x^{(k)})\, d^{(k)}}{\|d^{(k)}\|_M} \tag{5.39}$$

for all $k \in K$, and we have to modify the function ψ accordingly. (**Quiz 5.7:** Can you fill in the details?)

The transition to fast local convergence can be shown similarly as in Theorem 5.33. We can verify again that

$$d^{(k)} = d_N^{(k)} \quad \text{and} \quad \alpha^{(k)} = 1 \tag{5.29}$$

holds for sufficiently large indices k. Consequently, the convergence mode (Q-linear, Q-superlinear or even Q-quadratic) follows depending on the choice of forcing sequence, using Theorem 5.39; see also Geiger, Kanzow, 1999, Satz 10.8.

The combination of the inexact Newton method as outer algorithm with the truncated CG algorithm as inner solver is often referred to as **truncated Newton CG method**. Since we do not necessarily need to set up the full Hessian matrix $f''(x^{(k)})$, but matrix-vector products with $f''(x^{(k)})$ are sufficient, one also speaks of a **Hessian-free optimization**. Matrix-vector products with $f''(x^{(k)})$ can be realized, e. g., using algorithmic diffentiation techniques (Chapter 4).

End of Week 5

§ 5.7 Quasi-Newton Methods

In contrast to inexact Newton methods (§ 5.6), **quasi-Newton methods** make use of the freedom Newton-like methods offer in a different way. We discuss quasi-Newton methods only in the context of optimization. A quasi-Newton method determines symmetric (and often positive definite) approximations $H^{(k)}$ of the Hessians $f''(x^{(k)})$ but then solves the systems

$$H^{(k)} d^{(k)} = -\nabla f(x^{(k)}) \tag{5.40}$$

exactly before taking a step in the direction $d^{(k)}$. Quasi-Newton methods are also known as **variable metric methods** since, in contrast to gradient methods which maintain $H^{(k)} \equiv M$ constant, the model Hessian $H^{(k)}$ changes from iteration to iteration.

The model Hessian $H^{(k+1)}$ is constructed based on information gained in the k-th step. A consistent requirement in all quasi-Newton methods is the so-called **secant condition** or **quasi-Newton condition**,

$$H^{(k+1)}(x^{(k+1)} - x^{(k)}) = \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}). \tag{5.41}$$

This condition can be motivated in several ways:

(1) It follows from the fundamental theorem of calculus[30] that

$$\nabla f(x+d) = \nabla f(x) + f''(x+d)\, d + \int_0^1 [f''(x+t\,d) - f''(x+d)]\; d \, \mathrm{d}t \tag{5.42}$$

holds for all $x, d \in \mathbb{R}^n$. For the integral term, we have the following estimate, which uses the C^2 property of f: for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\left\| \int_0^1 [f''(x+t\,d) - f''(x+d)]\; d \, \mathrm{d}t \right\|_{M^{-1}} \le \varepsilon \, \|d\|_M$$

[30] The function $t \mapsto \nabla f(x+t\,d)$ is the integral of its derivative $t \mapsto f''(x+t\,d)\,d$, and thus $\int_0^1 f''(x+t\,d)\,d\,\mathrm{d}t = \nabla f(x+t\,d)\big|_0^1 = \nabla f(x+d) - \nabla f(x)$

holds for all $\|d\|_M \le \delta$.[31]

Omitting this integral term from (5.42), plugging in $x^{(k)}$ for x and $x^{(k+1)} - x^{(k)}$ for d, and replacing the true Hessian $f''(x+d)$ by its approximation $H^{(k+1)}$, we obtain

$$\nabla f(x^{(k+1)}) = \nabla f(x^{(k)}) + H^{(k+1)}(x^{(k+1)} - x^{(k)}),$$

which is the secant condition (5.41). Hence the secant condition approximately mimics the property (5.42) of the true Hessian.

(2) Let us consider the two quadratic models[32] at the consecutive iterates $x^{(k)}$ and $x^{(k+1)}$,

$$m^{(k)}(x) = f(x^{(k)}) \quad + f'(x^{(k)})(x - x^{(k)}) \quad + \frac{1}{2}(x - x^{(k)})^\intercal H^{(k)}(x - x^{(k)}), \tag{5.43a}$$

$$m^{(k+1)}(x) = f(x^{(k+1)}) + f'(x^{(k+1)})(x - x^{(k+1)}) + \frac{1}{2}(x - x^{(k+1)})^\intercal H^{(k+1)}(x - x^{(k+1)}). \tag{5.43b}$$

By construction, the derivative of the old model $m^{(k)}$ agrees with the derivative of f at $x^{(k)}$. Also, the derivative of the new model $m^{(k+1)}$ agrees with the derivative of f at $x^{(k+1)}$. We require now, in addition, that the *derivative of the new model* agrees with the derivative of f also *at the old iterate* $x^{(k)}$, i. e.,

$$\begin{aligned} & (m^{(k+1)})'(x^{(k)}) = f'(x^{(k)}) \\ \Leftrightarrow \quad & \nabla m^{(k+1)}(x^{(k)}) = \nabla f(x^{(k)}) \\ \Leftrightarrow \quad & \nabla f(x^{(k+1)}) + H^{(k+1)}(x^{(k)} - x^{(k+1)}) = \nabla f(x^{(k)}), \quad \text{since } H^{(k+1)} \text{ is symmetric} \\ \Leftrightarrow \quad & \text{secant condition (5.41).} \end{aligned}$$

Using the Dennis-Moré conditions (Corollary 5.37) for Newton-like methods, we can now characterize the fast (Q-superlinear) local convergence (without line search) of quasi-Newton methods.

Theorem 5.46 (Fast local convergence of quasi-Newton methods). *Suppose that $f: \mathbb{R}^n \to \mathbb{R}$ is a C^2 function. Suppose that the sequence $x^{(k)}$ is generated using the Newton-like method Algorithm 5.34 with model Hessians $H^{(k)}$ which are symmetric and satisfy the secant condition (5.41), and with zero residuals $\zeta^{(k)} = 0$. Suppose that $x^{(k)}$ converges to x^*, where $f''(x^*)$ is non-singular. Finally, suppose that for any $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\|(H^{(k+1)} - H^{(k)})\, d^{(k)}\|_{M^{-1}} \le \varepsilon\, \|d^{(k)}\|_M \tag{5.44}$$

holds for all $\|d^{(k)}\|_M \le \delta$.[33]. Then the conditions of Corollary 5.37 hold, and therefore $x^{(k)}$ converges to x^* Q-superlinearly w.r.t. the M-norm, and $f'(x^*) = 0$ holds.

[31] We could likewise say that the left-hand side belongs to $o(\|d\|_M)$.

[32] In contrast to (5.2), we write the model here in terms of x, not in terms of the direction d, which is notationally more convenient. That is, we write $m^{(k)}(x)$ rather than $q^{(k)}(d)$.

[33] In other words, $\|(H^{(k+1)} - H^{(k)})\, d^{(k)}\|_{M^{-1}} \in o(\|d^{(k)}\|_M)$

Proof. We can estimate

$$\begin{aligned}
&\|(f''(x^{(k)}) - H^{(k)})\, d^{(k)}\|_{M^{-1}} \\
&\quad \le \|(f''(x^{(k)}) - H^{(k+1)})\, d^{(k)}\|_{M^{-1}} + \|(H^{(k+1)} - H^{(k)})\, d^{(k)}\|_{M^{-1}} \quad \text{by the triangle inequality.}
\end{aligned}$$

By assumption, the second term is bounded by $\varepsilon \, \|d^{(k)}\|_M$, provided that $\|d^{(k)}\|_M \le \delta$:

$$\le \|f''(x^{(k)})\, d^{(k)} - \nabla f(x^{(k+1)}) + \nabla f(x^{(k)})\|_{M^{-1}} + \varepsilon \, \|d^{(k)}\|_M \qquad \text{by the secant condition.}$$

Using the uniform continuity of f'' "near the $\big(x^{(k)}\big)$" (as in the proof of Lemma 5.13), we can bound the first term by $\varepsilon \, \|x^{(k+1)} - x^{(k)}\|_M$, provided that $\|x^{(k+1)} - x^{(k)}\|_M$ is sufficiently small:

$$\begin{aligned}
&\le \varepsilon \, \|x^{(k+1)} - x^{(k)}\|_M + \varepsilon \, \|d^{(k)}\|_M \\
&= 2\, \varepsilon \, \|x^{(k+1)} - x^{(k)}\|_M.
\end{aligned}$$

We have shown that condition (5.34a) of Corollary 5.37 holds, which implies the Q-superlinear convergence and the stationarity of x^*. □

In view of

$$\|(H^{(k+1)} - H^{(k)})\, d^{(k)}\|_{M^{-1}} \le \|H^{(k+1)} - H^{(k)}\|_{M^{-1} \leftarrow M} \, \|d^{(k)}\|_M,$$

the convergence $H^{(k+1)} - H^{(k)} \longrightarrow 0$ is sufficient to ensure the prerequisite (5.44) of Theorem 5.46.

Let us now discuss how to construct quasi-Newton matrices $H^{(k)}$ in practice. From now on, we also include a step size $\alpha^{(k)} > 0$ in the update of the iterates

$$x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}.$$

To simplify notation, we introduce

$$s^{(k)} := x^{(k+1)} - x^{(k)} = \alpha^{(k)} d^{(k)} \quad \text{and} \quad y^{(k)} := \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}). \tag{5.45}$$

One important class of quasi-Newton methods constructs $H^{(k+1)}$ as a function of the previous matrix $H^{(k)}$ and the two vectors $s^{(k)}$ and $y^{(k)}$:

$$H^{(k+1)} := \Phi\big(H^{(k)}, s^{(k)}, y^{(k)}\big). \tag{5.46}$$

Due to the dependence of $H^{(k+1)}$ on $H^{(k)}$, we also speak of a **quasi-Newton update formula**.

For reference purposes, we state a generic line-search quasi-Newton method in Algorithm 5.47.

Algorithm 5.47 (Generic globalized quasi-Newton method for (**UP**)).

Input: *initial guess* $x^{(0)} \in \mathbb{R}^n$

Input: *routine to evaluate* f *and* f' *(or* ∇f*)*

Input: *initial symmetric model Hessian* $H^{(0)} \in \mathbb{R}^{n \times n}$ *(possibly s. p. d.)*

Input: *routine that implements the quasi-Newton update* Φ

Input: *s. p. d. matrix* M *(or matrix-vector products with* M^{-1}*)*

Output: *approximately stationary point of* (**UP**)

```
Set $k := 0$
Set $f^{(0)} := f(x^{(0)})$                                  // evaluate the initial objective value
Set $r^{(0)} := f'(x^{(0)})^\intercal = \nabla f(x^{(0)})$   // evaluate the initial residual
Set $d_G^{(0)} := -M^{-1} r^{(0)}$                           // evaluate the negative $M$-gradient
Set $\delta^{(0)} := -(r^{(0)})^\intercal d_G^{(0)}$          // $\delta^{(0)} = \|\nabla_M f(x^{(0)})\|_M^2 = \|d_G^{(0)}\|_M^2$
while stopping criterion not met do
    Determine the quasi-Newton direction $d^{(k)}$ by solving
        $H^{(k)} d^{(k)} = -r^{(k)}$                          (5.40)
    Determine a step size $\alpha^{(k)} > 0$ from a line search procedure with preferred step size $\alpha^{(k)} = 1$
    Set $s^{(k)} := \alpha^{(k)} d^{(k)}$
    Set $x^{(k+1)} := x^{(k)} + s^{(k)}$
    Set $f^{(k+1)} := f(x^{(k+1)})$
    Set $r^{(k+1)} := f'(x^{(k+1)})^\intercal = \nabla f(x^{(k+1)})$
    Set $d_G^{(k+1)} := -M^{-1} r^{(k+1)}$                   // evaluate the negative $M$-gradient
    Set $\delta^{(k+1)} := -(r^{(k+1)})^\intercal d_G^{(k+1)}$  // $\delta^{(k+1)} = \|\nabla_M f(x^{(k+1)})\|_M^2 = \|d_G^{(k+1)}\|_M^2$
    Set $y^{(k)} := r^{(k+1)} - r^{(k)}$
    Determine the quasi-Newton matrix $H^{(k+1)} := \Phi\big(H^{(k)}, s^{(k)}, y^{(k)}\big)$
    Set $k := k + 1$
end while
return $x^{(k)}$
```

We recall the following desirable properties:

(1) $H^{(k+1)}$ must be symmetric (provided that $H^{(k)}$ is symmetric).

(2) $H^{(k+1)}$ must satisfy the secant condition (5.41), now written in the form

$$H^{(k+1)} s^{(k)} = y^{(k)}. \tag{5.47}$$

(3) $H^{(k+1)}$ should be close to $H^{(k)}$ in the sense that $\|(H^{(k+1)} - H^{(k)})\, s^{(k)}\|_{M^{-1}} \in o(\|s^{(k)}\|_M)$ in order to satisfy the condition (5.44) for fast local convergence.

(4) Ideally, the matrix $H^{(k+1)}$ should be positive definite (provided that $H^{(k)}$ is positive definite).

Notice that condition (4) guarantees that $d^{(k)}$, and thus $s^{(k)} = \alpha^{(k)} d^{(k)}$, is a descent direction.

A necessary condition for the positive definiteness of $H^{(k+1)}$ is obtained by multiplying the secant condition (5.41) by $s^{(k)}$:

$$0 < \underbrace{(s^{(k)})^\intercal H^{(k+1)} s^{(k)}}_{\substack{\text{evaluation of the bilinear form } H^{(k+1)} \\ \text{in the particular direction } s^{(k)}}} = (y^{(k)})^\intercal\, s^{(k)} = \big(f'(x^{(k+1)}) - f'(x^{(k)})\big)(x^{(k+1)} - x^{(k)}). \tag{5.48}$$

For strictly convex functions, (5.48) is always satisfied in view of Theorem 2.9; see (2.26). Otherwise, we need a line search procedure to guarantee (5.48). The Wolfe-Powell line search lends itself for this purpose.

Lemma 5.48 (Wolfe-Powell line search ensures (5.48)). *Suppose that $d^{(k)}$ is a descent direction for f at $x^{(k)}$. If $\alpha^{(k)} > 0$ satisfies the curvature condition (5.17) for some $\tau < 1$, then the necessary condition (5.48) for positive definiteness of $H^{(k+1)}$ is satisfied.*

Proof. We can estimate

$$\begin{aligned} f'(x^{(k+1)})\, d^{(k)} &\ge \tau\, f'(x^{(k)})\, d^{(k)} && \text{by the curvature condition (5.17)} \\ &> f'(x^{(k)})\, d^{(k)} && \text{since } d^{(k)} \text{ is a descent condition.} \end{aligned}$$

This can be rewritten as

$$(y^{(k)})^\intercal d^{(k)} > 0,$$

and since $s^{(k)} = \alpha^{(k)} d^{(k)}$ holds with positive step size $\alpha^{(k)}$, (5.48) follows. □

There are infinitely many possibilities (in case $n > 1$) to satisfy the secant condition (5.47). (**Quiz 5.8:** Why?) We are now describing some of the most prominent quasi-Newton update formulas of the form

$$H^{(k+1)} := \Phi\big(H^{(k)}, s^{(k)}, y^{(k)}\big). \tag{5.46}$$

- **SR1** (**Symmetric rank-1**) update:
  There is only one symmetric rank-1 update formula that satisfies the secant condition (Nocedal, Wright, 2006, Chapter 6.2), and it is given by

  $$\Phi_{\text{SR1}}(H, s, y) = H + \frac{(y - H\, s)\,(y - H\, s)^\intercal}{(y - H\, s)^\intercal s}, \tag{5.49}$$

  which requires that $H^{(k)} s^{(k)} \neq y^{(k)}$ holds throughout the iterations. That, however, is problematic. Suppose that we have taken a full step ($\alpha^{(k)} = 1$), i. e., $s^{(k)} = d^{(k)}$ holds. Suppose, moreover, that the step size $\alpha^{(k)} = 1$ gave us an almost stationary point of the line search function φ. Then

  $$\begin{aligned} &(y^{(k)} - H^{(k)} s^{(k)})^\intercal s^{(k)} \\ &\quad = (y^{(k)} - H^{(k)} d^{(k)})^\intercal d^{(k)} && \text{due to } s^{(k)} = d^{(k)} \\ &\quad = \big(\nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) - H^{(k)} d^{(k)}\big)^\intercal d^{(k)} && \text{since } y^{(k)} = \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) \\ &\quad = f'(x^{(k+1)})\, d^{(k)} && \text{due to the quasi-Newton system (5.40)} \\ &\quad = \varphi'(\alpha^{(k)}), \end{aligned}$$

  which is almost zero.

  Moreover, the positive definiteness of $H_{\text{SR1}}^{(k+1)}$ cannot be guaranteed, even if $H^{(k)}$ was positive definite, since the denominator in (5.49) may be negative. (Still, the SR1 has its purpose, in particular in the context of trust-region methods.)

- **PSB** (**Powell-symmetric-Broyden**) update:
  Condition (3) from our list of desirable properties suggests to keep $H^{(k+1)}$ close to $H^{(k)}$. To this

end, we consider the auxiliary problem

$$\begin{aligned} &\text{Minimize} && \frac{1}{2}\|H^{(k+1)} - H^{(k)}\|_F^2, \quad \text{where } H^{(k+1)} \in \mathbb{R}^{n\times n}_{\text{sym}} \\ &\text{subject to} && \text{the secant condition (5.47).} \end{aligned} \tag{5.50}$$

Here $\|\cdot\|_F$ is the Frobenius norm, see (2.7). Problem (5.50) can be shown to have a unique solution[34], which is given in terms of the update formula

$$\Phi_{\text{PSB}}(H, s, y) = H + \frac{(y - H\,s)\,s^\intercal + s\,(y - H\,s)^\intercal}{s^\intercal s} - (y - H\,s)^\intercal s\,\frac{s\,s^\intercal}{(s^\intercal s)^2} \tag{5.51}$$

This formula is said to be a **rank-2 update** since the rank of $H^{(k+1)} - H^{(k)}$ is at most 2. To see this, consider that Φ_{PSB} is of the form

$$\begin{aligned} \Phi_{\text{PSB}}(H, s, y) &= H + v\,s^\intercal + s\,v^\intercal + \gamma\,s\,s^\intercal \quad \text{with } v = (y - H\,s)/\|s\|^2 \text{ and } \gamma = \frac{v^\intercal s}{\|s\|^2} \\ &= H + \underbrace{\gamma\,(s - \gamma^{-1}v)(s - \gamma^{-1}v)^\intercal}_{\text{symmetric, rank 1}} - \underbrace{\gamma^{-1}v\,v^\intercal}_{\text{symmetric, rank 1}}. \end{aligned}$$

Like SR1, the PSB update formula also cannot guarantee the positive definiteness.

- **DFP** (**Davidon-Fletcher-Powell**) update:
  The DFP update considers an auxiliary problem

  $$\begin{aligned} &\text{Minimize} && \frac{1}{2}\|W^{-\intercal}(H^{(k+1)} - H^{(k)})\,W^{-1}\|_F^2, \quad \text{where } H^{(k+1)} \in \mathbb{R}^{n\times n}_{\text{sym}} \\ &\text{subject to} && \text{the secant condition (5.47),} \end{aligned}$$

  where W is any non-singular matrix with the property $W^\intercal W s^{(k)} = y^{(k)}$. Using (2.6) and the property $\operatorname{trace}(ABC) = \operatorname{trace}(BCA)$ for products of matrices, and setting $M := W^\intercal W$ (which is s. p. d.), we can rewrite this problem as

  $$\begin{aligned} &\text{Minimize} && \frac{1}{2}\operatorname{trace}\big(M^{-1}(H^{(k+1)} - H^{(k)})\,M^{-1}(H^{(k+1)} - H^{(k)})\big), \quad \text{where } H^{(k+1)} \in \mathbb{R}^{n\times n}_{\text{sym}} \\ &\text{subject to} && \text{the secant condition (5.47)} \end{aligned} \tag{5.52}$$

  with data $M\,s^{(k)} = y^{(k)}$.[35]

  It can be shown that the unique solution of problem (5.52) is independent of M and it is given by

  $$\begin{aligned} \Phi_{\text{DFP}}(H, s, y) &= (\mathrm{Id} - \rho\,y\,s^\intercal)\,H\,(\mathrm{Id} - \rho\,s\,y^\intercal) + \rho\,y\,y^\intercal \\ &= H + \rho\,(y - H\,s)\,y^\intercal + \rho\,y\,(y - H\,s)^\intercal - \rho^2(y - H\,s)^\intercal s\,y\,y^\intercal, \end{aligned} \tag{5.53}$$

  where $\rho = 1/(y^\intercal s)$. This update formula can be seen to also be of rank 2. This time, positive definiteness can be guaranteed; see Lemma 5.49.

[34] see for instance Ulbrich, Ulbrich, 2012, p.76

[35] The average Hessian $\int_0^1 f''(x^{(k)} + t\,s^{(k)})\,\mathrm{d}t$ is one possible matrix M, provided it is positive definite.

- **BFGS** (**Broyden-Fletcher-Goldfarb-Shanno**) update:
  The BFGS formula starts from the optimization problem

$$\begin{aligned} &\text{Minimize} && \frac{1}{2}\|W[(H^{(k+1)})^{-1}-(H^{(k)})^{-1}]\,W^\intercal\|_F^2, \quad H^{(k+1)}\in\mathbb{R}^{n\times n}_{\text{sym}} \\ &\text{subject to} && \text{the secant condition (5.47)}, \end{aligned}$$

  where W is again any non-singular matrix with the property $W^\intercal W s^{(k)} = y^{(k)}$. Similarly as in (5.52), we can rewrite this problem as

$$\begin{aligned} &\text{Minimize} && \frac{1}{2}\operatorname{trace}\big(M[(H^{(k+1)})^{-1}-(H^{(k)})^{-1}]\,M\,[(H^{(k+1)})^{-1}-(H^{(k)})^{-1}]\big), \quad H^{(k+1)}\in\mathbb{R}^{n\times n}_{\text{sym}} \\ &\text{subject to} && \text{the secant condition (5.47)} \end{aligned} \tag{5.54}$$

  with data $M\,s^{(k)} = y^{(k)}$.

  The solution of this problem — once again independent of W — results in the rank-2 update formula

$$\Phi_{\text{BFGS}}(H,s,y) = H - \frac{H\,s\,s^\intercal H}{s^\intercal H\,s} + \rho\,y\,y^\intercal \tag{5.55}$$

  where $\rho = 1/(y^\intercal s)$. Also here, Lemma 5.49 will ensure the positive definiteness.

- **Broyden class** update:
  The update formulas of the Broyden class are the affine combinations of the DFP and BFGS formulas. For any parameter $\lambda\in\mathbb{R}$, we obtain

$$\Phi^{\lambda}_{\text{Broyden}}(H,s,y) = (1-\lambda)\,\Phi_{\text{BFGS}}(H,s,y) + \lambda\,\Phi_{\text{DFP}}(H,s,y). \tag{5.56}$$

  The formulas obtained by restricting $\lambda\in[0,1]$ are known as the **convex Broyden class**.

In (5.48) we had identified $y^\intercal s > 0$ as a necessary condition for any quasi-Newton update formula satisfying the secant condition to be positive definite. Indeed, this condition is already sufficient in case of the DFP and BFGS updates. Consequently, the positive definiteness can be ensured using a Wolfe-Powell line search as we proved in Lemma 5.48.

Lemma 5.49 (Positive definiteness of the DFP and BFGS updates)**.** *Suppose that H is symmetric and positive definite and that $y^\intercal s > 0$ holds. Then $H^+_{\text{BFGS}} := \Phi_{\text{BFGS}}(H,s,y)$ and $H^+_{\text{DFP}} := \Phi_{\text{DFP}}(H,s,y)$ are symmetric and positive definite as well.*

Note: One can show this result even for all members of the non-negative Broyden class ($\lambda \ge 0$); see Ulbrich, Ulbrich, 2012, Satz 13.4.

Proof. The symmetry of H^+_{BFGS} and H^+_{DFP} are obvious. For any $v\in\mathbb{R}^n$, $v\neq 0$, we have

$$v^\intercal H^+_{\text{DFP}}\,v = (v^\intercal - \rho\,(v^\intercal y)\,s^\intercal)\,H(v-\rho\,s\,(v^\intercal y)) + \rho\,(v^\intercal y)^2 \ge 0.$$

Equality in this inequality can hold only if both summands are zero. The first summand is zero precisely when v is a certain multiple of s, but in that case the second summand will be strictly positive due to $\rho = 1/(y^\mathsf{T} s) > 0$.

As for BFGS, we have

$$v^\mathsf{T} H^+_{\text{BFGS}}\, v = v^\mathsf{T} H\, v - \frac{(s^\mathsf{T} H\, v)^2}{s^\mathsf{T} H\, s} + \rho\, (y^\mathsf{T} v)^2$$

By the Cauchy-Schwarz inequality w.r.t. the H-inner product, we can estimate this as

$$\begin{aligned} v^\mathsf{T} H^+_{\text{BFGS}}\, v &\ge v^\mathsf{T} H\, v - \frac{(s^\mathsf{T} H\, s)\,(v\, H\, v)}{s^\mathsf{T} H\, s} + \frac{(y^\mathsf{T} v)^2}{y^\mathsf{T} s} \\ &= \frac{(y^\mathsf{T} v)^2}{y^\mathsf{T} s} \\ &\ge 0. \end{aligned}$$

If this expression were equal to zero, then both inequalities would need to be equalities. In the first inequality, this implies that v is a (non-zero) multiple of s. But then the second inequality is strict since $y^\mathsf{T} s > 0$ holds. □

While quasi-Newton methods avoid the evaluation of second-order derivatives of the objective f, we still need to solve a linear system (5.40) in every iteration. This brings up the question whether we could perhaps work with the *inverse matrix*[36] $B^{(k)} = (H^{(k)})^{-1}$ and avoid the solution of linear systems altogether, by evaluating

$$d^{(k)} = -B^{(k)} \nabla f(x^{(k)}) \quad \text{instead of solving} \quad H^{(k)} d^{(k)} = -\nabla f(x^{(k)}). \tag{5.57}$$

This is indeed possible, and we can find update formulas for the inverse matrices. The fact that common update formulas for $H^{(k)}$ are of low rank can be exploited, and it leads to low-rank update formulas for $B^{(k)}$. This is a consequence of the following key result, which has applications far beyond quasi-Newton methods.

Lemma 5.50 (Sherman-Morrison-Woodbury formula)**.** *Suppose that $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{r\times r}$ are non-singular matrices and that $U \in \mathbb{R}^{n\times r}$, $V \in \mathbb{R}^{r\times n}$ are arbitrary. Then $A + U\,C\,V$ is non-singular if and only if $C^{-1} + VA^{-1}U$ is non-singular. In this case,*

$$(A + U\,C\,V)^{-1} = A^{-1} - A^{-1}\,U\,(C^{-1} + VA^{-1}U)^{-1}\,VA^{-1}. \tag{5.58}$$

The significance of this lemma is the following. Typically we have $r \ll n$. Knowing the inverse A^{-1}, we can evaluate the inverse of the perturbed matrix $A + U\,C\,V$ with little effort, since the matrix $C^{-1} + VA^{-1}U$ on the right hand side, which is to be inverted, is only of size $r \times r$. In particular, the Sherman-Morrison-Woodbury lemma 5.50 states that the inverse of a rank-r update is a rank-r update of the inverse.

[36] We are pointing out that in part of the literature, notably in Nocedal, Wright, 2006, the notations for B and H are reversed.

Proof. Suppose first that $C^{-1} + V A^{-1} U$ is non-singular. Then it can be checked straightforwardly that the inverse of $A + U\,C\,V$ is given by the right-hand side in (5.58). For the converse statement, we can reverse the roles as follows:

$$A \leftrightsquigarrow C^{-1}, \quad C \leftrightsquigarrow A^{-1}, \quad V \leftrightsquigarrow U, \quad U \leftrightsquigarrow V.$$

□

Dropping indices, the update formulas we discussed in (5.46) are of the form

$$H^+ = \Phi(H, s, y).$$

We will obtain update formulas for the inverse of the form

$$B^+ = \Psi(B, s, y)$$

with the property that $B = H^{-1}$ implies $B^+ = (H^+)^{-1}$. Since Φ_{DFP} and Φ_{BFGS} are expressed in terms of rank-2 updates of the input H, the Sherman-Morrison-Woodbury lemma 5.50 allows us to express also Ψ_{DFP} and Ψ_{BFGS} in terms of rank-2 update formulas. Indeed, we can obtain the

- **inverse DFP quasi-Newton update**:
  The Sherman-Morrison-Woodbury formula applied to the DFP update formula (5.53) yields

$$\Psi_{\text{DFP}}(B, s, y) = B - \frac{B\,y\,y^\intercal B}{y^\intercal B\,y} + \rho\,s\,s^\intercal \tag{5.59}$$

  where, again, $\rho = 1/(y^\intercal s)$.

- **inverse BFGS quasi-Newton update**:
  Similarly, we can obtain

$$\begin{aligned}\Psi_{\text{BFGS}}(B, s, y) &= (\text{Id} - \rho\,s\,y^\intercal)\,B\,(\text{Id} - \rho\,y\,s^\intercal) + \rho\,s\,s^\intercal \\ &= B + \rho\,(s - B\,y)\,s^\intercal + \rho\,s\,(s - B\,y)^\intercal - \rho^2 (s - B\,y)^\intercal y\,s\,s^\intercal,\end{aligned} \tag{5.60}$$

cf. homework problem 6.4. Interestingly, it turns out that the DFP and BFGS updates are inverse to each other. More precisely, we have

$$\Psi_{\text{DFP}}(\cdot, s, y) = \Phi_{\text{BFGS}}(\cdot, y, s) \quad \text{and} \quad \Psi_{\text{BFGS}}(\cdot, s, y) = \Phi_{\text{DFP}}(\cdot, y, s). \tag{5.61}$$

Using an inverse quasi-Newton formula and obtaining the quasi-Newton direction from (5.57) solves both issues with Newton's method identified in the beginning of § 5.5. In the literature, the BFGS update is reported to be generally the most efficient among the quasi-Newton updates. For completeness, we therefore now state a globalized algorithm using the inverse BFGS update.

Algorithm 5.51 (Globalized quasi-Newton method with inverse BFGS update for (UP)).
Input: *initial guess* $x^{(0)} \in \mathbb{R}^n$

```
Input: routine to evaluate $f$ and $f'$ (or $\nabla f$)
Input: routine that implements the inverse BFGS update $\Psi_{\text{BFGS}}$
Input: s. p. d. matrix $M$ (or matrix-vector products with $M^{-1}$)
Input: Armijo and curvature parameters $0 < \sigma < \tau < 1$ with $\sigma \in (0, 1/2)$ // to be passed through to the Wolfe-Powell line search
Input: expansion parameter $\gamma > 1$    // to be passed through to the Wolfe-Powell line search
Input: nesting parameters $\underline{\gamma}, \overline{\gamma} \in (0, 1/2]$    // to be passed through to the Wolfe-Powell line search
Output: approximately stationary point of (UP)
Setze $k := 0$
Setze $f^{(0)} := f(x^{(0)})$
Setze $r^{(0)} := f'(x^{(0)})^\intercal = \nabla f(x^{(0)})$
Setze $d_G^{(0)} := -M^{-1} r^{(0)}$
Setze $\delta^{(0)} := -(r^{(0)})^\intercal d_G^{(0)}$    // $\delta^{(0)} = \|\nabla_M f(x^{(0)})\|_M^2 = \|d_G^{(0)}\|_M^2$
Input: Set $B_{\text{BFGS}}^{(0)} := M^{-1}$    // initial model Hessian equals the user-defined base metric
while stopping criterion not met do
    Determine the quasi-Newton direction $d^{(k)}$ from
        $$d^{(k)} := -B_{\text{BFGS}}^{(k)} \nabla f(x^{(k)})$$
    Determine a step size $\alpha^{(k)} > 0$ from a Wolfe-Powell line search procedure (Algorithm 5.18), applied to $\varphi(\alpha) := f(x^{(k)} + \alpha\, d^{(k)})$, with initial trial step size $\alpha^{(k,0)}$, Armijo parameter $\sigma$, curvature parameter $\tau$, expansion parameter $\gamma$ and nesting parameters $\underline{\gamma}, \overline{\gamma}$ // $\varphi(0) = f^{(k)}$ is already known and $\varphi'(0) = (r^{(k)})^\intercal d^{(k)}$ is easily evaluated
    Set $s^{(k)} := \alpha^{(k)} d^{(k)}$
    Set $x^{(k+1)} := x^{(k)} + s^{(k)}$
    Set $f^{(k+1)} := f(x^{(k+1)})$
    Set $r^{(k+1)} := f'(x^{(k+1)})^\intercal = \nabla f(x^{(k+1)})$
    Set $d_G^{(k+1)} := -M^{-1} r^{(k+1)}$    // evaluate the negative $M$-gradient
    Set $\delta^{(k+1)} := -(r^{(k+1)})^\intercal d_G^{(k+1)}$    // $\delta^{(k+1)} = \|\nabla_M f(x^{(k+1)})\|_M^2 = \|d_G^{(k+1)}\|_M^2$
    Set $y^{(k)} := r^{(k+1)} - r^{(k)}$
    Evaluate $B_{\text{BFGS}}^{(k+1)} := \Psi_{\text{BFGS}}(B_{\text{BFGS}}^{(k)}, s^{(k)}, y^{(k)})$
    Set $k := k + 1$
end while
return $x^{(k)}$
```

Remark 5.52 (on Algorithm 5.51).

(*i*) *The choice $B_{\text{BFGS}}^{(0)} = M^{-1}$ lends itself as the inverse of the initial model Hessian. In this way, the user-defined base metric M serves as the initial model Hessian, as in gradient descent methods. However, in contrast to gradient descent methods, the metric then evolves according to the data $s^{(k)}$ and $y^{(k)}$ acquired throughout the iterations.*

(*ii*) *Unfortunately, the convergence results for Algorithm 5.51 are not as rich as for other methods.*

- *One can show the local Q-superlinear convergence of $x^{(k)}$ to a point x^* satisfying the second-order sufficient optimality condition (see Theorem 3.3), provided that $x^{(0)}$ is sufficiently close* *to x^*, the step sizes are fixed to $\alpha^{(k)} = 1$, and the initial Hessian M is sufficiently close to $f''(x^*)$.*

- *Global convergence can be proved under the assumption that the generalized condition numbers of the inverse BFGS matrices $B^{(k)}_{\text{BFGS}}$ w.r.t. M^{-1} remains bounded. This is equivalent to the generalized condition numbers of the (non-inverse) BFGS matrices $H^{(k)}_{\text{BFGS}}$ w.r.t. M remaining bounded. (**Quiz 5.9:** Can you see why this is equivalent?)*

  Under this assumption, Lemma 5.5 ensures that the angle condition holds, and thus the search directions are admissible (Lemma 5.4). Moreover, under the assumption that the sublevel set $\mathcal{M}_f(x^{(0)}) \coloneqq \{x \in \mathbb{R}^n \mid f(x) \le f(x^{(0)})\}$ is compact, the Wolfe-Powell step sizes can be shown to be admissible (Ulbrich, Ulbrich, 2012, Satz 9.5). The global convergence, in the sense that every accumulation point of the sequence of iterates $x^{(k)}$ is a stationary point, then follows from the global convergence theorem 5.9. See Ulbrich, Ulbrich, 2012, Satz 13.11 for details.

- *Unfortunately, the boundedness of the (inverse) BFGS matrices' generalized condition numbers cannot be guaranteed a priori. One possible remedy is to introduce a generalized angle condition such as*

$$f'(x^{(k)})\, d^{(k)} \le -\min\{\eta,\ \rho\, \|d_G^{(k)}\|_M^p\}\, \|d_G^{(k)}\|_M\, \|d^{(k)}\|_M \tag{5.27}$$

  into Algorithm 5.51. Due to the estimate

$$-f'(x^{(k)})\, d \ge \frac{2\sqrt{\kappa}}{\kappa + 1} \|d_G^{(k)}\|_M\, \|d\|_M$$

  for all $d \in \mathbb{R}^n$ that we have from Lemma 5.5, a violation of condition (5.27) either means that the residual norm $\|d_G^{(k)}\|_M = \|f'(x^{(k)})\|_{M^{-1}}$ has already become small (so that further convergence can be entrusted to the local convergence result), or it otherwise indicates run-away generalized condition numbers of $H^{(k)}$ w.r.t. M.[37]

  When (5.27) is violated, we might reset the matrix to $B^{(k)}_{\text{BFGS}} \coloneqq B^{(0)}_{\text{BFGS}}$, discard and re-evaluate the current direction $d^{(k)}$, effectively resorting to a steepest descent step. Under these modifications of Algorithm 5.51, the a priori assumption on the boundedness of the (inverse) BFGS matrices can be dropped in the proof of global convergence. For further details, see Geiger, Kanzow, 1999, p.167 and the reference Kosmol, 1989, Kapitel 11.5 they are citing.

(iii) *In practice, Algorithm 5.51 often exhibits Q-superlinear convergence. This is remarkable since only first-order derivatives of f are being used.*

Limited-Memory BFGS Method

While the quasi-Newton methods we discussed so far successfully avoid second-order derivatives of the objective and replace the solutions of linear systems by matrix-vector products to obtain the search

[37] Notice that $\kappa \mapsto \frac{2\sqrt{\kappa}}{\kappa+1}$ is monotone decreasing and it goes to 0 if and only if $\kappa \to \infty$.

directions, one disadvantage remains. That issue is the high memory requirement to store the inverse quasi-Newton matrices $B^{(k)}$ (or their non-inverted counterparts $H^{(k)}$) when the problem dimension n is not small. For instance, a problem of the (still moderate) dimension $n = 10\,000$ already requires

$$\frac{n\,(n+1)}{2}\,\underbrace{8\,\mathrm{B}}_{\text{storage for one double precision number}} \approx 381\,\mathrm{MiB}$$

of storage, while a problem of size $n = 100\,000$ requires about 37 GiB.[38]!

Typically (but not always), the true Hessian f'' is sparse for large-scale problems. By contrast, the quasi-Newton matrices $H^{(k)}$ and their inverses $B^{(k)}$ are always fully populated, although their difference to their initial values $H^{(k)}$ and $B^{(k)}$ is at most of rank $2k$.

Two ideas can be used to solve the storage issue.

(1) Instead of storing the matrices $B^{(k)}$ entry by entry, we only store the pairs of vectors

$$(y^{(0)}, s^{(0)}), (y^{(1)}, s^{(1)}), \ldots$$

As we will see, this is sufficient to evalulate the matrix-vector products $B^{(k)}r$.

(2) The above modification still requires us to store two additional vectors of length n per iteration. We can, however, limit the storage by keeping only the most recent m pairs of vectors and dropping the previous.

The combination of these ideas leads to **limited-memory quasi-Newton methods**. While the limited-memory idea is generally applicable to all low-rank quasi-Newton update formulas, we concentrate here on the inverse BFGS update. We start by reviewing the respective update formula (5.60), which leads to the recursion

$$B^{(k+1)}_{\mathrm{BFGS}} = (V^{(k)})^\intercal B^{(k)}_{\mathrm{BFGS}} V^{(k)} + \rho^{(k)} s^{(k)} (s^{(k)})^\intercal,$$

where $\rho^{(k)} := 1/(y^{(k)})^\intercal s^{(k)}$ and $V^{(k)} := \mathrm{Id} - \rho^{(k)} y^{(k)} (s^{(k)})^\intercal$. Working out the first few elements of this sequence, we obtain

$$\begin{aligned}
B^{(1)}_{\mathrm{BFGS}} &= (V^{(0)})^\intercal B^{(0)}_{\mathrm{BFGS}} V^{(0)} + \rho^{(0)} s^{(0)} (s^{(0)})^\intercal \\
B^{(2)}_{\mathrm{BFGS}} &= (V^{(1)})^\intercal B^{(1)}_{\mathrm{BFGS}} V^{(1)} + \rho^{(1)} s^{(1)} (s^{(1)})^\intercal \\
&= (V^{(1)})^\intercal (V^{(0)})^\intercal B^{(0)}_{\mathrm{BFGS}} V^{(0)} V^{(1)} + \rho^{(0)} (V^{(1)})^\intercal s^{(0)} (s^{(0)})^\intercal V^{(1)} + \rho^{(1)} s^{(1)} (s^{(1)})^\intercal
\end{aligned}$$

etc. We only need to evaluate matrix-vector products such as

$$B^{(2)}_{\mathrm{BFGS}} r = (V^{(1)})^\intercal (V^{(0)})^\intercal B^{(0)}_{\mathrm{BFGS}} V^{(0)} V^{(1)} r + \rho^{(0)} (V^{(1)})^\intercal s^{(0)} (s^{(0)})^\intercal V^{(1)} r + \rho^{(1)} s^{(1)} (s^{(1)})^\intercal r,$$

which can be realized efficiently as follows.

Algorithm 5.53 (Recursive evaluation of $B^{(k)}_{\mathrm{BFGS}}\, r$).

[38] A Mebibyte (MiB) are 2^{20} bytes, a Gibibyte (GiB) are 2^{30} bytes. The prefixes "mebi" and gibi replace the former "mega" und "giga", which should however be reserved to mean 10^6 and 10^9, respectively.

```
Input: initial matrix $B^{(0)}_{\text{BFGS}}$ (or matrix-vector products with $B^{(0)}_{\text{BFGS}}$)
Input: pairs of vectors $(y^{(i)}, s^{(i)})$ and scalars $\rho^{(i)} = 1/(y^{(i)})^\intercal s^{(i)}$ for $i = 0, \ldots, k-1$
Input: vector $r \in \mathbb{R}^n$
Output: $B^{(k)}_{\text{BFGS}}\, r$
for $i := k-1, k-2, \ldots, 0$ do
    Set $\alpha^{(i)} := \rho^{(i)} (s^{(i)})^\intercal r$
    Set $r := r - \alpha^{(i)} y^{(i)}$
end for                                   // $r \rightsquigarrow V^{(0)} V^{(1)} \cdots V^{(k-1)} r$
Set $d := B^{(0)}_{\text{BFGS}} r$
for $i := 0, 1, \ldots, k-1$ do
    Set $\beta^{(i)} := \rho^{(i)} (y^{(i)})^\intercal d$
    Set $d := d + (\alpha^{(i)} - \beta^{(i)})\, s^{(i)}$
end for
return $d$                                // $d = B^{(k)}_{\text{BFGS}} r$
```

Remark 5.54 (on Algorithm 5.53).

(*i*) *We do not need $B^{(0)}_{\text{BFGS}}$ as a matrix since only matrix-vector products with $B^{(0)}_{\text{BFGS}}$ are required.*

(*ii*) *Using Algorithm 5.53, it is even possible to change the inverse base metric $B^{(0)}_{\text{BFGS}}$ during the run of the quasi-Newton algorithm. This is impossible when the update formula (5.60) is used to explicitly form the matrices $B^{(k)}_{\text{BFGS}}$.*

We now come back to the second idea of limiting the storage of the pairs of vectors $(y^{(i)}, s^{(i)})$ to the m most recent ones.[39] This idea is easily incorporated into Algorithm 5.53. The resulting update rule is called the **inverse limited-memory BFGS update rule** or briefly, the **(inverse) L-BFGS** or **(inverse) LM-BFGS** rule.

Algorithm 5.55 (Recursive evaluation of $B^{(k)}_{\text{LM-BFGS}}\, r$).

```
Input: initial matrix $B^{(0)}_{\text{BFGS}}$ (or matrix-vector products with $B^{(0)}_{\text{BFGS}}$)
Input: pairs of vectors $(y^{(i)}, s^{(i)})$ and scalars $\rho^{(i)} = 1/(y^{(i)})^\intercal s^{(i)}$ for $i = k-m, \ldots, k-1$
Input: vector $r \in \mathbb{R}^n$
Output: $B^{(k)}_{\text{BFGS}}\, r$
for $i := k-1, k-2, \ldots, k-m$ do
    Set $\alpha^{(i)} := \rho^{(i)} (s^{(i)})^\intercal r$
    Set $r := r - \alpha^{(i)} y^{(i)}$
end for                                   // $r \rightsquigarrow V^{(k-m)} V^{(k-m-1)} \cdots V^{(k-1)} r$
Set $d := B^{(0)}_{\text{BFGS}} r$
for $i := k-m, k-m-1, \ldots, k-1$ do
    Set $\beta^{(i)} := \rho^{(i)} (y^{(i)})^\intercal d$
    Set $d := d + (\alpha^{(i)} - \beta^{(i)})\, s^{(i)}$
end for
return $d$                                // $d = B^{(k)}_{\text{BFGS}} r$
```

[39] We do not relabel the vectors.

We conclude with some remarks on limited-memory quasi-Newton methods.

Remark 5.56 (on limited-memory quasi-Newton methods)**.**

(*i*) *During the first iterations, the number of vectors pairs is gradually increased until the desired size* $m_{\max}$ *of the storage window is reached. That is, we use* $m = \min\{k, m_{\max}\}$. *Typically,* $3 \le m_{\max} \le 10$ *holds.*

(*ii*) *The modifications in Algorithm 5.51 in order to use the inverse limited-memory BFGS update rather than the full (unlimited) inverse BFGS update are minor. In Line 7, we obtain the quasi-Newton direction by evaluating*

$$d^{(k)} = -B^{(k)}_{\text{LM-BFGS}} \nabla f(x^{(k)})$$

using Algorithm 5.55. The evaluation of the next inverse model Hessian $B^{(k+1)}_{\text{LM-BFGS}}$ *in Line 16 is replaced by adding the most recent vector pair* $(y^{(k)}, s^{(k)})$ *to the storage.*

(*iii*) *We cannot expect a limited-memory quasi-Newton method to converge Q-superlinearly in general.*

§ 5.8 Nonlinear Conjugate Gradient Methods

Let us recap the contents of § 5 up to here. After introducing the general framework of line search methods, we discussed a first example, the gradient descent method, in § 5.3. This makes do with first-order derivatives but does not yield Q-superlinear convergence in general. This led us to consider (inexact) Newton methods (§ 5.4, § 5.6), which achieve Q-superlinear or even Q-quadratic convergence but are more expensive due to the use of second-order derivatives and solving (albeit only inexactly) a linear systems with $f''(x^{(k)})$ in each iteration. As a compromise, we then introduced quasi-Newton methods (§ 5.7), which make do with first-order derivatives and are capable of achieving Q-superlinear convergence.

An alternative class of methods which also works with first-order derivatives only is based on the extension of the conjugate gradient (CG) method (§ 4.6) to nonlinear objective functions. These methods are known as **nonlinear conjugate gradient methods**.

The essential characteristics of the CG method for *quadratic* objectives were:

(1) Every new search direction $d^{(k+1)}$ was obtained from the current search direction $d^{(k)}$ and the direction of steepest descent, by forming a linear combination such that $d^{(k)}$ and $d^{(k+1)}$ became A-orthogonal. The A-orthogonality with all previous search directions was automatic.

(2) The Cauchy step size (exact minimizing step size) was taken along every search direction. This was possible due to the objective being a quadratic polynomial.

For nonlinear CG methods, we need to observe the following in comparison.

(1) The Cauchy step size is no longer available. Instead, a line search procedure is used, which is often a strong Wolfe-Powell line search.

(2) We continue to denote the residual as $r^{(k)} := \nabla f(x^{(k)})$.

(3) Since the Hessian of the objective is no longer a constant, s. p. d. matrix A, the requirement of A-conjugate (A-orthogonal) search directions does not make sense anymore. However, one maintains the construction principle that every new search direction $d^{(k+1)}$ is obtained from a linear combination of the current search direction $d^{(k)}$ and the direction of steepest descent:

$$\begin{aligned} d^{(0)} &:= -M^{-1} r^{(0)} && \text{for } k = 0, \\ d^{(k)} &:= -M^{-1} r^{(k)} + \beta^{(k)} \, d^{(k-1)} && \text{for } k \ge 1. \end{aligned} \tag{4.23}$$

The coefficients β are obtained using any of the formulas in (4.24'), which are no longer equivalent for nonlinear CG methods but yield distinct methods; see Table 5.1.

Algorithm 5.57 (Generic nonlinear conjugate gradient method; compare Algorithm 4.17).

```
Input: initial guess $x^{(0)} \in \mathbb{R}^n$
Input: right-hand side $b \in \mathbb{R}^n$
Input: s. p. d. matrix $M$ (or matrix-vector products with $M^{-1}$)
Output: approximately stationary point of (UP)
Set $k := 0$
Set $r^{(0)} := \nabla f(x^{(0)})$                // evaluate the initial residual
Set $d^{(0)} := -M^{-1} r^{(0)}$                 // evaluate the initial negative $M$-gradient
Set $\delta^{(0)} := -(r^{(0)})^\intercal d^{(0)}$                 // $\delta^{(0)} = \|\nabla_M f(x^{(0)})\|_M^2$
while stopping criterion not met do
    Determine a step size $\alpha^{(k)} > 0$ from an appropriate line search procedure
                    // the details depend on the type of method (rule for choosing $\beta$)
    Set $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$
    Set $r^{(k+1)} := \nabla f(x^{(k+1)})$            // updating the residual is not possible
    Set $d^{(k+1)} := -M^{-1} r_{k+1}$               // evaluate the negative $M$-gradient
    Set $\delta_{k+1} := -(r_{k+1})^\intercal d^{(k+1)}$     // $\delta^{(k+1)} = \|\nabla_M f(x^{(k+1)})\|_M^2 = \|r^{(k+1)}\|_{M^{-1}}^2$
    Set $y^{(k)} := r^{(k+1)} - r^{(k)}$              // some nonlinear CG methods use this
    Determine $\beta^{(k+1)}$                      // different nonlinear CG methods differ here
    Set $d^{(k+1)} := d^{(k+1)} + \beta^{(k+1)} d^{(k)}$      // obtain the new search direction
    Set $k := k + 1$
end while
return $x^{(k)}$
```

Different nonlinear CG methods differ with respect to the rule for choosing $\beta^{(k+1)}$. Altogether, we had seen in (4.24') the two expressions

$$(r^{(k+1)} - r^{(k)})^\intercal \, M^{-1} \, r^{(k+1)} \quad \text{and} \quad (r^{(k+1)})^\intercal \, M^{-1} \, r^{(k+1)} \tag{5.62}$$

for the numerator and the three expressions

$$(r^{(k+1)} - r^{(k)})^\intercal \, d^{(k)}, \quad -(r^{(k)})^\intercal \, d^{(k)} \quad \text{and} \quad (r^{(k)})^\intercal \, M^{-1} \, r^{(k)} \tag{5.63}$$

for the denominator.[40] All six combinations (as well as additional variants) appear in the literature and yield meaningful methods. Some prominent choices are summarized in Table 5.1. Some formulas use the abbreviation $y^{(k)} = \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) = r^{(k+1)} - r^{(k)}$ as in quasi-Newton methods.

The line search procedure should yield an approximately stationary point of the line search function $\varphi(\alpha) = f(x^{(k)} + \alpha\, d^{(k)})$. Typically, the strong Wolfe-Powell conditions are required for this purpose with a small curvature parameter such as $\tau = 10^{-1}$ or $\tau = 10^{-2}$. In the (theoretical) case of an exact line search, yielding an exactly stationary step size $\alpha^{(k)}$ of the line search function, we have $(r^{(k+1)})^\intercal d^{(k)} = f'(x^{(k+1)})\, d^{(k)} = \varphi'(\alpha^{(k)}) = 0$. Moreover, we also have $(r^{(k)})^\intercal d^{(k-1)} = 0$ from the previous iteration, and therefore

$$(r^{(k+1)} - r^{(k)})^\intercal\, d^{(k)} = -(r^{(k)})^\intercal\, d^{(k)} = (r^{(k)})^\intercal\, (M^{-1}\, r^{(k)} - \beta^{(k)}\, d^{(k-1)}) = (r^{(k)})^\intercal\, M^{-1}\, r^{(k)}.$$

In this case, we see that the three different expressions in (5.63) for the denominator of $\beta^{(k+1)}$ coincide.

Convergence proofs for nonlinear CG methods are siginificantly more technical than for other methods, and we do not go into the details. It can be generally stated that the methods using $(r^{(k)})^\intercal M^{-1}\, r^{(k)}$ as numerator admit better convergence theories but that the methods using $(r^{(k)} - r^{(k-1)})^\intercal M^{-1}\, r^{(k)}$ are often faster in practice.

As a stopping criterion we can use again a relative and/or absolute criterion involving $r^{(k)} = \nabla f(x^{(k)})$, see (4.14).

End of Week 6

[40] The expression $-(r^{(k)})^\intercal\, d^{(k)}$ was not explicitly given but can be derived immediately from (4.22), which implies $(r^{(k+1)} - r^{(k)})^\intercal\, d^{(k)} = -(r^{(k)})^\intercal\, d^{(k)}$.

Name	Choice of $\beta^{(k+1)}$	Remark
Hestenes–Stiefel (1952)	$\beta_{\text{HS}}^{(k+1)} = \dfrac{(y^{(k)})^\intercal M^{-1} r^{(k+1)}}{(y^{(k)})^\intercal d^{(k)}}$	
Fletcher–Reeves (1964)	$\beta_{\text{FR}}^{(k+1)} = \dfrac{\lVert r^{(k+1)} \rVert_{M^{-1}}^2}{\lVert r^{(k)} \rVert_{M^{-1}}^2}$	strong Wolfe-Powell conditions (5.12), (5.18) with $0 < \sigma < \tau < 1/2$
Polak–Ribière (1969)	$\beta_{\text{PR}}^{(k+1)} = \dfrac{(y^{(k)})^\intercal M^{-1} r^{(k+1)}}{\lVert r^{(k)} \rVert_{M^{-1}}^2}$	no descent guaranteed, therefore often $\beta_{\text{PR+}}^{(k+1)} := \max\{0, \beta_{\text{PR}}^{(k+1)}\}$, where $\beta_{\text{PR+}}^{(k+1)} = 0 \mathrel{\widehat{=}}$ gradient descent step
Powell (1985)	$\beta_{\text{PR+}}^{(k+1)} = \max\{0, \beta_{\text{PR}}^{(k+1)}\}$	refinement of the strong Wolfe-Powell conditions, see Gilbert, Nocedal, 1992, eq.(4.1) and section 6
Fletcher (1987)	$\beta_{\text{F}}^{(k+1)} = \dfrac{\lVert r^{(k+1)} \rVert_{M^{-1}}^2}{-(r^{(k)})^\intercal d^{(k)}}$	
Liu–Storey (1991)	$\beta_{\text{LS}}^{(k+1)} = \dfrac{(y^{(k)})^\intercal M^{-1} r^{(k+1)}}{-(r^{(k)})^\intercal d^{(k)}}$	
Gilbert–Nocedal (1992)	$\beta_{\text{GN}}^{(k+1)} = \begin{cases} -\beta_{\text{FR}}^{(k+1)}, & \text{if } \beta_{\text{PR}}^{(k+1)} < -\beta_{\text{FR}}^{(k+1)} \\ \beta_{\text{PR}}^{(k+1)}, & \text{if } \lvert\beta_{\text{PR}}^{(k+1)}\rvert \le \beta_{\text{FR}}^{(k+1)} \\ \beta_{\text{FR}}^{(k+1)}, & \text{if } \beta_{\text{PR}}^{(k+1)} > \beta_{\text{FR}}^{(k+1)} \end{cases}$	strong Wolfe-Powell conditions (5.12), (5.18) with $0 < \sigma < \tau < 1/2$
Dai–Yuan (1999)	$\beta_{\text{DY}}^{(k+1)} = \dfrac{\lVert r^{(k+1)} \rVert_{M^{-1}}^2}{(y^{(k)})^\intercal d^{(k)}}$	
Hager–Zhang (2005)	$\beta_{\text{HZ}}^{(k+1)} = \left(M^{-1} y^{(k)} - 2\, d^{(k)} \dfrac{\lVert y^{(k)} \rVert_{M^{-1}}^2}{(y^{(k)})^\intercal d^{(k)}} \right)^\intercal \dfrac{r^{(k+1)}}{(y^{(k)})^\intercal d^{(k)}}$	Wolfe-Powell conditions (5.12), (5.17) with $0 < \sigma < \tau < 1$

Table 5.1: Some common nonlinear conjugate gradient methods.

§ 6 Trust-Region Methods for Nonlinear Unconstrained Problems

The line search methods from § 5 proceed by first determining a search direction $d^{(k)}$, by (inexactly) minimizing a quadratic model of the objective

$$q^{(k)}(d) = f(x^{(k)}) + f'(x^{(k)})\, d + \frac{1}{2}\, d^\intercal H^{(k)} d \tag{5.2}$$

or by (inexactly) solving the linear system

$$H^{(k)} d^{(k)} = -\nabla f(x^{(k)}). \tag{5.4}$$

Subsequently, a suitable step size $\alpha^{(k)} > 0$ is determined and the iterate is updated according to

$$x^{(k+1)} \coloneqq x^{(k)} + \underbrace{\alpha^{(k)}\, d^{(k)}}_{s^{(k)}}.$$

Trust-region methods, by contrast, determine the direction and the step size simultaneously. They generate the combined step $s^{(k)}$ as a (usually inexact) solution of the **trust-region subproblem**

$$\begin{aligned} &\text{Minimize} && q^{(k)}(s) = f(x^{(k)}) + f'(x^{(k)})\, s + \frac{1}{2}\, s^\intercal H^{(k)} s, \quad \text{where } s \in \mathbb{R}^n \\ &\text{subject to} && \|s\|_M \le \Delta^{(k)}. \end{aligned} \tag{6.1}$$

As for line search methods, $H^{(k)}$ is the model Hessian, which we require to be symmetric but not necessarily positive definite. The quantity $\Delta^{(k)} > 0$ is the **trust-region radius** governing the **trust region**

$$\left\{ s \in \mathbb{R}^n \,\middle|\, \|s\|_M \le \Delta^{(k)} \right\}$$

attached to the quadratic model.

Note: At $s = 0$, the value of the model $q^{(k)}$ as well as its derivative agree with those of $s \mapsto f(x^{(k)} + s)$. In case of the Newton model ($H^{(k)} = f''(x^{(k)})$), the second derivatives agree as well. Either way, for small values of $\|s\|_M$, the model $q^{(k)}$ will be in good agreement with $s \mapsto f(x^{(k)} + s)$ by Taylor's theorem 2.3.

Since the trust region is a compact set and the objective is continuous (in fact, infinitely smooth), problem (6.1) always has a global minimizer even when $H^{(k)}$ is not positive definite.

Analogously as in line search methods, trust-region algorithms need to monitor the quality of the step $s^{(k)}$, in order to obtain sufficient descent. In fact, we should rather speak of the step proposal $s^{(k)}$ because trust-region methods may reject the proposal. The basis of evaluation of the quality of a tentative step $s^{(k)}$ at the point $x^{(k)}$ is the comparison of the **actual reduction** in objective values

$$\operatorname{ared}(x^{(k)}; s^{(k)}) \coloneqq f(x^{(k)}) - f(x^{(k)} + s^{(k)}) \tag{6.2}$$

to the **predicted reduction** based on the model $q^{(k)}$ associated with the iterate $x^{(k)}$

$$\begin{aligned} \operatorname{pred}(x^{(k)}; s^{(k)}) &\coloneqq q^{(k)}(0) - q^{(k)}(s^{(k)}) \\ &= f(x^{(k)}) - q^{(k)}(s^{(k)}) \\ &= -f'(x^{(k)})\, s^{(k)} - \frac{1}{2}\, (s^{(k)})^\intercal H^{(k)} s^{(k)}. \end{aligned} \tag{6.3}$$

This comparison is achieved in terms of the ratio of these two quantities,

$$\rho(x^{(k)}; s^{(k)}) := \frac{\operatorname{ared}(x^{(k)}; s^{(k)})}{\operatorname{pred}(x^{(k)}; s^{(k)})}. \tag{6.4}$$

In our algorithms, we are going to produce only proposals satisfying

$$\operatorname{pred}(x^{(k)}; s^{(k)}) > 0, \tag{6.5}$$

i. e., for which the model predicts a decrease. The actural decrease (6.2), however, can take either sign.

Suppose now that $s^{(k)}$ is an (inexact) solution of the trust-region subproblem (6.1). Based on the value of $\rho(x^{(k)}; s^{(k)})$, two decisions need to be taken:

- whether to accept or reject the step proposal $s^{(k)}$,
- how to choose the next the next trust-region radius.

These decisions usually depend on two algorithmic parameters $0 < \eta_1 < \eta_2 < 1$:

(1) In case $\rho(x^{(k)}; s^{(k)}) < \eta_1$, the step proposal $s^{(k)}$ is considered unsatisfactory. The reason for this must be that the model $q^{(k)}$ does not coincide well with the true objective function f within the current trust region. In other words, the current trust region is too large.

We therefore reject and discard the step by setting $x^{(k+1)} := x^{(k)}$. We label this iterate unsuccessful.[41] We also choose a new trust-region radius $\Delta^{(k+1)} < \Delta^{(k)}$. In fact, the new trust-region radius should even satisfy $\Delta^{(k+1)} < \|s^{(k)}\|_M$ in order to avoid computing the same unsuccessful step proposal again in the subsequent iteration.

(2) In case $\rho(x^{(k)}; s^{(k)}) \ge \eta_1$, the step proposal $s^{(k)}$ is considered satisfactory and we accept it by setting $x^{(k+1)} := x^{(k)} + s^{(k)}$. The step is labeled successful. The trust region radius $\Delta^{(k+1)}$ for the subsequent step is chosen as follows.

(a) In case $\rho(x^{(k)}; s^{(k)}) \ge \eta_2$, the coincidence between the predicted and actual reductions is considered exceptionally good. We can therefore allow the trust region for the next step to grow. However, this is sensible only if the current step $s^{(k)}$ actually did lie on the boundary of the trust region, i. e., when $\|s^{(k)}\|_M = \Delta^{(k)}$ holds.

(b) Otherwise we keep the trust-region radius: $\Delta^{(k+1)} := \Delta^{(k)}$.

We may call η_1 the **acceptance threshold** and η_2 the **quality threshold**.

The above guidelines lead to the following generic trust-region method (Algorithm 6.1).

Algorithm 6.1 (Generic trust-region method).
Input: *initial guess* $x^{(0)} \in \mathbb{R}^n$

[41] We still count this as an iterate since essentially the same amount of work has been carried out as in a successful iterate.

```
Input: routine to evaluate $f$ and $f'$ (or $\nabla f$)
Input: routine to construct the model Hessians $H^{(k)}$
Input: s. p. d. matrix $M$ (or matrix-vector products with $M^{-1}$)
Input: initial trust-region radius $\Delta^{(0)} > 0$
Input: trust-region step quality parameters $0 < \eta_1 < \eta_2 < 1$
Input: trust-region growth parameters $0 < \gamma_1 < 1 < \gamma_2$
Output: approximately stationary point of (UP)
Set $k := 0$
Set $f^{(0)} := f(x^{(0)})$                                  // evaluate the initial objective value
while stopping criterion not met do
    Determine a step proposal $s^{(k)}$ by an inexact solution of the trust-region subproblem (6.1)
    Evaluate the ratio $\rho(x^{(k)}; s^{(k)})$ according to (6.4)
    if $\rho(x^{(k)}; s^{(k)}) \ge \eta_1$ then                  // satisfactory step proposal
        Set $x^{(k+1)} := x^{(k)} + s^{(k)}$                     // accept the step proposal
        Set $f^{(k+1)} := f(x^{(k+1)})$
        if $\rho(x^{(k)}; s^{(k)}) \ge \eta_2$ and $\|s^{(k)}\|_M = \Delta^{(k)}$ then
                                  // exceptionally good step proposal and trust region too small
            Set $\Delta^{(k+1)} := \gamma_2 \, \Delta^{(k)}$               // grow the trust region
        else      // satisfactory but not exceptionally good step, or trust region sufficiently large
            Set $\Delta^{(k+1)} := \Delta^{(k)}$                     // keep the trust region
        end if
    else                                                     // unsatisfactory step proposal
        Set $x^{(k+1)} := x^{(k)}$                               // reject the step proposal
        Set $f^{(k+1)} := f^{(k)}$
        Set $\Delta^{(k+1)} := \gamma_1 \, \|s^{(k)}\|_M$                 // shrink the trust region
    end if
    Set $k := k + 1$
end while
return $x^{(k)}$
```

Remark 6.2 (on Algorithm 6.1). *The evaluation of the ratio $\rho(x^{(k)}; s^{(k)})$ requires*

- *one function evaluation $f(x^{(k)} + s^{(k)})$ and*
- *one model evaluation $q^{(k)}(s^{(k)}) = f(x^{(k)}) + f'(x^{(k)}) \, s^{(k)} + \frac{1}{2}(s^{(k)})^\intercal H^{(k)} s^{(k)}$*

per iteration. If the step proposal is accepted, then $f(x^{(k)} + s^{(k)})$ becomes $f(x^{(k+1)})$, so that really only one evaluation of the objective f is required per iteration. The evaluation of $q^{(k)}(s^{(k)})$ is usually a by-product of the computation of the (inexact) solution of the trust-region subproblem (6.1).

In the remainder of this section we will consider the following questions.

(1) Which requirements do we have to impose on Algorithm 6.1, in particular concerning the choice of model Hessians $H^{(k)}$ and the inexactness of the trust-region subproblem solves, in order to obtain global convergence? (§ 6.1)

(2) How can we obtain fast local convergence? (§ 6.2)

(3) What is a good algorithmic approach to solving the trust-region subproblems (6.1) with adjustable accuracy? (§ 6.3)

Assumption 6.3. *Throughout § 6 we are assuming that $f \colon \mathbb{R}^n \to \mathbb{R}$ is a C^1 function.*

§ 6.1 Global Convergence

In line search methods, we used the angle condition to compare candidate search directions, such as the Newton direction in Algorithm 5.30 or the inexact Newton direction in Algorithm 5.44, to a reference direction. The steepest descent direction $d_G^{(k)}$ served as that reference and simultaneously as the fallback search direction. This was essential in establishing the global convergence.

A similar idea for trust-region methods uses as reference the **Cauchy point** or **Cauchy step** s_C. The Cauchy point is defined as the unique solution of the trust-region subproblem (6.1), but restricted to the subspace generated by the steepest descent direction. Dropping the iteration index for the time being, the **Cauchy point problem** reads

$$\begin{aligned} \text{Minimize} \quad & q(s) = f(x) + f'(x)\, s + \frac{1}{2}\, s^\intercal H\, s, \quad \text{where } s \in \mathbb{R}^n, \tau \in \mathbb{R} \\ \text{subject to} \quad & \|s\|_M \le \Delta^{(k)} \\ \text{and} \quad & s = -\tau\, \nabla_M f(x). \end{aligned} \tag{6.6}$$

We assume $f'(x) \neq 0$. (**Quiz 6.1:** Why?) Abbreviating

$$g \coloneqq \nabla_M f(x)$$

and reducing the problem to the variable τ by plugging in the constraint $s = -\tau\, \nabla_M f(x)$, we obtain the reduced Cauchy point problem

$$\begin{aligned} \text{Minimize} \quad & q(-\tau\, g) = f(x) - \tau\, \|g\|_M^2 + \frac{\tau^2}{2}\, g^\intercal H\, g, \quad \text{where } \tau \in \mathbb{R} \\ \text{subject to} \quad & |\tau| \le \frac{\Delta}{\|g\|_M}. \end{aligned} \tag{6.7}$$

This is the minimization of a univariate quadratic polynomial over a compact interval that is symmetric about 0. The solution of this problem is given in the following lemma.

Lemma 6.4 (Evaluation of the Cauchy point)**.** *Suppose that $g \neq 0$ and $\Delta \ge 0$ hold. Then the unique solution $s_C = -\tau_C\, g$ of (6.6), respectively the unique solution τ_C of the reduced problem (6.7), is given by*

$$\tau_C = \begin{cases} \min\left\{ \dfrac{\|g\|_M^2}{g^\intercal H\, g}, \dfrac{\Delta}{\|g\|_M} \right\}, & \textit{if } g^\intercal H\, g > 0, \\ \dfrac{\Delta}{\|g\|_M} & \textit{otherwise.} \end{cases} \tag{6.8}$$

Therefore, the decrease predicted by the model at the Cauchy point s_C satisfies

$$\begin{aligned} \operatorname{pred}(x; s_C) &= f(x) - q(s_C) \\ &\ge \frac{1}{2} \, \|g\|_M \, \min\Big\{\Delta, \; \frac{\|g\|_M^3}{\max\{0, g^\mathsf{T} H \, g\}}\Big\} \\ &\ge \frac{1}{2} \, \|g\|_M \, \min\Big\{\Delta, \; \frac{\|g\|_M}{\max\{0, \lambda_{\max}(H; M)\}}\Big\}, \end{aligned} \tag{6.9}$$

where we interpret $\frac{\|g\|_M^3}{0}$ *and* $\frac{\|g\|_M}{0}$ *as* $+\infty$.

Proof. We denote the objective in (6.7), which is a univariate quadratic polynomial, by

$$\varphi(\tau) \coloneqq f(x) - \tau \, \|g\|_M^2 + \frac{\tau^2}{2} \, g^\mathsf{T} H \, g.$$

For any $\tau \in \mathbb{R}$ and $s = -\tau \, g$, we have

$$\operatorname{pred}(x; s) = f(x) - q(-\tau \, g) = f(x) - \varphi(\tau) = \tau \, \|g\|_M^2 - \frac{\tau^2}{2} \, g^\mathsf{T} H \, g. \tag{$*$}$$

We need to distinguish two cases to find the optimal value for τ.

Case 1: $g^\mathsf{T} H \, g > 0$ (φ is strongly convex)
The derivative of φ is equal to zero precisely at

$$\tau^* = \frac{\|g\|_M^2}{g^\mathsf{T} H \, g} > 0.$$

In case this value is feasible, it is the unique solution of (6.7). Otherwise, due to $\varphi'(0) = -\tau \, \|g\|_M^2 < 0$, φ' is negative on the entire interval $[0, \frac{\Delta}{\|g\|_M}]$, and therefore φ is decreasing on this interval. Consequently, the maximal feasible value $\tau = \frac{\Delta}{\|g\|_M}$ is the unique solution of (6.7). To summarize this case:

$$\tau_C = \min\Big\{\frac{\|g\|_M^2}{g^\mathsf{T} H \, g}, \; \frac{\Delta}{\|g\|_M}\Big\}.$$

In order to evaluate $\operatorname{pred}(x; s_C)$, we obtain from ($*$):

$$\begin{aligned} \operatorname{pred}(x; s_C) &= \tau_C \, \|g\|_M^2 - \frac{\tau_C^2}{2} \, g^\mathsf{T} H \, g \\ &= \begin{cases} \frac{1}{2} \frac{\|g\|_M^4}{g^\mathsf{T} H \, g} = \frac{1}{2} \|g\|_M \frac{\|g\|_M^3}{\max\{0, g^\mathsf{T} H \, g\}} & \text{if } \tau_C = \dfrac{\|g\|_M^2}{g^\mathsf{T} H \, g} \le \dfrac{\Delta}{\|g\|_M}, \\ \tau_C \left[\|g\|_M^2 - \frac{1}{2} \left[\frac{\Delta}{\|g\|_M} \right] g^\mathsf{T} H \, g \right] & \text{if } \tau_C = \dfrac{\Delta}{\|g\|_M} \le \dfrac{\|g\|_M^2}{g^\mathsf{T} H \, g}. \end{cases} \end{aligned}$$

In the first case, (6.9) is satisfied since clearly

$$\frac{1}{2} \|g\|_M \frac{\|g\|_M^3}{\max\{0, g^\mathsf{T} H \, g\}} \ge \frac{1}{2} \|g\|_M \, \min\Big\{\Delta, \; \frac{\|g\|_M^3}{\max\{0, g^\mathsf{T} H \, g\}}\Big\}$$

holds. Moreover, due to (2.12), we have $\lambda_{\max}(H;M) \ge \frac{g^\intercal H g}{\|g\|_M^2} > 0$ and therefore

$$\cdots \ge \frac{1}{2}\, \|g\|_M \, \min\Big\{\Delta, \; \frac{\|g\|_M}{\max\{0, \lambda_{\max}(H;M)\}}\Big\}.$$

In the second case, we have

$$\begin{aligned}
\tau_C \Big[\|g\|_M^2 - \frac{1}{2}\Big[\frac{\Delta}{\|g\|_M}\Big]\, g^\intercal H g\Big] &\ge \tau_C \Big[\|g\|_M^2 - \frac{1}{2}\Big[\frac{\|g\|_M^2}{g^\intercal H g}\Big]\, g^\intercal H g\Big] \\
&= \frac{1}{2}\, \tau_C \, \|g\|_M^2 \\
&= \frac{1}{2}\, \Delta \, \|g\|_M \\
&\ge \frac{1}{2}\|g\|_M \, \min\Big\{\Delta, \; \frac{\|g\|_M^3}{\max\{0, g^\intercal H g\}}\Big\} \\
&= \frac{1}{2}\, \|g\|_M \, \min\Big\{\Delta, \; \frac{\|g\|_M}{\max\{0, \lambda_{\max}(H;M)\}}\Big\},
\end{aligned}$$

so (6.9) also holds here.

Case 2: $g^\intercal H g \le 0$ (φ is concave)
Since φ is now concave, the solution τ_C of (6.7) must lie on the boundary of the feasible interval. In view of $\varphi'(0) = -\tau \, \|g\|_M^2 < 0$, we have

$$\tau_C = + \frac{\Delta}{\|g\|_M}.$$

Therefore, we obtain

$$\begin{aligned}
\operatorname{pred}(x; s_C) &= \Delta \, \|g\|_M + \frac{1}{2}\, \Big[\frac{\Delta}{\|g\|_M}\Big]^2 \, |g^\intercal H g| \\
&\ge \Delta \, \|g\|_M \\
&\ge \frac{1}{2}\|g\|_M \, \min\Big\{\Delta, \; \frac{\|g\|_M^3}{\max\{0, g^\intercal H g\}}\Big\} \\
&\ge \frac{1}{2}\, \|g\|_M \, \min\big\{\Delta, \; \infty\big\} \\
&= \frac{1}{2}\, \|g\|_M \, \Delta \\
&\ge \frac{1}{2}\, \|g\|_M \, \min\Big\{\Delta, \; \frac{\|g\|_M}{\max\{0, \lambda_{\max}(H;M)\}}\Big\},
\end{aligned}$$

so (6.9) holds again. □

In order to formulate a condition on the step proposals $s^{(k)}$ that ensures the global convergence of the general trust-region method (Algorithm 6.1), we compare

- the decrease in the current model $\operatorname{pred}(x^{(k)}; s^{(k)})$ obtained by the proposed step $s^{(k)}$
- with the decrease in the model $\operatorname{pred}(x^{(k)}; s_C^{(k)})$ obtained by the Cauchy step $s_C^{(k)}$.

We require that $s^{(k)}$ realizes at least a fixed fraction of the decrease obtained by $s_C^{(k)}$. In fact, it will be sufficient that $s^{(k)}$ realizes a fixed fraction of the lower bound (6.9).

Definition 6.5 (Fraction of Cauchy decrease condition)**.**
Consider the trust-region subproblem

$$\begin{aligned} &\textit{Minimize} && q(s) = f(x) + f'(x)\, s + \frac{1}{2}\, s^\top H\, s, \quad \textit{where } s \in \mathbb{R}^n \\ &\textit{subject to} && \|s\|_M \le \Delta \end{aligned} \tag{6.10}$$

with $f'(x) \neq 0$ and thus also $g \coloneqq \nabla_M f(x) \neq 0$. The Cauchy step for (6.10) *is denoted by s_C.*

(*i*) *A vector $s \in \mathbb{R}^n$ satisfies the* ***fraction of Cauchy decrease condition*** *for* (6.10) *if there exists a constant $\underline{C} \in (0,1]$ such that*

$$\operatorname{pred}(x; s) \ge \underline{C} \operatorname{pred}(x; s_C) \tag{6.11}$$

holds.

(*ii*) *A vector $s \in \mathbb{R}^n$ satisfies the* ***weak fraction of Cauchy decrease condition*** *for* (6.10) *if there exists a constant $\underline{C} \in (0,1]$ such that*

$$\operatorname{pred}(x; s) \ge \underline{C}\, \frac{1}{2} \|g\|_M \min\Big\{\Delta, \frac{\|g\|_M}{\max\{0, \lambda_{\max}(H; M)\}}\Big\} \tag{6.12}$$

holds.

Note: By (6.9), the fraction of Cauchy decrease condition (6.11) implies the weak fraction of Cauchy decrease condition (6.12). Moreover, the weak fraction of Cauchy decrease condition (6.12) implies that an accepted step size proposal satisfies

$$\begin{aligned} f(x) - f(x+s) &= \operatorname{ared}(x; s) \\ &= \rho(x; s) \operatorname{pred}(x; s) && \text{by definition of the ratio } \rho(\cdot\,;\cdot) \\ &\ge \eta_1 \operatorname{pred}(x; s) && \text{since the step proposal was accepted} \\ &> 0 && \text{by (6.12).} \end{aligned}$$

In particular, any trust-region method that falls into the framework of Algorithm 6.1 is a **descent method**.

Our proof of global convergence requires (roughly) the following auxiliary results to be shown:

(1) The weak fraction of Cauchy decrease condition implies that Algorithm 6.1 cannot get stuck in an infinite sequence of consecutively rejected steps (Lemma 6.6 and Corollary 6.7).

(2) If the sequence of function values $f(x^{(k)})$ is bounded below, then any subsequence of successful steps, whose associated trust-region radii $\Delta^{(k)}$ sum up to ∞, is a sequence of vanishing gradients (Lemma 6.8).

We now show that for sufficiently small trust-region radius, step proposals satisfying the (weak) fraction of Cauchy decrease condition will always be successful. This result even holds uniformly for an entire class of trust-region subproblems.

Lemma 6.6 (Any acceptance threshold is achievable for small trust-region radius)**.** *Suppose that $\overline{x} \in \mathbb{R}^n$ is a point satisfying $f'(\overline{x}) \neq 0$. Suppose, moreover, that $\eta_1 \in (0,1)$ and $\underline{C} > 0$ and $\overline{H} > 0$ are given. Then there exist $\delta > 0$ and $\overline{\Delta} > 0$ such that the following holds: for any trust-region subproblem*

$$\begin{aligned} &\textit{Minimize} \quad f(x) + f'(x)\, s + \frac{1}{2} s^\intercal H\, s, \quad \textit{where } s \in \mathbb{R}^n \\ &\textit{subject to} \quad \|s\|_M \le \Delta \end{aligned}$$

with data $x \in B_\delta^M(\overline{x})$ and $\Delta \in (0, \overline{\Delta}]$ and H symmetric with $\|H\|_{M^{-1} \leftarrow M} \le \overline{H}$, any step proposal s that is feasible and satisfies the weak fraction of Cauchy decrease condition

$$\operatorname{pred}(x;s) \ge \underline{C} \frac{1}{2} \|g\|_M \min\Big\{\Delta, \frac{\|g\|_M}{\max\{0, \lambda_{\max}(H;M)\}}\Big\} \tag{6.12}$$

achieves a ratio $\rho(x;s) \ge \eta_1$.

Proof. The ratio under consideration satisfies

$$\rho(x;s) = \frac{\operatorname{ared}(x;s)}{\operatorname{pred}(x;s)} = 1 - \frac{\operatorname{pred}(x;s) - \operatorname{ared}(x;s)}{\operatorname{pred}(x;s)}.$$

In order to show that it is $> \eta_1$ we need to estimate the numerator of the fraction from above and the denominator from below.

We begin with the denominator. Since f' is continuous, we can find $\delta > 0$ such that

$$\|g\|_M = \|\nabla_M f(x)\|_M \ge \frac{\|\nabla_M f(\overline{x})\|_M}{2} =: \varepsilon$$

holds for all $x \in B_\delta^M(\overline{x})$. We now set $\overline{\Delta} \coloneqq \varepsilon / \overline{H}$ and consider a trust-region problem with data as specified in the statement of the lemma. Then we have

$$\Delta \le \overline{\Delta} = \frac{\varepsilon}{\overline{H}} = \frac{\|\nabla_M f(\overline{x})\|_M}{2\,\overline{H}} \le \frac{\|\nabla_M f(x)\|_M}{\overline{H}} = \frac{\|g\|_M}{\overline{H}} \le \begin{cases} \frac{\|g\|_M}{\lambda_{\max}(H;M)}, & \text{in case } \lambda_{\max}(H;M) > 0 \\ \infty & \text{in case } \lambda_{\max}(H;M) \le 0. \end{cases}$$

Notice that we have used (2.13) to infer $\lambda_{\max}(H;M) \le \|H\|_{M^{-1} \leftarrow M} \le \overline{H}$ in the last inequality. By the weak fraction of Cauchy decrease condition, we therefore conclude

$$\begin{aligned} \operatorname{pred}(x;s) &\ge \underline{C} \frac{1}{2} \|g\|_M \min\Big\{\Delta, \frac{\|g\|_M}{\max\{0, \lambda_{\max}(H;M)\}}\Big\} \\ &\ge \underline{C} \frac{1}{2} \|g\|_M \min\Big\{\Delta, \frac{\|g\|_M}{\overline{H}}\Big\} && (6.13a) \\ &= \underline{C} \frac{1}{2} \|g\|_M \,\Delta. && (6.13b) \end{aligned}$$

For the numerator, we can estimate

$$
\begin{aligned}
&\operatorname{pred}(x;s) - \operatorname{ared}(x;s) \\
&\quad = -f'(x)\,s - \frac{1}{2}\,s^{\mathsf{T}} H\,s - \bigl[f(x) - f(x+s)\bigr] && \text{see (6.3)} \\
&\quad = -f'(x)\,s - \frac{1}{2}\,s^{\mathsf{T}} H\,s + f'(x+\xi\,s)\,s && \text{by Taylor's theorem 2.4} \\
&\quad \le \|f'(x+\xi\,s) - f'(x)\|_{M^{-1}}\,\|s\|_M - \frac{1}{2}\,\lambda_{\min}(H;M)\,\|s\|_M^2 && \text{by (2.3) and (2.12)} \\
&\quad \le \|f'(x+\xi\,s) - f'(x)\|_{M^{-1}}\,\Delta + \frac{1}{2}\,\overline{H}\,\Delta^2 && \text{by (2.13) and } \|s\|_M \le \Delta.
\end{aligned}
$$

Notice that $x + \xi\,s$ and x are both close to $\overline{x}$, namely

$$
\|x + \xi\,s - \overline{x}\|_M \le \delta + \Delta \le \delta + \overline{\Delta} \quad \text{and} \quad \|x - \overline{x}\|_M \le \delta.
$$

Using the continuity of f' and reducing δ and $\overline{\Delta}$ if necessary, we can thus achieve

$$
\|f'(x + \xi_k\,s) - f'(x)\|_{M^{-1}} + \frac{1}{2}\,\overline{H}\,\Delta \le \frac{1}{2}(1-\eta_1)\,\underline{C}\,\varepsilon.
$$

This allows us to continue the estimate above as follows,

$$
\begin{aligned}
\operatorname{pred}(x;s) - \operatorname{ared}(x;s) &\le \|f'(x + \xi_k\,s) - f'(x)\|_{M^{-1}}\,\Delta + \frac{1}{2}\,\overline{H}\,\Delta^2 \\
&\le \frac{1}{2}(1-\eta_1)\,\underline{C}\,\varepsilon\,\Delta \\
&\le \frac{1}{2}(1-\eta_1)\,\underline{C}\,\|g\|_M\,\Delta.
\end{aligned}
$$

Combining this with the estimate (6.13) of the denominator, we conclude

$$
\rho(x;s) = 1 - \frac{\operatorname{pred}(x;s) - \operatorname{ared}(x;s)}{\operatorname{pred}(x;s)} \ge 1 - \frac{\frac{1}{2}(1-\eta_1)\,\underline{C}\,\|g\|_M\,\Delta}{\underline{C}\,\frac{1}{2}\|g\|_M\,\Delta} = \eta_1. \qquad \square
$$

The result of Lemma 6.6 immediately implies that Algorithm 6.1 produces infinitely many successful step proposals, provided that it does not stop with $f'(x^{(k)}) = 0$ and that the model Hessians remain bounded.

Corollary 6.7 (Infinitely many successful step proposals). *Suppose that the iterates $x^{(k)}$ of Algorithm 6.1 satisfy $f'(x^{(k)}) \neq 0$ and that the model Hessians $H^{(k)}$ are symmetric with $\|H^{(k)}\|_{M^{-1} \leftarrow M} \le \overline{H}$. Moverover, suppose that the step proposals $s^{(k)}$ are feasible and satisfy the weak fraction of Cauchy decrease condition* (6.12)*. Then there exist infinitely many indices k satisfying $\rho(x^{(k)}; s^{(k)}) \ge \eta_1$, i. e., the step proposal will be accepted.*

Note: In particular, a successful step can only be followed by finitely many unsuccessful steps.

Proof. We proceed by way of contradiction. Suppose the opposite, i. e., that $\rho(x^{(k)}; s^{(k)}) < \eta_1$ holds for all $k \ge k_0$. This means that from a certain iterate onwards, all step proposals will be rejected. By the logic of Algorithm 6.1, see Line 18, this implies

$$\Delta^{(k+1)} := \gamma_1 \, \|s^{(k)}\|_M \le \gamma_1 \, \Delta^{(k)}$$

for all $k \ge k_0$ and thus $\Delta^{(k)} \longrightarrow 0$. Moreover, the iterates $x^{(k)}$ and model Hessians $H^{(k)}$ remain constant for $k \ge k_0$. In particular, the entire sequence $\big(H^{(k)}\big)$ is bounded. The subproblems (6.1) thus satisfy the prerequisites of Lemma 6.6.[42] This entails that, no matter how close the acceptance threshold η_1 is to 1, as soon as the trust-region radius $\Delta^{(k)}$ has become sufficiently small, we will have $\rho(x^{(k)}; s^{(k)}) \ge \eta_1$ and the step proposal will be accepted. This is in contradiction with our assumption. □

We proceed to show Statement (2).

Lemma 6.8. *Suppose that the iterates $x^{(k)}$ of Algorithm 6.1 satisfy $f'(x^{(k)}) \neq 0$ and that the model Hessians $H^{(k)}$ are symmetric with $\|H^{(k)}\|_{M^{-1} \leftarrow M} \le \overline{H}$. Assume that the sequence of objective values $f(x^{(k)})$ is bounded below by $\underline{f}$. Moverover, suppose that the step proposals $s^{(k)}$ satisfy the weak fraction of Cauchy decrease condition (6.12). If $K \subseteq \mathbb{N}_0$ is any (finite or infinite) index set of successful steps such that $\|g^{(k)}\|_M \ge \varepsilon > 0$ holds, then the associated trust-region radii are summable:*

$$\sum_{k \in K} \Delta^{(k)} < \infty.$$

Proof. For any index $k \in K$ we have by assumption $\rho(x^{(k)}; s^{(k)}) \ge \eta_1$ and thus

$$\begin{aligned} f(x^{(k)}) - f(x^{(k+1)}) &= \operatorname{ared}(x^{(k)}; s^{(k)}) \\ &\ge \eta_1 \operatorname{pred}(x^{(k)}; s^{(k)}) \\ &\ge \eta_1 \underline{C} \, \frac{1}{2} \|g^{(k)}\|_M \min\Big\{\Delta^{(k)}, \ \frac{\|g^{(k)}\|_M}{\overline{H}}\Big\} && \text{as in (6.13a)} \\ &\ge \eta_1 \underline{C} \, \frac{1}{2} \varepsilon \min\Big\{\Delta^{(k)}, \ \frac{\varepsilon}{\overline{H}}\Big\} && \text{by assumption.} \end{aligned}$$

Since the function values $f(x^{(k)})$ are monotone decreasing and, by assumption, bounded below, we must have

$$\sum_{k \in K} \eta_1 \underline{C} \, \frac{1}{2} \varepsilon \min\Big\{\Delta^{(k)}, \ \frac{\varepsilon}{\overline{H}}\Big\} < \infty.$$

[42] We can use $\overline{x} = x^{(k_0)}$ and we do not need the variation in x.

More precisely, let $k_{\min}$ be the smallest index in K, then we have

$$\begin{aligned}
&\sum_{k \in K} \eta_1 \underline{C} \frac{1}{2} \varepsilon \min\Big\{\Delta^{(k)}, \frac{\varepsilon}{\overline{H}}\Big\} \\
&\qquad \le \sum_{k \in K} f(x^{(k)}) - f(x^{(k+1)}) \\
&\qquad \le \sum_{k=k_{\min}}^{\infty} f(x^{(k)}) - f(x^{(k+1)}) && \text{all summands are } \ge 0 \text{ and a superset of the above} \\
&\qquad \le f(x^{(k_{\min})}) - \underline{f} && \text{since } f(x^{(k)}) \ge \underline{f} \text{ for all } k \in \mathbb{N}_0 \\
&\qquad < \infty.
\end{aligned}$$

This implies the claim. □

With Statements (1) and (2) in place, we can now prove a global convergence theorem for Algorithm 6.1.

Theorem 6.9 (Global convergence of model Algorithm 6.1). *Suppose that the model Hessians $H^{(k)}$ in Algorithm 6.1 are symmetric with $\|H^{(k)}\|_{M^{-1} \leftarrow M} \le \overline{H}$. Assume that the sequence of objective values $f(x^{(k)})$ is bounded below by $\underline{f}$. Moverover, suppose that the step proposals $s^{(k)}$ are feasible and satisfy the weak fraction of Cauchy decrease condition (6.12). Then the following holds.*

(1) Algorithm 6.1 either terminates after finitely many iterations due to $f'(x^{(k)}) = 0$, or else we have[43]

$$\liminf_{k \to \infty} \|f'(x^{(k)})\|_{M^{-1}} = 0. \tag{6.14}$$

(2) Suppose in addition that f' is uniformly continuous on the set of iterates $\{x^{(k)} \mid k \in \mathbb{N}_0\}$.[44] *Then Algorithm 6.1 either terminates after finitely many iterations due to $f'(x^{(k)}) = 0$, or else we have*

$$\lim_{k \to \infty} f'(x^{(k)}) = 0. \tag{6.15}$$

Note: Statement (6.15) implies that all accumulation points of $\big(x^{(k)}\big)$ are stationary points. **Quiz 6.2:** Details?

Proof. Statement (1): Assume that Algorithm 6.1 does not terminate, i. e., we have $f'(x^{(k)}) \neq 0$ for all iterates. Suppose that (6.14) does not hold. This means that there exists $\varepsilon_1 > 0$ such that $g^{(k)} := \nabla_M f(x^{(k)})$ satisfies $\|g^{(k)}\|_M \ge \varepsilon_1$ for all $k \ge k_0$. Let $S \subseteq \mathbb{N}_0$ denote the index set of successful step proposals. This index set is infinite by Corollary 6.7. From Lemma 6.8, we further obtain

$$\sum_{k \in S} \Delta^{(k)} < \infty. \tag{$*$}$$

[43] Statement (6.14) means that there exists a subsequence $\big(x^{(k)}\big)_{k \in K}$ such that $f'(x^{(k)}) \xrightarrow{k \in K} 0$.

[44] Unlike in the proof of, e. g., Lemma 5.13, we cannot *deduce* the uniform continuity since the sequence $x^{(k)}$ is not necessarily bounded.

This implies in particular $\Delta^{(k)} \xrightarrow{k\in S} 0$.

Step (1) We show that $x^{(k)}$ is a Cauchy sequence.
Let $k > \ell \ge 0$ be any two indices. Then

$$\begin{aligned} \|x^{(k)} - x^{(\ell)}\|_M &\le \sum_{\substack{j=\ell\\ j\in S}}^{k-1} \|s^{(j)}\|_M \quad \text{unsuccessful steps do not move } x \\ &\le \sum_{\substack{j=\ell\\ j\in S}}^{k-1} \Delta^{(j)} \le \sum_{\substack{j=\ell\\ j\in S}}^{\infty} \Delta^{(j)} \le \sum_{j\in S} \Delta^{(j)}. \end{aligned}$$

The last sum is finite by (∗). Therefore, the tails of this series ("Reihenreste") $\sum_{\substack{j=\ell\\ j\in S}}^{\infty} \Delta^{(j)}$ converge to 0 as $\ell \longrightarrow \infty$. This shows that $x^{(k)}$ is a Cauchy sequence.

Step (2) We show that $\big(\Delta^{(k)}\big)_{k\ge L}$ is bounded away from 0.
$x^{(k)}$ has been shown to be a Cauchy sequence, hence it converges to some $\overline{x}$. The continuity of f' and $\|g^{(k)}\|_M \ge \varepsilon_1$ for all $k \ge k_0$ imply $\|\nabla_M f'(\overline{x})\|_M \ge \varepsilon_1$. We now apply Lemma 6.6 with η_2 (the quality threshold) in place of η_1 (the acceptance threshold). It shows that there exists $\overline{\Delta}$ and an index $L \in \mathbb{N}_0$ such that $\rho(x^{(k)}; s^{(k)}) \ge \eta_2$ holds for all indices $k \ge L$ (such that $x^{(k)}$ is close enough to $\overline{x}$) satisfying $\Delta^{(k)} \le \overline{\Delta}$.

We now show, by way of induction, that

$$\Delta^{(k)} \ge \min\big\{\Delta^{(L)},\ \gamma_1 \overline{\Delta}\big\} \quad \text{for all } k \ge L \tag{$**$}$$

holds, so that indeed the trust-region radii are bounded away from 0. Claim (∗∗) holds trivially for $k = L$. As induction hypothesis, suppose that (∗∗) is true for some $k \ge L$. In the induction step, we distinguish two cases. In case $\Delta^{(k)} \le \overline{\Delta}$, we have $\rho(x^{(k)}; s^{(k)}) \ge \eta_2$ as noted above. Therefore,

$$\Delta^{(k+1)} = \gamma_2\, \Delta^{(k)} > \Delta^{(k)} \ge \min\big\{\Delta^{(L)},\ \gamma_1 \overline{\Delta}\big\}.$$

In the opposite case $\Delta^{(k)} > \overline{\Delta}$, the step might not have been successful and the trust-region radius might have been reduced, but we can estimate it as

$$\Delta^{(k+1)} \ge \gamma_1\, \Delta^{(k)} > \gamma_1 \overline{\Delta} \ge \min\big\{\Delta^{(L)},\ \gamma_1 \overline{\Delta}\big\}.$$

To summarize, we have shown (∗∗) for $k+1$ and the induction is complete.

We now have reached the contradiction that $\Delta^{(k)}$ is bounded away from zero for $k \ge L$ and simultaneously, the subsequence $\Delta^{(k)} \xrightarrow{k\in S} 0$; see above **Step** (1).

Statement (2): Suppose now that f' is uniformly continuous on the set of iterates $\{x^{(k)} \mid k \in \mathbb{N}_0\}$, i. e., for every $\varepsilon > 0$, there exists $\delta > 0$ such that $\|x^{(k)} - x^{(\ell)}\|_M \le \delta$ implies $\|f'(x^{(k)}) - f'(x^{(\ell)})\|_{M^{-1}} \le \varepsilon$. We proceed again by contradiction and assume that (6.15) does *not* hold. That is, there exists $\varepsilon_2 > 0$ such that the index set

$$K_{2\,\varepsilon_2} \coloneqq \big\{k \in \mathbb{N}_0 \,\big|\, \|f'(x^{(k)})\|_{M^{-1}} \ge 2\,\varepsilon_2\big\}$$

is infinite. The uniform continuity of f' implies the existence of $\delta > 0$ such that

$$\|f'(x^{(\ell)})\|_{M^{-1}} \ge \|f'(x^{(k)})\|_{M^{-1}} - \|f'(x^{(k)}) - f'(x^{(\ell)})\|_{M^{-1}} \ge \varepsilon_2 \tag{***}$$

holds for all $x^{(\ell)}$ with the property $\|x^{(\ell)} - x^{(k)}\|_M \le \delta$ for some $k \in K_{2\,\varepsilon_2}$.

Consider now also the index set

$$K_{\varepsilon_2} \coloneqq \big\{k \in \mathbb{N}_0 \,\big|\, \|f'(x^{(k)})\|_{M^{-1}} \ge \varepsilon_2\big\},$$

which is infinite as well due to $K_{2\,\varepsilon_2} \subseteq K_{\varepsilon_2}$. Recall that $S \subseteq \mathbb{N}_0$ denotes the index set of successful step proposals. The set $S \cap K_{\varepsilon_2}$ may be finite or infinite. In any case, Lemma 6.8 implies

$$\sum_{k \in S \cap K_{\varepsilon_2}} \Delta^{(k)} < \infty.$$

Therefore, there exists $k_0 \in S \cap K_{\varepsilon_2}$ such that the remainder of this series ("Reihenrest") that comes after k_0 is small. More precisely,

$$\sum_{\substack{j=k_0 \\ j \in S \cap K_{\varepsilon_2}}}^{\infty} \Delta^{(j)} < \delta. \tag{****}$$

In fact, by making k_0 larger, if necessary, k_0 can be chosen to lie in $k_0 \in K_{2\,\varepsilon_2}$. (**Quiz 6.3:** Why?)

We now show by induction that $\|x^{(\ell)} - x^{(k_0)}\|_M < \delta$ holds for all $\ell \ge k_0$. The claim holds trivially for $\ell = k_0$. As induction hypothesis, suppose that there exists $\ell \ge k_0$ such that $\|x^{(j)} - x^{(k_0)}\|_M < \delta$ is true for $j = k_0, \ldots, \ell$. We can invoke (***) to see that $\|f'(x^{(j)})\|_{M^{-1}} \ge \varepsilon_2$ holds for all $j = k_0, \ldots, \ell$. In other words, all indices $j = k_0, \ldots, \ell$ belong to K_{ε_2}. We can therefore estimate

$$\begin{aligned}
\|x^{(\ell+1)} - x^{(k_0)}\|_M &\le \sum_{\substack{j=k_0 \\ j \in S}}^{\ell} \Delta^{(j)} && \text{unsuccessful steps do not move } x \\
&= \sum_{\substack{j=k_0 \\ j \in S \cap K_{\varepsilon_2}}}^{\ell} \Delta^{(j)} && \text{all indices } j = k_0, \ldots, \ell \text{ belong to } K_{\varepsilon_2} \\
&\le \sum_{\substack{j=k_0 \\ j \in S \cap K_{\varepsilon_2}}}^{\infty} \Delta^{(j)} && \text{summands are positive} \\
&< \delta && \text{by (****).}
\end{aligned}$$

This concludes the induction step and we have shown that indeed $\|x^{(\ell)} - x^{(k_0)}\|_M < \delta$ holds for all $\ell \ge k_0$. Invoking again (***), we infer that $\|f'(x^{(\ell)})\|_{M^{-1}} \ge \varepsilon_2$ holds for all $\ell \ge k_0$, which is in contradiction to (6.14). □

Remark 6.10 (on the global convergence theorem 6.9). *For the global convergence theorem 6.9 to hold, it is not strictly necessary that the step proposal satisfy $\|s^{(k)}\|_M \le \Delta^{(k)}$. A relaxed constraint of the form $\|s^{(k)}\|_M \le \beta\,\Delta^{(k)}$ is enough and allows us to gain some more flexibility in a practical algorithm; see for instance Ulbrich, Ulbrich, 2012, Satz 14.10 and eq.(14.50). However, in this class we are not going to exploit this algorithmically.*

§ 6.2 Fast Local Convergence

In this section we show that the generic trust-region method Algorithm 6.1 will transition to a local, inexact Newton method. This means that the trust-region constraint $\|s^{(k)}\|_M \le \Delta^{(k)}$ will be inactive from a certain interation index onwards. If one then solves the trust-region subproblems to sufficient accuracy (compare the forcing sequence in inexact Newton methods in § 5.6), one can obtain Q-superlinear or even Q-quadratic convergence of the iterates.

We restrict the discussion here to **trust region Newton methods**, which are defined by the choice $H^{(k)} = f''(x^{(k)})$. However, model Hessians based on quasi-Newton updates are very important in practice as well.[45].

Theorem 6.11 (Transition to fast local convergence in Algorithm 6.1)**.** *Suppose that f is of class C^2 and that the model Hessians $H^{(k)}$ in Algorithm 6.1 are the exact Hessians $f''(x^{(k)})$. Assume that the sequence of objective values $f(x^{(k)})$ is bounded below by $\underline{f}$. Moverover, suppose that the step proposals $s^{(k)}$ are feasible and satisfy the weak fraction of Cauchy decrease condition (6.12). Suppose further that the sublevel set $\mathcal{M}_f(x^{(0)}) \coloneqq \{x \in \mathbb{R}^n \mid f(x) \le f(x^{(0)})\}$ is compact.*

(*i*) *Suppose that x^* is an accumulation point of $x^{(k)}$ and that $f''(x^*)$ is positive semidefinite. Then the entire sequence convergences to x^*, i. e., x^* is indeed the unique limit point of $x^{(k)}$.*

(*ii*) *There exists an index $k_0 \in \mathbb{N}$, such that*

(*a*) *$\rho(x^{(k)}; s^{(k)}) > \eta_1$ holds for all $k \ge k_0$, i. e., the step proposal will be accepted.*

(*b*) *$f''(x^{(k)})$ is positive semidefinite and $\|f''(x^{(k)})^{-1}\,\nabla f(x^{(k)})\|_M \le \dfrac{\Delta^{(k)}}{2}$ holds for all $k \ge k_0$.*

(*iii*) *If, in addition, the step proposals $s^{(k)}$ satisfy*

$$\|f''(x^{(k)})\,s^{(k)} + \nabla f(x^{(k)})\|_{M^{-1}} \le \eta^{(k)}\,\|\nabla f(x^{(k)})\|_{M^{-1}} \tag{6.16}$$

with a forcing sequence $\eta^{(k)} \searrow 0$, then $x^{(k)}$ converges to x^* Q-superlinearly w.r.t. the M-norm.

Statement (b) means that from a certain index k_0 onwards, full exact Newton steps would be useful step proposals that are not only feasible w.r.t. the constraint $\|s^{(k)}\|_M \le \Delta^{(k)}$ but lie well inside the trust region.

Condition (6.16) requires that the step proposal $s^{(k)}$ will eventually be close to the full exact Newton step in the sense that the relative residual norm is bounded by the forcing sequence $\eta^{(k)}$; compare the condition (5.36) for inexact Newton methods for root finding.

[45] In contrast to the quasi-Newton line search methods from § 5.7, the positive definiteness of $H^{(k)}$ is no longer required. A further distinction is that one works with quasi-Newton update formulas for the model Hessian $H^{(k)}$ and not for the inverse $B^{(k)}$, since the inexact iterative solution of the trust-region subproblems (6.1) is based on matrix-vector products with $H^{(k)}$.

§ 6.3 Solution of the Trust-Region Subproblem

We now address options for the inexact numerical solution of the trust-region subproblem (6.1), i. e., with a problem of the form

$$\begin{aligned} &\text{Minimize} \quad q(s) = f - b^\intercal s + \frac{1}{2}\, s^\intercal H\, s, \quad \text{where } s \in \mathbb{R}^n \\ &\text{subject to} \quad \|s\|_M \le \Delta. \end{aligned} \tag{6.17}$$

The data of the problem are the model offset $q(0) = f \in \mathbb{R}$, (negative) model derivative $-q'(0) = b \in \mathbb{R}^n$, symmetric model Hessian $H \in \mathbb{R}^{n\times n}$ and trust-region radius $\Delta > 0$.

Although the Cauchy point is a sufficiently accurate solution of the trust-region subproblem in order to achieve global convergence (Theorem 6.9), and fast local convergence can only take effect as soon as the trust-region constraint $\|s\|_M \le \Delta$ is inactive and no longer plays a role (Theorem 6.11), we will characterize the exact global solution(s) of (6.17) first.

Note: Due to the possibility of negative eigenvalues in H, problem (6.17) is not convex in general.

Theorem 6.12 (Characterization of global solutions of the trust-region subproblem (6.17))**.**

(*i*) *Suppose that $s \in \mathbb{R}^n$ is a global minimizer of* (6.17)*. Then there exists $\mu \in \mathbb{R}$ such that the following holds:*

$$\mu \ge 0, \quad \|s\|_M - \Delta \le 0, \quad \mu\,(\|s\|_M - \Delta) = 0 \tag{6.18a}$$

$$(H + \mu\, M)\, s = b \tag{6.18b}$$

$$H + \mu\, M \textit{ is positive semidefinite.} \tag{6.18c}$$

The number μ is uniquely determined.

(*ii*) *Now suppose that $(s, \mu) \in \mathbb{R}^n \times \mathbb{R}$ is such that* (6.18) *is satisfied. Then s is a global minimizer of* (6.17)*.*

(*iii*) *If $(s, \mu) \in \mathbb{R}^n \times \mathbb{R}$ is such that* (6.18) *is satisfied and in addition, $H + \mu\, M$ is positive definite, then s is the unique global minimizer of* (6.17)*.*

Proof. Statement (*i*): Suppose that s is a global minimizer of (6.17). Then $\|s\|_M \le \Delta$ is obvious.

Case 1: $\|s\|_M < \Delta$ holds.
Then s is also a local minimizer of the unconstrained problem

$$\text{Minimize} \quad q(s), \quad \text{where } s \in \mathbb{R}^n.$$

Consequently, the necessary optimality condition of first order (Theorem 3.1) holds, i. e.,

$$\nabla q(s) = H\, s - b = 0,$$

as well as the necessary optimality condition of second order (Theorem 3.2), i. e., H is positive semidefinite. This shows that (6.18) holds for the choice $\mu = 0$.[46]

Owing to the complementarity in (6.18a), $\mu = 0$ is the only possible choice.

Case 2: $\|s\|_M = \Delta$ holds, and in particular, $s \neq 0$.
We first show that there exists $\mu \ge 0$ such that $(H + \mu\, M)\, s = b$ holds. We proceed by contradiction and therefore suppose that for all $\mu \ge 0$, we have $(H + \mu\, M)\, s \neq b$.

The choice $\mu = 0$ implies $y := \nabla q(s) = H\, s - b \neq 0$. The choices $\mu > 0$ imply that the vectors y and $M\, s$ cannot be anti-parallel. Therefore, $M^{-1} y$ and s cannot be anti-parallel either. Concerning the angle α (w.r.t. the M-inner product) between these two vectors, we thus have

$$\cos\alpha = \frac{(M^{-1}y)^\mathsf{T} M\, s}{\|M^{-1}y\|_M\, \|s\|_M} = \frac{y^\mathsf{T} s}{\|y\|_{M^{-1}}\, \|s\|_M} > -1.$$

Let v be a vector in the direction of the angle bisector between $-M^{-1}y$ and $-s$, e. g.,

$$v := -\frac{M^{-1}y}{\|y\|_{M^{-1}}} - \frac{s}{\|s\|_M} \neq 0.$$

Then we have

$$\begin{aligned} y^\mathsf{T} v &= y^\mathsf{T} \left(-\frac{M^{-1}y}{\|y\|_{M^{-1}}} - \frac{s}{\|s\|_M} \right) \\ &= -\|y\|_{M^{-1}} - \frac{y^\mathsf{T} s}{\|s\|_M} \frac{\|y\|_{M^{-1}}}{\|y\|_{M^{-1}}} \\ &= -\|y\|_{M^{-1}} \big(1 + \cos\alpha\big) \\ &< 0. \end{aligned}$$

Hence we conclude $q'(s)\, v = y^\mathsf{T} v < 0$ and therefore v is a descent direction for q at s. Due to

$$\begin{aligned} \left[\frac{\mathrm{d}}{\mathrm{d}t} \frac{1}{2} \|s + t\, v\|_M^2 \right]_{t=0} &= v^\mathsf{T} M\, s \\ &= \left(-\frac{M^{-1}y}{\|y\|_{M^{-1}}} - \frac{s}{\|s\|_M} \right)^\mathsf{T} M\, s \\ &= -\frac{y^\mathsf{T} s}{\|y\|_{M^{-1}}} \frac{\|s\|_M}{\|s\|_M} - \|s\|_M \\ &= -\|s\|_M \big(\cos\alpha + 1\big) \\ &< 0 \end{aligned}$$

we have $\|s + t\, v\|_M < \|s\|_M = \Delta$ for small enough $t > 0$. Hence we can obtain a feasible point with a strictly smaller objective value than s, which is in contradiction to the optimality of s.

Consequently, it must be possible to satisfy the conditions (6.18a) and (6.18b) with some $\mu \ge 0$. It is not difficult to see that, in fact, μ must be unique. Indeed, the assumption $H\, s + \mu_1\, M\, s = b = H\, s + \mu_2\, M\, s$ leads to $(\mu_1 - \mu_2)\, M\, s = 0$, which is only possible for $\mu_1 = \mu_2$ due to $s \neq 0$ and the invertibility of M.

[46] In this case q is convex, and thus s is also a a global minimizer of the unconstrained problem.

It remains to show that $H + \mu M$ is positive semidefinite, i. e., (6.18c). To this end, we consider a direction $d \in \mathbb{R}^n$ with the property $d^\intercal M s < 0$. We need to show

$$d^\intercal (H + \mu M)\, d \ge 0.$$

Let us define $t := \dfrac{-2\, d^\intercal M s}{\|d\|_M^2} > 0$. Then

$$\|s + t\, d\|_M^2 = \|s\|_M^2 + 2\, t\, d^\intercal M s + t^2\, \|d\|_M^2 = \|s\|_M^2 \le \Delta^2.$$

The global optimality of s implies

$$\begin{aligned}
0 &\le q(s + t\, d) - q(s) \\
&= q'(s)\,(t\, d) + \frac{1}{2}(t\, d)^\intercal H\,(t\, d) \\
&= t\, y^\intercal d + \frac{t^2}{2} d^\intercal H d \\
&= -t\, \mu\, s^\intercal M d + \frac{t^2}{2} d^\intercal H d \quad \text{due to } y = H s - b = -\mu M s \\
&= \frac{t^2}{2}\mu\, \|d\|_M^2 + \frac{t^2}{2} d^\intercal H d \quad \text{due to } \frac{t}{2}\,\|d\|_M^2 = -d^\intercal M s \\
&= \frac{t^2}{2} d^\intercal (H + \mu M)\, d.
\end{aligned}$$

Thus we have confirmed that $d^\intercal (H + \mu M)\, d \ge 0$ holds for all directions $d \in \mathbb{R}^n$ satisfying $d^\intercal M s < 0$. Since the sign of d is irrelevant, the same result likewise holds for directions $d^\intercal M s > 0$. The remaining case of directions $d^\intercal M s = 0$ follows by continuity.

Statement (*ii*): Suppose that $(s, \mu) \in \mathbb{R}^n \times \mathbb{R}$ is such that (6.18) is satisfied. Let $\bar{s} \in \mathbb{R}^n$ with $\|\bar{s}\|_M \le \Delta$ be an arbitrary comparison point, $d := \bar{s} - s$ and $y := \nabla q(s) = H s - b$ as above. We estimate

$$\begin{aligned}
q(\bar{s}) - q(s) &= y^\intercal d + \frac{1}{2} d^\intercal H d \\
&= -\mu\, s^\intercal M d + \frac{1}{2} d^\intercal H d \quad \text{due to } y = H s - b = -\mu M s \text{ by (6.18b)} \\
&\ge -\mu\, s^\intercal M d - \frac{1}{2}\mu\, \|d\|_M^2 \quad \text{by (6.18c)} \\
&= -\frac{\mu}{2}\big(2\, s^\intercal M d + \|d\|_M^2\big) \\
&= -\frac{\mu}{2}\big(\|d + s\|_M^2 - \|s\|_M^2\big) \\
&= -\frac{\mu}{2}\big(\|\bar{s}\|_M^2 - \|s\|_M^2\big) \\
&\ge 0.
\end{aligned}$$

The previous inequality requires a comment. It holds trivially in case $\mu = 0$. For $\mu > 0$, (6.18a) implies $\|s\|_M^2 = \Delta^2$, and by assumption we have $\|\bar{s}\|_M^2 \le \Delta^2$. This shows the final inequality and thus the global optimality of s for (6.17).

Statement (iii): If $H + \mu M$ is even positive definite, then the estimate above can be sharpened for $\bar{s} \neq s$, i. e., $d \neq 0$:

$$\begin{aligned} q(\bar{s}) - q(s) &= -\mu\, s^\intercal M\, d + \frac{1}{2} d^\intercal H\, d \\ &> \ -\mu\, s^\intercal M\, d - \frac{1}{2}\mu\, \|d\|_M^2 \\ &= \cdots \\ &= -\frac{\mu}{2}\left(\|\bar{s}\|_M^2 - \|s\|_M^2\right) \\ &\geq 0. \end{aligned}$$

This shows that s is even the unique global minimizer of (6.17). □

Remark 6.13 (on the characterization of global minimizers)**.**

(i) *The remarkable fact about Theorem 6.12 is that the optimality condition* (6.18) *is simultaneously necessary and sufficient, even though problem* (6.17) *is generally non-convex.*

(ii) *As an alternative to the proof in Theorem 6.12 we might use an optimality condition of Karush-Kuhn-Tucker type (see Chapter 2) to prove that* (6.18a) *and* (6.18b) *are* necessary *conditions for any* local *minimizer and, in particular, for global minimizers of* (6.17)*. In this way, we will see that the number* μ *can be viewed as the unique Lagrange multiplier pertaining to the constraint*

$$\frac{1}{2}\left(\|s\|_M^2 - \Delta^2\right) \leq 0,$$

which is of course equivalent to $\|s\|_M \leq \Delta$*. As we can see easily, the linear independence constraint qualification (LICQ) always holds.*[47] *Proceeding in this way, however, we will not find the condition* (6.18c) *that is — as we saw in the proof — characteristic for* global *minimizers of* (6.17)*.*

(iii) *It was proved in Martínez, 1994 that* (6.17) *can have, besides its global minimizers, at most one additional local minimizer which is not a global minimizer.*

Based on the characterization (6.18), one can devise methods to find an exact global minimizer of the trust-region subproblem (6.17). The most prominent method utilizes a one-dimensional Newton method for the equation

$$\frac{1}{\|(H + \mu\, M)^{-1} b\|_M} - \frac{1}{\Delta} = 0$$

to determine the value of μ, unless $\mu = 0$ and $s = H^{-1}b$ already solve (6.18); see Nocedal, Wright, 2006, Chapter 4.3 for details if you are interested.

In the remainder of this section we discuss a practical method to find an inexact solution s to the trust-region subproblem (6.17). Our goal is to satisfy

[47] See Definition 8.16 for the definition of LICQ. In case $\|s\|_M < \Delta$, the only constraint present is inactive, sothe LICQ holds. Otherwise we have $\|s\|_M = \Delta > 0$ and thus $s \neq 0$. This implies that the gradient of the only active constraint, $M\,s$, is non-zero and thus linearly independent, so the LICQ holds also in this case.

(i) the conditions for global convergence imposed in Theorem 6.9, i. e., the trust-region constraint $\|s\|_M \le \Delta$ and the (weak) fraction of Cauchy decrease condition (6.11),

(ii) the conditions for fast local convergence imposed in Theorem 6.11.

It will turn out that a clever modification of the conjugate gradient method — the **Steihaug(-Toint) conjugate gradient method**[48] — is capable of satisfying these requirements.[49]

This variant of the CG method will be applied to the linear system $H\,s = b$. In case that H is s. p. d., this system constitutes the necessary and sufficient optimality condition for an unconstrained trust-region subproblem, i. e., (6.17) with $\Delta = \infty$. The Steihaug-Toint variant of the CG algorithm has two significant modifications compared to the base version (Algorithm 4.17).[50]

(1) Similarly as in the truncated conjugate method (Algorithm 5.41), we start the method with $s^{(0)} = 0$ and we detect the occurrence of a search direction $p^{(\ell)}$ with non-positive curvature $(p^{(\ell)})^\intercal H p^{(\ell)} \le 0$. In constrat to the truncated CG method, where we would stop immediately, we proceed with the current step and produce $s^{(\ell+1)} = s^{(\ell)} + \alpha^* p^{(\ell)}$, where α^* is not the usual Cauchy step size but is chosen such that $s^{(\ell+1)}$ lies on the boundary of the trust region. We then stop the method.

(2) In case that the current step would leave the trust region, i. e., in case $\|s^{(\ell)} + \alpha^{(\ell)}\, p^{(\ell)}\|_M > \Delta$, we do not utilize the full step size $\alpha^{(\ell)}$ but, again, proceed only to the boundary of the trust region and then stop the algorithm.

If none of these two situations occur, then the Steihaug-Toint CG method stops as soon as an approximate solution to $H\,s = -b$ has been found with sufficiently small residual norm. With regards to Theorem 6.11 we utilize the relative residual norm and a tolerance given by a suitable forcing sequence, compare (5.37),

$$\frac{\|\text{residual associated with } s^{(\ell)}\|_{M^{-1}}}{\|\text{residual associated with } 0\|_{M^{-1}}} = \frac{\|\zeta^{(\ell)}\|_{M^{-1}}}{\|b\|_{M^{-1}}} = \frac{\|H\,s^{(\ell)} - b\|_{M^{-1}}}{\|b\|_{M^{-1}}} \le \eta. \tag{6.19}$$

In case the outer iterate $x^{(k)}$ is already close to a point x^* satisfying the second-order sufficient optimality condition, one can guarantee that the Steihaug-Toint CG method will *not* stop for any of the two reasons above but it stops as soon as (6.19) holds. Therefore, the condition (6.16) concerning the accuracy of the step proposal is satisfied, and Theorem 6.11 yields the Q-superlinear convergence, provided that the further requisites of that theorem hold.

The Steihaug-Toint CG algorithm for the inexact solution of (6.17) is given in Algorithm 6.14.

Algorithm 6.14 (Steihaug-Toint conjugate gradient method for the trust-region subproblem (6.17)).
Input: *negative model derivative* $-q'(0) = b \in \mathbb{R}^n$

[48] after Steihaug, 1983 and Toint, 1981

[49] Other, slighly more elaborate approaches, can be found in the literature, e. g., the **dogleg method** and the **two-dimensional subspace minimization approach**; see Nocedal, Wright, 2006, Chapter 4.1 if you are interested.

[50] As we did with the truncated conjugate method (Algorithm 5.41) in the context of inexact Newton methods, we switch to problem adapted variable names. That is, we use iterates $s^{(\ell)}$, search directions $p^{(\ell)}$ and residuals $\zeta^{(\ell)}$.

```
Input: symmetric matrix $H$ (or matrix-vector products with $H$)
Input: s. p. d. matrix $M$ (or matrix-vector products with $M^{-1}$)
Input: relative residual $\varepsilon_{\text{rel}}$
Input: trust-region radius $\Delta > 0$
Output: approximate solution of the trust-region subproblem (6.17)
Set $\ell := 0$
Set $s^{(0)} := 0$                                           // zero initial guess
Set $\zeta^{(0)} := -b$                                      // evaluate the initial residual
Set $p^{(0)} := -M^{-1}\zeta^{(0)}$
Set $\delta^{(0)} := -(\zeta^{(0)})^\intercal p^{(0)}$      // $\delta^{(0)} = \|\zeta^{(0)}\|^2_{M^{-1}}$
while $\delta^{(\ell)} \ge \varepsilon_{\text{rel}}^2 \delta^{(0)}$ do        // check stopping criterion (6.19)
    Set $q^{(\ell)} := H\,p^{(\ell)}$
    Set $\theta^{(\ell)} := (q^{(\ell)})^\intercal p^{(\ell)}$
    if $\theta^{(\ell)} > 0$ then
        Set $\alpha^{(\ell)} := \delta^{(\ell)}/\theta^{(\ell)}$
        Set $s^{(\ell+1)} := s^{(\ell)} + \alpha^{(\ell)}\,p^{(\ell)}$
        if $\|s^{(\ell+1)}\|_M > \Delta$ then                // iterate would leave the trust region
            Determine $\alpha^*$ as the positive solution of $\|s^{(\ell)} + \alpha\,p^{(\ell)}\|_M = \Delta$
            Set $s^{(\ell+1)} := s^{(\ell)} + \alpha^*\,p^{(\ell)}$     // go to the boundary of the trust region
            Set $\ell := \ell + 1$
            Abort the while loop
        else
            Set $\zeta^{(\ell+1)} := \zeta^{(\ell)} + \alpha^{(\ell)} q^{(\ell)}$
            Set $p^{(\ell+1)} := -M^{-1}\zeta^{(\ell+1)}$
            Set $\delta^{(\ell+1)} := -(\zeta^{(\ell+1)})^\intercal p^{(\ell+1)}$      // $\delta^{(\ell+1)} = \|\zeta^{(\ell+1)}\|^2_{M^{-1}}$
            Set $\beta^{(\ell+1)} := \delta^{(\ell+1)}/\delta^{(\ell)}$
            Set $p^{(\ell+1)} := p^{(\ell+1)} + \beta^{(\ell+1)}\,p^{(\ell)}$
            Set $\ell := \ell + 1$
        end if
    else
        Determine $\alpha^*$ as the positive solution of $\|s^{(\ell)} + \alpha\,p^{(\ell)}\|_M = \Delta$
        Set $s^{(\ell+1)} := s^{(\ell)} + \alpha^*\,p^{(\ell)}$         // go to the boundary of the trust region
        Set $\ell := \ell + 1$
        Abort the while loop
    end if
end while
return $s^{(\ell)}$
```

Remark 6.15 (on Algorithm 6.14).

(i) *The first iterate $s^{(1)}$ is the Cauchy point of problem (6.17).*

(ii) *We consider three cases which may occur in an iteration.*

(a) *In case of a “usual” CG step (identical to what the plain CG algorithm 4.17 would produce),* *then*

$$q(s^{(\ell+1)}) = q(s^{(\ell)}) - \frac{1}{2}\,\underbrace{\alpha^{(\ell)}}_{>0}\,\underbrace{\delta^{(\ell)}}_{>0}$$

holds, see (4.12).

(b) *If $\theta^{(\ell)} > 0$ holds (the current search direction $p^{(\ell)}$ is a direction of positive curvature) but the full Cauchy step $s^{(\ell)} + \alpha^{(\ell)}\, p^{(\ell)}$ lands outside the trust region, then it is useful follow the convex function*

$$\alpha \mapsto \varphi(\alpha) := f + b^\top(s^{(\ell)} + \alpha\, p^{(\ell)}) + \frac{1}{2}(s^{(\ell)} + \alpha\, p^{(\ell)})^\top H\,(s^{(\ell)} + \alpha\, p^{(\ell)})$$

in the direction $\alpha > 0$ to the boundary of the trust region in order to obtain the largest possible descent, since

$$\varphi'(0) = b^\top p^{(\ell)} + (s^{(\ell)})^\top H\, p^{(\ell)} = (\zeta^{(\ell)})^\top p^{(\ell)} = -\delta_\ell = -\|\zeta^{(\ell)}\|^2_{M^{-1}} < 0$$

ist.

(c) *If, by contrast, $\theta^{(\ell)} \le 0$ holds (the current search direction $p^{(\ell)}$ is a direction of non-positive curvature), then the function φ is concave. This suggests, again, to proceed to the boundary of the trust region in the direction $\alpha > 0$.*

(*iii*) *The considerations above show: in case the Steihaug-Toint CG algorithm does not stop with $s^{(1)}$ but continues, the further iterates continue to reduce the objective q monotonically. Therefore, the inexact solution $s^{(\ell)}$ returned by Algorithm 6.14 always satisfies the fraction of Cauchy decrease condition (6.11) with $\underline{C} = 1$.*

(*iv*) *The reason for terminating the algorithm when an iterate $s^{(\ell+1)}$ is about to leave the trust region is the following. As we know from Lemma 4.22, the sequence of norms $\|s^{(\ell)} - 0\|_M$ is strictly monotonically increasing. If we would let the algorithm continue, we would never return into the trust region.*

(*v*) *Concerning the recursive update of the quantity $\|s^{(\ell)}\|_M$ without using the matrix M, we can refer to (4.33)–(4.34) to obtain the formulas*

$$\omega^{(0)} := 0, \qquad \omega^{(k+1)} := \omega^{(k)} + 2\,\alpha^{(k)}\xi^{(k)} + (\alpha^{(k)})^2\,\gamma^{(k)} \tag{6.20a}$$

$$\xi^{(0)} := 0, \qquad \xi^{(k+1)} := \beta^{(k+1)}\,(\xi^{(k)} + \alpha^{(k)}\gamma^{(k)}) \tag{6.20b}$$

$$\gamma^{(0)} := \delta^{(0)}, \qquad \gamma^{(k+1)} := \delta^{(k+1)} + (\beta^{(k+1)})^2\,\gamma^{(k)} \tag{6.20c}$$

for the quantities

$$\omega^{(\ell)} := \|s^{(\ell)} - 0\|^2_M, \quad \xi^{(\ell)} := (s^{(\ell)} - 0)^\top M\, p^{(\ell)}, \quad \gamma^{(\ell)} := \|p^{(\ell)}\|^2_M.$$

(*vi*) *Using these quantities, we can evaluate the step size α^* required to reach the boundary by solving the quadratic equation*

$$\|s^{(\ell)} + \alpha\, p^{(\ell)}\|^2_M = \omega^{(\ell)} + \alpha\,\xi^{(\ell)} + \alpha^2\gamma^{(\ell)} = \Delta^2.$$

Due to $\omega^{(\ell)} = \|s^{(\ell)}\|_M^2 < \Delta$*, this equation has exactly one positive solution, which is*

$$\alpha^* := -\frac{\xi^{(\ell)}}{2\,\gamma^{(\ell)}} + \frac{\xi^{(\ell)}}{2\,\gamma^{(\ell)}} \left(1 - 4\,\gamma^{(\ell)}\,(\omega^{(\ell)} - \Delta^2)\right)^{1/2}.$$

End of Week 7

Chapter 2 Theory for Constrained Optimization Problems

§ 7 Introduction

In this chapter we come back to the generic constrained nonlinear optimization problem (**nonlinear program** or **NLP**) with basic set $\mathbb{R}^n$ from (1.1),

$$\left.\begin{array}{rll} \text{Minimize} & f(x) & \text{where } x \in \mathbb{R}^n \\ \text{subject to} & g_i(x) \le 0 & \text{for } i = 1, \ldots, n_{\text{ineq}} \\ \text{and} & h_j(x) = 0 & \text{for } j = 1, \ldots, n_{\text{eq}}. \end{array}\right\} \tag{7.1}$$

We have finitely many (possibly zero) inequality constraints $g_i \colon \mathbb{R}^n \to \mathbb{R}$ and equality constraints $h_j \colon \mathbb{R}^n \to \mathbb{R}$, i. e., $n_{\text{ineq}}, n_{\text{eq}} \in \mathbb{N}_0$. The set

$$F := \left\{ x \in \mathbb{R}^n \,\middle|\, g_i(x) \le 0 \text{ for all } i = 1, \ldots, n_{\text{ineq}},\ h_j(x) = 0 \text{ for all } j = 1, \ldots, n_{\text{eq}} \right\} \tag{7.2}$$

associated with problem (7.1) is termed the **feasible set**. Any $x \in F$ is termed a **feasible point**.

Our goal is to derive optimality conditions for (7.1) which are verifiable numerically and which will serve as the basis for numerical algorithms.

Assumption 7.1. *Throughout Chapter 2 we are assuming that $f \colon \mathbb{R}^n \to \mathbb{R}$, $g_i \colon \mathbb{R}^n \to \mathbb{R}$ and $h_j \colon \mathbb{R}^n \to \mathbb{R}$ are all C^1 functions, $i = 1, \ldots, n_{\text{ineq}}$ and $j = 1, \ldots, n_{\text{eq}}$.*

The derivation of a first-order optimality condition for problem (7.1) is significantly more complex than the derivation for the unconstrained problem (**UP**) achieved in Theorem 3.1. Recall that the proof was based on considering curves $\gamma(t) = x^* + t\,d$ which run through the local minimizer x^*. Since problem (**UP**) was unconstrained, we were free to choose any direction d and we obtained $f'(x^*)\,d = 0$ for all $d \in \mathbb{R}^n$. In the present situation, however, $\gamma(t)$ may not be feasible even for arbitrarily small $t > 0$. Therefore, we will need to restrict the directions d.

It will turn out that the appropriate set of directions is the tangent cone.

§ 7.1 The Tangent Cone

Definition 7.2 (Tangent cone). *Suppose that $M \subseteq \mathbb{R}^n$ is an arbitrary set and $x \in M$. Then*

$$\mathcal{T}_M(x) := \left\{ d \in \mathbb{R}^n \,\middle|\, \text{there exist sequences } x^{(k)} \in M \text{ and } t^{(k)} \searrow 0 \text{ such that } d = \lim_{k \to \infty} \frac{x^{(k)} - x}{t^{(k)}} \right\} \tag{7.3}$$

*is termed the **tangent cone** of M at x. A vector $d \in \mathcal{T}_M(x)$ is termed a **tangent direction** of M at x. We also define $\mathcal{T}_M(x) := \emptyset$ for $x \notin M$.*

Note: The sequence $x^{(k)}$ in (7.3) necessarily converges to x.

The tangent cone is also known as **contingent cone** or **Bouligand cone** in the literature. By the way, a set $K \subseteq \mathbb{R}^n$ is termed a **cone** if $x \in K$ and $\beta > 0$ imply that $\beta x \in K$ holds.

Lemma 7.3 (Properties of the tangent cone). *Suppose that $M \subseteq \mathbb{R}^n$ is an arbitrary set. Then $\mathcal{T}_M(x)$ is a closed cone.*

Proof. We can assume $x \in M$ since otherwise $\mathcal{T}_M(x) = \emptyset$ holds, which is a closed cone.

We first show the cone property. Suppose that $d \in \mathcal{T}_M(x)$ and $\beta > 0$. That is, there exist sequences $x^{(k)} \in M$ and $t^{(k)} \searrow 0$ such that $d = \lim_{k \to \infty} \frac{x^{(k)} - x}{t^{(k)}}$. Replacing $t^{(k)}$ by $\beta^{-1} t^{(k)}$, the limit becomes $\beta\, d$. This shows $\beta\, d \in \mathcal{T}_M(x)$.

It remains to show the closeness of $\mathcal{T}_M(x)$. To this end, let $d^{(\ell)} \in \mathcal{T}_M(x)$ be a sequence which converges to some $d \in \mathbb{R}^n$. We need to show $d \in \mathcal{T}_M(x)$.

For each $\ell \in \mathbb{N}$, there exist sequences $\big(x^{(\ell,k)}\big)_k \in M$ and $\big(t^{(\ell,k)}\big)_k$ satisfying

$$x^{(\ell,k)} \to x, \quad t^{(\ell,k)} \searrow 0 \quad \text{and} \quad \frac{x^{(\ell,k)} - x}{t^{(\ell,k)}} \to d^{(\ell)} \quad \text{as } k \to \infty.$$

This implies that for each $\ell \in \mathbb{N}$, there exists an index $k^{(\ell)}$ such that

$$\|x^{(\ell,k^{(\ell)})} - x\| \le \frac{1}{\ell}, \quad 0 < t^{(\ell,k^{(\ell)})} \le \frac{1}{\ell} \quad \text{and} \quad \left\| \frac{x^{(\ell,k^{(\ell)})} - x}{t^{(\ell,k^{(\ell)})}} - d^{(\ell)} \right\| \le \frac{1}{\ell}.$$

We now consider the “diagonal” sequences

$$\widehat{x}^{(\ell)} := x^{(\ell,k^{(\ell)})} \quad \text{and} \quad \widehat{t}^{(\ell)} := t^{(\ell,k^{(\ell)})}.$$

These satisfy $\widehat{x}^{(\ell)} \to x$ and $\widehat{t}^{(\ell)} \searrow 0$ as well as

$$\left\| \frac{\widehat{x}^{(\ell)} - x}{\widehat{t}^{(\ell)}} - d \right\| \le \left\| \frac{\widehat{x}^{(\ell)} - x}{\widehat{t}^{(\ell)}} - d^{(\ell)} \right\| + \|d^{(\ell)} - d\| \le \frac{1}{\ell} + \|d^{(\ell)} - d\| \to 0,$$

which shows that d belongs to $\mathcal{T}_M(x)$. □

We can now formulate and prove a first version of the first-order necessary optimality condition for local minimizers of (7.1).

Theorem 7.4 (First-order necessary optimality condition).
Suppose that x^* is a local minimizer of (7.1). Then

$$f'(x^*)\,d \ge 0 \quad \textit{for all } d \in \mathcal{T}_F(x^*). \tag{7.4}$$

This theorem states that in a local minimizer, f is growing to first order in any direction tangent to the feasible set.

Proof. Suppose that $d \in \mathcal{T}_F(x^*)$. Then there exist sequences $x^{(k)} \in F$ and $t^{(k)} \searrow 0$ such that

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

holds. Since f is continuously differentiable, the mean value theorem 2.4 implies that there exist numbers $\xi^{(k)} \in (0,1)$ such that

$$f(x^{(k)}) = f(x^*) + f'\big(x^* + \xi^{(k)}\,(x^{(k)} - x^*)\big)(x^{(k)} - x^*).$$

Using the local optimality of x^* and the fact that $x^{(k)} \to x^*$, we obtain

$$\begin{aligned} 0 &\le f(x^{(k)}) - f(x^*) \\ &= f'\big(x^* + \xi^{(k)}\,(x^{(k)} - x^*)\big)(x^{(k)} - x^*). \end{aligned}$$

Division by $t^{(k)}$ leaves us with

$$0 \le f'\big(\underbrace{x^* + \xi^{(k)}\,(x^{(k)} - x^*)}_{\to x^*}\big)\underbrace{\Big(\frac{x^{(k)} - x^*}{t^{(k)}}\Big)}_{\to d}.$$

Using again the C^1-property of f, we infer that the expression on the right converges, so we conclude

$$0 \le f'(x^*)\,d. \qquad \square$$

In analogy to § 3, we refer to points $x \in F$ satisfying (7.4) as **stationary points**.

There are two obstacles that make the first-order necessary optimality condition (7.4) generally difficult to use in practice:

(1) (7.4) comes in the form of a variational inequality. Even when a candidate point x^* is given, it may be difficult to verify (7.4) due to the generally infinite number of tangent directions $d \in \mathcal{T}_F(x^*)$. *Finding* a candidate point x^* on the basis of (7.4) is even more difficult.

(2) The tangent cone's implicit definition can render its explicit characterization difficult. After all, $\mathcal{T}_F(x)$ only refers to the feasible set F itself but it does not depend on the description (7.2) of F in terms of the inequality and equality constraints g_i and h_j.

Therefore we now seek to replace the tangent cone $\mathcal{T}_F(x)$ by a cone which is simpler to handle algorithmically. More precisely, we will try and replace an analytical/geometrical object (the tangent cone) by an algebraic object (the linearizing cone).

§ 7.2 The Linearizing Cone

Definition 7.5 (Linearizing cone).
Consider the feasible set F as in (7.2).

(*i*) *For $x \in F$, we define*

$$\begin{aligned} \mathcal{A}(x) &:= \{i \in \{1, \ldots, n_{\text{ineq}}\} \mid g_i(x) = 0\} && \textit{the set of active indices at } x, \\ \mathcal{I}(x) &:= \{i \in \{1, \ldots, n_{\text{ineq}}\} \mid g_i(x) < 0\} && \textit{the set of inactive indices at } x. \end{aligned}$$

(*ii*) *For $x \in F$, the set*

$$\mathcal{T}_F^{\text{lin}}(x) := \left\{ d \in \mathbb{R}^n \;\middle|\; \begin{aligned} g_i'(x)\, d &\le 0 && \textit{for all } i \in \mathcal{A}(x) \\ h_j'(x)\, d &= 0 && \textit{for all } j = 1, \ldots, n_{\text{eq}} \end{aligned} \right\} \tag{7.5}$$

is termed the ***linearizing cone*** *or* ***cone of first-order feasible directions*** *to the feasible set at x. We set $\mathcal{T}_F^{\text{lin}}(x) := \emptyset$ when $x \notin F$.*

Note: Inactive inequalities do not play a role in the definition of the linearizing cone.

Remark 7.6 (on the linearizing cone).

(*i*) *Consider the closed convex polyhedral set*[1]

$$F^{\text{lin}}(x) = \left\{ y \in \mathbb{R}^n \;\middle|\; \begin{aligned} g_i(x) + g_i'(x)\,(y - x) &\le 0 && \textit{for all } i = 1, \ldots, n_{\text{ineq}} \\ h_j(x) + h_j'(x)\,(y - x) &= 0 && \textit{for all } j = 1, \ldots, n_{\text{eq}} \end{aligned} \right\}$$

for a given point $x \in F$. One can show that the linearizing cone is also the tangent cone to $F^{\text{lin}}(x)$ at the point $x \in F \cap F^{\text{lin}}(x)$, i. e.,

$$\mathcal{T}_F^{\text{lin}}(x) = \mathcal{T}_{F^{\text{lin}}(x)}(x).$$

(*ii*) *$\mathcal{T}_F^{\text{lin}}(x)$ is a closed convex cone.*

[1] A convex polyhedral set is the intersection of finitely many half spaces.

(*iii*) *In contrast to the tangent cone $\mathcal{T}_F(x)$, the linearizing cone $\mathcal{T}_F^{\text{lin}}(x)$ does depend on the description of the feasible set F in terms of the inequality and equality constraints.*

The details of this remark are worked out in homework problem 8.3 and homework problem 8.4.

Lemma 7.7 (Relation between the cones)**.**
Consider the feasible set F as in (7.2) and $x \in F$. Then $\mathcal{T}_F(x) \subseteq \mathcal{T}_F^{\text{lin}}(x)$.

Proof. Suppose that $d \in \mathcal{T}_F(x)$. Then there exist sequences $x^{(k)} \in F$ and $t^{(k)} \searrow 0$ such that

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

We prove that $g_i'(x)\, d \le 0$ holds for all $i \in \mathcal{A}(x)$. Let us fix $i \in \mathcal{A}(x)$. The mean value theorem 2.4 implies

$$\begin{aligned} 0 &\ge g_i(x^{(k)}) \quad \text{since } x^{(k)} \text{ is feasible} \\ &= \underbrace{g_i(x)}_{=0} + g_i'\big(x + \xi^{(k)}\,(x^{(k)} - x)\big)(x^{(k)} - x) \end{aligned}$$

with some $\xi^{(k)} \in (0,1)$. Division by $t^{(k)}$ and passage to the limit $k \to \infty$ yields $0 \ge g_i'(x)\, d$.

Analogously we can show $h_j'(x)\, d = 0$ for all $j = 1, \ldots, n_{\text{eq}}$, and thus we obtain $d \in \mathcal{T}_F^{\text{lin}}(x)$. □

§ 7.3 Polar Cones

The formulation of optimality conditions in an algebraic form amenable to solution algorithms is going to require the introduction of a dual description of the linearizing cone $\mathcal{T}_F^{\text{lin}}(x)$. We now introduce such a dual description for arbitrary sets.

Definition 7.8 (Polar cone)**.** *Suppose that $M \subseteq \mathbb{R}^n$ is an arbitrary set. The* ***polar cone*** *to M is defined as*

$$M^\circ \coloneqq \{s \in \mathbb{R}^n \mid s^\mathsf{T} x \le 0 \text{ for all } x \in M\}. \tag{7.6}$$

Note: The polar cone consists of the normal vectors $s \in \mathbb{R}^n$ of hyperplanes through the origin such that M is entirely contained in the negative half space $H^-(s,0) = \{y \in \mathbb{R}^n \mid s^\mathsf{T} y \le 0\}$. One also says that the hyperplane $s^\mathsf{T} y = 0$ **separates** 0 and M.

Lemma 7.9 (Properties of the polar cone)**.** *Suppose that $M, M_1, M_2 \subseteq \mathbb{R}^n$ are arbitrary sets.*

(*i*) *M° is a closed convex cone.*

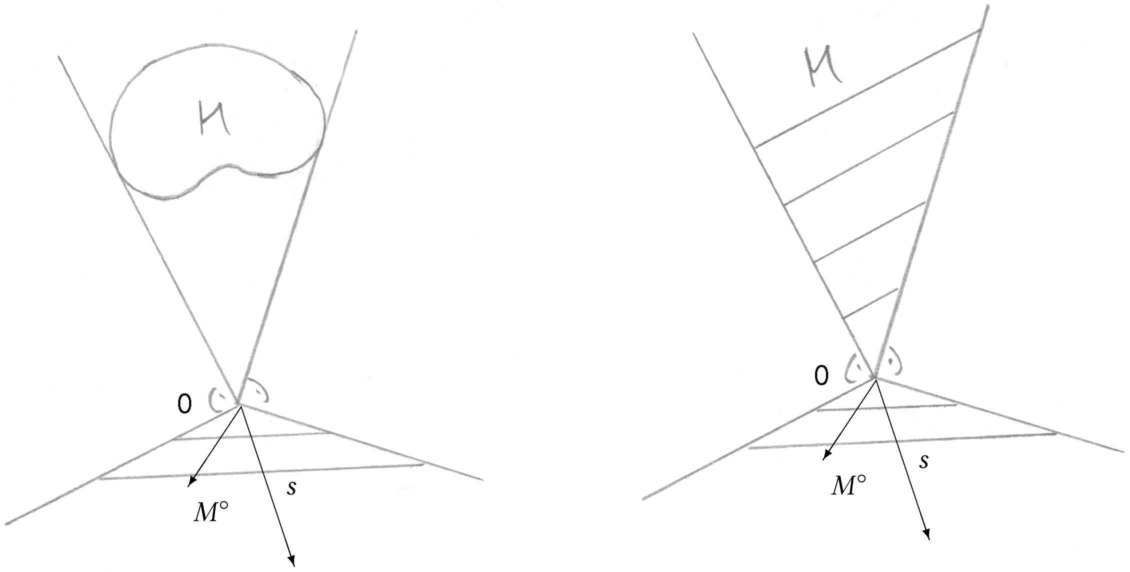


Figure 7.1: Illustration of two sets M which have the same polar cone M°.

(ii) *$M_1 \subseteq M_2$ implies $M_2^\circ \subseteq M_1^\circ$.*

Proof. The proof is part of homework problem 8.2. □

Example 7.10 (Polar cones).

(i) *Suppose that A is an affine subspace of $\mathbb{R}^n$, i. e., $A = U + \{\overline{x}\}$ where U is a subspace of $\mathbb{R}^n$ and $\overline{x} \in \mathbb{R}^n$. Then $A^\circ = \{\overline{x}\}^\circ \cap U^\perp$ (the intersection of the polar cone of $\{\overline{x}\}$ with the orthogonal complement of U w.r.t. the Euclidean inner product).*

(ii) *In the absence of inequality constraints ($n_{\text{ineq}} = 0$), the polar of the linearizing cone has the representation*

$$\begin{aligned}\mathcal{T}_F^{\text{lin}}(x)^\circ &= \{s \in \mathbb{R}^n \mid s \text{ is some linear combination of } h_j'(x)^\mathsf{T},\ j = 1, \ldots, n_{\text{eq}}\} \\ &= \operatorname{range} h'(x)^\mathsf{T}\end{aligned}$$

for $x \in F$.

(iii) *Let $N \coloneqq (\mathbb{R}_{\ge 0})^n$ denote the non-negative orthant in $\mathbb{R}^n$. Then $N^\circ = (\mathbb{R}_{\le 0})^n$ is the non-positive orthant.*

(iv) *For $0 \in \mathbb{R}^n$, we have $\{0\}^\circ = \mathbb{R}^n$ and $(\mathbb{R}^n)^\circ = \{0\}$.*

The proof is part of homework problem 8.2.

§ 7.4 The Farkas Lemma and the Representation of $\mathcal{T}_F^{\text{lin}}(x)^\circ$

We recall that our plan is to understand the dual description $\mathcal{T}_F^{\text{lin}}(x)^\circ$ of the linearizing cone $\mathcal{T}_F^{\text{lin}}(x)$. In the absence of inequality constraints, this polar has already been characterized in Statement (ii) of Example 7.10. For the general case, we require the Farkas lemma.

Lemma 7.11 (Farkas lemma (1902)).
Suppose that $B \in \mathbb{R}^{m\times n}$ and $c \in \mathbb{R}^n$. Then the following are equivalent.

(i) *The linear system $B^\mathsf{T}\xi = c$ has a non-negative solution $\xi \ge 0$.*

(ii) *All elements of $\{d \in \mathbb{R}^n \mid B\,d \ge 0\}$ satisfy $c^\mathsf{T} d \ge 0$.*

Statement (i) means that c belongs to the closed convex cone[2]

$$K := \{B^\mathsf{T}\,\xi \mid \xi \in \mathbb{R}^m,\ \xi \ge 0\}.$$

In order to understand Statement (ii), we reason as follows:

$$\begin{aligned} B\,d \ge 0 \quad &\Leftrightarrow \quad \xi^\mathsf{T} B\,d \ge 0 \quad \text{for all } \xi \ge 0 \quad \text{(by Statement } (iii) \text{ of Example 7.10)} \\ &\Leftrightarrow \quad d^\mathsf{T}(B^\mathsf{T}\xi) \ge 0 \quad \text{for all } \xi \ge 0 \\ &\Leftrightarrow \quad K \text{ is a subset of the half space } H^+(d,0) = \{x \in \mathbb{R}^n \mid d^\mathsf{T} x \ge 0\}. \end{aligned}$$

We can thus read Statement (ii) as follows. Any half space $H^+(d,0)$ that contains the cone K also contains the point c. Therefore, the negation of Statement (ii) means that there exists a vector d such that K is contained in $H^+(d,0)$ but the point c is not. In this case, we say that the hyperplane $H(d,0)$ **separates** separates K and c.

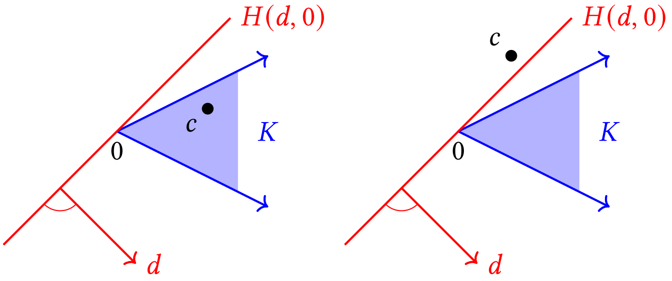


Figure 7.2: Illustration of both cases in the Farkas lemma 7.11. Left: Statements (i) and (ii) are both true. Right: Statements (i) and (ii) are both false.

Proof of Lemma 7.11. We first show Statement (i) ⇒ Statement (ii): Suppose that $\xi \ge 0$ satisfies $B^\mathsf{T}\xi = c$. Moreover, let $d \in \mathbb{R}^n$ be such that $B\,d \ge 0$. Then we obtain

$$c^\mathsf{T} d = (B^\mathsf{T}\xi)^\mathsf{T} d = \xi^\mathsf{T}(B\,d) \ge 0.$$

[2] It is very easy to check that K is a convex cone. We do not prove here that it is closed. Please see Herzog, 2022, Lemma 6.10 if you are interested in the proof.

In order to prove Statement $(ii) \Rightarrow$ Statement (i), we consider the contraposition, i. e., $\neg$ Statement $(i) \Rightarrow \neg$ Statement (ii). Thus we assume $c \notin K$. Due to $0 \in K$ we must have $c \neq 0$. Let $\operatorname{cl} B_R(c)$ be the closed ball with radius $R = \|c\|$. We consider the orthogonal projection of c onto $K \cap \operatorname{cl} B_R(c)$, i. e.,

$$\begin{aligned} \text{Minimize} \quad & \|x - c\|, \quad \text{where } x \in \mathbb{R}^n \\ \text{subject to} \quad & x \in K \cap \operatorname{cl} B_R(c). \end{aligned} \tag{$*$}$$

Since K is closed and $\operatorname{cl} B_R(c)$ is compact, $K \cap \operatorname{cl} B_R(c)$ is compact as well. By the Weierstraß' extreme value theorem[3], problem $(*)$ has a global minimizer w. The point w is also a global minimizer of the relaxed problem

$$\begin{aligned} \text{Minimize} \quad & \|x - c\|, \quad \text{where } x \in \mathbb{R}^n \\ \text{subject to} \quad & x \in K, \end{aligned} \tag{$**$}$$

since points outside of $\operatorname{cl} B_R(c)$ are definitely not global minimizers of $(**)$. (**Quiz 7.1:** Why is it that points outside of $\operatorname{cl} B_R(c)$ cannot be global minimizers of $(**)$?)

We now prove that $d = w - c$ serves as the normal vector of a hyperplane which separates K from the point c. The construction is illustrated in Figure 7.3. Notice that we have $d \neq 0$ due to $K \ni w \neq c \notin K$.

Suppose that y is an arbitrary point in K. All points on the line segment between y and w, i. e., $\alpha\, y + (1 - \alpha)\, w$ for $\alpha \in [0, 1]$, also belong to K due to convexity. We obtain

$$\begin{aligned} \|w - c\|^2 &\le \|\alpha\, y + (1 - \alpha)\, w - c\|^2 \quad \text{since } w \text{ is optimal for } (**) \\ &= \|\alpha\, (y - w) + (w - c)\|^2 \\ &= \alpha^2 \|y - w\|^2 + 2\,\alpha (y - w)^\top (w - c) + \|w - c\|^2. \end{aligned}$$

This implies

$$2\,(y - w)^\top \underbrace{(w - c)}_{=d} \ge -\alpha\, \|y - w\|^2$$

for all $\alpha \in [0, 1]$. Passage to the limit $\alpha \searrow 0$ implies

$$(y - w)^\top d \ge 0 \text{ for all } y \in K. \tag{$***$}$$

Inserting $y = 2\,w$ and $y = 0$ (both belong to K), we conclude $w^\top d \ge 0$ and simultaneously $w^\top d \le 0$, hence

$$w^\top d = 0. \tag{$****$}$$

Moreover,

$$c^\top d = (c - w)^\top d + w^\top d = -\|\underbrace{w - c}_{=d \neq 0}\|^2 + \underbrace{w^\top d}_{=0} < 0. \tag{$*****$}$$

Overall we obtain

$$\begin{aligned} y^\top d &\ge w^\top d && \text{by } (***) \\ &= 0 && \text{by } (****) \\ &> c^\top d && \text{by } (*****) \end{aligned}$$

for all $y \in K$. This shows that $K \subseteq H^+(d, 0)$ but $c \notin H^+(d, 0)$. In other words, we obtain that indeed, $d = w - c$ serves as the normal vector of a hyperplane which separates K from the point c. That is, Statement (ii) does not hold, which was to be proved. □

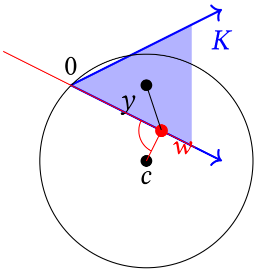


Figure 7.3: Illustration of the construction of the normal vector $d = w - c$ of the separating hyperplane (red) in the proof of the Farkas-Lemmas 7.11.

Remark 7.12. *Statement (i) of the Farkas-Lemmas 7.11 can be verified numerically by checking whether or not the linear optimization problem*

$$\begin{aligned} \text{Minimize} \quad & 0^\top \xi, \quad \text{where } \xi \in \mathbb{R}^m \\ \text{subect to} \quad & B^\top \xi = c \\ \text{and} \quad & \xi \ge 0 \end{aligned}$$

has a feasible point, e. g., using phase I of the simplex method; see for instance Herzog, 2022, *§ 7.2.*

Lemma 7.13 (Representation of $\mathcal{T}_F^{\text{lin}}(x)^\circ$).
Consider the feasible set F as in (7.2) and $x \in F$. Then the polar $\mathcal{T}_F^{\text{lin}}(x)^\circ$ of the linearizing cone (7.5) has the representation

$$\begin{aligned} \mathcal{T}_F^{\text{lin}}(x)^\circ &= \left\{ \sum_{i \in \mathcal{A}(x)} \mu_i \nabla g_i(x) + \sum_{j=1}^{n_{\text{eq}}} \lambda_j \nabla h_j(x) \;\middle|\; \begin{matrix} \mu_i \ge 0 & \textit{for } i \in \mathcal{A}(x) \\ \lambda_j \in \mathbb{R} & \textit{for } j = 1, \ldots, n_{\text{eq}} \end{matrix} \right\} \\ &= \left\{ \sum_{i=1}^{n_{\text{ineq}}} \mu_i \nabla g_i(x) + \sum_{j=1}^{n_{\text{eq}}} \lambda_j \nabla h_j(x) \;\middle|\; \begin{matrix} \mu_i \ge 0 & \textit{for } i = 1, \ldots, n_{\text{ineq}} \\ \mu_i = 0 & \textit{for } i \in \mathcal{I}(x) \\ \lambda_j \in \mathbb{R} & \textit{for } j = 1, \ldots, n_{\text{eq}} \end{matrix} \right\}. \end{aligned} \tag{7.7}$$

Proof. We first show that the set on the right-hand side is a subset of $\mathcal{T}_F^{\text{lin}}(x)^\circ$. To this end, let y be a vector from the right-hand side of (7.7) and let $d \in \mathcal{T}_F^{\text{lin}}(x)$ be arbitrary, i. e., we have $g_i'(x)\, d \le 0$ for all $i \in \mathcal{A}(x)$ and $h_j'(x)\, d = 0$ for all $j = 1, \ldots, n_{\text{eq}}$. We need to show $y^\top d \le 0$. This follows immediately from

$$\begin{aligned} y^\top d &= \sum_{i \in \mathcal{A}(x)} \underbrace{\mu_i}_{\ge 0} \underbrace{g_i'(x)\, d}_{\le 0} \quad + \quad \sum_{j=1}^{n_{\text{eq}}} \lambda_j \underbrace{h_j'(x)\, d}_{=0} \\ &\le 0 + 0. \end{aligned}$$

It remains to show that, conversely, $\mathcal{T}_F^{\text{lin}}(x)^\circ$ is also a subset of the set on the right-hand side. Here we will use the Farkas lemma 7.11 and set

$$B := \begin{bmatrix} -g_i'(x)|_{i \in \mathcal{A}(x)} \\ -h_j'(x)|_{j=1,\ldots,n_{\text{eq}}} \\ h_j'(x)|_{j=1,\ldots,n_{\text{eq}}} \end{bmatrix}.$$

[3] A continuous function attains its global minimum (and its global maximum) on any compact set.

This choice implies

$$d \in \mathcal{T}_F^{\text{lin}}(x) \quad \Leftrightarrow \quad B\,d \ge 0.$$

Now let $y \in \mathcal{T}_F^{\text{lin}}(x)^\circ$, hence

$$y^\intercal d \le 0 \quad \text{for all } d \in \mathbb{R}^n \text{ satisfying } B\,d \ge 0.$$

We need to show that y has a representation as the elements in the right-hand side of (7.7). Using the Farkas lemma 7.11 with $c = -y$ yields that the system

$$B^\intercal \xi = -y \text{ has a solution } \xi \ge 0.$$

We split ξ according to

$$\xi =: \begin{pmatrix} \mu_i \big|_{i \in \mathcal{A}(x)} \\ \lambda_j^+ \big|_{j=1,\ldots,n_{\text{eq}}} \\ \lambda_j^- \big|_{j=1,\ldots,n_{\text{eq}}} \end{pmatrix} \quad \text{and set} \quad \lambda := \lambda^+ - \lambda^-.$$

Hence we can write $B^\intercal \xi = -y$ or $-B^\intercal \xi = y$ as

$$\sum_{i \in \mathcal{A}(x)} \mu_i \nabla g_i(x) + \sum_{j=1}^{n_{\text{eq}}} \lambda_j \nabla h_j(x) = y,$$

which confirms that y has indeed a representation as in (7.7). □

§ 8 First-Order Necessary Optimality Conditions

We are now in the position to revisit first-order optimality conditions for (7.1). We already know from Theorem 7.4 that we have

$$f'(x^*)\,d \ge 0 \quad \text{for all } d \in \mathcal{T}_F(x^*) \tag{7.4}$$

at a local minimizer x^*. Equivalently, we could write

$$-\nabla f(x^*) \in \mathcal{T}_F(x^*)^\circ. \tag{8.1}$$

However, the tangent cone $\mathcal{T}_F(x^*)$ and its polar $\mathcal{T}_F(x^*)^\circ$ are generally difficult to characterize. By contrast, the linearizing cone $\mathcal{T}_F^{\text{lin}}(x^*)$ and its polar $\mathcal{T}_F^{\text{lin}}(x^*)^\circ$ have an explicit description in terms of the constraint functions' derivatives and are much easier to handle.

We have the relation $\mathcal{T}_F(x) \subseteq \mathcal{T}_F^{\text{lin}}(x)$ and thus, by Lemma 7.9, $\mathcal{T}_F^{\text{lin}}(x)^\circ \subseteq \mathcal{T}_F(x)^\circ$. This means that, unfortunately, $\mathcal{T}_F^{\text{lin}}(x)^\circ$ may be "too small" and we cannot expect, in general,

$$-\nabla f(x^*) \in \mathcal{T}_F^{\text{lin}}(x^*)^\circ. \tag{8.2}$$

Before we address this gap and how to overcome it, let us see what (8.2) means explicitly.

Definition 8.1 (Lagrange function). *The function*

$$\begin{aligned}\mathcal{L}(x,\mu,\lambda) &\coloneqq f(x) + \sum_{i=1}^{n_{\text{ineq}}} \mu_i\, g_i(x) + \sum_{j=1}^{n_{\text{eq}}} \lambda_j\, h_j(x) \\ &= f(x) + \mu^\mathsf{T} g(x) + \lambda^\mathsf{T} h(x) \end{aligned} \tag{8.3}$$

is termed the ***Lagrange function*** *or* ***Lagrangian*** *associated with problem* (7.1).

One says that the constraints g and h are **adjoined** to the objective f.

Definition 8.2 (KKT conditions, Lagrange multipliers).

(i) *The conditions*

$$\nabla_x \mathcal{L}(x,\mu,\lambda) = \left\{\begin{matrix} \nabla f(x) + \displaystyle\sum_{i=1}^{n_{\text{ineq}}} \mu_i \nabla g_i(x) + \sum_{j=1}^{n_{\text{eq}}} \lambda_j \nabla h_j(x) \\ \nabla f(x) + \quad g'(x)^\mathsf{T}\mu \quad + \quad h'(x)^\mathsf{T}\lambda \end{matrix}\right\} = 0, \tag{8.4a}$$

$$h(x) = 0, \tag{8.4b}$$

$$\mu \ge 0, \quad g(x) \le 0, \quad \mu^\mathsf{T} g(x) = 0 \tag{8.4c}$$

are termed the ***Karush-Kuhn-Tucker conditions*** *or* ***KKT conditions***[4] *associated with problem* (7.1).

(ii) *Given a point $x \in F$, any vectors μ and λ satisfying* (8.4) *are called* ***Lagrange multipliers*** *associated with the inequality and equality constraints, respectively, at x.*

(iii) *We denote the set of all Lagrange multipliers at $x \in F$ by*

$$\Lambda(x) \coloneqq \left\{ (\mu,\lambda) \;\middle|\; \begin{matrix} g'(x)^\mathsf{T}\mu + h'(x)^\mathsf{T}\lambda = -\nabla f(x) \\ \mu \ge 0, \quad \mu^\mathsf{T} g(x) = 0 \end{matrix} \right\}. \tag{8.5}$$

We set $\Lambda(x) \coloneqq \emptyset$ for $x \notin F$.

(iv) *A point x is called a* ***KKT point*** *for problem* (7.1) *if $\Lambda(x) \neq \emptyset$, i. e., if there exist Lagrange multipliers (μ,λ) such that (x,μ,λ) satisfy the KKT conditions* (8.4).

Remark 8.3 (on the KKT conditions).

(i) *Condition* (8.4c) *is known as a* ***complementarity condition***.[5] *It can be written equivalently as*

$$\mu_i \ge 0, \quad g_i(x) \le 0, \quad \mu_i\, g_i(x) = 0 \quad \textit{for all } i = 1,\ldots,n_{\text{ineq}},$$

[4] Karush, 1939; Kuhn, Tucker, 1951

[5] A complementarity system is any system of inequalities and equalities in the form $a \gtrless 0$, $b \gtrless 0$ and $a^\mathsf{T} b = 0$.

i. e., for every i, at least one of the two numbers μ_i and $g_i(x)$ is equal to 0. (This explains the name ***complementarity****.)*

In particular, a Lagrange multiplier μ_i associated with an inactive inequality constraint $g_i(x) < 0$ must be equal to 0. Sometimes, one writes (8.4c) in the form

$$0 \le \mu \quad \perp \quad g(x) \le 0.$$

(*ii*) *Given $x \in F$, the conditions (8.5) on the Lagrange multipliers are linear equations and inequalities. Therefore, the set $\Lambda(x)$ is a polyhedron. Whether or not $\Lambda(x)$ is empty can be verified numerically by checking whether or not the linear optimization problem*

$$\begin{aligned} \textit{Minimize} \quad & 0^\top \mu + 0^\top \lambda \\ \textit{subject to} \quad & g'(x)^\top \mu + h'(x)^\top \lambda = -\nabla f(x) \\ \textit{and} \quad & \mu \ge 0 \\ \textit{and} \quad & \mu^\top g(x) = 0 \end{aligned} \tag{8.6}$$

has a feasible point, e. g., using phase I of the simplex method; see for instance Herzog, 2022, § 7.2. The associated dual problem reads

$$\begin{aligned} \textit{Minimize} \quad & f'(x)\, d, \quad && \textit{where } d \in \mathbb{R}^n \\ \textit{subject to} \quad & g_i'(x)\, d \le 0 \quad && \textit{for } i \in \mathcal{A}(x) \\ \textit{and} \quad & h_j'(x)\, d = 0 \quad && \textit{for } j = 1, \ldots, n_{\text{eq}}. \end{aligned} \tag{8.7}$$

That is, the dual problem considers the minimization of the linearized objective on the linearizing cone $\mathcal{T}_F^{\text{lin}}(x)$.

The primal problem (8.6) cannot be unbounded. The dual problem (8.7) always has the feasible point $d = 0$. Therefore, only the following two situations can occur:

(1) *Both problems (8.6) and (8.7) are solvable.*
This means that Lagrange multipliers exist, i. e., $\Lambda(x) \neq \emptyset$.

(2) *The primal problem (8.6) is infeasible, and the dual problem (8.7) is unbounded.*
No Lagrange multipliers exist, i. e., $\Lambda(x) = \emptyset$.

Situation (2) means that there exists a descent direction for the objective f at x, which belongs to the linearizing cone. By contrast, in situation (1), all directions in the linearizing cone as non-descent directions.

We will now see that the KKT conditions are just an explicit way of writing (8.2).

Lemma 8.4 (Significance of the KKT conditions)**.**
Consider the optimization problem (7.1). The following statements are equivalent.

(*i*) *x is a KKT point.*
That is, there exist Lagrange multipliers (μ, λ) such that the KKT conditions (8.4) hold.

(*ii*) *x is a feasible point that satisfies (8.2).*
That is, $-\nabla f(x) \in \mathcal{T}_F^{\text{lin}}(x)^\circ$.

Proof. Equation (8.4a) of the KKT conditions state that

$$-\nabla f(x) = \sum_{i=1}^{n_{\text{ineq}}} \mu_i \, \nabla g_i(x) + \sum_{j=1}^{n_{\text{eq}}} \lambda_j \, \nabla h_j(x) \tag{$*$}$$

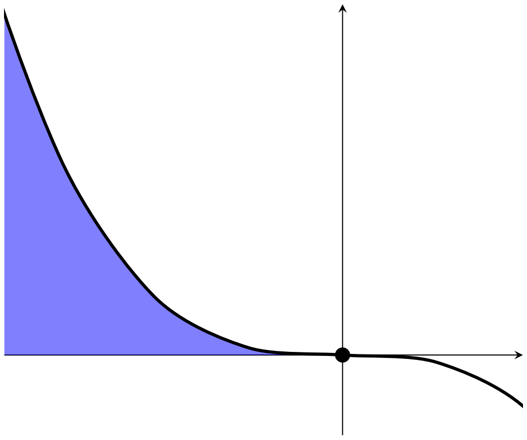
holds with scalars λ_j and μ_i, for which (8.4c) implies: $\mu_i \ge 0$ for all $i = 1, \ldots, n_{\text{ineq}}$ and even $\mu_i = 0$ for $i \in \mathcal{I}(x)$. Owing to the representation (7.7) of the polar $\mathcal{T}_F^{\text{lin}}(x)^\circ$ of the linearizing cone, ($*$) together with the conditions on μ is equivalent to $-\nabla f(x) \in \mathcal{T}_F^{\text{lin}}(x)^\circ$. Moreover, the conditions (8.4b) and (8.4c) ensure the feasibility of x. □

The KKT conditions (8.4) associate with an optimization problem (7.1) a (generally nonlinear) system of equations and inequalities. (**Quiz 8.1:** In which case are the KKT conditions affine conditions?) They are the basis of most numerical constrained optimization solvers. Unfortunately, the KKT conditions are not, in general, necessary optimality conditions, since they express the condition $-\nabla f(x) \in \mathcal{T}_F^{\text{lin}}(x)^\circ$ instead of $-\nabla f(x) \in \mathcal{T}_F(x)^\circ$. The following example shows that this may actually be an issue in practice.

Example 8.5 ($\mathcal{T}_F^{\text{lin}}(x) \subsetneq \mathcal{T}_F(x)$)**.** *Consider the optimization problem*

$$\begin{aligned} \textit{Minimize} \quad & -x_1 \\ \textit{subject to} \quad & x_2 + x_1^3 \le 0 \\ \textit{and} \quad & -x_2 \le 0. \end{aligned} \tag{8.8}$$

The feasible set has the following shape:



Obviously, $x^* = (0,0)^\intercal$ is the unique global minimizer of (8.8), and there are no further local minimizers. The inequalities g_1 and g_2 are both active in x^*, and we find

$$
\begin{aligned}
\nabla g_1(x) &= \begin{pmatrix} 3\,x_1^2 \\ 1 \end{pmatrix} &&\Rightarrow& \nabla g_1(x^*) &= \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\
\nabla g_2(x) &= \begin{pmatrix} 0 \\ -1 \end{pmatrix} &&\Rightarrow& \nabla g_2(x^*) &= \begin{pmatrix} 0 \\ -1 \end{pmatrix}.
\end{aligned}
$$

We evaluate

$$
\begin{aligned}
\mathcal{T}_F(x^*) &= \{d \in \mathbb{R}^2 \mid d_1 \le 0,\ d_2 = 0\} = \mathbb{R}_{\le 0} \times \{0\}, \\
\mathcal{T}_F^{\text{lin}}(x^*) &= \{d \in \mathbb{R}^2 \mid g_i'(x^*)\, d \le 0,\ i = 1,2\} \\
&= \{d \in \mathbb{R}^2 \mid d_2 = 0\} = \mathbb{R} \times \{0\},
\end{aligned}
$$

so that we infer $\mathcal{T}_F^{\text{lin}}(x^*) \supsetneq \mathcal{T}_F(x^*)$.

The optimality condition (7.4) is satisfied as expected from Theorem 7.4:

$$
f'(x^*)\, d = \begin{pmatrix} -1 & 0 \end{pmatrix} d \ge 0 \quad \text{for all } d \in \mathcal{T}_F(x^*) = \mathbb{R}_{\le 0} \times \{0\}.
$$

The same variational inequality is not *satisfied with the larger cone $\mathcal{T}_F^{\text{lin}}(x^*)$. Consequently, $-\nabla f(x^*)$ belongs to $\mathcal{T}_F(x^*)^\circ$ but it does not belong to $\mathcal{T}_F^{\text{lin}}(x^*)^\circ$:*

$$
\begin{aligned}
-\nabla f(x^*) &\in \mathcal{T}_F(x^*)^\circ = \{s \in \mathbb{R}^2 \mid s_1 \ge 0\} = \mathbb{R}_{\ge 0} \times \mathbb{R}, \\
-\nabla f(x^*) &\notin \mathcal{T}_F^{\text{lin}}(x^*)^\circ = \{s \in \mathbb{R}^2 \mid s_1 = 0\} = \{0\} \times \mathbb{R}.
\end{aligned}
$$

Consequently, the KKT conditions cannot be satisfied at x^*. Here they boil down to

$$
\begin{pmatrix} -1 \\ 0 \end{pmatrix} + \mu_1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \mu_2 \begin{pmatrix} 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}
$$

and they cannot be satisfied for any $\mu \in \mathbb{R}^2$, $\mu \ge 0$.

We are going to bridge the gap between $-\nabla f(x) \in \mathcal{T}_F^{\text{lin}}(x)^\circ$ and $-\nabla f(x) \in \mathcal{T}_F(x)^\circ$ by means of **constraint qualifications** (**CQs**).

§ 8.1 First-Order Necessary Optimality Conditions Under the Abadie and Guignard CQ

Definition 8.6 (Abadie CQ and Guignard CQ[6]).

(i) *A feasible point $x \in F$ satisfies the* ***Abadie constraint qualification*** *(**ACQ**) in case $\mathcal{T}_F^{\text{lin}}(x) = \mathcal{T}_F(x)$.*[7]

(ii) *A feasible point $x \in F$ satisfies the* ***Guignard constraint qualification*** *(**GCQ**) in case $\mathcal{T}_F^{\text{lin}}(x)^\circ = \mathcal{T}_F(x)^\circ$.*

[6] see Guignard, 1969 and Flegel, Kanzow, 2005, Corollary 3.8

[7] Some authors also speak of a **quasiregular point**; see Bertsekas, 1999, p.345–349.

Note: The ACQ implies the GCQ.

Theorem 8.7 (KKT conditions are necessary optimality conditions under the Abadie CQ and the Guignard CQ)**.**
Suppose that x^* is a local minimizer of (7.1) which satisfies the Abadie CQ or the Guignard CQ. Then $\Lambda(x^*) \neq \emptyset$, i. e., there exist Lagrange multipliers μ^* and λ^* (not necessarily unique) such that the KKT conditions (8.4) are satisfied.

Proof. As a local minimizer of (7.1), x^* is feasible of course. From Theorem 7.4 and the ACQ or GCQ we infer

$$-\nabla f(x^*) \in \mathcal{T}_F^{\text{lin}}(x^*)^\circ = \mathcal{T}_F(x^*)^\circ.$$

Lemma 8.4 confirms that this is equivalent to the KKT conditions. □

Remark 8.8 (on the significance of constraint qualifications)**.**

(*i*) *Constraint qualifications (CQs) render the KKT conditions necessary optimality conditions. Therefore, the KKT conditions (8.4) are sometimes called a* ***qualified optimality system****.*

(*ii*) *If the GCQ does* not *hold at a local minimizer x^*, then Lagrange multipliers* may *fail to exist, as we saw in Example 8.5. It is also possible that they do exist, but it depends on the objective.*

(*iii*) *CQs tell us whether the tangent cone to the feasible set admits an algebraic description in terms of the linearizing cone. The latter depends on our description of the feasible set F in terms of inequality and equality constraints. Two different descriptions of the same feasible set may differ w.r.t. the CQs they satisfy.*

(*iv*) *A general weakness of CQs is that they need to be checked at the point x^*, which is usually a priori unknown. Algorithms searching for KKT points may miss minimizers which fail to satisfy a CQ.*

(*v*) *The verification of the ACQ or GCQ is often cumbersome. Therefore, we will later introduce stronger CQs which are easier to handle and which imply the ACQ.*

(*vi*) *In the absence of CQs, one arrives at* ***Fritz-John conditions (FJ conditions)***[8] **Note:** *Was bedeuten denn die FJ-Bedingungen wirklich, vor allem natürlich für den Fall $\mu_0 = 0$? Das wird eine Aussage über den Linearisierungskegel sein.*

$$\mu_0 \nabla f(x) + g'(x)^\top \mu + h'(x)^\top \lambda = 0, \tag{8.9a}$$

$$h(x) = 0, \tag{8.9b}$$

$$\mu \ge 0, \quad g(x) \le 0, \quad \mu^\top g(x) = 0, \tag{8.9c}$$

$$\mu_0 \ge 0. \tag{8.9d}$$

A point x is termed a ***Fritz-John point*** *for problem (7.1) if there exist (μ_0, μ, λ) which are not all zero, such that (x, μ_0, μ, λ) satisfy the FJ conditions (8.9).*

[8] John, 1948

*One calls (8.9) an **unqualified optimality system** since one can show—without assuming any CQ—that any local minimizer of (7.1) is an FJ point; see for instance Geiger, Kanzow, 2002, p.72. In case we happen to have $\mu_0 > 0$, we can re-normalize the vector (μ_0, μ, λ) so that $\mu_0 = 1$ holds, and we have indeed a KKT point.*

§ 8.2 First-Order Necessary Optimality Conditions Under Affine Constraints

We consider a special case of problem (7.1) which has only affine ("linear") constraints.

$$\left.\begin{aligned} \text{Minimize} \quad & f(x), && \text{where } x \in \mathbb{R}^n \\ \text{subject to} \quad & A_{\text{ineq}}\, x \le b_{\text{ineq}} \\ \text{and} \quad & A_{\text{eq}}\, x = b_{\text{eq}} \end{aligned}\right\} \tag{8.10}$$

with $A_{\text{ineq}} \in \mathbb{R}^{n_{\text{ineq}} \times n}$ and $A_{\text{eq}} \in \mathbb{R}^{n_{\text{eq}} \times n}$ as well as $b_{\text{ineq}} \in \mathbb{R}^{n_{\text{ineq}}}$ and $b_{\text{eq}} \in \mathbb{R}^{n_{\text{eq}}}$. The KKT conditions associated with (8.10) read

$$\nabla f(x) + A_{\text{ineq}}^\mathsf{T} \mu + A_{\text{eq}}^\mathsf{T} \lambda = 0, \tag{8.11a}$$

$$A_{\text{eq}}\, x - b_{\text{eq}} = 0, \tag{8.11b}$$

$$\mu \ge 0, \quad A_{\text{ineq}}\, x - b_{\text{ineq}} \le 0, \quad \mu^\mathsf{T}(A_{\text{ineq}}\, x - b_{\text{ineq}}) = 0. \tag{8.11c}$$

Theorem 8.9 (KKT conditions are necessary optimality conditions under affine constraints)**.**
Suppose that x^* is a local minimizer of (8.10). Then $\Lambda(x^*) \neq \emptyset$, i. e., there exist Lagrange multipliers μ^* and λ^* (not necessarily unique) such that the KKT conditions (8.11) are satisfied.

Proof. When the feasible set F is described exclusively in terms of affine constraints, one can show that the Abadie CQ holds at every feasible point; see homework problem 8.4. □

The observation that affine constraints imply the ACQ everywhere is the reason why one does not explicitly require CQs in linear optimization; compare Herzog, 2022, § 8.

End of Week 8

§ 8.3 First-Order Necessary Optimality Conditions Under MFCQ

Definition 8.10 (Mangasarian-Fromovitz Constraint Qualification)**.**
*A feasible point $x \in F$ satisfies the **Mangasarian-Fromovitz constraint qualification** (**MFCQ**) if the following conditions hold:*

(*i*) *The gradients $\{\nabla h_j(x)\}_{j=1}^{n_{\text{eq}}}$ are linearly independent. In other words, the Jacobian $h'(x)$ has full row rank (is surjective).*

(*ii*) *There exists a vector* $d \in \mathbb{R}^n$ *such that*

$$\begin{aligned} g_i'(x)\, d &< 0 \quad \text{for all } i \in \mathcal{A}(x), \\ h_j'(x)\, d &= 0 \quad \text{for all } j = 1, \ldots, n_{\text{eq}}. \end{aligned} \tag{8.12}$$

The vector d from (8.12) is sometimes called an **MFCQ vector**. Notice that condition (*i*) implies $0 \le n_{\text{eq}} \le n$, i. e., there cannot be too many equality constraints.

Lemma 8.11 (alternative formulation of condition (*ii*)). *Suppose that* x *is a feasible point for* (7.1)*. Then the following are equivalent to condition* (*ii*)*:*

(*iii*) *There exists a vector* $d \in \mathbb{R}^n$ *such that*

$$\begin{aligned} g(x) + g'(x)\, d &< 0, \\ h(x) + h'(x)\, d &= 0. \end{aligned} \tag{8.13}$$

(*iv*) *There exists a vector* $d \in \mathbb{R}^n$ *such that*

$$\begin{aligned} g_i'(x)\, d &\le -1 \quad \text{for all } i \in \mathcal{A}(x), \\ h_j'(x)\, d &= 0 \quad \text{for all } j = 1, \ldots, n_{\text{eq}}. \end{aligned} \tag{8.14}$$

Proof. □

Remark 8.12 (on the Mangasarian-Fromovitz CQs). *We can numerically verify the MFCQ at a feasible point* x *as follow.*

(*i*) *We determine the row rank of* $h'(x)$*, e. g., by computing its singular values.*

(*ii*) *We verify whether or not the linear optimization problem with zero objective and constraints as in* (8.14) *is feasible.*[9]

Lemma 8.13 (MFCQ implies existence of a feasible curve).
Suppose that x *is a feasible point for* (7.1) *which satisfies the MFCQ with the MFCQ vector* d*. Then there exist* $\varepsilon > 0$ *and a curve* $\gamma\colon (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ *with the following properties:*

(*i*) γ *is of class* C^1 *on* $(-\varepsilon, \varepsilon)$.

(*ii*) γ *satisfies* $\begin{cases} h(\gamma(t)) = 0 & \text{for } t \in (-\varepsilon, \varepsilon), \\ g(\gamma(t)) \le 0 & \text{for } t \in [0, \varepsilon), \\ g(\gamma(t)) < 0 & \text{for } t \in (0, \varepsilon). \end{cases}$

[9] Further reading: Burke, 2014, Ch.7.2, p.76, who also discusses the dual of this LP.

(iii) $\gamma(0) = x$ *and* $\dot\gamma(0) = d$.

Proof. In the absence of equality constraints, we simply set $\gamma(t) \coloneqq x + t\,d$ on some interval $(-\varepsilon_0, \varepsilon_0)$; then $\gamma(0) = x$ and $\dot\gamma(0) = d$ hold. Otherwise, we will apply the implicit function theorem to show the existence of the curve γ with the desired properties. The curve will be constructed using only the equality constraints. However, we cannot apply the implicit function theorem directly to $h(\gamma(t)) = 0$ since the number of equality constraints n_{eq} is generally smaller than the dimension n. As a remedy, we consider the function

$$H(y, t) \coloneqq h\big(x + t\,d + h'(x)^\intercal y\big). \tag{$*$}$$

This function is of class C^1 and it satisfies $H(0, 0) = h(x) = 0$. The partial Jacobian w.r.t. y is

$$H_y(y, t) = h'\big(x + t\,d + h'(x)^\intercal y\big)\, h'(x)^\intercal.$$

In particular,

$$H_y(0, 0) = h'(x)\, h'(x)^\intercal$$

holds, which is an s. p. d. matrix in $\mathbb{R}^{n_{\text{eq}} \times n_{\text{eq}}}$. **(Quiz 8.2:** Can you argue why?)

We now apply the implicit function theorem to $H(y, t) = 0$. There exists $\varepsilon_0 > 0$ and a C^1-curve $\xi\colon (-\varepsilon_0, \varepsilon_0) \to \mathbb{R}^{n_{\text{eq}}}$ such that

$$H(\xi(t), t) = 0 \quad \text{holds for } t \in (-\varepsilon_0, \varepsilon_0) \quad \text{and} \quad \xi(0) = 0.$$

The derivative of ξ can be obtained from the linear system

$$H_y(\xi(t), t)\, \dot\xi(t) = -H_t(\xi(t), t) = -h'\big(x + t\,d + h'(x)^\intercal \xi(t)\big)\, d.$$

In particular, at $t = 0$, we obtain

$$H_y(0, 0)\, \dot\xi(0) = -H_t(0, 0) = -h'(x)\, d = 0,$$

which shows $\dot\xi(0) = 0$. Using ξ, we now define the C^1-curve

$$\gamma(t) \coloneqq x + t\,d + h'(x)^\intercal \xi(t) \quad \text{for } t \in (-\varepsilon_0, \varepsilon_0).$$

This function has the properties

$$\gamma(0) = x + 0\,d + h'(x)^\intercal \xi(0) = x + h'(x)^\intercal 0 = x$$

and $\dot\gamma(t) = d + h'(x)^\intercal \dot\xi(t)$ and thus

$$\dot\gamma(0) = d + h'(x)^\intercal \dot\xi(0) = d + h'(x)^\intercal 0 = d,$$

as desired. Moreover, by construction, we have

$$h(\gamma(t)) = H(\xi(t), t) = 0 \quad \text{for all } t \in (-\varepsilon_0, \varepsilon_0).$$

It remains to show $g(\gamma(t)) \le 0$ for $t \in [0, \varepsilon_0)$ and the strict inequality when $t \in (0, \varepsilon_0)$. When $i \in \{1, \ldots, n_{\text{ineq}}\}$ is an inactive index at x, i. e., $g_i(x) < 0$, then by continuity, there exists $\varepsilon_i > 0$ such that

$$g_i(\gamma(t)) < 0 \quad \text{for all } t \in (-\varepsilon_i, \varepsilon_i).$$

By contrast, when i is an active index at x, i. e., $g_i(x) = 0$, let us abbreviate $\varphi(t) \coloneqq g_i(\gamma(t))$. By the chain rule, this function is of class C^1. Then $\varphi(0) = g_i(\gamma(0)) = g_i(x) = 0$ holds as well as

$$\begin{aligned} \dot\varphi(t) &= g_i'(\gamma(t))\,\dot\gamma(t) \\ \text{and thus } \dot\varphi(0) &= g_i'(\gamma(0))\,\dot\gamma(0) \\ &= g_i'(x)\,d < 0. \end{aligned}$$

The properties $\varphi(0) = 0$ and $\dot\varphi(0) < 0$ and the continuity of $\dot\varphi$ imply that there exists $\varepsilon_i > 0$ such that $\varphi(t) \le 0$ holds for all $t \in [0, \varepsilon_i)$ and even $\varphi(t) < 0$ for all $t \in (0, \varepsilon_i)$.

Choosing $\varepsilon \coloneqq \min\{\varepsilon_0, \varepsilon_1, \dots, \varepsilon_{n_\text{ineq}}\}$ yields Statement *(ii)*. □

Corollary 8.14 (MFCQ implies ACQ)**.**
Suppose that x is a feasible point for (7.1) *which satisfies the MFCQ. Then the ACQ holds at x.*

Proof. We need to show $\mathcal{T}_F^\text{lin}(x) = \mathcal{T}_F(x)$. The inclusion $\mathcal{T}_F(x) \subseteq \mathcal{T}_F^\text{lin}(x)$ follows from Lemma 7.7. We only need to show the reverse inclusion $\mathcal{T}_F^\text{lin}(x) \subseteq \mathcal{T}_F(x)$.

To this end, let $d_0 \in \mathcal{T}_F^\text{lin}(x)$ be arbitrary and let d be an MFCQ vector, i. e., we have

$$\begin{aligned} g_i'(x)\,d_0 &\le 0 \text{ and } g_i'(x)\,d < 0 \quad \text{for all } i \in \mathcal{A}(x), \\ h_j'(x)\,d_0 &= 0 \text{ and } h_j'(x)\,d = 0 \quad \text{for all } j = 1, \dots, n_\text{eq}. \end{aligned}$$

We set $d(\tau) \coloneqq d_0 + \tau\, d$ for $\tau > 0$. Then

$$\begin{aligned} g_i'(x)\,d(\tau) &< 0 \qquad \text{for all } i \in \mathcal{A}(x), \\ h_j'(x)\,d(\tau) &= 0 \qquad \text{for all } j = 1, \dots, n_\text{eq}. \end{aligned}$$

That is, $d(\tau)$ is also an MFCQ vector.

We show $d(\tau) \in \mathcal{T}_F(x)$ for any fixed $\tau > 0$. Lemma 8.13 implies the existence of a curve $\gamma\colon (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ such that $\gamma(t) \in F$ holds for $t \in [0, \varepsilon)$ and $\gamma(0) = x$ as well as $\dot\gamma(0) = d(\tau)$. To show that $d(\tau) \in \mathcal{T}_F(x)$ holds, let $t^{(k)} \in (0, \varepsilon)$ be any sequence with $t^{(k)} \searrow 0$ and $x^{(k)} \coloneqq \gamma(t^{(k)}) \in F$.

Then we have $x^{(k)} \to \gamma(0) = x$ and

$$d(\tau) = \dot\gamma(0) = \lim_{k\to\infty} \frac{\gamma(t^{(k)}) - \gamma(0)}{t^{(k)}} = \lim_{k\to\infty} \frac{x^{(k)} - x}{t^{(k)}},$$

which confirms $d(\tau) \in \mathcal{T}_F(x)$ for any $\tau > 0$. Since $\mathcal{T}_F(x)$ is closed by Lemma 7.3, the limit point $d_0 = \lim\limits_{\tau\searrow 0} d(\tau)$ belongs to $\mathcal{T}_F(x)$ as well. □

Theorem 8.15 (KKT conditions are necessary optimality conditions under the MFCQ)**.**
Suppose that x^* is a local minimizer of (7.1) *which satisfies the MFCQ. Then $\Lambda(x^*) \neq \emptyset$, i. e., there exist Lagrange multipliers μ^* and λ^* (not necessarily unique) such that the KKT conditions* (8.4) *are satisfied. The set of Lagrange multipliers $\Lambda(x^*)$ is compact.*

Proof. The existence of Lagrange multipliers follows directly from the combination of Corollary 8.14 and Theorem 8.7. We do not give a proof of compactness here. □

§ 8.4 First-Order Necessary Optimality Conditions Under LICQ

Definition 8.16 (Linear Independence Constraint Qualification).
*A feasible point $x \in F$ satisfies the **linear independence constraint qualification (LICQ)** if the gradients $\left\{\nabla h_j(x)\right\}_{j=1}^{n_{\text{eq}}} \cup \left\{\nabla g_i(x)\right\}_{i \in \mathcal{A}(x)}$ are linearly independent.*[10]

Notice that the LICQ implies $0 \le n_{\text{eq}} + |\mathcal{A}(x)| \le n$, i. e., there cannot be too many equality plus active inequality constraints.

Lemma 8.17 (LICQ implies MFCQ).
Suppose that x is a feasible point for (7.1) which satisfies the LICQ. Then the MFCQ holds at x.

Proof. The LICQ directly implies condition (i) in Definition 8.10 since any subset of linearly independent vectors is linearly independent as well. Define

$$A := \begin{bmatrix} g_i'(x)|_{i \in \mathcal{A}(x)} \\ h_j'(x)|_{j=1,\ldots,n_{\text{eq}}} \end{bmatrix} \quad \text{and} \quad b := \big(\underbrace{-1, \ldots, -1}_{|\mathcal{A}(x_0)|},\ \underbrace{0, \ldots, 0}_{n_{\text{eq}}}\big)^{\mathsf{T}}.$$

The rows of A are linearly independent, hence A is surjective so that $A\,d = b$ has a solution $d \in \mathbb{R}^n$. This vector d satisfies (8.14) and thus it also satisfies condition (ii) of Definition 8.10. □

Theorem 8.18 (KKT conditions are necessary optimality conditions under the LICQ).
Suppose that x^* is a local minimizer of (7.1) which satisfies the LICQ. Then $\Lambda(x^*)$ is a singleton, i. e., there exist uniquely defined Lagrange multipliers μ^* and λ^* such that the KKT conditions (8.4) are satisfied.

Proof. By Lemma 8.17, the MFCQ holds at x^*, so that the set $\Lambda(x^*)$ of Lagrange multipliers is non-empty (and compact) by Theorem 8.15. It remains to show the uniqueness. Due to the complementarity system (8.4c), the components of μ^* corresponding to inactive constraints must be 0. These components can therefore be removed from the linear system, and (8.5) reduces to

$$\begin{bmatrix} g_i'(x)|_{i \in \mathcal{A}(x)} \\ h_j'(x)|_{j=1,\ldots,n_{\text{eq}}} \end{bmatrix}^{\mathsf{T}} \begin{pmatrix} \mu_i^*|_{i \in \mathcal{A}(x)} \\ \lambda_j^*|_{j=1,\ldots,n_{\text{eq}}} \end{pmatrix} = -\nabla f(x).$$

The LICQ states that the matrix (without the transpose sign) has full row rank and hence the system matrix (including the transpose sign) has full column rank. This implies that μ^* and λ^* must be unique. □

[10] Some authors also speak of a **regular point**; see Bertsekas, 1999, p.349.

We summarize the relations between the various constraint qualifications:

$$\boxed{\text{LICQ}} \xRightarrow{\text{Lemma 8.17}} \boxed{\text{MFCQ}} \xRightarrow{\text{Corollary 8.14}} \boxed{\text{ACQ}} \xRightarrow{\text{Definition 8.6}} \boxed{\text{GCQ}} \tag{8.15}$$

The reverse relations do not hold in general. However, we can note that for problems without inequality constraints, LICQ and MFCQ are identical (and $d = 0$ is an MFCQ vector).

All CQs ensure that the KKT conditions are necessary optimality conditions for local minimizers, i. e., CQ ensure the existence of Lagrange multipliers. Under the MFCQ, the set of multipliers is compact, and the multipliers are unique under the LICQ.

§ 8.5 First-Order Sufficient Optimality Conditions for Convex Problems

The KKT conditions so far were shown to be necessary optimality conditions of first order, provided that some constraint qualification was satisfied. There is one important class of problems where the KKT conditions also serve as sufficient optimality conditions.

$$\left.\begin{aligned} \text{Minimize} \quad & f(x) \qquad \text{where } x \in \mathbb{R}^n \\ \text{subject to} \quad & g_i(x) \le 0 \\ \text{and} \quad & A\,x = b. \end{aligned}\right\} \tag{8.16}$$

For problem (8.16) we assume that

(1) f is of class C^1 and convex on $\mathbb{R}^n$,

(2) g_i is of class C^1 and convex on $\mathbb{R}^n$,

(3) $A \in \mathbb{R}^{n_{\text{eq}} \times n}$ and $b \in \mathbb{R}^{n_{\text{eq}}}$.

Note: Under these conditions, the feasible set F is convex. We refer to this type of problem as a **convex nonlinear optimization problem** or **convex nonlinear program** (**convex NLP**). To be more precise, we can speak of a **convex description of a convex NLP**.

Problems such as (8.16), which seek to minimize a convex objective over a convex feasible set have the remarkable property that every local minimizer is already a global minimizer; see for instance Herzog, 2022, Satz 14.2.

There are customized constraint qualifications which apply only to convex NLPs. They are known as **Slater constraint qualifications** and they can be shown to imply the ACQ, but we do not discuss them here. However, we do point out that the KKT conditions are sufficient optimality conditions for convex NLPs.

Theorem 8.19 (KKT conditions are sufficient optimality conditions for convex problems)**.**
Suppose that x^* is a KKT point for problem (8.16). Then x^* is a global minimizer of (8.16).

Proof. Let $(\mu^*, \lambda^*) \in \Lambda(x^*)$ be any Lagrange multiplier for (8.16) at x^*, and let $x \in \mathbb{R}^n$ be an arbitrary feasible point. We estimate

$$\begin{aligned} f(x) &\ge f(x^*) + f'(x^*)\,(x - x^*) && \text{by Statement (a) of Theorem 2.9} \\ &= f(x^*) - (\mu^*)^\intercal g'(x^*)\,(x - x^*) - (\lambda^*)^\intercal \underbrace{A\,(x - x^*)}_{=b-b=0} && \text{by the KKT condition (8.4a)} \\ &= f(x^*) - \sum_{i=1}^{n_{\text{ineq}}} \mu_i^*\, g_i'(x^*)\,(x - x^*) \\ &\ge f(x^*) - \sum_{i=1}^{n_{\text{ineq}}} \underbrace{\mu_i^*}_{\substack{=0 \text{ on } \mathcal{I}(x^*) \\ \ge 0 \text{ on } \mathcal{A}(x^*)}} \Big(\underbrace{g_i(x)}_{\substack{\le 0 \text{ on } \mathcal{I}(x^*) \\ \le 0 \text{ on } \mathcal{A}(x^*)}} - \underbrace{g_i(x^*)}_{\substack{<0 \text{ on } \mathcal{I}(x^*) \\ =0 \text{ on } \mathcal{A}(x^*)}} \Big) && \text{again by Statement (a) of Theorem 2.9} \\ &\ge f(x^*), \end{aligned}$$

which shows that x^* is indeed a global minimizer. □

Theorem 8.19 implies that convex problems do not require second-order optimality conditions. For non-convex problems, however, second-order conditions bring additional information. Moreover, Theorem 5.33 We address them in the following section.

§ 9 Second-Order Optimality Conditions

We begin by motivating which directions $d \in \mathcal{T}_F^{\text{lin}}(x)$ should play a role in second-order optimality conditions. Suppose that x is a KKT point for (7.1) with multipliers (μ, λ) and $d \in \mathcal{T}_F^{\text{lin}}(x)$. We have

$$\begin{aligned} 0 &= \big[\nabla f(x) + g'(x)^\intercal \mu + h'(x)^\intercal \lambda\big]^\intercal d && \text{by the KKT condition (8.4a)} \\ &= f'(x)\,d + \sum_{i \in \mathcal{A}(x)} \underbrace{\mu_i}_{\ge 0}\, \underbrace{g_i'(x)\,d}_{\le 0} + \sum_{j=1}^{n_{\text{eq}}} \lambda_j\, \underbrace{h_j'(x)\,d}_{=0} && \text{by (8.4c) and the definition of } \mathcal{T}_F^{\text{lin}}(x) \\ &\le f'(x)\,d. \end{aligned}$$

Directions $d \in \mathcal{T}_F^{\text{lin}}(x)$ satisfying $\nabla f(x)^\intercal d > 0$ do not require second-order information in a second-order sufficient condition, nor do they contribute information in a second-order necessary condition. The remaining directions

$$d \in \mathcal{T}_F^{\text{lin}}(x) \quad \text{with} \quad f'(x)\,d = 0 \tag{9.1}$$

are termed **critical directions** since second-order derivatives do matter in those directions.

By our calculation above, the following two conditions are equivalent for $d \in \mathcal{T}_F^{\text{lin}}(x)$.

(1) $f'(x)\,d = 0$.

(2) $g_i'(x)\,d = 0$ for all $i \in \mathcal{A}(x)$ with $\mu_i > 0$.

This motivates the following definition.

Definition 9.1 (strongly and weakly active inequalities, critical cone).
Suppose that x is a KKT point for (7.1) with multipliers (μ, λ).

(*i*) *We further distinguish the indices in the active set as follows:*

$$\begin{aligned}\mathcal{A}_0(x,\mu) &:= \{i \in \mathcal{A}(x) \mid \mu_i = 0\} && \text{the set of weakly active indices at } x,\\ \mathcal{A}_>(x,\mu) &:= \{i \in \mathcal{A}(x) \mid \mu_i > 0\} && \text{the set of strongly active indices at } x.\end{aligned}$$

(*ii*) *The set*

$$\begin{aligned}\mathcal{T}^{\text{critical}}_{\text{NLP}}(x) &:= \left\{d \in \mathcal{T}^{\text{lin}}_F(x) \,\middle|\, f'(x)\,d = 0\right\}\\ &= \left\{d \in \mathcal{T}^{\text{lin}}_F(x) \,\middle|\, g_i'(x)\,d = 0 \quad \text{for all } i \in \mathcal{A}_>(x,\mu)\right\}\\ &= \left\{ d \in \mathbb{R}^n \,\middle|\, \begin{aligned} g_i'(x)\,d &= 0 && \text{for all } i \in \mathcal{A}_>(x,\mu)\\ g_i'(x)\,d &\le 0 && \text{for all } i \in \mathcal{A}_0(x,\mu)\\ h_j'(x)\,d &= 0 && \text{for all } j = 1,\ldots,n_{\text{eq}} \end{aligned} \right\}\end{aligned} \tag{9.2}$$

*is termed the **critical cone** to the problem (7.1) at x. When x is not a KKT point, then we set $\mathcal{T}^{\text{critical}}_{\text{NLP}}(x) := \emptyset$.*

Remark 9.2 (on the critical cone $\mathcal{T}^{\text{critical}}_{\text{NLP}}(x)$).

(*i*) *The critical cone is a closed convex cone. Just like the linearizing cone, it has an algebraic description in terms of the constraint (and objective) gradients.*

(*ii*) *The multipliers (μ, λ) for a KKT point may not be unique. The weakly and strongly active indices at x depend on the inequality constraint multiplier μ and thus they may differ for different multipliers at the same KKT point x. Nevertheless, the critical cone $\mathcal{T}^{\text{critical}}_{\text{NLP}}(x)$ does not depend on the multipliers. (**Quiz 9.1:** Can you explain how this is possible?)*

(*iii*) *Unlike the tangent cone and the linearizing cone, the critical cone does not only depend on the feasible set (or its description in terms of the constraints) but in fact depends also on the objective f. We therefore do not say "critical cone to F" but instead "critical cone to the problem".*

(*iv*) *For any specific choice of the multipliers (μ, λ), $\mathcal{T}^{\text{critical}}_{\text{NLP}}(x)$ can also be viewed as the usual linearizing cone for a modified problem, that is obtained by replacing inequalities which are active at x by equations.*

§ 9.1 Second-Order Necessary Optimality Conditions

We only consider second-order necessary optimality conditions for local minimizers satisfying the LICQ.

Theorem 9.3 (Second-order necessary optimality condition under the LICQ).

Suppose that f, g and h are C^2 functions. Suppose that x^* is a local minimizer of (7.1) which satisfies the LICQ. Let (μ^*, λ^*) be the associated unique Lagrange multipliers; see Theorem 8.18. Then

$$d^\intercal \mathcal{L}_{xx}(x^*, \mu^*, \lambda^*)\, d \ge 0 \quad \text{for all } d \in \mathcal{T}_{\text{NLP}}^{\text{critical}}(x^*). \tag{9.3}$$

Note: Besides second derivatives of the objective, the second-order optimality conditions also involve the second derivatives of all constraints except those where $\mu_i^* = 0$.

Proof. Let d be any fixed element of $\mathcal{T}_{\text{NLP}}^{\text{critical}}(x^*)$, $d \neq 0$. The proof is based on constructing a curve γ which satisfies $\gamma(0) = x^*$ and $\dot\gamma(0) = d$, and along which we are going to run through the Lagrangian. The construction of the curve will be based on Lemma 8.13, which requires the MFCQ to be satisfied. We will therefore define an auxiliary feasible set with slightly tightened constraints compared to (7.1), so that d becomes an MFCQ vector for the auxiliary set. More precisely, we will have to treat some of the active inequality constraints as equality constraints.

To define the modified constraints, we decompose the weakly active indices according to

$$\begin{aligned}
\mathcal{A}_0^<(x^*, \mu^*, d) &:= \{i \in \mathcal{A}_0(x^*, \mu^*) \mid g_i'(x^*)\, d < 0\},\\
\mathcal{A}_0^=(x^*, \mu^*, d) &:= \{i \in \mathcal{A}_0(x^*, \mu^*) \mid g_i'(x^*)\, d = 0\}.
\end{aligned}$$

Clearly, the inequalities g_i with $i \in \mathcal{A}_0^=(x^*, \mu^*, d)$ have to become equalities in order for d to be an MFCQ vector. In addition, we also convert the inequalities which are strongly active at x^* into equalities. Hence we define the modified constraints as

$$\overline{g}(x) := \begin{pmatrix} g_i(x)|_{i \in \mathcal{I}(x^*)} \\ g_i(x)|_{i \in \mathcal{A}_0^<(x^*, \mu^*, d)} \end{pmatrix} \quad \begin{array}{l} \text{the corresponding inactive set at } x^* \text{ is } \overline{\mathcal{I}}(x^*) = \mathcal{I}(x^*) \\ \text{the corresponding active set at } x^* \text{ is } \overline{\mathcal{A}}(x^*) = \mathcal{A}_0^<(x^*, \mu^*, d) \end{array}$$

$$\overline{h}(x) := \begin{pmatrix} h_j(x)|_{j=1,\ldots,n_{\text{eq}}} \\ g_i(x)|_{i \in \mathcal{A}_>(x^*, \mu^*)} \\ g_i(x)|_{i \in \mathcal{A}_0^=(x^*, \mu^*, d)} \end{pmatrix}$$

The new feasible set

$$\overline{F} := \{x \in \mathbb{R}^n \mid \overline{g}(x) \le 0,\ \overline{h}(x) = 0\}$$

satisfies $\overline{F} \subseteq F$ (**Quiz 9.2:** Why?), and $x^* \in \overline{F}$ holds.

We now verify that the chosen vector $d \in \mathcal{T}_{\text{NLP}}^{\text{critical}}(x^*)$ indeed satisfies the MFCQ at x^* for the modified problem. Following Definition 8.10, we show:

(i) The gradients $\{\nabla \overline{h}_j(x^*)\}_j$ are linearly independent.
This is true since we assumed the LICQ. (We could even have included the gradients $\{\nabla g_i(x^*)\}_{i \in \mathcal{A}(x^*)}$ of *all* active inequalities, not only those in $\mathcal{A}_0^<(x^*, \mu^*, d)$.)

(ii) The given vector d is an MFCQ vector.
Indeed, it satisfies

$$\begin{aligned}
\overline{g}_i'(x^*)\, d &< 0 \quad \text{for all } i \in \overline{\mathcal{A}}(x^*),\\
\overline{h}_j'(x^*)\, d &= 0 \quad \text{for all } j.
\end{aligned}$$

> The first statement is true by $\big\{\overline{g}_i'(x^*)\,d\big\}_{i\in\overline{\mathcal{A}}(x^*)} = \big\{g_k'(x^*)\,d\big\}_{k\in\mathcal{A}_0^<(x^*,\mu^*,d)}$ and the definition of $\mathcal{A}_0^<(x^*,\mu^*,d)$. The second statement holds since the constraints collected in $\overline{h}$ are composed of former equality constraints and some of the former active inequality constraints. Since d belongs to the critical cone and thus, in particular, to the linearizing cone $\mathcal{T}_F^{\text{lin}}(x^*)$, all directional derivatives are zero.

Since the MFCQ holds, we can apply Lemma 8.13. We infer that there exists $\varepsilon > 0$ and a curve $\gamma\colon(-\varepsilon,\varepsilon)\to\mathbb{R}^n$ with the following properties:

(i) γ is of class C^2 on $(-\varepsilon,\varepsilon)$.

(ii) $\gamma(t)\in\overline{F}$ for $t\in[0,\varepsilon)$ and even $\overline{g}(\gamma(t))<0$ for $t\in(0,\varepsilon)$.

(iii) $\gamma(0)=x^*$ and $\dot{\gamma}(0)=d$.

The C^2-property of γ is easy to see when we revisit the proof of Lemma 8.13. Since $H(y,t) = h\big(x + t\,d + h'(x)^\intercal y\big)$ is now of class C^2, the implicit function theorem implies that ξ and thus γ are C^2 curves as well.

We finally define the function to run through the Lagrangian along γ as

$$\varphi(t) \coloneqq \mathcal{L}(\gamma(t),\mu^*,\lambda^*).$$

Then φ is also of class C^2 on $(-\varepsilon,\varepsilon)$ and has the following derivatives:

$$\begin{aligned}
\dot{\varphi}(t) &= \dot{\gamma}(t)^\intercal\nabla_x\mathcal{L}(\gamma(t),\mu^*,\lambda^*)\\
\ddot{\varphi}(t) &= \ddot{\gamma}(t)^\intercal\nabla_x\mathcal{L}(\gamma(t),\mu^*,\lambda^*) + \dot{\gamma}(t)^\intercal\mathcal{L}_{xx}(\gamma(t),\mu^*,\lambda^*)\,\dot{\gamma}(t).
\end{aligned}$$

In particular, for $t=0$ we obtain

$$\begin{aligned}
\dot{\varphi}(0) &= \dot{\gamma}(0)^\intercal\nabla_x\mathcal{L}(x^*,\mu^*,\lambda^*) = 0 \quad \text{by the KKT conditions}\\
\ddot{\varphi}(0) &= \ddot{\gamma}(0)^\intercal\nabla_x\mathcal{L}(x^*,\mu^*,\lambda^*) + \dot{\gamma}(0)^\intercal\mathcal{L}_{xx}(x^*,\mu^*,\lambda^*)\,\dot{\gamma}(0)\\
&= 0 + d^\intercal\mathcal{L}_{xx}(x^*,\mu^*,\lambda^*)\,d.
\end{aligned}$$

Furthermore, for $t\in[0,\varepsilon)$, we have

$$\begin{aligned}
\varphi(t) = \mathcal{L}(\gamma(t),\mu^*,\lambda^*) &= f(\gamma(t)) + \sum_{i=1}^{n_{\text{ineq}}}\mu_i^*\,g_i(\gamma(t)) + \sum_{j=1}^{n_{\text{eq}}}\lambda_j^*\,h_j(\gamma(t))\\
&= f(\gamma(t)) + \sum_{i\in\mathcal{A}_>(x^*)}\overbrace{\mu_i^*}^{>0}\;\overbrace{g_i(\gamma(t))}^{=0,\text{ since }\gamma(t)\in\overline{F}} + \sum_{i\in\mathcal{A}_0(x^*)}\overbrace{\mu_i^*}^{=0}\,g_i(\gamma(t)) + \sum_{i\in\mathcal{I}(x^*)}\overbrace{\mu_i^*}^{=0}\,g_i(\gamma(t)) + \sum_{j=1}^{n_{\text{eq}}}\overbrace{\lambda_j^*}^{?}\;\overbrace{h_j(\gamma(t))}^{=0,\text{ since }\gamma(t)\in\overline{F}}\\
&= f(\gamma(t)).
\end{aligned}$$

By assumption, $\gamma(0)=x^*$ is a (global) minimizer of f, restricted to a neighborhood $U(x^*)$ and intersected with F. If necessary, we make $\varepsilon>0$ smaller so that $\gamma(t)\in U(x^*)\cap\overline{F}\subseteq U(x^*)\cap F$ holds for all $t\in[0,\varepsilon)$. It follows that $t=0$ is a (global) minimizer of the C^2-function φ, restricted to $[0,\varepsilon)$.

We already know $\dot{\varphi}(0) = 0$. If we had $\ddot{\varphi}(0) < 0$, then by continuity, $\ddot{\varphi}$ would be negative on an entire interval $[0, \varepsilon')$ with $0 < \varepsilon' \le \varepsilon$. Therefore, $\dot{\varphi}$ would be negative on $(0, \varepsilon')$. This would be a contradiction to $t = 0$ being a minimizer of φ on $[0, \varepsilon)$. Therefore, we must have

$$\ddot{\varphi}(0) = d^\mathsf{T} \mathcal{L}_{xx}(x^*, \mu^*, \lambda^*)\, d \ge 0. \qquad \square$$

Remark 9.4 (on Theorem 9.3).

(*i*) *Theorem 9.3 states that, at a local minimizer that satisfies the LICQ, the Hessian of the Lagrangian is necessarily positive semidefinite on the critical cone. Since $\mathcal{T}_{\text{NLP}}^{\text{critical}}(x^*)$ has a simple algebraic description, this condition can be verified numerically. (**Quiz 9.3:** How could this be done?)*

(*ii*) *Since the critical cone $\mathcal{T}_{\text{NLP}}^{\text{critical}}(x^*)$ is in general not a subspace, eigenvalues of $\mathcal{L}_{xx}(x^*, \mu^*, \lambda^*)$ do not provide a precise representation of (9.3) in general.*

(*iii*) *The critical cone is a subspace in case all inequality constraints (if any) are either inactive or strongly active. This situation is called **strict complementarity**.[11] In this case, (9.3) is equivalent to the positive semidefiniteness of the Hessian $\mathcal{L}_{xx}(x^*, \mu^*, \lambda^*)$ projected to that subspace. (**Quiz 9.4:** Is there another case when (9.3) is characterized by eigenvalues?)*

(*iv*) *In the absence of any constraints, Theorem 9.3 reduces to the second-order necessary optimality condition of Theorem 3.2. (**Quiz 9.5:** Can you provide the details?)*

§ 9.2 Second-Order Sufficient Optimality Conditions

We would like to obtain a second-order sufficient optimality condition which are as close a possible to the second-order necessary condition from Theorem 9.3.

Theorem 9.5 (**Second-order sufficient optimality condition**).
Suppose that f, g and h are C^2 functions. Suppose that x^* is a KKT point for problem (7.1).[12] Moreover, assume that there exists $\alpha > 0$ such that[13]

$$d^\mathsf{T} \mathcal{L}_{xx}(x^*, \lambda^*, \mu^*)\, d \ge \alpha\, \|d\|^2 \quad \text{holds for all } d \in \mathcal{T}_{\text{NLP}}^{\text{critical}}(x^*). \tag{9.4}$$

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

$$f(x) \ge f(x^*) + \frac{\beta}{2} \|x - x^*\|^2 \quad \text{for all } x \in U(x^*) \cap F. \tag{9.5}$$

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

[11] In other words, for all components i, either $\mu_i^* = 0$ or $g_i(x^*) = 0$ holds but not both.
[12] We do not require a constraint qualification to be satisfied at x^*, but we still need Lagrange multipliers to exist.
[13] That is, the Hessian of the Lagrangian is positive definite on the critical cone $\mathcal{T}_{\text{NLP}}^{\text{critical}}(x^*)$.

Proof. We assume the contrary, i. e., there exists $\beta \in (0, \alpha)$ and a sequence $x^{(k)} \subseteq F$ with the properties $x^{(k)} \to x^*$, $x^{(k)} \neq x^*$ and $f(x^{(k)}) < f(x^*) + \frac{\beta}{2}\|x^{(k)} - x^*\|^2$. From this, we are going to construct a direction $d^* \in \mathcal{T}_{\text{NLP}}^{\text{critical}}(x^*)$ which violates (9.4).

To this end we define

$$d^{(k)} := \frac{x^{(k)} - x^*}{\|x^{(k)} - x^*\|}.$$

Due to $\|d^{(k)}\| = 1$, we can find a convergent subsequence

$$d^{(k^{(\ell)})} \to d^* \text{ as } \ell \to \infty \quad \text{with some } d^* \in \mathbb{R}^n \text{ and } \|d^*\| = 1.$$

Step (1) We begin by confirming $d^* \in \mathcal{T}_F^{\text{lin}}(x^*)$.
By construction, we have $d^* \in \mathcal{T}_F(x^*)$, and by Lemma 7.7, we have $\mathcal{T}_F(x^*) \subseteq \mathcal{T}_F^{\text{lin}}(x^*)$.

Step (2) We show $f'(x^*)\, d^* \le 0$ and $d^* \in \mathcal{T}_{\text{NLP}}^{\text{critical}}(x^*)$.
We can estimate

$$\begin{aligned} f(x^*) + \frac{\beta}{2}\|x^{(k^{(\ell)})} - x^*\|^2 &> f(x^{(k^{(\ell)})}) \quad \text{by assumption} \\ &= f(x^*) + f'(x^* + \xi^{(k^{(\ell)})}\,(x^{(k^{(\ell)})} - x^*))\,(x^{(k^{(\ell)})} - x^*) \end{aligned}$$

by the mean value theorem 2.4. This further implies

$$\frac{\beta}{2}\|x^{(k^{(\ell)})} - x^*\|^2 > f'(x^* + \xi^{(k^{(\ell)})}\,(x^{(k^{(\ell)})} - x^*))\,(x^{(k^{(\ell)})} - x^*).$$

We divide by $\|x^{(k^{(\ell)})} - x^*\|$:

$$\frac{\beta}{2}\|x^{(k^{(\ell)})} - x^*\| > f'(x^* + \xi^{(k^{(\ell)})}\,(x^{(k^{(\ell)})} - x^*))\,\frac{x^{(k^{(\ell)})} - x^*}{\|x^{(k^{(\ell)})} - x^*\|}.$$

Passing to the limit yields

$$0 \ge f'(x^*)\, d^*. \tag{9.6}$$

Moreover, the KKT conditions imply $-\nabla f(x^*) \in \mathcal{T}_F^{\text{lin}}(x^*)^\circ$, hence $f'(x^*)\, d \ge 0$ for all $d \in \mathcal{T}_F^{\text{lin}}(x^*)$ and in particular for d^* by the result of **Step** (1). Together with (9.6), we obtain $f'(x^*)\, d = 0$. By definition (9.2), this means $d^* \in \mathcal{T}_{\text{NLP}}^{\text{critical}}(x^*)$.

Step (3) We show that d^* violates (9.4).
We estimate

$$\begin{aligned} & f(x^*) + \frac{\beta}{2}\|x^{(k)} - x^*\|^2 \\ &\quad > f(x^{(k)}) \quad \text{by assumption} \\ &\quad \ge f(x^{(k)}) + \sum_{i=1}^{n_{\text{ineq}}} \underbrace{\mu_i^*}_{\ge 0}\, \underbrace{g_i(x^{(k)})}_{\le 0} + \sum_{j=1}^{n_{\text{eq}}} \lambda_j^*\, \underbrace{h_j(x^{(k)})}_{=0} \\ &\quad = \mathcal{L}(x^{(k)}, \mu^*, \lambda^*) \end{aligned}$$

By Taylor's theorem 2.4, we find $\eta^{(k)} \in (0,1)$ such that

$$= \underbrace{\mathcal{L}(x^*, \mu^*, \lambda^*)}_{=f(x^*)} + \underbrace{\nabla_x \mathcal{L}(x^*, \mu^*, \lambda^*)^\mathsf{T}}_{=\,0 \text{ by the KKT conditions}} (x^{(k)} - x^*)$$
$$+ \frac{1}{2}(x^{(k)} - x^*)^\mathsf{T} \mathcal{L}_{xx}\big(x^* + \eta^{(k)}\,(x^{(k)} - x^*), \mu^*, \lambda^*\big)(x^{(k)} - x^*).$$

The division by $\|x^{(k)} - x^*\|^2$ and passage to the limit on the subsequence $k^{(\ell)}$ yields

$$(d^*)^\mathsf{T} \mathcal{L}_{xx}(x^*, \mu^*, \lambda^*)\, d^* \le \beta\, \|d^*\|^2 < \alpha\, \|d^*\|^2 \quad \text{since } \beta \in (0, \alpha).$$

We have thus reached a contradiction to (9.4). □

Note: In the absence of any constraints, Theorem 9.5 reduces to the second-order sufficient optimality condition of Theorem 3.3.

End of Week 9

Chapter 3 Numerical Techniques for Constrained Optimization Problems

§ 10 Introduction

We will discuss in this chapter numerical algorithms for the solution of nonlinear programs (7.1). Similarly as in Chapter 1, where we considered algorithms that generally went after stationary points, we will now aim to find KKT points and associated Lagrange multipliers.

The class of methods we discuss are akin to those of § 5, which build a sequence of quadratic models of the objective. In the constrained case, these models are obtained by formulating a quadratic model of the Lagrangian and by linearizing the constraints about the current iterate $(x^{(k)}, \mu^{(k)}, \lambda^{(k)})$. This leads to the following type of subproblems:

$$\begin{aligned} \text{Minimize} \quad & \underbrace{\mathcal{L}(x^{(k)}, \mu^{(k)}, \lambda^{(k)})}_{\text{constant}} + \mathcal{L}_x(x^{(k)}, \mu^{(k)}, \lambda^{(k)})\, d + \frac{1}{2}\, d^\intercal H^{(k)} d, \quad \text{where } d \in \mathbb{R}^n \\ \text{subject to} \quad & g(x^{(k)}) + g'(x^{(k)})\, d \le 0 \\ \text{and} \quad & h(x^{(k)}) + h'(x^{(k)})\, d = 0. \end{aligned} \tag{10.1}$$

The solution $d^{(k)}$ of (10.1) (provided it exists and is ideally unique) then serves to update the iterate according to

$$x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)},$$

where the step size $\alpha^{(k)}$ is determined by a line search in an effort to globalize the method. Moreover, the Lagrange multipliers need to be updated as well.

Since (10.1) is a **quadratic program** (**QP**), the ensuing class of methods is known as **sequential quadratic programming method** (**SQP method**). The symmetric Hessian $H^{(k)}$ of the subproblem's objective can be taken to be equal to $\mathcal{L}_{xx}(x^{(k)}, \mu^{(k)}, \lambda^{(k)})$, but in view of the effort generally expected in evaluating this matrix (or its matrix-vector products), it is more customary in practice to resort to a substitute as we did with quasi-Newton methods in § 5.5. It is worthwhile noticing that SQP methods are **infeasible methods**, i. e., they allow the iterates $x^{(k)}$ to violate the constraints and only obtain feasibility in the limit.

An important property for QPs such as (10.1) is the following result. It states that a QP already has a solution when it is neither infeasible nor unbounded.[1] The result was originally stated for convex QPs

[1] The special case of LPs was considered in Herzog, 2022, Satz 6.9.

in Frank, Wolfe, 1956. A direct proof which includes the non-convex case was given in Blum, Oettli, 1972.

Lemma 10.1 (Frank-Wolfe lemma). *A QP has a global minimizer if and only if the feasible set is nonempty and the objective is bounded below on the feasible set.*

§ 11 Local SQP as Local Newton's Method

We saw in § 5.4 that the sequential minimization of second-order Taylor models of the objective in unconstrained problems agrees with Newton's method applied to the first-order optimality condition; see around (5.24). A similar observation holds here. To make this more precise, we first consider in § 11.1 a problem (7.1) with no inequality constraints before returning to the general case in § 11.2.

§ 11.1 Equality Constrained Problems

Consider a problem (7.1) with no inequality constraints. Consequently, we also omit the Lagrange multipliers μ. At a point $(\overline{x}, \overline{\lambda})$, the QP (10.1) with exact second-order Taylor polynomial reads

$$\begin{aligned} \text{Minimize} \quad & \mathcal{L}(\overline{x},\overline{\lambda}) + \mathcal{L}_x(\overline{x},\overline{\lambda})\, d + \frac{1}{2}\, d^\intercal \mathcal{L}_{xx}(\overline{x},\overline{\lambda})\, d, \quad \text{where } d \in \mathbb{R}^n \\ \text{subject to} \quad & h(\overline{x}) + h'(\overline{x})\, d = 0. \end{aligned} \tag{11.1}$$

Since the constraints in any QP are linear, the Abadie CQ holds everywhere, and the associated KKT conditions become first-order necessary optimality conditions for the QP. In case of (11.1), the KKT conditions read

$$\begin{aligned} \nabla_x \mathcal{L}(\overline{x},\overline{\lambda}) + \mathcal{L}_{xx}(\overline{x},\overline{\lambda})\, d + h'(\overline{x})^\intercal \lambda &= 0, \\ h(\overline{x}) + h'(\overline{x})\, d &= 0, \end{aligned}$$

or written more compactly,

$$\begin{bmatrix} \mathcal{L}_{xx}(\overline{x},\overline{\lambda}) & h'(\overline{x})^\intercal \\ h'(\overline{x}) & 0 \end{bmatrix} \begin{pmatrix} d \\ \lambda \end{pmatrix} = - \begin{pmatrix} \nabla_x \mathcal{L}(\overline{x},\overline{\lambda}) \\ h(\overline{x}) \end{pmatrix}. \tag{11.2}$$

An equality constrained QP has a very similar characterization of solvability as an unconstrained quadratic problem. First we notice that every local minimizer is already a global minimizer.

Lemma 11.1. *Suppose that d^* is a local minimizer of the equality constrained QP (11.1). Then d^* is already a global minimizer of (11.1).*

Note: Simple examples show that this statement is not true for the case of QPs with inequality constraints!

Proof. Let d be any feasible point for (11.1) in the neighborhood of optimality of d^*. Let us denote the objective in (11.1) by $q(d)$. We obtain

$$\begin{aligned}
0 &\le q(d) - q(d^*) \\
&= \underbrace{\mathcal{L}(\overline{x},\overline{\lambda})}_{\text{cancel}} + \mathcal{L}_x(\overline{x},\overline{\lambda})\,d + \frac{1}{2}\,d^\top \mathcal{L}_{xx}(\overline{x},\overline{\lambda})\,d - \Big[\underbrace{\mathcal{L}(\overline{x},\overline{\lambda})}_{\text{cancel}} + \mathcal{L}_x(\overline{x},\overline{\lambda})\,d^* + \frac{1}{2}\,d^{*\top} \mathcal{L}_{xx}(\overline{x},\overline{\lambda})\,d^*\Big] \\
&= \mathcal{L}_x(\overline{x},\overline{\lambda})\,d + \frac{1}{2}\,d^\top \mathcal{L}_{xx}(\overline{x},\overline{\lambda})\,d + (\lambda^*)^\top \underbrace{\big(h(\overline{x}) + h'(\overline{x})\,d\big)}_{=0} \\
&\quad - \Big[\mathcal{L}_x(\overline{x},\overline{\lambda})\,d^* + \frac{1}{2}\,d^{*\top} \mathcal{L}_{xx}(\overline{x},\overline{\lambda})\,d^* + (\lambda^*)^\top \underbrace{\big(h(\overline{x}) + h'(\overline{x})\,d^*\big)}_{=0}\Big] \\
&= \mathcal{L}_x(\overline{x},\overline{\lambda})\,d + \frac{1}{2}\,d^\top \mathcal{L}_{xx}(\overline{x},\overline{\lambda})\,d + (\lambda^*)^\top h'(\overline{x})\,d \\
&\quad - \Big[\mathcal{L}_x(\overline{x},\overline{\lambda})\,d^* + \frac{1}{2}\,d^{*\top} \mathcal{L}_{xx}(\overline{x},\overline{\lambda})\,d^* + (\lambda^*)^\top h'(\overline{x})\,d^*\Big].
\end{aligned}$$

Since (d^*, λ^*) satisfies the KKT system (11.2), we can reduce the term inside the bracket:

$$= \mathcal{L}_x(\overline{x},\overline{\lambda})\,d + \frac{1}{2}\,d^\top \mathcal{L}_{xx}(\overline{x},\overline{\lambda})\,d + (\lambda^*)^\top h'(\overline{x})\,d + \Big[\frac{1}{2}\,d^{*\top} \mathcal{L}_{xx}(\overline{x},\overline{\lambda})\,d^*\Big].$$

We complete the square:

$$= \mathcal{L}_x(\overline{x},\overline{\lambda})\,d + (\lambda^*)^\top h'(\overline{x})\,d + \frac{1}{2}\,(d - d^*)^\top \mathcal{L}_{xx}(\overline{x},\overline{\lambda})\,(d - d^*) + d^{*\top} \mathcal{L}_{xx}(\overline{x},\overline{\lambda})\,d.$$

Using the KKT system (11.2) for (d^*, λ^*) again, we can simplify this to

$$0 \le q(d) - q(d^*) = \frac{1}{2}\,(d - d^*)^\top \mathcal{L}_{xx}(\overline{x},\overline{\lambda})\,(d - d^*). \tag{$*$}$$

By varying d in the feasible neighborhood of d^*, $d - d^*$ ranges over $\ker h'(\overline{x})$ intersected with a neighborhood of 0. Since the right-hand term in ($*$) is invariant to scaling, i. e., it continues to hold when $d - d^*$ is multiplied by a scalar, we indeed get ($*$) for all d such that $d - d^* \in \ker h'(\overline{x})$. Due to the linearity of the constraints, this means that ($*$) indeed holds for all d which are feasible for (11.1). Consequently, d^* is a global minimizer. □

Without additional assumptions, the equality constrained QP (11.1) may be infeasible or unbounded. To discuss its solvability, it is going to be helpful to consider the **reduced QP**. To this end, suppose that $h'(\overline{x})$ has rank r, where $0 \le r \le \min\{n, n_{\text{eq}}\}$.[2] The rank of a matrix coincides with the dimension of its range. By the dimension formula $n = \dim\ker h'(\overline{x}) + \dim\operatorname{range} h'(\overline{x})$, we must have $\dim\ker h'(\overline{x}) = n - r$, which has values in $\{\max\{n - n_{\text{eq}}, 0\}, \ldots, n\}$.

Let us denote by $Z \in \mathbb{R}^{n \times (n-r)}$ a matrix whose columns constitute a basis of $\ker h'(\overline{x})$.[3] We can thus express any vector d satisfying the linear system $h(\overline{x}) + h'(\overline{x})\,d = 0$ in the form $d = d_{\text{hom}} + d_{\text{part}}$, where d_{part} is any fixed ("particular") solution of $h(\overline{x}) + h'(\overline{x})\,d = 0$ and $d_{\text{hom}} \in \ker h'(\overline{x})$. Using the basis Z,

[2] When the LICQ holds, then $r = n_{\text{eq}}$, but we do not require this in general.

[3] In case of $r = n$, we have a matrix of zero width, which does not cause a problem.

we can thus express any vector d solving the linear system in the form $d = Z y + d_{\text{part}}$ with unique coefficient vector $y \in \mathbb{R}^{n-r}$. Plugging this representation into (11.1), we can express it equivalently as the **reduced QP**

$$\text{Minimize} \quad \mathcal{L}(\overline{x}, \overline{\lambda}) + \mathcal{L}_x(\overline{x}, \overline{\lambda}) \, (Z y + d_{\text{part}}) + \frac{1}{2} \, (Z y + d_{\text{part}})^{\mathsf{T}} \mathcal{L}_{xx}(\overline{x}, \overline{\lambda}) \, (Z y + d_{\text{part}})$$

Dropping all constant terms in the objective, we obtain the equivalent form

$$\text{Minimize} \quad \left[\mathcal{L}_x(\overline{x}, \overline{\lambda}) + d_{\text{part}}^{\mathsf{T}} \mathcal{L}_{xx}(\overline{x}, \overline{\lambda})\right] Z y + \frac{1}{2} \, y^{\mathsf{T}} Z^{\mathsf{T}} \mathcal{L}_{xx}(\overline{x}, \overline{\lambda}) \, Z y, \quad \text{where } y \in \mathbb{R}^{n-r}. \tag{11.3}$$

We refer to the matrix $Z^{\mathsf{T}} \mathcal{L}_{xx}(\overline{x}, \overline{\lambda}) \, Z$ as the **reduced Hessian**. The first-order necessary optimality conditions for (11.3) are

$$Z^{\mathsf{T}} \mathcal{L}_{xx}(\overline{x}, \overline{\lambda}) \, Z y = -Z^{\mathsf{T}} \left[\nabla_x \mathcal{L}(\overline{x}, \overline{\lambda}) + \mathcal{L}_{xx}(\overline{x}, \overline{\lambda}) \, d_{\text{part}}\right]. \tag{11.4}$$

We are now in a position to discuss the solvability of (11.1).

Lemma 11.2 (Solvability and global solutions of the equality constrained QP (11.1); compare Lemma 4.1)**.**

(i) *Suppose that the linear system $h(\overline{x}) + h'(\overline{x}) \, d = 0$ is solvable, and that d_{part} is some particular solution. Suppose, moreover, that the reduced Hessian $Z^{\mathsf{T}} \mathcal{L}_{xx}(\overline{x}, \overline{\lambda}) \, Z$ is positive semidefinite. Then the objective in the reduced QP (11.3) is convex. In this case, the following are equivalent:*

- *(a) The QP (11.1) possesses at least one (global) minimizer.*
- *(b) The QP (11.1) is neither unbounded nor infeasible.*
- *(c) The KKT conditions (11.2) are solvable.*
- *(d) The reduced QP (11.3) possesses at least one (global) minimizer.*
- *(e) The reduced QP (11.3) is not unbounded.*
- *(f) The first-order optimality condition (11.4) is solvable.*

The global minimizers of (11.1) are precisely the KKT points, i. e., the d-components of solutions (d, λ) to the KKT system (11.2).

(ii) *Suppose that the linear system $h(\overline{x}) + h'(\overline{x}) \, d = 0$ is solvable, and that d_{part} is some particular solution. Suppose now that the reduced Hessian $Z^{\mathsf{T}} \mathcal{L}_{xx}(\overline{x}, \overline{\lambda}) \, Z$ is not positive semidefinite. Then the QP (11.1) and the reduced QP (11.3) are unbounded.*

(iii) *Suppose that the linear system $h(\overline{x}) + h'(\overline{x}) \, d = 0$ is not solvable. Then the QP (11.1) is infeasible and the reduced QP cannot be formulated for lack of a particular solution d_{part}.*

Proof. The proof is part of homework problem 10.1. □

Note: In the situation of Statement (ii), the KKT system may still be solvable, but clearly a KKT point is not a (local or global) minimizer. (**Quiz 11.1:** Can you explain this?)

As a consequence of Lemma 11.2, we can characterize the unique solvability of equality constrained QPs now.

Corollary 11.3 (Unique solvability of the equality constrained QP (11.1))**.** *The following are equivalent:*

(i) *The QP* (11.1) *possesses a unique* (*global*) *minimizer.*

(ii) *The reduced QP* (11.3) *possesses a unique* (*global*) *minimizer.*

(iii) *The linear system* $h(\overline{x}) + h'(\overline{x})\, d = 0$ *is solvable, and the reduced Hessian* $Z^\mathsf{T} \mathcal{L}_{xx}(\overline{x}, \overline{\lambda})\, Z$ *is positive definite.*

If the above hold, let x^* *denote the unique global minimizer of* (11.1)*. Then the following are equivalent:*

(i) $h'(\overline{x})$ *has full row rank, i. e., the LICQ is satisfied at* x^**.*

(ii) $\Lambda(x^*)$ *is a singleton, i. e., there is a unique multiplier* λ^* *corresponding to* x^**.*

Let us now see how, in the absence of inequalities, the local SQP method is equivalent to local Newton's method applied to the KKT conditions. To be precise, by the local SQP method for an equality constrained NLP (7.1) we mean the following iteration. Given an iterate $(x^{(k)}, \lambda^{(k)})$, consider the QP

$$\begin{aligned} &\text{Minimize} \quad \underbrace{\mathcal{L}(x^{(k)}, \lambda^{(k)})}_{\text{constant}} + \mathcal{L}_x(x^{(k)}, \lambda^{(k)})\, d + \frac{1}{2} d^\mathsf{T} \mathcal{L}_{xx}(x^{(k)}, \lambda^{(k)})\, d, \quad \text{where } d \in \mathbb{R}^n \\ &\text{subject to} \quad h(x^{(k)}) + h'(x^{(k)})\, d = 0. \end{aligned} \tag{11.5}$$

under the assumption that $\mathcal{L}_{xx}(x^{(k)}, \lambda^{(k)})$ is positive definite on $\ker h'(x^{(k)})$, so that (11.5) is uniquely solvable. Suppose moreover, that $h'(x^{(k)})$ has full row rank (i. e., the LICQ holds), then the unique solution of (11.5) is characterized by the KKT conditions

$$\begin{bmatrix} \mathcal{L}_{xx}(x^{(k)}, \lambda^{(k)}) & h'(x^{(k)})^\mathsf{T} \\ h'(x^{(k)}) & 0 \end{bmatrix} \begin{pmatrix} d \\ \lambda \end{pmatrix} = - \begin{pmatrix} \nabla_x \mathcal{L}(x^{(k)}, \lambda^{(k)}) \\ h(x^{(k)}) \end{pmatrix} \tag{11.6}$$

with unique Lagrange multiplier. We then set

$$\begin{pmatrix} x^{(k+1)} \\ \lambda^{(k+1)} \end{pmatrix} \coloneqq \begin{pmatrix} x^{(k)} \\ \lambda^{(k)} \end{pmatrix} + \begin{pmatrix} d \\ \lambda \end{pmatrix}.$$

On the other hand, the KKT conditions for an equality constrained NLP (7.1) read

$$\begin{aligned} \nabla_x \mathcal{L}(x, \lambda) &= 0, \\ h(x) &= 0. \end{aligned}$$

This is a (nonlinear) system of equations for the unknowns (x, λ). A step of local Newton's method at an iterate $(x^{(k)}, \lambda^{(k)})$ is given by

$$\begin{bmatrix} \mathcal{L}_{xx}(x^{(k)}, \lambda^{(k)}) & h'(x^{(k)})^\intercal \\ h'(x^{(k)}) & 0 \end{bmatrix} \begin{pmatrix} d \\ \lambda \end{pmatrix} = - \begin{pmatrix} \nabla_x \mathcal{L}(x^{(k)}, \lambda^{(k)}) \\ h(x^{(k)}) \end{pmatrix} \tag{11.6}$$

followed by

$$\begin{pmatrix} x^{(k+1)} \\ \lambda^{(k+1)} \end{pmatrix} \coloneqq \begin{pmatrix} x^{(k)} \\ \lambda^{(k)} \end{pmatrix} + \begin{pmatrix} d \\ \lambda \end{pmatrix}.$$

This is exactly the same iteration as above. Therefore, the SQP method is sometimes also called the **Lagrange Newton method**.

In practice, the QP (11.5) is usually formulated equivalently as

$$\begin{aligned} &\text{Minimize} \quad \underbrace{f(x^{(k)})}_{\text{constant}} + f'(x^{(k)})\, d + \frac{1}{2} d^\intercal \mathcal{L}_{xx}(x^{(k)}, \lambda^{(k)})\, d, \quad \text{where } d \in \mathbb{R}^n \\ &\text{subject to} \quad h(x^{(k)}) + h'(x^{(k)})\, d = 0 \end{aligned} \tag{11.7}$$

whose KKT conditions are

$$\begin{bmatrix} \mathcal{L}_{xx}(x^{(k)}, \lambda^{(k)}) & h'(x^{(k)})^\intercal \\ h'(x^{(k)}) & 0 \end{bmatrix} \begin{pmatrix} d \\ \lambda^{(k+1)} \end{pmatrix} = - \begin{pmatrix} \nabla_x f(x^{(k)}) \\ h(x^{(k)}) \end{pmatrix} \tag{11.8}$$

Note: When we use the Lagrange multiplier associated with (11.7) as the next iterate (as indicated in (11.8)), then the iteration is again the same as above. (**Quiz 11.2:** Can you show this?)

§ 11.2 Problems with General Constraints

We now address the relation of the local SQP method and Newton's method for general NLPs (7.1). The formulation of the sequence of QPs (compare (10.1))

$$\begin{aligned} &\text{Minimize} \quad \underbrace{\mathcal{L}(x^{(k)}, \mu^{(k)}, \lambda^{(k)})}_{\text{constant}} + \mathcal{L}_x(x^{(k)}, \mu^{(k)}, \lambda^{(k)})\, d + \frac{1}{2} d^\intercal \mathcal{L}_{xx}(x^{(k)}, \mu^{(k)}, \lambda^{(k)})\, d, \quad \text{where } d \in \mathbb{R}^n \\ &\text{subject to} \quad g(x^{(k)}) + g'(x^{(k)})\, d \le 0 \\ &\qquad \text{and} \quad h(x^{(k)}) + h'(x^{(k)})\, d = 0, \end{aligned} \tag{11.9}$$

the associated KKT system

$$\nabla_x \mathcal{L}(x^{(k)}, \mu^{(k)}, \lambda^{(k)}) + \mathcal{L}_{xx}(x^{(k)}, \mu^{(k)}, \lambda^{(k)})\, d + g'(x^{(k)})^\intercal \mu + h'(x^{(k)})^\intercal \lambda = 0, \tag{11.10a}$$

$$\mu \ge 0, \quad g(x^{(k)}) + g'(x^{(k)})\, d \le 0, \quad \mu^\intercal \big(g(x^{(k)}) + g'(x^{(k)})\, d\big) = 0, \tag{11.10b}$$

$$h(x^{(k)}) + h'(x^{(k)})\, d = 0, \tag{11.10c}$$

and the update

$$\begin{pmatrix} x^{(k+1)} \\ \mu^{(k+1)} \\ \lambda^{(k+1)} \end{pmatrix} := \begin{pmatrix} x^{(k)} \\ \mu^{(k)} \\ \lambda^{(k)} \end{pmatrix} + \begin{pmatrix} d \\ \mu \\ \lambda \end{pmatrix}$$

are straightforward.

The characterization of solvability and unique solvability of QPs with inequality constraints, however, is more intricate than in Lemma 11.2. We do not attempt such a characterization here. Moreover, a QP with inequality constraints may have local minimizers which are not global minimizers.

Finally, it is not obvious how to interpret this scheme as a Newton method for the KKT conditions (8.4) of (7.1), i. e.,

$$\begin{aligned} \nabla_x \mathcal{L}(x, \mu, \lambda) &= 0, && \text{(11.11a)} \\ \mu \ge 0, \quad g(x) \le 0, \quad \mu^\intercal g(x) &= 0, && \text{(11.11b)} \\ h(x) &= 0. && \text{(11.11c)} \end{aligned}$$

To this end, we rewrite the complementarity condition (11.11b) in the equivalent form

$$\mu \in K \quad \text{and} \quad g(x)^\intercal (\nu - \mu) \le 0 \quad \text{for all } \nu \in K \tag{11.12}$$

with the closed convex cone $K := \mathbb{R}^{n_{\text{ineq}}}_{\ge 0}$ (the non-negative orthant).

Lemma 11.4 (Complementarity is equivalent to variational inequality).
(11.11b) *and* (11.12) *are equivalent.*

Proof. The proof is part of homework problem 10.3. □

Definition 11.5 (Normal cone). *Suppose that $M \subseteq \mathbb{R}^n$ is an arbitrary set and $x \in M$. The set*

$$\mathcal{N}_M(x) := \{s \in \mathbb{R}^n \mid s^\intercal (y - x) \le 0 \textit{ for all } y \in M\} \tag{11.13}$$

is termed the ***normal cone*** *of M at x. A vector $s \in \mathcal{N}_M(x)$ is termed a* ***normal direction*** *of M at x. We set $\mathcal{N}_M(x) := \emptyset$ for $x \notin M$.*

Lemma 11.6 (Properties of the normal cone). *Suppose that $M \subseteq \mathbb{R}^n$ is a set and $x \in M$. Then the following holds.*

(*i*) *The normal cone is a closed convex cone.*

(*ii*) *$\mathcal{N}_M(x) = (M - \{x\})^\circ$ holds.*

Proof. The proof is part of homework problem 10.4. □

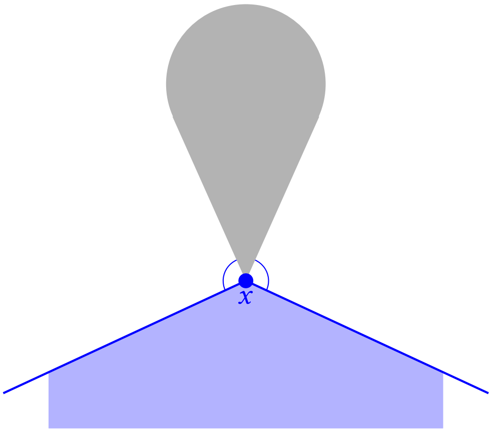


Figure 11.1: Normal cone $\mathcal{N}_M(x)$ (blue) of a (convex) set M at a point x. We are showing the shifted cone $\{x\} + \mathcal{N}_M(x)$.

Using the notion of normal cone, we can express the variational inequality (11.12) and thus the complementarity condition (11.11b) in the equivalent form

$$g(x) \in \mathcal{N}_K(\mu). \tag{11.14}$$

Finally, we see that the entire KKT system (11.11) takes the form

$$0 \in \underbrace{\begin{pmatrix} +\nabla_x \mathcal{L}(x,\mu,\lambda) \\ -g(x) \\ -h(x) \end{pmatrix}}_{=:F(x,\mu,\lambda)} + \underbrace{\mathcal{N}_{\mathbb{R}^n \times K \times \mathbb{R}^{n_{\text{eq}}}}(x,\mu,\lambda)}_{=:N(x,\mu,\lambda)}. \tag{11.15}$$

Due to the Cartesian structure, the set $\mathcal{N}_{\mathbb{R}^n \times K \times \mathbb{R}^{n_{\text{eq}}}}(x,\mu,\lambda)$ is simply

$$\mathcal{N}_{\mathbb{R}^n \times K \times \mathbb{R}^{n_{\text{eq}}}}(x,\mu,\lambda) = \left\{ \begin{pmatrix} 0 \\ s \\ 0 \end{pmatrix} \;\middle|\; s \in \mathcal{N}_K(\mu) \right\}$$

Equation (11.15) is known as a **generalized equation**, for which an extension of the classical Newton method exists.

§ 11.3 Local Convergence of a Generalized Newton Method

In this section, we discuss the local convergence of a generalized form of Newton's method, applied to the **generalized equation**

$$0 \in F(z) + N(z), \tag{11.16}$$

where $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is a function of class C^1 and $N\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a set-valued map.[4] Apparently Newton's method for generalized equations, referred to as **generalized Newton method** or **Josephy-Newton method**, was introduced in the Ph.D. dissertation Josephy, 1979b, supervised by Stephen

[4] In other words, N is a function mapping $\mathbb{R}^n$ into $\mathcal{P}(\mathbb{R}^n)$, the power set of $\mathbb{R}^n$.

M. Robinson. At an iterate $z^{(k)}$, the (local) generalized Newton method produces its next iterate as a solution of

$$0 \in F(z^{(k)}) + F'(z^{(k)})(z^{(k+1)} - z^{(k)}) + N(z^{(k+1)}). \tag{11.17}$$

This iteration coincides with the usual Newton method in case of $N(z) = \{0\}$. The local convergence analysis of the classical Newton method (Theorem 5.27) is based on the assumption of invertibility of the Jacobian $F'(z^*)$ at a root z^* of F. This condition is not quite the correct one for the generalized Newton method. We notice, however, that the invertibility condition for the classical Newton method is equivalent to the map

$$\Delta \mapsto \text{unique solution } z \text{ of } \Delta = F(z^*) + F'(z^*)\,(z - z^*)$$

being affine and thus Lipschitz continuous (with Lipschitz constant equal to the operator norm of the inverse of $F'(z^*)$). This Lipschitz continuity turns out to be the condition applicable to generalized equations as well.

Note: Here and throughout the remainder of § 11.3 we will frequently need (generalized) equations with parameters, which we highlight in this color. By contrast, unknowns will be highlighted in this color. Also notice that we carry out our analysis w.r.t. a general inner product M. The results, however, are qualitatively independent of the specific choice of M.

Definition 11.7 (Strong regularity, introduced in Robinson, 1980)**.** *Suppose that z^* is a solution of the generalized equation*

$$0 \in F(z) + N(z). \tag{11.16}$$

That generalized equation is said to be ***strongly regular*** *at z^* if there exist open balls $B_\delta^{M^{-1}}(0)$ and $B_\varepsilon^M(z^*)$ such that*

(i) for every $\Delta \in B_\delta^{M^{-1}}(0)$, the linearized and perturbed generalized equation

$$\Delta \in F(z^*) + F'(z^*)\,(w - z^*) + N(w) \tag{11.18}$$

has a solution $w(\Delta) \in B_\varepsilon^M(z^*)$,

(ii) there is no other solution in $B_\varepsilon^M(z^*)$,

(iii) the map $B_\delta^{M^{-1}}(0) \ni \Delta \mapsto w(\Delta) \in B_\varepsilon^M(z^*)$ is Lipschitz continuous, i. e.,

$$\|w(\Delta) - w(\Delta')\|_M \le L\,\|\Delta - \Delta'\|_{M^{-1}} \tag{11.19}$$

holds for all $\Delta, \Delta' \in B_\delta^{M^{-1}}(0)$ with some Lipschitz constant $L \ge 0$.

Just like the invertibility of F' extends from a point into an entire neighborhood (Lemma 5.25), strong regularity at a point z^* implies that it holds in an entire neighborhood $U(z^*)$. Although the Lipschitz constant may grow, we can allow the ball $B_\delta^{M^{-1}}(0)$ containing the perturbation Δ to be uniform.

Theorem 11.8 (Strong regularity extends to a neighborhood)**.** *Suppose that z^* is a solution of the generalized equation (11.16) and that (11.16) is strongly regular at z^* with Lipschitz constant L in (11.19). Then there exist open balls $B_\rho^M(z^*)$, $B_R^{M^{-1}}(0)$ and $B_r^M(z^*)$ such that the following holds:*

(*i*) *for every* $z \in B_\rho^M(z^*)$ *and every* $\Delta \in B_R^{M^{-1}}(0)$, *the linearized and perturbed generalized equation*

$$\Delta \in F(z) + F'(z)\,(w - z) + N(w) \tag{11.20}$$

has a solution $w(\Delta) \in B_r^M(z^*)$,

(*ii*) *there is no other solution in* $B_r^M(z^*)$,

(*iii*) *the map* $B_R^{M^{-1}}(0) \ni \Delta \mapsto w(\Delta) \in B_r^M(z^*)$ *— which now depends on* z *— is Lipschitz continuous, i. e.,*

$$\|w(\Delta) - w(\Delta')\|_M \le L(z)\,\|\Delta - \Delta'\|_{M^{-1}} \tag{11.21}$$

holds for all $\Delta, \Delta' \in B_R^{M^{-1}}(0)$, *with Lipschitz constant*[5]

$$L(z) = \frac{L}{1 - L\,\|F'(z) - F'(z^*)\|_{M^{-1}\leftarrow M}}.$$

Proof. The result and its proof are given in Robinson, 1980, Theorem 2.4 and Josephy, 1979a, Corollary 1. □

We can now prove a local convergence result for the generalized Newton method.

Theorem 11.9 (Local convergence of the generalized Newton method; compare Josephy, 1979a; Izmailov, Kurennoy, Solodov, 2012 and Theorem 5.27)**.** *Suppose that* $F\colon \mathbb{R}^n \to \mathbb{R}^n$ *is a* C^1 *function. Suppose, moreover, that* z^* *is a solution of the generalized equation* (11.16) *and that* (11.16) *is strongly regular at* z^*. *Then there exists a neighborhood* $B_\varepsilon^M(z^*)$ *such that*

(*i*) z^* *is the unique solution of* $0 \in F(\cdot) + N(\cdot)$ *in* $B_\varepsilon^M(z^*)$.

(*ii*) *For any initial guess* $z^{(0)} \in B_\varepsilon^M(z^*)$, *the generalized Newton method is well-defined, and it generates a sequence* $z^{(k)}$ *which converges to* z^*.

(*iii*) $\big(z^{(k)}\big)$ *converges to* z^* *Q-superlinearly w.r.t. the* M*-norm.*

(*iv*) *If* F' *is Lipschitz continuous in* $B_\varepsilon^M(z^*)$, *then this convergence is even Q-quadratic.*

Proof. Let L denote the Lipschitz constant associated with the strong regularity of (11.16) at z^*; see (11.19). Owing to the continuity of F', we can find, for any $\varepsilon_1 > 0$, a radius $r_1(\varepsilon_1) > 0$ such that

$$\|F'(z) - F'(z^*)\|_{M^{-1}\leftarrow M} \le \varepsilon_1 \quad \text{holds for all } z \in B^M_{r_1(\varepsilon_1)}(z^*). \tag{$*$}$$

[5] Compare this to the statement

$$\|F'(x)^{-1}\|_{M\leftarrow M^{-1}} \le \frac{\|F'(x^*)^{-1}\|_{M\leftarrow M^{-1}}}{1 - \|\mathrm{Id} - F'(x^*)^{-1}F'(x)\|_{M\leftarrow M}} \le \frac{\|F'(x^*)^{-1}\|_{M\leftarrow M^{-1}}}{1 - \|F'(x^*)^{-1}_{M\leftarrow M^{-1}}\|\|F'(x^*) - F'(x)\|_{M\leftarrow M}}$$

in the proof of Lemma 5.25.

Moreover, due to the continuity of F' we can show (**Quiz 11.3:** Can you do it?) that, for any $\varepsilon_2 > 0$, there exists a radius $r_2(\varepsilon_2) > 0$ such that

$$\|F(z) + F'(z)(z^* - z) - F(z^*)\|_{M^{-1}} \le \varepsilon_2 \, \|z - z^*\|_M \quad \text{holds for all } z \in B^M_{r_2(\varepsilon_2)}(z^*). \tag{$**$}$$

Let R, r and ρ denote the radii from Theorem 11.8. We require that the initial guess $z^{(0)}$ belongs to $B^M_\varepsilon(z^*)$ with $\varepsilon := \min\{\rho,\ r,\ r_1(\frac{1}{2L}),\ r_2(\frac{1}{4L}),\ 4\,L\,R\}$.

Step (1) Local uniqueness of z^*:
Since we can write $0 \in F(z^*) + N(z^*)$ also as

$$0 \in F(z^*) + F'(z^*)(z^* - z^*) + N(z^*),$$

condition (ii) of Theorem 11.8 shows that z^* is the only solution in $B^M_r(z^*)$. This proves Statement (i).

Step (2) Well-posedness of the Newton iteration (11.17):
We proceed by induction. Since $z^{(0)} \in B^M_\rho(z^*)$ holds, the initial Newton step

$$0 \in F(z^{(0)}) + F'(z^{(0)})(z^{(1)} - z^{(0)}) + N(z^{(1)}) \tag{$***$}$$

has a unique solution $z^{(1)} \in B^M_r(z^*)$ by Theorem 11.8. We need to estimate $\|z^{(1)} - z^*\|_M$. To this end, we regard z^* as the unique solution in $B^M_r(z^*)$ of a perturbed version of ($***$), namely

$$\Delta^{(0)} \in F(z^{(0)}) + F'(z^{(0)})(z^* - z^{(0)}) + N(z^*). \tag{$****$}$$

with $\Delta^{(0)} := F(z^{(0)}) + F'(z^{(0)})(z^* - z^{(0)}) - F(z^*)$. The norm of $\Delta^{(0)}$ can be estimated by ($**$) as

$$\|\Delta^{(0)}\|_{M^{-1}} \le \frac{1}{4L} \, \|z^{(0)} - z^*\|_M < \frac{\varepsilon}{4L} \le R.$$

We can therefore invoke Theorem 11.8 to obtain

$$\begin{aligned}
\|z^{(1)} - z^*\|_M &= \|w(0) - w(\Delta^{(0)})\|_M \\
&\le \frac{L}{1 - L\,\|F'(z^{(0)}) - F'(z^*)\|_{M^{-1}\leftarrow M}} \, \|0 - \Delta^{(0)}\|_{M^{-1}} \quad \text{by (11.21)} \\
&\le \frac{L}{1 - L\frac{1}{2L}} \, \frac{1}{4L} \, \|z^{(0)} - z^*\|_M \\
&= \frac{1}{2}\|z^{(0)} - z^*\|_M.
\end{aligned}$$

The induction step merely requires a repetition of this argument so we do not expand on it. We only note that

$$\|z^{(k+1)} - z^*\|_M \le \frac{1}{2} \, \|\Delta^{(k)}\|_{M^{-1}} \tag{$*****$}$$

holds for all $k \in \mathbb{N}_0$. This concludes Statement (ii), and in fact we have shown the Q-linear convergence of $z^{(k)} \to z^*$.

Step (3) We now verify that the sequence indeed converges Q-superlinearly. To this end, we infer from ($**$) that, for any $\varepsilon_2 > 0$, we have

$$\|\Delta^{(k)}\|_{M^{-1}} = \|F(z^{(k)}) + F'(z^{(k)})(z^* - z^{(k)}) - F(z^*)\|_{M^{-1}} \le \varepsilon_2 \, \|z^{(k)} - z^*\|_M$$

for $k \in \mathbb{N}_0$ sufficiently large. Plugging this into ($*****$) proves Statement (*iii*).

Step (4) In case F' is Lipschitz in $B^M_\varepsilon(z^*)$ with Lipschitz constant K, then we can obtain

$$\begin{aligned}\|\Delta^{(k)}\|_{M^{-1}} &= \|F(z^{(k)}) + F'(z^{(k)})(z^* - z^{(k)}) - F(z^*)\|_{M^{-1}} \\ &= \left\| \int_0^1 \left[F'(z^{(k)}) - F'(z^* + s\,(z^{(k)} - z^*))\right] \mathrm{d}s\,(z^* - z^{(k)}) \right\|_{M^{-1}} \\ &\le \int_0^1 \left\|F'(z^{(k)}) - F'(z^* + s\,(z^{(k)} - z^*))\right\|_{M^{-1} \leftarrow M} \|z^* - z^{(k)}\|_M \\ &\le \int_0^1 K\, s \,\mathrm{d}s \, \|z^* - z^{(k)}\|_M^2 \\ &= \frac{1}{2} K \, \|z^{(k)} - z^*\|_M^2.\end{aligned}$$

From there, the Q-quadratic convergence follows using ($*****$), and Statement (*iv*) is proved. □

End of Week 10

§ 11.4 Local Convergence of the SQP Method

In this section we establish the local fast convergence of the SQP method under appropriate assumptions. We begin by confirming that the Josephy-Newton method (11.17) applied to the KKT system in the form of the generalized equation

$$0 \in \underbrace{\begin{pmatrix} \nabla_x \mathcal{L}(x,\mu,\lambda) \\ -g(x) \\ -h(x) \end{pmatrix}}_{=:F(x,\mu,\lambda)} + \underbrace{\mathcal{N}_{\mathbb{R}^n \times K \times \mathbb{R}^{n_{\text{eq}}}}(x,\mu,\lambda)}_{=:N(x,\mu,\lambda)} \tag{11.15}$$

is identical to the local SQP method described in (11.10). One step of the Josephy-Newton method (11.17) applied to (11.15) at iterate $z^{(k)} := (x^{(k)}, \mu^{(k)}, \lambda^{(k)})$ reads

$$\begin{aligned}\nabla_x \mathcal{L}(x^{(k)}, \mu^{(k)}, \lambda^{(k)}) + \mathcal{L}_{xx}(x^{(k)}, \mu^{(k)}, \lambda^{(k)})\, d + g'(x^{(k)})^\intercal \mu + h'(x^{(k)})^\intercal \lambda = 0,& \\ g(x^{(k)}) + g'(x^{(k)})\, d \in \mathcal{N}_K(\mu) = \mathcal{N}_{\mathbb{R}^{n_{\text{ineq}}}_{\ge 0}}(\mu),& \\ h(x^{(k)}) + h'(x^{(k)})\, d = 0.&\end{aligned}$$

By Lemma 11.4, we can rewrite the inclusion as

$$\mu \ge 0, \quad g(x^{(k)}) + g'(x^{(k)})\, d \le 0, \quad \mu^\intercal \big(g(x^{(k)}) + g'(x^{(k)})\, d\big) = 0.$$

A comparison with (11.10) shows that, indeed, the local SQP method is the Josephy-Newton method applied to (11.15).

The natural next questions are:

(1) What does the strong regularity of the KKT system (11.15) mean when spelled out?

(2) Under which conditions is the KKT system (11.15) strongly regular at a given KKT triplet (x^*, μ^*, λ^*)?

To address question (1), we consider the linearized and perturbed generalized equation $\Delta \in F(z^*) + F'(z^*)\,(z - z^*) + N(z)$ specifically for the case (11.15). We obtain

$$\begin{pmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \end{pmatrix} \in \begin{pmatrix} \nabla_x \mathcal{L}(x^*, \mu^*, \lambda^*) \\ -g(x^*) \\ -h(x^*) \end{pmatrix} + \begin{bmatrix} \mathcal{L}_{xx}(x^*, \mu^*, \lambda^*) & g'(x^*)^\top & h'(x^*)^\top \\ -g'(x^*) & 0 & 0 \\ -h'(x^*) & 0 & 0 \end{bmatrix} \begin{pmatrix} x - x^* \\ \mu - \mu^* \\ \lambda - \lambda^* \end{pmatrix} \begin{matrix} + \{0\} \\ + \mathcal{N}_K(\mu) \\ + \{0\}. \end{matrix} \tag{11.22}$$

This can be verified to be the KKT conditions for the QP

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2}(x - x^*)^\top \mathcal{L}_{xx}(x^*, \mu^*, \lambda^*)\,(x - x^*) + [f'(x^*) - \Delta_1^\top](x - x^*), \quad \text{where } x \in \mathbb{R}^n \\ \text{subject to} \quad & g(x^*) + g'(x^*)\,(x - x^*) + \Delta_2 \le 0 \\ \text{and} \quad & h(x^*) + h'(x^*)\,(x - x^*) + \Delta_3 = 0, \end{aligned}$$

and (μ, λ) are the Lagrange multipliers. If we wanted to study the strong regularity of the KKT system (11.15) "by hand", we would need to show — under appropriate assumptions — that the linearized and perturbed KKT conditions (11.22) have a unique solution near (x^*, μ^*, λ^*) when $(\Delta_1, \Delta_2, \Delta_3)$ is sufficiently small. It is, however, more convenient to resort to the answer of question (2), which is given in the following theorem.

Theorem 11.10 (Strong regularity of the KKT conditions, see Dontchev, Rockafellar, 1996, Theorem 5)**.** *Suppose that f, g and h are C^2 functions. Moreover, suppose that x^* is a KKT point for problem (7.1) with associated Lagrange multipliers μ^* and λ^*. Then the following are equivalent:*

(*i*) *The KKT system (11.15) is strongly regular at (x^*, μ^*, λ^*).*

(*ii*) *The LICQ holds at x^*, and for each partition of the weakly active set $\mathcal{A}_0(x^*, \mu^*)$ into $\mathcal{A}_0^+(x^*, \mu^*)$, $\mathcal{A}_0^0(x^*, \mu^*)$ and $\mathcal{A}_0^-(x^*, \mu^*)$, and the associated cone*[6]

$$\overline{K} := \left\{ d \in \mathbb{R}^n \;\middle|\; \begin{aligned} & g_i'(x)\,d = 0 && \text{for all } i \in \mathcal{A}_>(x^*, \mu^*) \cup \mathcal{A}_0^+(x^*, \mu^*) \\ & g_i'(x)\,d \le 0 && \text{for all } i \in \mathcal{A}_0^0(x^*, \mu^*) \\ & h_j'(x)\,d = 0 && \text{for all } j = 1, \ldots, n_{\text{eq}} \end{aligned} \right\},$$

we have

$$\left[\mathcal{L}_{xx}(x^*, \mu^*, \lambda^*)\,\overline{K}\right] \cap \overline{K}^\circ = \{0\}.$$

[6] Note that $\overline{K}$ is similar to the critical cone (9.2). However, the weakly active constraints in $\mathcal{A}_0^+(x^*, \mu^*)$ are treated as strongly active ones, those in $\mathcal{A}_0^0(x^*, \mu^*)$ continue to be treated as weakly active, while those in $\mathcal{A}_0^-(x^*, \mu^*)$ are treated as inactive.

Admittedly, this characterization of strong regularity of a KKT system is not particularly handy, not even in the presence of strict complementarity (when the set $\mathcal{A}_0(x^*, \mu^*)$ of weakly active constraints is empty).

The characterization would become more tractable if some sort of second-order sufficient optimality condition could be used. The following example due to Robinson, 1980, Section 4, however, shows that the standard second-order sufficient optimality condition (Theorem 9.5) is not sufficient to establish the strong regularity.

Example 11.11 (The standard second-order sufficient optimality condition is not sufficient for strong regularity)**.**

$$\begin{aligned} \textit{Minimize} \quad & \frac{1}{2}(x_1^2 - x_2^2) - \Delta_1 x_1, \quad \textit{where } x \in \mathbb{R}^2 \\ \textit{subject to} \quad & -x_1 + 2\,x_2 \le 0 \\ \textit{and} \quad & -x_1 - 2\,x_2 \le 0. \end{aligned}$$

Here $\Delta_1 \in \mathbb{R}$ represents a perturbation.[7] *The KKT conditions associated with this QP read*

$$\begin{pmatrix} \Delta_1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \in \begin{bmatrix} 1 & 0 & -1 & -1 \\ 0 & -1 & 2 & -2 \\ 1 & -2 & 0 & 0 \\ 1 & 2 & 0 & 0 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \mu_1 \\ \mu_2 \end{pmatrix} + \mathcal{N}_{\mathbb{R}^2 \times \mathbb{R}^2_{\ge 0}}(x^*, \mu^*).$$

(The problem is chosen deliberately as a QP so that the linearization of the KKT conditions does not change them.) The unperturbed problem ($\Delta_1 = 0$) has a unique global minimizer at $x^* = (0, 0)^\intercal$, which is also the only local minimizer. The LICQ holds at any point, and the unique Lagrange multipliers at x^* are given by $\mu^* = (0, 0)^\intercal$. Therefore, according to (9.2), the critical cone is $\mathcal{T}_{\text{NLP}}^{\text{critical}}(x^*) = \mathcal{T}_F^{\text{lin}}(x^*) = \mathbb{R}_{\le 0} \times \{0\}$. The second-order sufficient optimality condition (Theorem 9.5) holds since we have

$$d^\intercal \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} d > 0 \quad \textit{for all } d \in \mathcal{T}_{\text{NLP}}^{\text{critical}}(x^*) = \mathbb{R}_{\le 0} \times \{0\}, \; d \ne 0.$$

Let us now consider the perturbed problem with arbitrary $\Delta_1 > 0$. Then the problem under consideration has three KKT points:

(i) $x^* = (\Delta_1, 0)^\intercal$ *with unique multiplier* $\mu^* = (0, 0)^\intercal$.
This is a saddle point of the problem. Both constraints are inactive (so the problem is locally unconstrained) and the Hessian has eigenvalues of either sign.

(ii) $x^* = \frac{2}{3}(2\,\Delta_1, \Delta_1)^\intercal$ *with unique multiplier* $\mu^* = \frac{1}{3}(1, 0)^\intercal$.
*This is a local minimizer since the second-order sufficient optimality condition holds. (**Quiz 11.4:** Can you check this?)*

(iii) $x^* = \frac{2}{3}(2\,\Delta_1, -\Delta_1)^\intercal$ *with unique multiplier* $\mu^* = \frac{1}{3}(0, 1)^\intercal$.
This is another local minimizer for the same reason.

[7] We do not need to consider the most general perturbations $\Delta_1 \in \mathbb{R}^2$ and $\Delta_2 \in \mathbb{R}^2$ here to show that strong regularity fails to hold.

Taking $\Delta_1 > 0$ arbitrarily close to 0, the existence of these three solutions of the KKT conditions imply that the KKT system does not have a solution which is unique in a neighborhood of $x^* = (0,0)^\intercal$. Therefore, the KKT system cannot be strongly regular at (x^*, μ^*).

As a remedy to the situation, Robinson, 1980 considered the **strong second-order sufficient condition**.

Definition 11.12 (Strong second-order sufficient condition)**.** *Suppose that f, g and h are C^2 functions. Suppose that x^* is a KKT point for problem (7.1) with associated Lagrange multipliers μ^* and λ^*. Define the subspace*

$$\mathcal{T}_F^{\text{strong}}(x^*, \mu^*) := \left\{ d \in \mathbb{R}^n \;\middle|\; \begin{array}{ll} g_i'(x)\, d = 0 & \text{for all } i \in \mathcal{A}_>(x^*, \mu^*) \\ h_j'(x^*)\, d = 0 & \text{for all } j = 1, \ldots, n_{\text{eq}} \end{array} \right\}. \tag{11.23}$$

We say that the ***strong second-order sufficient optimality condition*** *holds at x^*, provided that there exists $\alpha > 0$ such that*

$$d^\intercal \mathcal{L}_{xx}(x^*, \mu^*, \lambda^*)\, d \ge \alpha \, \|d\|^2 \quad \text{holds for all } d \in \mathcal{T}_F^{\text{strong}}(x^*, \mu^*). \tag{11.24}$$

Remark 11.13 (on the strong second-order sufficient optimality condition)**.**

(*i*) *Since $\mathcal{T}_F^{\text{strong}}(x^*, \mu^*)$ is a subspace, the verification of (11.24) can be performed by verifying the positivity of the smallest eigenvalue of the Hessian $\mathcal{L}_{xx}(x^*, \mu^*, \lambda^*)$, reduced to this subspace.*

(*ii*) *The subspace $\mathcal{T}_F^{\text{strong}}(x^*, \mu^*)$ is larger than the critical cone $\mathcal{T}_{\text{NLP}}^{\text{critical}}(x^*)$, see (9.2). Therefore, the strong second-order sufficient optimality condition implies the (standard) second-order sufficient optimality condition (9.4).*

(*iii*) *The subspace $\mathcal{T}_F^{\text{strong}}(x^*, \mu^*)$ is not necessarily contained in the linearizing cone $\mathcal{T}_F^{\text{lin}}(x^*)$.*

(*iv*) *In the absence of weakly active constraints, the strong and the standard second-order sufficient conditions agree.*

We have the following result.

Theorem 11.14 (Strong regularity of the KKT conditions under strong second-order optimality conditions and LICQ, see Robinson, 1980, Theorem 4.1)**.** *Suppose that f, g and h are C^2 functions. Moreover, suppose that x^* is a KKT point for problem (7.1) with associated Lagrange multipliers μ^* and λ^*. If the LICQ and the strong second-order sufficient optimality condition holds at x^*, then the KKT system (11.15) is strongly regular at (x^*, μ^*, λ^*).*

Moreover, it can be shown under the assumptions of the previous theorem, that the solution $x(\Delta)$ of the linearized and perturbed KKT conditions (11.22) are indeed local minimizers of the QP (11.22) and that the LICQ continues to hold for this QP, so that the multipliers (μ, λ) are unique. This addition is relatively easy to show since the LICQ and the strong second-order sufficient optimality condition are stable in a neighborhood of (x^*, μ^*, λ^*); see also Dontchev, Rockafellar, 1996, Theorem 6.

Corollary 11.15 (Local convergence of the local SQP method). *Suppose that the assumptions of Theorem 11.14 hold. Consider the local SQP method described in the beginning of § 11.2. More precisely, suppose that the iterates* $(x^{(k)}, \mu^{(k)}, \lambda^{(k)})$ *are defined by choosing the solution to the linearized KKT conditions* (11.10) *with* d *having minimal norm* $\|d\|_M$*. Then there exists a neighborhood* $B_\varepsilon(x^*, \mu^*, \lambda^*)$ *(w.r.t. the* M*-norm in* $\mathbb{R}^n$ *and the Euclidean norms in* $\mathbb{R}^{n_{\text{eq}}}$ *and* $\mathbb{R}^{n_{\text{ineq}}}$*) such that the following holds:*

(*i*) *The local SQP method is well-defined in the sense that the solution* d *having minimal norm* $\|d\|_M$ *is unique. Moreover, the LICQ holds for all QPs, so that the Lagrange multipliers* (μ, λ) *in* (11.10) *associated with* d *are unique.*

(*ii*) *In every iteration, the unknown* d *is in fact a local minimizer of the QP* (11.9).

(*iii*) *The convergence* $(x^{(k)}, \mu^{(k)}, \lambda^{(k)}) \to (x^*, \mu^*, \lambda^*)$ *is Q-superlinear.*

(*iv*) *If the second derivatives of* f*,* g_i *and* h_j *are Lipschitz continuous in a neighborhood of* x^**, then the convergence is Q-quadratic.*

§ 12 Reformulation using Slack Variables

In the remainder of Chapter 3, we will be discussing practical aspects of the solution of NLPs (7.1) by SQP methods.

As we saw in the previous sections, the presence of inequality constraints renders the theory of NLPs, but also numerical algorithms, more difficult. It is therefore customary to convert the inequality constraints into simpler bound constraints using **slack variables** $s \in \mathbb{R}^{n_{\text{ineq}}}$. Using slacks, we can reformulate (7.1) equivalently as

$$\left.\begin{aligned} \text{Minimize} \quad & f(x), && \text{where } x \in \mathbb{R}^n, s \in \mathbb{R}^{n_{\text{ineq}}} \\ \text{subject to} \quad & g(x) + s = 0 \\ \text{and} \quad & h(x) = 0 \\ \text{as well as} \quad & s \geq 0. \end{aligned}\right\} \tag{12.1}$$

The only inequality constraints are now simple sign constraints on some of the optimization variables, which are somewhat easier to handle algorithmically than general inequality constraints. It is important to note that this conversion is usually carried out internally by the solution algorithm and need not be done by the user.

It should be clear that there is a bijective relation between the feasible points x of (7.1) and the feasible points (x, s) of (12.1). Moreover, one can show that the slack reformulation (12.1) does not alter the satisfaction of the GCQ, ACQ, MFCQ, or LICQ. More precisely, the GCQ for (7.1) at x implies the GCQ for (12.1) at (x, s). Moreover, the ACQ, MFCQ, LICQ hold for (7.1) at x if and only if the respective condition holds for (12.1) at (x, s). See homework problem 9.3.

Owing to the reformulation (12.1), we can assume from now on that problem (7.1) is of the simplified form

$$\left.\begin{aligned} \text{Minimize} \quad & f(x), \qquad \text{where } x \in \mathbb{R}^n \\ \text{subject to} \quad & h(x) = 0 \\ \text{and} \quad & x \geq \ell. \end{aligned}\right\} \tag{12.2}$$

The lower bound ℓ has values in $\big[\mathbb{R} \cup \{-\infty\}\big]^n$, where $-\infty$ means that the bound is effectively not present for the respective component. The Lagrangian for (12.2) is

$$\mathcal{L}(x, \mu, \lambda) \coloneqq f(x) + \mu^\top(\ell - x) + \lambda^\top h(x)$$

and the QP associated with (12.2), in the form following (11.7), reads

$$\begin{aligned} \text{Minimize} \quad & f'(x^{(k)})\, d + \frac{1}{2} d^\top H^{(k)} d, \quad \text{where } d \in \mathbb{R}^n \\ \text{subject to} \quad & h(x^{(k)}) + h'(x^{(k)})\, d = 0 \\ \text{and} \quad & x^{(k)} + d \geq \ell. \end{aligned} \tag{12.3}$$

§ 13 Numerical Solution of QPs

In this section we discuss numerical approaches to solve equality and inequality constrained QPs. Due to the possibility to use slack variables, we can assume inequality constrained QPs to be of the form (13.6).

§ 13.1 Solution of Equality Constrained QPs

We begin with QPs with equality constraints only. These are of the form

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2} d^\top A\, d - b^\top d \\ \text{subject to} \quad & B\, d = c \end{aligned} \tag{13.1}$$

where $A \in \mathbb{R}^{n \times n}$ is symmetric, $B \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}^m$.[8] We already know from Lemma 11.1 that all minimizers of (13.1) (if any exist) are global. The KKT conditions associated with problem (13.1) are given by the linear system

$$\begin{bmatrix} A & B^\top \\ B & 0 \end{bmatrix} \begin{pmatrix} d \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}. \tag{13.2}$$

Assumption 13.1. *For simplicity we assume that B has full row rank, i. e., the LICQ holds.*

Lemma 11.2 suggests the following strategy for a possible algorithm to solve (13.1).

[8] The roles are $A = \mathcal{L}_{xx}(x^{(k)}, \mu^{(k)}, \lambda^{(k)})$, $B = h'(x^{(k)})$, $b = -\nabla_x \mathcal{L}(x^{(k)}, \mu^{(k)}, \lambda^{(k)})$ or $b = -\nabla f(x^{(k)})$, see (11.7), and $c = -h(x^{(k)})$.

(1) Determine a particular solution d_{part} of $B\,d = c$.[9]

(2) Determine any matrix $Z \in \mathbb{R}^{n\times(n-r)}$ whose columns form a basis of $\ker B$. Then consider the reduced problem

$$\text{Minimize} \quad \left[-b + d_{\text{part}}^\mathsf{T} A\right] Z\,y + \frac{1}{2}\,y^\mathsf{T} Z^\mathsf{T} A\,Z\,y, \quad \text{where } y \in \mathbb{R}^{n-r}, \tag{13.3}$$

whose first-order necessary optimality conditions are

$$Z^\mathsf{T} A\,Z\,y = Z^\mathsf{T}\left[b^\mathsf{T} - A\,d_{\text{part}}\right]. \tag{13.4}$$

Provided that A is positive definite on the kernel of B, i. e., the reduced Hessian $Z^\mathsf{T} A\,Z$ is s. p. d., a unique minimizer exists for (13.3). In this case, the unique minimizer of (13.1) is $d = d_{\text{part}} + Z\,y$, which is also the unique d-component of any solution (d, λ) to (13.2).

(3) Recover the unique Lagrange multiplier λ as any solution of $B^\mathsf{T}\lambda = b - A\,d$.

We recall from § 4 the conjugate gradient Algorithm 4.17 as our preferred solver for s. p. d. systems such as (13.4). Using the CG method to solve (13.3) has several advantages:

(i) We can take advantage of the possibility to solve the QP only inexactly by using a relative stopping criterion.

(ii) We can equip the CG method with a detector for search directions of non-positive curvature, which previously led to the truncated conjugate gradient Algorithm 5.41.

Interestingly, Gould, Hribar, Nocedal, 2001 discovered that the CG algorithm (w.r.t. the M-inner product) for the reduced problem (13.4) can actually be formulated in the “full space” and without making explicit reference to a null space basis matrix Z. The main idea is to apply the CG method directly to (13.2) and to employ a preconditioner of the form

$$\begin{bmatrix} M & B^\mathsf{T} \\ B & 0 \end{bmatrix}, \tag{13.5}$$

where M is still a user-defined inner product on $\mathbb{R}^n$. This is known as a **constraint preconditioner**.[10] Although neither the system matrix in (13.2) nor the preconditioner are positive definite, one can show that the CG method is still well-defined, as it is equivalent to the CG method for the reduced problem. One refers to the CG algorithm applied to a symmetric saddle-point system (13.2) as the **subspace CG method** or **projected CG method**. In fact, the solution of a linear system governed by the matrix (13.5) can be viewed as an orthogonal projection problem w.r.t. the norm induced by M onto $\ker B$. More precisely, since M is s. p. d.,

$$\begin{bmatrix} M & B^\mathsf{T} \\ B & 0 \end{bmatrix} \begin{pmatrix} p \\ \lambda \end{pmatrix} = -\begin{pmatrix} \zeta \\ 0 \end{pmatrix}$$

[9] This is always possible under our assumption of full row rank, i. e., surjectivity of B. In general, when B is not surjective, then the constraints $B\,d = c$ may be impossible to satisfy. In that case, we would declare the problem infeasible and stop.

[10] The name derives from the fact that the preconditioner exactly replicates the constraint block B and the block B^T for the associated Lagrange multiplier.

holds if and only if p solves the problem

$$\begin{aligned} &\text{Minimize} && \frac{1}{2}\|p - (-M^{-1}\zeta)\|_M^2 \\ &\text{subject to} && B\,p = 0. \end{aligned}$$

We refer to the unique solution as $p = \operatorname{proj}^M_{\ker B}(-M^{-1}\zeta)$, the M-orthogonal projection of $-M^{-1}\zeta$ to $\ker B$.

For completeness, we provide the projected CG algorithm, equipped with a detector of non-positive curvature, i. e., failure of positive definiteness of A on $\ker B$.

Algorithm 13.2 (Truncated projected conjugate gradient method for symmetric systems (13.2) w.r.t. the constraint preconditioner defined by the M-inner product; compare Algorithm 5.41)**.**

```
Input: right-hand side $b \in \mathbb{R}^n$
Input: particular solution $d_{\text{part}}$ of $B\,d_{\text{part}} = c$
Input: symmetric matrix $A$ (or matrix-vector products with $A$)
Input: s. p. d. matrix $M$ and matrix $B \in \mathbb{R}^{m \times n}$ of full row rank (or matrix-vector products with $\left[\begin{smallmatrix} M & B^\intercal \\ B & 0 \end{smallmatrix}\right]^{-1}$)
Input: relative residual $\varepsilon_{\text{rel}}$
Output: approximate solution of (13.2)
Set $\ell := 0$
Set $d^{(0)} := d_{\text{part}}$                         // partial solution as initial guess
Set $\zeta^{(0)} := A\,d^{(0)} - b$                      // evaluate the initial residual
Solve
    $\begin{bmatrix} M & B^\intercal \\ B & 0 \end{bmatrix} \begin{pmatrix} p^{(0)} \\ \cdot \end{pmatrix} = - \begin{pmatrix} \zeta^{(0)} \\ 0 \end{pmatrix}$
                  // steepest descent direction, projected to $\ker B$, w.r.t. $M$-inner product
Set $\delta^{(0)} := -(\zeta^{(0)})^\intercal p^{(0)}$    // $\delta^{(0)} = \|\operatorname{proj}^M_{\ker B}(M^{-1}\zeta^{(0)})\|_M^2$
while $\delta^{(\ell)} \ge \varepsilon_{\text{rel}}^2 \delta^{(0)}$ do    // check stopping criterion
    Set $q^{(\ell)} := A\,p^{(\ell)}$
    Set $\theta^{(\ell)} := (q^{(\ell)})^\intercal p^{(\ell)}$
    if $\theta^{(\ell)} > 0$ then
        Set $\alpha^{(\ell)} := \delta^{(\ell)} / \theta^{(\ell)}$
        Set $d^{(\ell+1)} := d^{(\ell)} + \alpha^{(\ell)} p^{(\ell)}$
        Set $\zeta^{(\ell+1)} := \zeta^{(\ell)} + \alpha^{(\ell)} q^{(\ell)}$
        Solve
            $\begin{bmatrix} M & B^\intercal \\ B & 0 \end{bmatrix} \begin{pmatrix} p^{(\ell+1)} \\ \cdot \end{pmatrix} = - \begin{pmatrix} \zeta^{(\ell+1)} \\ 0 \end{pmatrix}$
                  // steepest descent direction, projected to $\ker B$, w.r.t. $M$-inner product
        Set $\delta^{(\ell+1)} := -(\zeta^{(\ell+1)})^\intercal p^{(\ell+1)}$    // $\delta^{(\ell+1)} = \|\operatorname{proj}^M_{\ker B}(M^{-1}\zeta^{(\ell+1)})\|_M^2$
        Set $\beta^{(\ell+1)} := \delta^{(\ell+1)} / \delta^{(\ell)}$
        Set $p^{(\ell+1)} := p^{(\ell+1)} + \beta^{(\ell+1)}\,p^{(\ell)}$
        Set $\ell := \ell + 1$
    else
        Abort the while loop
    end if
    end while
    return $d^{(\ell)}$
```

Remark 13.3 (on Algorithm 13.2).

(*i*) *One can show that Algorithm 13.2 is equivalent to the classical truncated conjugate gradient Algorithm 5.41 applied to the reduced problem* (13.3)*. When the initial guesses for Algorithm 5.41 and Algorithm 13.2 are related via*

$$d^{(0)} = \mathrm{d}_{\text{part}} + Z y^{(0)},$$

then this relation also holds for all subsequent iterates.

(*ii*) *The constraint preconditioner in Algorithm 13.2, which acts as an M-orthogonal projector onto* $\ker B$*, ensures that all directions $p^{(\ell)}$ belong to* $\ker B$*. Consequently, all iterates will satisfy $d^{(\ell)} \in d_{\text{part}} + \ker B$ and are therefore feasible, i. e., they satisfy $B\, d^{(\ell)} = c$.*

(*iii*) *When B is formally not present, i. e., its row dimension is $m = 0$, and when $d_{\text{part}} = 0$ is chosen, then Algorithm 13.2 reduces to the classical truncated conjugate gradient Algorithm 5.41.*

(*iv*) *Algorithm 13.2 as written does not return or maintain an estimate of the Lagrange multiplier λ of* (13.2)*, but it could be extended to provide that as well.*

(*v*) *Remark 5.43, which refers to the classical truncated conjugate gradient Algorithm 5.41, can be transferred to Algorithm 13.2. In particular, the first search direction is $\operatorname{proj}^M_{\ker B}(M^{-1}b)$, which is equal to the steepest descent direction $-M^{-1}\nabla_x \mathcal{L}(x^{(k)}, \mu^{(k)}, \lambda^{(k)})$ M-orthogonally projected onto* $\ker B$ *in the optimization context. When $p^{(0)}$ is a direction of positive curvature (if $\theta^{(0)} > 0$), then $d^{(1)}$ is the same as though we had applied a projected steepest descent method with Cauchy step size.*

(*vi*) *Moreover, the sequence $b^\intercal d^{(\ell)}$, corresponding to*

$$b^\intercal d^{(\ell)} = -\mathcal{L}_x(x^{(k)}, \lambda^{(k)})\, d^{(\ell)} = -f'(x^{(k)})\, d^{(\ell)} - (\lambda^{(k)})^\intercal h'(x^{(k)})\, d^{(\ell)}$$

is strictly monotonically decreasing as long as the search directions $p^{(\ell)}$ remain directions of positive curvature for A. Therefore, it is reasonable to continue performing projected CG iterations until either the desired tolerance (dictated by the outer SQP iteration, to be discussed) is reached, or a direction of non-positive curvature is encountered.

End of Week 11

§ 13.2 Solution of QPs with Additional Lower Bound Constraints

We now come to the solution of QPs with equality constraints and lower bounds, i. e., problems of the form

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2} d^\top A\, d - b^\top d \\ \text{subject to} \quad & B\, d = c \\ \text{and} \quad & d \ge \ell \end{aligned} \tag{13.6}$$

with lower bounds $\ell \in \big[\mathbb{R} \cup \{-\infty\}\big]^n$. The KKT conditions associated with (13.6) are

$$A\, d - \mu + B^\top \lambda = b \tag{13.7a}$$

$$\mu \ge 0, \quad \ell - d \le 0, \quad \mu^\top(\ell - d) = 0 \tag{13.7b}$$

$$B\, d = c. \tag{13.7c}$$

Recall that an inequality constrained QP may have many local minimizers. How can we find such a local minimizer for (13.6) or, more generally, a KKT point with associated Lagrange multipliers? We cannot use the Josephy-Newton method since in each iteration, we would have to solve precisely a problem of type (13.6).

There are various methods available in the literature for bound constrained (and more general inequality constrained) QPs. We will consider here a **semismooth Newton method** to solve the QP (13.6) respectively its first-order optimality system (13.7).[11] The semismooth Newton method addresses a non-smooth (in fact, semismooth) reformulation of the complementarity system (13.7c), achieved by means of a **nonlinear complementarity function**.

Definition 13.4 (Nonlinear complementarity function)**.** *A function $\Phi\colon \mathbb{R}^2 \to \mathbb{R}$ is said to be a **nonlinear complementarity function** (NCP) if the scalar complementarity condition*

$$a \ge 0, \quad b \ge 0, \quad a\, b = 0 \quad \textit{for } a, b \in \mathbb{R}$$

is equivalent to $\Phi(a, b) = 0$.

Prominent examples of NCP functions are

$$\Phi_{\text{min}}(a, b) := \min\{a, b\} \qquad \text{“min” function,} \tag{13.8a}$$

$$\Phi_{\text{FB}}(a, b) := \sqrt{a^2 + b^2} - a - b \qquad \text{Fischer-Burmeister function (Fischer, 1992).} \tag{13.8b}$$

Note: NCP functions must either be nonsmooth at the origin or have vanishing derivative there. (**Quiz 13.1:** What is the precise formulation of this statement, and why is this true?)

We will use the NCP function $\Phi_{\text{min}}(a, b)$. It allows us to represent the complementarity condition (13.7c) equivalently as

$$\min\{\mu,\ d - \ell\} = 0, \tag{13.9}$$

[11] Alternative methods comprise active set methods, penalty methods, Augmented Lagrangian methods, and interior point methods.

where the "min" is understood componentwise.

We now investigate the generalized differentiability properties of such a function.

Consider a function $F\colon \mathbb{R}^n \to \mathbb{R}^m$ which is Lipschitz continuous on some neighborhood $U(x)$ of $x \in \mathbb{R}^n$. Then Rademacher's theorem states that F is differentiable on $U(x)$ except on a set of Lebesgue measure zero; see, for instance Ziemer, 1989, Theorem 2.2.1.[12] We denote by D_F the set of points in $\mathbb{R}^n$ where F is differentiable.

Definition 13.5 (Bouligand and Clarke generalized derivatives). *Suppose that $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is a function which is Lipschitz continuous on some neighborhood $U(x)$ of $x \in \mathbb{R}^n$.*

(*i*) *The* ***Bouligand generalized derivative*** *is defined as the collection of limit points of derivatives of F near x:*

$$\partial_B F(x) := \left\{ M \in \mathbb{R}^{m \times n} \;\middle|\; \begin{array}{l} \text{there exists a sequence } (x^{(k)}) \subset D_F \\ \text{such that } x^{(k)} \to x \text{ and } F'(x^{(k)}) \to M \end{array} \right\}. \tag{13.10}$$

(*ii*) *The* ***Clarke generalized derivative*** *at x is defined as the convex hull[13] of $\partial_B f(x)$:*

$$\partial F(x) := \operatorname{conv} \partial_B F(x). \tag{13.11}$$

Example 13.6 (Bouligand and Clarke generalized derivatives of $\Phi_{\min}$).
The function $\Phi_{\min}(a,b) := \min\{a,b\}$ is globally Lipschitz (with Lipschitz constant 1). It is differentiable everywhere in $\mathbb{R}^2$ except on the diagonal $H := \{(a,b) \in \mathbb{R}^n \mid a = b\}$. On the open half space $H^+ := \{(a,b) \in \mathbb{R}^n \mid a > b\}$, we have $\Phi_{\min}(a,b) = b$, while $\Phi_{\min}(a,b) = a$ holds on $H^- := \{(a,b) \in \mathbb{R}^n \mid a < b\}$. Consequently, we have

$$\Phi'_{\min}(a,b) = \begin{cases} (0,1) & \text{on } H^+, \\ (1,0) & \text{on } H^-. \end{cases}$$

Since H^+ and H^- are open, the Bouligand generalized derivative agrees with the derivative there. Points on the diagonal can be approximated by points of differentiability from either side, and hence we obtain

$$\partial_B \Phi_{\min}(a,b) = \begin{cases} \{(0,1)\} & \text{for } (a,b) \in H^+, \\ \{(0,1)\} \cup \{(1,0)\} & \text{for } (a,b) \in H, \\ \{(1,0)\} & \text{for } (a,b) \in H^-. \end{cases}$$

for the Bouligand generalized derivative and

$$\partial \Phi_{\min}(a,b) = \begin{cases} \{(0,1)\} & \text{for } (a,b) \in H^+, \\ \{(\alpha, 1-\alpha) \mid \alpha \in [0,1]\} & \text{for } (a,b) \in H, \\ \{(1,0)\} & \text{for } (a,b) \in H^- \end{cases}$$

for the Clarke generalized derivative.

[12] One also says that F is **almost everywhere differentiable** on $U(x)$.

[13] The **convex hull** of a set is the smallest convex set containing it.

Based on the Clarke generalized derivative, we can formulate a generalized Newton iteration

$$0 = F(z^{(m)}) + V^{(m)}(z^{(m+1)} - z^{(m)}). \tag{13.12}$$

As a substitute $V^{(m)}$ for the Jacobian, we can use any invertible element of $\partial F(z^{(m)}) \subseteq \mathbb{R}^{n\times n}$. The convergence analysis of this generalized Newton scheme is based on the notion of semismoothness. Hence we also refer to (13.12) as a **semismooth Newton method**.

Definition 13.7 (Semismooth function; see Mifflin, 1977; Qi, 1993; Qi, Sun, 1993; Pang, Qi, 1993)**.** *Suppose that $F\colon \mathbb{R}^n \to \mathbb{R}^m$ is a function which is Lipschitz continuous on some neighborhood $U(z)$ of $z \in \mathbb{R}^n$. F is said to be **semismooth** at z if, for any $d \in \mathbb{R}^n$ and all sequences $d^{(k)} \to d$, $t^{(k)} \searrow 0$ and $M^{(k)} \in \partial F(z + t^{(k)} d^{(k)})$, the limit*

$$\lim_{k\to\infty} M^{(k)} d^{(k)}$$

exists.

There are equivalent characterizations of semismoothness available in the literature that may be easier to verify, but we do not discuss them here since we are primarily interested in the semismoothness of the Φ_{min} complementarity function.

Example 13.8 (Φ_{min} is semismooth)**.**
The function $\Phi_{\text{min}}\colon \mathbb{R}^2 \to \mathbb{R}$ defined by $\Phi_{\text{min}}(a,b) \coloneqq \min\{a,b\}$ is semismooth everywhere.

There is a rich calculus available for semismooth function. For our purpose, it is relevant that

(1) C^1 functions are semismooth (Mifflin, 1977, Proposition 4) and their Clarke generalized derivative equals the classical derivative.

(2) componentwise semismoothness implies semismoothness (Qi, Sun, 1993, Corollary 2.4). This can be used to show that the "vectorized" version $\Phi_{\text{min}}\colon \mathbb{R}^{2n} \to \mathbb{R}^n$ defined by $\Phi_{\text{min}}(a,b) \coloneqq \min\{a,b\}$, with the "min" understood componentwise, is semismooth everywhere.

(3) a chain rule holds, showing that the composition of semismooth functions is semismooth (Mifflin, 1977, Theorem 5).

Theorem 13.9 (Convergence of the local semismooth Newton method)**.** *Suppose that $F\colon \mathbb{R}^n \to \mathbb{R}^n$ and that $z^* \in \mathbb{R}^n$ is a point where $F(z^*) = 0$. Suppose further that F is Lipschitz continuous on some neighborhood $U(z^*)$ of $z^* \in \mathbb{R}^n$, that F is semismooth at z^* and that all $V \in \partial F(z^*)$ are non-singular. Then there exists a neighborhood $B_\delta^M(z^*)$ such that*

(i) z^* is the unique zero of F in $B_\delta^M(z^*)$.

(ii) For any initial guess $z^{(0)} \in B_\delta^M(z^*)$, the local semismooth Newton method is well-defined (indeed, any $V^{(k)} \in \partial F(z^{(k)})$ will be invertible), and it generates a sequence $z^{(k)}$ which converges to z^*.

(iii) $\left(z^{(k)}\right)$ *converges to* z^* *Q-superlinearly w.r.t. the* M*-norm.*

The proof of Theorem 13.9 can be found, for instance, in Qi, Sun, 1993, Theorem 3.2. The order of convergence can be shown to be of Q-order $1 + p$ for $p \in [0, 1]$, i. e., up to Q-quadratic, provided that F is semismooth of order p, but we do not pursue this any further. Notice that Theorem 13.9 contains the local convergence theorem 5.27 for the classical Newton method as a special case since C^1 functions are semismooth and their Clarke generalized derivative equals the classical derivative.

Let us now address the application of the semismooth Newton method to the solution of bound constrained QPs, respectively their KKT conditions (13.7) written in semismooth form:

$$A\,d - \mu + B^\top \lambda - b = 0 \tag{13.13a}$$

$$\min\{\mu,\ d - \ell\} = 0 \tag{13.13b}$$

$$B\,d - c = 0. \tag{13.13c}$$

We can thus write this system as a root finding problem for the function $F\colon \mathbb{R}^n \times \mathbb{R}^{n_{\text{eq}}} \times \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^{n_{\text{eq}}} \times \mathbb{R}^n$ defined by

$$F(d, \mu, \lambda) \coloneqq \begin{pmatrix} A\,d - \mu + B^\top \lambda - b \\ \min\{\mu,\ d - \ell\} \\ B\,d - c \end{pmatrix}$$

Since the components F_1 and F_3 are linear and thus C^1 and the function inside the “min” term is also linear and thus C^1, we conclude that F is everywhere semismooth.

At points satisfying $\mu = d - \ell$, the Clarke generalized derivative is not unique. Given a point (x, μ, λ), we define the **primal-dual index sets**

$$\mathcal{A}(d, \mu) \coloneqq \{i \in \{1, \ldots, n_{\text{ineq}}\} \mid \mu_i \ge d_i - \ell_i\} \quad \textbf{primal-dual active indices} \text{ at } (d, \mu), \tag{13.14a}$$

$$\mathcal{I}(d, \mu) \coloneqq \{i \in \{1, \ldots, n_{\text{ineq}}\} \mid \mu_i < d_i - \ell_i\} \quad \textbf{primal-dual inactive indices} \text{ at } (d, \mu). \tag{13.14b}$$

Also, let the diagonal matrices $D_{\mathcal{I}(d,\mu)}$ and $D_{\mathcal{A}(d,\mu)}$ be defined by

$$[D_{\mathcal{I}(d,\mu)}]_{ii} = \begin{cases} 1 & \text{if } i \in \mathcal{I}(d, \mu), \\ 0 & \text{if } i \in \mathcal{A}(d, \mu), \end{cases} \quad \text{and} \quad D_{\mathcal{A}(d,\mu)} = \operatorname{Id} - D_{\mathcal{I}(d,\mu)}.$$

One choice of the Clarke generalized derivative of F is then

$$V(d, \mu, \lambda) = \begin{bmatrix} A & -\operatorname{Id} & B^\top \\ D_{\mathcal{A}(d,\mu)} & D_{\mathcal{I}(d,\mu)} & 0 \\ B & 0 & 0 \end{bmatrix}.$$

At an iterate $(d^{(m)}, \mu^{(m)}, \lambda^{(m)})$, a semismooth Newton (13.12) step reads

$$\begin{pmatrix} A\,d^{(m)} - \mu^{(m)} + B^\top \lambda^{(m)} - b \\ \min\{\mu^{(m)},\ d^{(m)} - \ell\} \\ B\,d^{(m)} - c \end{pmatrix} + \begin{bmatrix} A & -\operatorname{Id} & B^\top \\ D_{\mathcal{A}(d^{(m)},\mu^{(m)})} & D_{\mathcal{I}(d^{(m)},\mu^{(m)})} & 0 \\ B & 0 & 0 \end{bmatrix} \begin{pmatrix} d^{(m+1)} - d^{(m)} \\ \mu^{(m+1)} - \mu^{(m)} \\ \lambda^{(m+1)} - \lambda^{(m)} \end{pmatrix} = 0.$$

Due to most terms in F being linear in the unknowns, and using

$$\min\{\mu^{(m)},\ d^{(m)} - \ell\} = D_{\mathcal{I}(d^{(m)},\mu^{(m)})}\,\mu^{(m)} + D_{\mathcal{A}(d^{(m)},\mu^{(m)})}\,(d^{(m)} - \ell),$$

we can write the semismooth Newton step more concisely as

$$\begin{bmatrix} A & -\mathrm{Id} & B^\mathsf{T} \\ D_{\mathcal{A}(d^{(m)},\mu^{(m)})} & D_{\mathcal{I}(d^{(m)},\mu^{(m)})} & 0 \\ B & 0 & 0 \end{bmatrix} \begin{pmatrix} d^{(m+1)} \\ \mu^{(m+1)} \\ \lambda^{(m+1)} \end{pmatrix} = \begin{pmatrix} b \\ D_{\mathcal{A}(d^{(m)},\mu^{(m)})}\,\ell \\ c \end{pmatrix} \tag{13.15}$$

Notice that the previous iterate enters solely through the active and inactive sets. Also notice that the matrix — unlike in classical Newton methods — is not symmetric. The Schwarz theorem does not apply since the matrix is not the Hessian of a twice differentiable function but rather the Clarke generalized derivative of a nonsmooth reformulation of the first-order optimality conditions in KKT form. We could, however, easily replace problem (13.15) by an equivalent one with a symmetric matrix. (**Quiz 13.2:** Do you see how?)

The middle block row of (13.15) reads:

$$\begin{aligned} d^{(m+1)} &= \ell \quad \text{on } \mathcal{A}(d^{(m)},\mu^{(m)}), \\ \mu^{(m+1)} &= 0 \quad \text{on } \mathcal{I}(d^{(m)},\mu^{(m)}). \end{aligned}$$

We may therefore remove these unknowns from (13.15). For simplicity of notation, we drop the iteration indices from the unknowns and the active/inactive sets for now and introduce selection matrices

$$\begin{aligned} Z_\mathcal{A} &\coloneqq \text{rows of } \mathrm{Id} \in \mathbb{R}^{n\times n} \text{ pertaining to active indices} \\ Z_\mathcal{I} &\coloneqq \text{rows of } \mathrm{Id} \in \mathbb{R}^{n\times n} \text{ pertaining to inactive indices} \end{aligned}$$

so that we can write

$$d = Z_\mathcal{A}^\mathsf{T} d_\mathcal{A} + Z_\mathcal{I}^\mathsf{T} d_\mathcal{I} \quad \text{and} \quad \mu = Z_\mathcal{A}^\mathsf{T} \mu_\mathcal{A} + Z_\mathcal{I}^\mathsf{T} \mu_\mathcal{I}$$

with subvectors $d_\mathcal{A} = Z_\mathcal{A} d$ and $d_\mathcal{I} = Z_\mathcal{I} d$ of d, and similarly for μ. Notice that $Z_\mathcal{A}^\mathsf{T} Z_\mathcal{A} = D_\mathcal{A}$ holds and also $Z_\mathcal{I}^\mathsf{T} Z_\mathcal{I} = D_\mathcal{I}$. Moreover, $Z_\mathcal{A} Z_\mathcal{A}^\mathsf{T}$ and $Z_\mathcal{I} Z_\mathcal{I}^\mathsf{T}$ are identity matrices of appropriate dimensions.

Plugging in the information $d_\mathcal{A} = Z_\mathcal{A}\,\ell$ and $\mu_\mathcal{I} = 0$ obtained from the middle block row of (13.15), we arrive at the equivalent reduced problem

$$\begin{bmatrix} Z_\mathcal{I} A Z_\mathcal{I}^\mathsf{T} & Z_\mathcal{I} B^\mathsf{T} \\ B Z_\mathcal{I}^\mathsf{T} & 0 \end{bmatrix} \begin{pmatrix} d_\mathcal{I} \\ \lambda \end{pmatrix} = \begin{pmatrix} Z_\mathcal{I}(b - A D_\mathcal{A} \ell) \\ c - B D_\mathcal{A} \ell \end{pmatrix} \tag{13.16}$$

together with $d_\mathcal{A} = Z_\mathcal{A}\ell$, $\mu_\mathcal{I} = 0$ and

$$\mu_\mathcal{A} = Z_\mathcal{A} A Z_\mathcal{I}^\mathsf{T} d_\mathcal{I} + Z_\mathcal{A} B^\mathsf{T} \lambda - Z_\mathcal{A} b + Z_\mathcal{A} A D_\mathcal{A} \ell = Z_\mathcal{A}(A\,d + B^\mathsf{T}\lambda - b).$$

(**Quiz 13.3:** Can you confirm this?)

We are now in a position to state the semismooth Newton method for the lower-bound constrained QP (13.6).

Algorithm 13.10 (Semismooth Newton method for the KKT conditions (13.13) of the lower-bound constrained QP (13.6))**.**

Input: *initial guess* $d^{(0)} \in \mathbb{R}^n$

Input: *initial guess* $\mu^{(0)} \in \mathbb{R}^n$

```
Input: right-hand side $b \in \mathbb{R}^n$
Input: right-hand side $c \in \mathbb{R}^{n_{\text{eq}}}$
Input: lower bound $\ell \in \big[\mathbb{R} \cup \{-\infty\}\big]^n$
Input: symmetric matrix $A$ (or matrix-vector products with $A$)
Input: matrix $B$ of full row rank (or matrix-vector products with $B$ and $B^\top$)
Output: approximate solution of (13.13)
 1: Set $m := 0$
 2: Determine the active and inactive sets
      $\mathcal{A}^{(m)} := \big\{i \in \{1,\ldots,n\} \,\big|\, \mu_i^{(m)} \ge d_i^{(m)} - \ell_i\big\}$
      $\mathcal{I}^{(m)} := \big\{i \in \{1,\ldots,n\} \,\big|\, \mu_i^{(m)} < d_i^{(m)} - \ell_i\big\}$
 3: while $m = 0$ or $\mathcal{A}^{(m)}$ is different from $\mathcal{A}^{(m-1)}$ do
 4:     Set $d^{(m+1)} := D_{\mathcal{A}^{(m)}}\,\ell$        // Set the active components of $d^{(m+1)}$
 5:     Solve the linear system        // Solve for the inactive components of $d^{(m+1)}$ as well as for $\lambda^{(m+1)}$
          $\begin{bmatrix} Z_{\mathcal{I}^{(m)}} A Z_{\mathcal{I}^{(m)}}^\top & Z_{\mathcal{I}^{(m)}} B^\top \\ B Z_{\mathcal{I}^{(m)}}^\top & 0 \end{bmatrix} \begin{pmatrix} d_{\mathcal{I}^{(m)}}^{(m+1)} \\ \lambda^{(m+1)} \end{pmatrix} = \begin{pmatrix} Z_{\mathcal{I}^{(m)}} (b - A\, d^{(m+1)}) \\ c - B\, d^{(m+1)} \end{pmatrix}$
 6:     Set $\mu_{\mathcal{I}^{(m)}}^{(m+1)} := 0$        // Set the inactive multiplier components
 7:     Set $\mu_{\mathcal{A}^{(m)}}^{(m+1)} := Z_{\mathcal{A}^{(m)}} [A\, d^{(m+1)} + B^\top \lambda^{(m+1)} - b]$        // Set the active multiplier components
 8:     Determine the active and inactive sets
          $\mathcal{A}^{(m+1)} := \big\{i \in \{1,\ldots,n\} \,\big|\, \mu_i^{(m+1)} \ge d_i^{(m+1)} - \ell_i\big\}$
          $\mathcal{I}^{(m+1)} := \big\{i \in \{1,\ldots,n\} \,\big|\, \mu_i^{(m+1)} < d_i^{(m+1)} - \ell_i\big\}$
 9:     Set $m := m + 1$
10: end while
11: return $d^{(m)}$, $\mu^{(m)}$ and $\lambda^{(m)}$
```

Remark 13.11 (on Algorithm 13.10).

(*i*) *You can convince yourself that the solution $(d^{(m+1)}, \mu^{(m+1)}, \lambda^{(m+1)})$ to the semismooth Newton system (13.15), achieved through Lines 4 to 7 in Algorithm 4.6, satisfies the necessary optimality conditions for the equality constrained QP*

$$\begin{aligned} &\text{Minimize} && \frac{1}{2} d^\top A\, d - b^\top d \\ &\text{subject to} && B\, d = c \\ &\text{and} && d = \ell \quad \text{on } \mathcal{A}^{(m)} \\ &\text{and} && d \text{ is free on } \mathcal{I}^{(m)}. \end{aligned} \tag{13.17}$$

The quantities $\mu_{\mathcal{A}^{(m)}}^{(m+1)}$ and $\lambda^{(m+1)}$ serve as the unique Lagrange multipliers associated with the constraints $\ell - d = 0$ and $B\, d - c = 0$, respectively, and $\mu^{(m+1)}$ is then padded with zeros on $\mathcal{I}^{(m)}$.

(*ii*) *In case A is positive definite on $\ker B$, the semismooth Newton system (13.15) will be uniquely solvable, regardless of the active set. (**Quiz 13.4:** Can you explain why?)*

(*iii*) *From iteration to iteration, the only quantity that changes is the active set (and therefore also the complementary, inactive set). Once the active sets agree between two consecutive iterations, a solution of the KKT conditions has been found.*

(*iv*) *In a numerical realization of Algorithm 13.10, we do not actually have to form the selection matrices $Z_{\mathcal{I}^{(m)}}$ and $Z_{\mathcal{A}^{(m)}}$. Instead of (13.16), we can formally consider the system*

$$\begin{bmatrix} A & B^\mathsf{T} \\ B & 0 \end{bmatrix} \begin{pmatrix} d \\ \lambda \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}$$

and implement the projected conjugate gradient method (Algorithm 13.2) to work only on the inactive indices of the unknown d, and to consider exclusively the inactive indices of the first block row. When initialized with $d = D_{\mathcal{A}} \ell$, the iterations will be identical to those of the projected conjugate gradient method applied to the reduced problem (13.16) and initialized with $d_{\mathcal{I}} = 0$. The details are beyond our scope here.

(*v*) *Algorithm 13.10 is is known as a* ***semismooth Newton method*** *and as a* ***primal-dual active set method****. The latter derives from the fact that*

(1) *it is an active set method, which means that it maintains and updates a working set of indices (into the components of the inequality constraints) deemed active, enforces the inequalities as equalities there, and ignores the inequalities deemed inactive,*

(2) *the active sets are determined from a combination of primal (d) and dual (μ) quantities; see (13.14). The equality constraint multiplier λ does not play a role here, therefore we do not require an initial guess for it.*

(*vi*) *In the absence of a better initial guess, an initialization of Algorithm 13.10 with $d^{(0)} = 0$ and $\mu_i^{(0)} = -\infty$ for all $i = 1, \ldots, n_{\text{ineq}}$ leads to an initially empty active set $\mathcal{A}^{(0)} = \emptyset$ and will result in $d^{(1)}$ solving the QP with the inequality constraints removed.*

(*vii*) *We recall that the lower bound ℓ can have components equal to $-\infty$. Those components will belong to the inactive set in all iterations.*

(*viii*) *In the presence of inequality constraints in (7.1), an SQP method may result in three levels of iterations:*

(1) *The outermost iteration (with iteration counter k) is the SQP iteration. In every iteration, we have to solve an inequality constrained QP. Using the slack reformulation (12.1), the inequality constraints in all QPs will be lower bound constraints.*

(2) *The middle iteration (with iteration counter m) is the semismooth Newton iteration which deals with the inequality constraints in each QP. In every iteration, we have to solve an equality constrained QP (13.17).*

(3) *The innermost iteration (with iteration counter ℓ) arises when we apply the projected conjugate gradient to the solution of this equality constrained QP, respectively to the linear system representing its KKT conditions (13.16).*

End of Week 12

§ 14 Further Practical Aspects of SQP Methods

So far we have discussed only local aspects of SQP methods. In order to devise a practical algorithm, many further questions need to be addressed. Let us mention some of them:

(1) How can we achieve global convergence of an SQP method, in the sense that accumulation points are KKT points?

(2) How can we ensure that the fast local convergence properties of the local SQP method (see Corollary 11.15) are preserved under the globalization?

(3) Can we use quasi-Newton updates of the Hessian of the Lagrangian instead of forming it exactly?

(4) How can we deal with potentially infeasible QP subproblems?

(5) How can we deal with non-unique Lagrange multipliers in the QPs, i. e., lack of LICQ?

We will discuss some of these aspects in this section. However, given the wealth of these and further questions, it can be expected that there are countless useful varieties of practical SQP methods.

§ 14.1 Globalization of SQP Methods by Line Search

We recall that globalization is an effort to ensure that accumulation points of the sequence of iterates are KKT points. For constrained problems, the two goals "feasibility" and "stationarity" need to be taken into account simultaneously. Therefore, a line search solely with respect to the objective as in § 5 is not appropriate. Instead, one considers line search with respect to a **merit function**.

Various merit functions are in use in the literature. We consider the ℓ_1 **penalty function** as merit function, which measures the constraint violation in terms of the ℓ_1-norm. It is defined as

$$\begin{aligned}\phi_1(x;\gamma) &\coloneqq f(x) + \gamma\,\pi_1(x)\\ &= f(x) + \gamma \sum_{i=1}^{n_{\text{ineq}}} \max\{0, g_i(x)\} + \gamma \sum_{j=1}^{n_{\text{eq}}} |h_j(x)|\\ &= f(x) + \gamma\,\|\max\{0, g(x)\}\|_1 + \gamma\,\|h(x)\|_1 \end{aligned} \tag{14.1}$$

with the ℓ_1-norm $\|\cdot\|_1$. Here γ is called the **penalty parameter**. The function π_1 and thus ϕ_1 are not everywhere differentiable. However, the (one-sided) directional derivatives

$$\pi_1'(x;d) = \lim_{t\searrow 0} \frac{\pi_1(x+t\,d) - \pi_1(x)}{t}$$

exist everywhere and they are given by

$$\begin{aligned}\pi_1'(x;d) = \sum_{\substack{i=1\\ g_i(x)<0}}^{n_{\text{ineq}}} 0 \quad + \sum_{\substack{i=1\\ g_i(x)=0}}^{n_{\text{ineq}}} \max\{0, g_i'(x)\,d\} \quad + \sum_{\substack{i=1\\ g_i(x)>0}}^{n_{\text{ineq}}} g_i'(x)\,d \\ + \sum_{\substack{j=1\\ h_j(x)<0}}^{n_{\text{eq}}} -h_j'(x)\,d \quad + \sum_{\substack{j=1\\ h_j(x)=0}}^{n_{\text{eq}}} |h_j'(x)\,d| \quad + \sum_{\substack{j=1\\ h_j(x)>0}}^{n_{\text{eq}}} h_j'(x)\,d.\end{aligned} \tag{14.2}$$

Lemma 14.1 (descent direction for ϕ_1; compare Ulbrich, Ulbrich, 2012, Satz 19.11). *Suppose that (d, μ, λ) is a KKT point of*

$$\begin{aligned} \text{Minimize} \quad & f'(x^{(k)})\,d + \frac{1}{2}\,d^\intercal H^{(k)} d, \quad \text{where } d \in \mathbb{R}^n \\ \text{subject to} \quad & g(x^{(k)}) + g'(x^{(k)})\,d \le 0 \\ \text{and} \quad & h(x^{(k)}) + h'(x^{(k)})\,d = 0. \end{aligned} \tag{14.3}$$

Moreover, suppose that

$$\gamma \ge \max\{\|\mu\|_\infty, \|\lambda\|_\infty\}$$

holds. Then we have

$$\phi_1'(x;d) \le -d^\intercal H^{(k)} d.$$

When $H^{(k)}$ is positive definite and $d \neq 0$, we therefore have

$$\phi_1'(x;d) < 0,$$

i. e., d is a descent direction for the ℓ_1 penalty function.

Proof. We estimate

$$\begin{aligned} \mu^\intercal g'(x^{(k)})\,d &= \sum_{i=1}^{n_{\text{ineq}}} \mu_i\, g_i'(x^{(k)})\,d \\ &= \sum_{\substack{i=1\\ g_i(x^{(k)})>0}}^{n_{\text{ineq}}} \mu_i\, g_i'(x^{(k)})\,d \quad - \sum_{\substack{i=1\\ g_i(x^{(k)})\le 0}}^{n_{\text{ineq}}} \mu_i\, g_i(x^{(k)}) && \text{since } \mu_i\,\big(g_i(x^{(k)}) + g_i'(x^{(k)})\,d\big) = 0 \\ &\ge \sum_{\substack{i=1\\ g_i(x^{(k)})>0}}^{n_{\text{ineq}}} \mu_i\, g_i'(x^{(k)})\,d && \text{since } \mu_i \ge 0 \\ &\ge \gamma \sum_{\substack{i=1\\ g_i(x^{(k)})>0}}^{n_{\text{ineq}}} g_i'(x^{(k)})\,d && \text{since } \gamma \ge \mu_i \text{ and } g_i'(x^{(k)})\,d \le -g_i(x^{(k)}) < 0. \end{aligned}$$

For the equality constraints, we obtain similarly

$$\begin{aligned}
\lambda^\intercal h'(x^{(k)})\, d &= \sum_{j=1}^{n_{\text{eq}}} \lambda_j\, h_j'(x^{(k)})\, d \\
&= \sum_{\substack{j=1\\ h_j(x^{(k)})\neq 0}}^{n_{\text{eq}}} \lambda_j\, h_j'(x^{(k)})\, d \quad + \sum_{\substack{j=1\\ h_j(x^{(k)})=0}}^{n_{\text{eq}}} \lambda_j\, h_j'(x^{(k)})\, d \\
&= \sum_{\substack{j=1\\ h_j(x^{(k)})>0}}^{n_{\text{eq}}} \lambda_j\, h_j'(x^{(k)})\, d \quad + \sum_{\substack{j=1\\ h_j(x^{(k)})<0}}^{n_{\text{eq}}} \lambda_j\, h_j'(x^{(k)})\, d \quad && \text{since } h(x^{(k)}) + h'(x^{(k)})\, d = 0 \\
&\ge \gamma \sum_{\substack{j=1\\ h_j(x^{(k)})>0}}^{n_{\text{eq}}} h_j'(x^{(k)})\, d \quad - \quad \gamma \sum_{\substack{j=1\\ h_j(x^{(k)})<0}}^{n_{\text{eq}}} h_j'(x^{(k)})\, d \quad && \text{since } \gamma \ge |\lambda_j|.
\end{aligned}$$

From the stationarity of the Lagrangian for (14.3), i. e., $\nabla f(x^{(k)}) + H^{(k)} d + g'(x^{(k)})^\intercal \mu + h'(x^{(k)})^\intercal \lambda = 0$, we thus have

$$\begin{aligned}
f'(x^{(k)})\, d &= -d^\intercal H^{(k)} d - \mu^\intercal g'(x^{(k)})\, d - \lambda^\intercal h'(x^{(k)})\, d \\
&\le -d^\intercal H^{(k)} d - \gamma \sum_{\substack{i=1\\ g_i(x^{(k)})>0}}^{n_{\text{ineq}}} g_i'(x^{(k)})\, d - \gamma \sum_{\substack{j=1\\ h_j(x^{(k)})>0}}^{n_{\text{eq}}} h_j'(x^{(k)})\, d + \gamma \sum_{\substack{j=1\\ h_j(x^{(k)})<0}}^{n_{\text{eq}}} h_j'(x^{(k)})\, d. \qquad (*)
\end{aligned}$$

Plugging this into the formula (14.2) for the directional derivative, we find

$$\begin{aligned}
&\phi_1'(x^{(k)}; d) \\
&\quad = f'(x^{(k)})\, d + \gamma\, \pi_1'(x^{(k)}; d) \\
&\quad = f'(x^{(k)})\, d + \gamma \sum_{\substack{i=1\\ g_i(x^{(k)})=0}}^{n_{\text{ineq}}} \max\{0, g_i'(x^{(k)})\, d\} + \gamma \sum_{\substack{i=1\\ g_i(x^{(k)})>0}}^{n_{\text{ineq}}} g_i'(x^{(k)})\, d \\
&\qquad - \gamma \sum_{\substack{j=1\\ h_j(x^{(k)})<0}}^{n_{\text{eq}}} h_j'(x^{(k)})\, d + \gamma \sum_{\substack{j=1\\ h_j(x^{(k)})=0}}^{n_{\text{eq}}} |h_j'(x^{(k)})\, d| + \gamma \sum_{\substack{j=1\\ h_j(x^{(k)})>0}}^{n_{\text{eq}}} h_j'(x^{(k)})\, d \qquad \text{due to (14.2)} \\
&\quad \le -d^\intercal H^{(k)} d + \gamma \sum_{\substack{i=1\\ g_i(x^{(k)})=0}}^{n_{\text{ineq}}} \max\{0, g_i'(x^{(k)})\, d\} + \gamma \sum_{\substack{j=1\\ h_j(x^{(k)})=0}}^{n_{\text{eq}}} |h_j'(x^{(k)})\, d| \qquad \text{due to } (*) \\
&\quad = -d^\intercal H^{(k)} d.
\end{aligned}$$

The last equality holds since, due to feasibility, we have $g(x^{(k)}) + g'(x^{(k)})\, d \le 0$ and $h(x^{(k)}) + h'(x^{(k)})\, d = 0$. This shows

$$\phi_1'(x; d) \le -d^\intercal H^{(k)} d$$

as claimed. When H is positive definite[14] and $d \neq 0$ holds, then clearly

$$\phi_1'(x; d) \le -d^\intercal H^{(k)} d < 0. \qquad \square$$

[14] In fact, it is enough for H to be positive definite on $\ker h'(x^{(k)})$.

When all inequality constraints in an NLP (7.1) are converted into lower bound constraints by means of the slack reformulation (12.2), we use the ℓ_1 penalty function in the form

$$\phi_1(x;\gamma) \coloneqq f(x) + \gamma \, \|h(x)\|_1. \tag{14.4}$$

That is, the lower bound constraints are not included by the penalty function because they can be taken into account exactly, through the QP solver. The directional derivative (14.2) then simplifies to

$$\pi_1'(x;d) = \sum_{\substack{j=1\\ h_j(x)<0}}^{n_{\text{eq}}} -h_j'(x)\,d \quad + \sum_{\substack{j=1\\ h_j(x)=0}}^{n_{\text{eq}}} |h_j'(x)\,d| \quad + \sum_{\substack{j=1\\ h_j(x)>0}}^{n_{\text{eq}}} h_j'(x)\,d. \tag{14.5}$$

We now specify an SQP line search model algorithm for the slack reformulation

$$\left.\begin{aligned} \text{Minimize} \quad & f(x), \qquad \text{where } x \in \mathbb{R}^n \\ \text{subject to} \quad & h(x) = 0 \\ \text{and} \quad & x \ge \ell. \end{aligned}\right\} \tag{12.2}$$

of an NLP (7.1) with lower bound $\ell \in \big[\mathbb{R} \cup \{-\infty\}\big]^n$.

Algorithm 14.2 (SQP Model Algorithm with Line Search; compare Geiger, Kanzow, 2002, Algorithmus 5.37, Nocedal, Wright, 2006, Algorithm 18.3 and Ulbrich, Ulbrich, 2012, Algorithmus 19.12)**.**

```
Input: initial guess $x^{(0)} \in \mathbb{R}^n$
Input: routine to evaluate $f$ and $f'$ (or $\nabla f$)
Input: routine to evaluate $h$ and $h'$
Input: lower bound $\ell \in \big[\mathbb{R} \cup \{-\infty\}\big]^n$
Input: initial symmetric model Hessian $H^{(0)} \in \mathbb{R}^{n \times n}$ (possibly s. p. d.)
Input: routine to determine the symmetric model Hessians $H^{(k)}$
Input: Armijo parameter $\sigma \in (0, 1/2)$
Input: backtracking parameter $\beta \in (0, 1)$
Input: penalty rule parameter $\rho \in (0, 1)$
Output: approximate KKT point of (12.2)
Set $k \coloneqq 0$
Set the initial penalty parameter $\gamma^{(-1)} \coloneqq -\infty$
while stopping criterion not met do
    Determine a solution $d^{(k)}$ of
        $$\begin{aligned} \text{Minimize} \quad & f'(x^{(k)})\,d + \frac{1}{2} d^\top H^{(k)} d, \qquad \text{where } d \in \mathbb{R}^n \\ \text{subject to} \quad & h(x^{(k)}) + h'(x^{(k)})\,d = 0 \\ \text{and} \quad & x^{(k)} + d \ge \ell \end{aligned} \tag{12.3}$$
    with associated Lagrange multipliers $\mu$ and $\lambda$
    Set $\chi^{(k)} \coloneqq \begin{cases} 1 & \text{if } (d^{(k)})^\top H^{(k)} d^{(k)} \ge 0 \\ 0 & \text{if } (d^{(k)})^\top H^{(k)} d^{(k)} < 0 \end{cases}$    // curvature indicator
    if $\gamma^{(k-1)}$ satisfies $\gamma^{(k-1)} \ge \dfrac{f'(x^{(k)})\,d^{(k)} + \chi^{(k)}\,\frac{1}{2}(d^{(k)})^\intercal H^{(k)} d^{(k)}}{(1-\rho)\,\|h(x^{(k)})\|_1}$ then
        Set $\gamma^{(k)} := \gamma^{(k-1)}$                                  // penalty parameter is already large enough
    else
        Set $\gamma^{(k)} := \dfrac{f'(x^{(k)})\,d^{(k)} + \chi^{(k)}\,\frac{1}{2}(d^{(k)})^\intercal H^{(k)} d^{(k)}}{(1-\rho)\,\|h(x^{(k)})\|_1}$        // increase the penalty parameter
    end if
    Determine a step size $\alpha^{(k)} > 0$ from an Armijo line search procedure (Algorithm 5.11), applied
      to $\varphi(\alpha) := f(x^{(k)} + \alpha\, d^{(k)}) + \gamma^{(k)} \|h(x^{(k)} + \alpha\, d^{(k)})\|_1$, with initial trial step size $\alpha^{(k,0)} = 1$, Armijo
      parameter $\sigma$ and backtracking parameter $\beta$
    Set $x^{(k+1)} := x^{(k)} + \alpha^{(k)} d^{(k)}$                               // update the iterate
    Set $\mu^{(k+1)} := \mu^{(k)} + \alpha^{(k)} (\mu - \mu^{(k)})$            // update the bound constraint Lagrange multiplier
    Set $\lambda^{(k+1)} := \lambda^{(k)} + \alpha^{(k)} (\lambda - \lambda^{(k)})$        // update the equality constraint Lagrange multiplier
    Set $k := k + 1$
end while
return $d^{(k)}$, $\mu^{(k)}$ and $\lambda^{(k)}$
```

Remark 14.3 (on Algorithm 14.2).

(*i*) *The choice of the penalty parameter is motivated as follows. We can use the model*

$$q^{(k)}(d) := f(x^{(k)}) + f'(x^{(k)})\,d + \chi^{(k)}(d)\,\frac{1}{2} d^\intercal H^{(k)} d + \gamma\,\|h(x^{(k)}) + h'(x^{(k)})\,d\|_1 \tag{14.6}$$

for the ℓ_1 merit function. Here $\chi^{(k)}$ is a curvature indicator, defined as

$$\chi^{(k)}(d) := \begin{cases} 1 & \text{if } d^\intercal H^{(k)} d \ge 0, \\ 0 & \text{if } d^\intercal H^{(k)} d < 0. \end{cases}$$

The goal is to make the penalty parameter γ large enough so that a certain fraction in the decrease of $q^{(k)}$ in the current step can be attributed to a decrease in the penalty term. In terms of formulas, one requires

$$\underbrace{q^{(k)}(0) - q^{(k)}(d^{(k)})}_{\text{decrease in the model of the merit function}} \ge \overbrace{\rho}^{\text{fraction}}\, \underbrace{\gamma \left[\|h(x^{(k)})\|_1 - \|h(x^{(k)}) + h'(x^{(k)})\,d\|_1\right]}_{\text{decrease in the penalty term of the model}}.$$

Plugging in the model (14.5), we can rewrite this as

$$\begin{aligned} &\gamma\,\|h(x^{(k)})\|_1 - f'(x^{(k)})\,d^{(k)} - \chi^{(k)}(d^{(k)})\,\frac{1}{2}(d^{(k)})^\intercal H^{(k)} d^{(k)} - \gamma\,\|h(x^{(k)}) + h'(x^{(k)})\,d^{(k)}\|_1 \\ &\qquad \ge \rho\,\gamma \left[\|h(x^{(k)})\|_1 - \|h(x^{(k)}) + h'(x^{(k)})\,d^{(k)}\|_1\right]. \end{aligned}$$

Since $d^{(k)}$ is feasible for the QP (12.3), this is equivalent to

$$\begin{aligned} &\gamma\,\|h(x^{(k)})\|_1 - f'(x^{(k)})\,d^{(k)} - \chi^{(k)}(d^{(k)})\,\frac{1}{2}(d^{(k)})^\intercal H^{(k)} d^{(k)} \\ &\qquad \ge \rho\,\gamma\,\|h(x^{(k)})\|_1 \end{aligned}$$

and finally to

$$\gamma \ge \frac{f'(x^{(k)})\, d^{(k)} + \chi^{(k)}(d^{(k)})\, \frac{1}{2}(d^{(k)})^\intercal H^{(k)} d^{(k)}}{(1-\rho)\, \|h(x^{(k)})\|_1}.$$

This is the condition enforced through Lines 6 to 10.

(*ii*) *The step size $\alpha^{(k)}$ obtained from the Armijo line search is used to update both the optimization variable and the Lagrange multipliers; see Lines 12 to 14.*

(*iii*) *The model Algorithm 14.2 still leaves a number of important details open, which are touched upon in the coming subsections.*

§ 14.2 Quasi-Newton SQP Methods

The benefits of using quasi-Newton approximations of the Hessian of the Lagrangian $\mathcal{L}_{xx}(x^{(k)}, \mu^{(k)}, \lambda^{(k)})$ as Hessians $H^{(k)}$ in the sequence of QPs are the same as in unconstrained optimization (§ 5.7). In contrast to (5.45), quasi-Newton updates are based on the data

$$s^{(k)} := x^{(k+1)} - x^{(k)} = \alpha^{(k)} d^{(k)} \quad \text{and} \quad y^{(k)} := \nabla_x \mathcal{L}(x^{(k+1)}, \mu^{(k)}, \lambda^{(k)}) - \nabla_x \mathcal{L}(x^{(k)}, \mu^{(k)}, \lambda^{(k)}). \tag{14.7}$$

In the unconstrained optimization context, we used the fact that $d^{(k)}$ was a descent direction for the objective. A Wolfe-Powell line search then ensured $(y^{(k)})^\intercal s^{(k)} > 0$, which is a necessary condition for the positive definiteness of any quasi-Newton formula and is also sufficient for DFP and BFGS; see Lemma 5.49. Unfortunately, we cannot expect this result to be applicable here since $d^{(k)}$ is not necessarily a descent direction for $\mathcal{L}(x^{(k)}, \mu^{(k)}, \lambda^{(k)})$. Therefore, a Wolfe-Powell line search w.r.t. to $\mathcal{L}$ is generally not applicable. Even if it were successful, it might not yield $(y^{(k)})^\intercal s^{(k)} > 0$; see the proof of Lemma 5.48. In any case, the line search applied in Algorithm 14.2 is w.r.t. the merit function, not the Lagrangian, and one usually goes with the simpler Armijo backtracking procedure.

We describe here an often used modification of the BFGS update proposed by Powell, 1978, which is still able to maintain positive definiteness. This strategy replaces the BFGS update

$$\Phi_{\text{BFGS}}(H, s, y) = H - \frac{H\, s\, s^\intercal H}{s^\intercal H\, s} + \rho\, y\, y^\intercal \quad \text{with } \rho = \frac{1}{y^\intercal s} \tag{5.55}$$

by the **damped BFGS**

$$\Phi_{\text{damped BFGS}}(H, s, y) = H - \frac{H\, s\, s^\intercal H}{s^\intercal H\, s} + \rho\, \overline{y}\, \overline{y}^\intercal \quad \text{with } \rho = \frac{1}{\overline{y}^\intercal s}. \tag{14.8a}$$

Here $\overline{y}$ is a convex combination of y and $H\, s$, namely

$$\overline{y} := \theta\, y + (1-\theta)\, H\, s, \tag{14.8b}$$

where the scalar θ is defined as

$$\theta := \begin{cases} 1 & \text{if } y^\intercal s \ge 0.2\, s^\intercal H\, s, \\ \frac{0.8\, s^\intercal H\, s}{s^\intercal H\, s - y^\intercal s} & \text{if } y^\intercal s < 0.2\, s^\intercal H\, s. \end{cases} \tag{14.8c}$$

When H is symmetric positive definite, then it can be easily shown that $\theta \in [0, 1]$. (**Quiz 14.1:** Details?) One of the extreme cases is $\theta = 0$, which means $\Phi_{\text{damped BFGS}}(H, s, y) = H$ and results in $H^{(k+1)} = H^{(k)}$ in an algorithm. The other extreme case is $\theta = 1$, which means $\Phi_{\text{damped BFGS}}(H, s, y) = \Phi_{\text{BFGS}}(H, s, y)$, i. e., a regular BFGS update step.

Lemma 14.4 (Positive definiteness of the damped BFGS update; see Geiger, Kanzow, 2002, Lemma 5.38)**.** *Suppose that H is symmetric and positive definite and that y, s are any vectors, $s \neq 0$. Then $H^+_{\text{damped BFGS}} := \Phi_{\text{damped BFGS}}(H, s, y)$ is symmetric and positive definite as well.*

Proof. The proof is part of **??**. □

§ 14.3 Infeasible QP Subproblems

The linearization $h(x^{(k)}) + h'(x^{(k)})\, d$ of the constraints $h(x) = 0$ may lead to infeasible QP subproblems. This phenomenon is also known as **inconsistent linearization**.

Example 14.5 (from Geiger, Kanzow, 2002, p.264)**.** *Consider the problem*

$$\begin{aligned} \textit{Minimize} \quad & x^2, \quad \textit{where } x \in \mathbb{R} \\ \textit{subject to} \quad & 1 - x^2 \le 0. \end{aligned}$$

The feasible set is $(-\infty, -1] \cup [1, \infty)$. When linearized at $x^{(k)} = 0$, the constraint $g(x^{(k)}) + g'(x^{(k)})\, d \le 0$ reads

$$1 + 0\, d \le 0,$$

which is impossible to satisfy.

Quiz 14.2: Would the reformulation using slack variables solve the issue?

A prominent idea to deal with potentially infeasible QPs is to relax and penalize their constraints. We describe this technique for a QP with general constraints (14.3). To simplify the notation, we write this here as

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2} d^\mathsf{T} A\, d - b^\mathsf{T} d \\ \text{subject to} \quad & B_{\text{eq}}\, d - c_{\text{eq}} = 0 \\ \text{and} \quad & B_{\text{ineq}}\, d - c_{\text{ineq}} \le 0. \end{aligned} \tag{14.9}$$

The slack reformulation is contained as a special case.

The penalty reformulation of (14.9) reads

$$\begin{aligned}
\text{Minimize} \quad & \frac{1}{2} d^\top A\, d - b^\top d + \gamma \left[\mathbf{1}^\top v + \mathbf{1}^\top w + \mathbf{1}^\top t \right], \quad \text{where } (d, v, w, t) \in \mathbb{R}^n \times \mathbb{R}^{n_{\text{eq}}} \times \mathbb{R}^{n_{\text{eq}}} \times \mathbb{R}^{n_{\text{ineq}}} \\
\text{subject to} \quad & B_{\text{eq}}\, d - c_{\text{eq}} \;= v - w \\
\text{and} \quad & B_{\text{ineq}}\, d - c_{\text{ineq}} \le t \\
\text{as well as} \quad & v \ge 0, \ w \ge 0, \ t \ge 0.
\end{aligned} \tag{14.10}$$

This penalized QP is always feasible (**Quiz 14.3:** Why?).

A convergence analysis of a modified SQP algorithm utilizing this penalty approach can be found in Geiger, Kanzow, 2002, Abschnitt 5.5.7 and 5.5.8.

§ 14.4 Fast Local Convergence and the Maratos Effect

We already emphasized in the context of unconstrained optimization methods that a globalization mechanism is not supposed to interfere with the fast local convergence we typically obtain in a local version of the method; see for instance Theorem 5.33. Unfortunately, the use of a line search w.r.t. a merit function can cause the step size $\alpha^{(k)} = 1$ to become unacceptable, thus impeding fast local convergence. This phenomenon is known as the **Maratos effect**, first described in the dissertation Maratos, 1978.

Example 14.6 (Maratos effect; compare Geiger, Kanzow, 2002, Beispiel 5.39 and Powell, 1986).
Consider the problem

$$\begin{aligned}
\textit{Minimize} \quad & 2\,(x_1^2 + x_2^2 - 1) - x_1, \quad \textit{where } x \in \mathbb{R}^2 \\
\textit{subject to} \quad & x_1^2 + x_2^2 - 1 = 0.
\end{aligned} \tag{14.11}$$

We find the following expressions for the derivative of the objective f and the equality constraint h:

$$\begin{aligned}
\nabla f(x) &= 4\,x - \begin{pmatrix} 1 \\ 0 \end{pmatrix}, & f''(x) &= 4\,\mathrm{Id}, & \mathcal{L}_{xx}(x, \lambda) &= (4 + 2\,\lambda)\,\mathrm{Id}, \\
\nabla h(x) &= 2\,x, & h''(x) &= 2\,\mathrm{Id}.
\end{aligned}$$

It is easy to see that $x^* = (1, 0)^\top$ is the unique global minimizer and that $\lambda^* = -3/2$ is the unique associated Lagrange multiplier. The second-order sufficient condition (Theorem 9.5) and the LICQ hold. Due to the absence of inequality constraints, the strong second-order sufficient condition (Definition 11.12) coincides with the usual second-order sufficient condition and thus holds as well. Therefore, Corollary 11.15 guarantees the local Q-quadratic convergence of a local SQP method.

We consider an iterate $(x^{(k)}, \lambda^{(k)})$ and the associated QP

$$\begin{aligned}
\textit{Minimize} \quad & f'(x^{(k)})\, d + \frac{1}{2}\,(4 + 2\,\lambda^{(k)})\, \|d\|^2 \\
\textit{subject to} \quad & h(x^{(k)}) + 2\,(x^{(k)})^\top d = 0.
\end{aligned}$$

The KKT conditions of the associated QP read

$$\nabla f(x^{(k)}) + (4 + 2\,\lambda^{(k)})\,d + 2\,\lambda\,x^{(k)} = 0, \tag{14.12a}$$

$$h(x^{(k)}) + 2\,(x^{(k)})^\intercal d = 0. \tag{14.12b}$$

Since f and h are quadratic, their second-order Taylor expansion is exact. We obtain

$$\begin{aligned} f(x^{(k)} + d) - f(x^{(k)}) &= f'(x^{(k)})\,d + \frac{1}{2} d^\intercal f''(x^{(k)})\,d \\ &= -(4 + 2\,\lambda^{(k)})\,\|d\|^2 - 2\,\lambda\,(x^{(k)})^\intercal d + \frac{1}{2}\,4\,\|d\|^2 && \textit{by } (14.12a) \\ &= -(2 + 2\,\lambda^{(k)})\,\|d\|^2 + \lambda\,h(x^{(k)}) && \textit{by } (14.12b). \end{aligned}$$

We also have

$$\begin{aligned} h(x^{(k)} + d) - h(x^{(k)}) &= h'(x^{(k)})\,d + \frac{1}{2} d^\intercal h''(x^{(k)})\,d \\ &= -h(x^{(k)}) + \|d\|^2 && \textit{by } (14.12b). \end{aligned}$$

*We now consider the case where $x^{(k)}$ is already admissible, i. e., $h(x^{(k)}) = 0$, but $x^{(k)} \neq (\pm 1, 0)^\intercal$. Moreover, we assume $\lambda^{(k)} < -1$. Then the unique solution d of the QP satisfies $d \neq 0$ (**Quiz 14.4:** Why?) Thus we have*

$$f(x^{(k)} + d) - f(x^{(k)}) = -(2 + 2\,\lambda^{(k)})\,\|d\|^2 + \lambda\,\underbrace{h(x^{(k)})}_{=0} > 0$$

and also

$$h(x^{(k)} + d) - \underbrace{h(x^{(k)})}_{=0} = -\underbrace{h(x^{(k)})}_{=0} + \|d\|^2 > 0.$$

This means that the full QP step d would increase both the value of the objective and the value of the constraint penalty, i. e.,

$$\phi_1(x^{(k)} + d) = f(x^{(k)} + d) + \gamma^{(k)}\,|h(x^{(k)} + d)| > \phi_1(x^{(k)}),$$

for any value of the penalty parameter $\gamma^{(k)}$! Therefore, the Armijo condition will not hold for the full step size $\alpha^{(k)} = 1$. Consequently, a step size $\alpha^{(k)} < 1$ will be chosen in Algorithm 14.2, resulting in the loss of Q-superlinear convergence. Starting close to the solution does not help.

Another concrete example for the Maratos effect is given in Nocedal, Wright, 2006, Example 18.1.

One strategy to overcome the Maratos effect is known as **second-order correction**. In a nutshell, one modifies the QP direction d in case the step size $\alpha^{(k)} = 1$ is not acceptable. This modification requires the solution of a second QP. We refer the reader to Geiger, Kanzow, 2002, Abschnitt 5.5.6 or Nocedal, Wright, 2006, p.543 for details.

§ 14.5 Globalization Strategies Different from Line Search

Apart from line search, two other globalization strategies are known in constrained optimization. These are trust-region and filter approaches.

Trust-Region Approach

We already saw trust-region methods for unconstrained optimization in § 6. An advantage is that we can dispense with the positive definiteness of the model Hessians since the trust-region constraint ensures that the QP cannot be unbounded. However, a direct transfer of the technique by considering trust-region constrained QPs such as

$$\begin{aligned} \text{Minimize} \quad & f'(x^{(k)})\, d + \frac{1}{2}\, d^\intercal H^{(k)}\, d, \quad \text{where } d \in \mathbb{R}^n \\ \text{subject to} \quad & h(x^{(k)}) + h'(x^{(k)})\, d = 0 \\ \text{and} \quad & \ell - x^{(k)} - d \le 0 \\ \text{plus} \quad & \|d\|_M \le \Delta^{(k)} \end{aligned} \tag{14.13}$$

is impossible. (**Quiz 14.5:** What might be the problem with (14.13)?)

One possible fix was proposed in the dissertation Omojokun, 1989 supervised by Richard Byrd; see also Conn, Gould, Toint, 2000, Chapter 15.4. The resulting class of methods is known as **composite-step trust-region methods** or **Byrd-Omojokun methods**. They decompose the step d into a **normal step** and a **tangential step** according to $d = n + t$.

The **normal step** serves to improve the linearized feasibility. It is defined as the solution to

$$\begin{aligned} \text{Minimize} \quad & \|h(x^{(k)}) + h'(x^{(k)})\, n\|^2 + \|\max\{0,\ \ell - x^{(k)} - n\}\|^2 \\ \text{subject to} \quad & \|n\|_M \le \zeta\, \Delta^{(k)}. \end{aligned} \tag{14.14}$$

Here $\zeta \in (0,1)$ is the fraction of the trust region radius reserved for the normal step. Problem (14.14) is a convex problem with a C^1, piecewise quadratic objective. In the absence of inequality constraints, we can approximately solve (14.14) by the dogleg method, which considers the Cauchy point n_C of (14.14) as well as the unique least-norm solution n_{UC} of (14.14) with the trust-region constraint removed:

$$\begin{aligned} \text{Minimize} \quad & \frac{1}{2}\|n\|_M^2 \\ \text{subject to} \quad & h(x^{(k)}) + h'(x^{(k)})\, n = 0. \end{aligned} \tag{14.15}$$

Problem (14.15) can be solved using the projected conjugate gradient method (Algorithm 13.2). The dogleg method then either returns n_{UC} (if it satisfies the constraint $\|n_{UC}\|_M \le \zeta\, \Delta^{(k)}$), or else it finds the unique intersection of the path

$$0 \longrightarrow \text{Cauchy point } n_C \longrightarrow \text{least-norm solution } n_{UC}$$

with the trust-region ball $\|n\|_M \le \zeta\, \Delta^{(k)}$.

The purpose of the **tangential step** is to reduce the value of the objective in (14.13) while maintaining the achieved level of linearized feasibility. The tangential step may use up the remainder of the trust-region ball. This results in the problem

$$\begin{aligned} \text{Minimize} \quad & f'(x^{(k)})\,(n+t) + \frac{1}{2}\,(n+t)^\intercal \mathcal{L}_{xx}(x^{(k)}, \mu^{(k)}, \lambda^{(k)})\,(n+t) \\ \text{subject to} \quad & h'(x^{(k)})\, t = 0 \\ \text{and} \quad & t \ge \min\{0,\ \ell - x^{(k)} - n\} \\ \text{as well as} \quad & \|n + t\|_M \le \Delta^{(k)}. \end{aligned} \tag{14.16}$$

In practice, the trust-region constraint in (14.16) is often replaced by the simpler $\|t\|_M^2 \le (\Delta^{(k)})^2 - \|n\|_M^2$. In the absence of inequality constraints, problem (14.16) too can be solved using the using the projected conjugate gradient method (Algorithm 13.2), endowed with the Steihaug-Toint modifications (Algorithm 6.14) to monitor the curvature of search directions and the trust-region constraint.

Finally, there is typically a Lagrange multiplier update step. In the absence of inequality constraints, this can be formulated as

$$\text{Minimize} \quad \frac{1}{2}\|\nabla f(x^{(k+1)}) + h'(x^{(k+1)})^\intercal \lambda\|_{M^{-1}}^2 \tag{14.17}$$

Filter Approach

A filter approach for the solution of nonlinear optimization problems was first introduced in Fletcher, Leyffer, 2002. It uses a measure of infeasibility, e. g.,

$$\pi_1(x) = \sum_{i=1}^{n_{\text{ineq}}} \max\{0, g_i(x)\} + \sum_{j=1}^{n_{\text{eq}}} |h_j(x)|,$$

as in the ℓ_1 penalty function. We can then consider the problem

$$\left.\begin{aligned} \text{Minimize} \quad & f(x), && \text{where } x \in \mathbb{R}^n \\ \text{subject to} \quad & g_i(x) \le 0 && \text{for } i = 1, \ldots, n_{\text{ineq}} \\ \text{and} \quad & h_j(x) = 0 && \text{for } j = 1, \ldots, n_{\text{eq}}. \end{aligned}\right\} \tag{7.1}$$

as a minimization problem with two goals, the minimization of both f and π_1. Such problems are known as **multiobjective problem**.

Filter methods maintain a collection of "interesting" points $x^{(k)}$, called the **filter**. All subsequent iterates compete against points in this filter. More precisely, we say that a pair $\big(f(x^{(k)}), \pi_1(x^{(k)})\big)$ **dominates** another pair $\big(f(x^{(\ell)}), \pi_1(x^{(\ell)})\big)$ if both $f(x^{(k)}) \le f(x^{(\ell)})$ and $h(x^{(k)}) \le h(x^{(\ell)})$ hold, i. e., if $x^{(k)}$ is better than $x^{(\ell)}$ both with respect to objective value and feasibility violation. A **filter** is any collection of points $x^{(k)}$ such that no pair dominates any other pair, i. e.,

$$\mathcal{F} = \Big\{x^{(1)}, \ldots, x^{(N)} \;\Big|\; \big(f(x^{(k)}), \pi_1(x^{(k)})\big) \text{ is not dominated by } \big(f(x^{(\ell)}), \pi_1(x^{(\ell)})\big) \text{ for any } \ell \ne k\Big\}. \tag{14.18}$$

A trial iteration, obtained from the solution of a trust-region constrained QP, is then accepted into the filter if it provides a certain improvement over the points already present. Points which are dominated by the new member are then removed from the filter. The details of a practical method are beyond our scope here.

End of Week 13

Chapter 4 Differentiation Techniques

All algorithms we considered require at least first-order derivatives of the objective and the constraint functions. Providing these derivatives can be an extra burden to the user. Therefore, the question arises whether these derivatives can somehow be obtained automatically. This of course also of interest outside of applications in optimization.

We generally in this chapter consider first-order derivatives for differentiable functions $F \colon \mathbb{R}^n \to \mathbb{R}^m$. We recall the structure of the Jacobian

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

We denote the partial derivatives of F, i. e., the columns of the Jacobian, by $\frac{\partial F}{\partial x_j}$.

§ 15 Finite Difference and Complex-Step Approximation

§ 15.1 First- and Second-Order Finite Differences

Finite differencing (also known as **numerical differentiation**) is the most elementary approach to evaluate derivatives for a given function. Recalling the differentiation formula

$$F'(x)\, d = \lim_{t \to 0} \frac{F(x + t\, d) - F(x)}{t},$$

it is an obvious idea to approximate this directional derivative by the **finite difference**

$$F'(x)\, d \approx \frac{F(x + t\, d) - F(x)}{t} \tag{15.1}$$

for some value of $t \neq 0$. In case of $t > 0$, we speak of a **forward difference**, while $t < 0$ is referred to as a **backward difference**. Using (15.1), we can thus approximate derivatives using only function values. Every column of the Jacobian requires one additonal function evaluation with $v = e_j$, the j-unit vector.

A Taylor analysis shows that the accuracy of (15.1) is of order t (provided that F is of class C^2), in short:

$$F'(x)\, d - \frac{F(x + t\, d) - F(x)}{t} \in O(|t|) \quad \text{as } t \to 0.$$

(**Quiz 15.1:** Can you fill in the details?) Thus, in principle, the approximation can be made arbitrarily precise by choosing t small enough. In numerical practice, however, the accuracy is limited by round-off effects, which dominate the error when t is small, since taking the difference between two similar numbers is ill-conditioned ("catastrophic cancellation").

In order to achieve higher accuracy, one resorts to a **central difference** formula

$$F'(x)\, d \approx \frac{F(x + t\, d) - F(x - t\, d)}{2t} \tag{15.2}$$

which is twice as expensive but yields $O(|t|^2)$ accuracy (provided that F is of class C^3). (**Quiz 15.2:** Can you fill in the details here as well?) It thus (often) allows us to use larger values of t, which cause less round-off error. In any case, the optimal value for the step size t depends on the function F and the direction v and is generally unknown.

Many optimization solvers provide a finite difference functionality and fall back to it in case the user does not provide the respective derivatives.

§ 15.2 Complex-Step Differentiation

A lesser known approach is **complex-step differentiation**; see Lyness, Moler, 1967; Squire, Trapp, 1998. It requires $F\colon \mathbb{R}^n \to \mathbb{R}^m$ to be a function which smoothly extends into the complex space to a function $F\colon \mathbb{C}^n \to \mathbb{C}^m$. A sufficient condition is that F is analytic.

The main idea is to use the Taylor expansion

$$F(x + i\, t\, d) \in F(x) + i\, t\, F'(x)\, d - \frac{1}{2} t^2\, F''(x)[d, d] + O(|t|^3) \tag{15.3}$$

for $x, d \in \mathbb{R}^n$, where i is the imaginary unit. This means that

$$F(x) \in \operatorname{Re} F(x + i\, t\, d) + O(|t|^2) \quad \text{and} \quad F'(x)\, d \in \operatorname{Im} \frac{F(x + i\, t\, d)}{t} + O(|t|^2)$$

holds for $t \in \mathbb{R}$, $t \neq 0$. This has the advantage that we can obtain a second-order accurate approximation of $F'(x)\, d$ with just one evaluation of F, albeit for a complex argument. Since no differences need to be taken, catastrophic cancellation does not take place.

Example 15.1 (finite differences vs. complex-step differentiation[1]). *We consider the function $F\colon \mathbb{R} \to \mathbb{R}$, given by*

$$F(x) = \frac{\exp(x)}{(\cos x)^3 + (\sin x)^3}.$$

*Its true **symbolic derivative** (in the direction $d = 1$) is given by*

$$F'(x) = \frac{\exp(x)\left[(\cos x)^3 + (\sin x)^3 - 3\left((\sin x)^2 \cos x - (\cos x)^2 \sin x\right)\right]}{\left[(\cos x)^3 + (\sin x)^3\right]^2}.$$

[1] This example is from Cleve Moler's blog https://blogs.mathworks.com/cleve/2013/10/14/complex-step-differentiation/.

F can be extended to a function $\mathbb{C} \to \mathbb{C}$ and the formulas for F and F' remain the same.

We numerically compare

$$\begin{array}{ll} \textit{the first-order finite difference} & \dfrac{F(x+t)-F(x)}{t} \\ \textit{the second-order central finite difference} & \dfrac{F(x+t)-F(x-t)}{2t} \\ \textit{the complex-step approximation} & \operatorname{Im}\dfrac{F(x+i\,t)}{t} \end{array}$$

for a range of values $t > 0$ to approximate the derivative $F'(x)$ at $x = \frac{\pi}{4}$. It exact value is $F'(x) \approx 3.1018$. The results are shown in Figure 15.1.

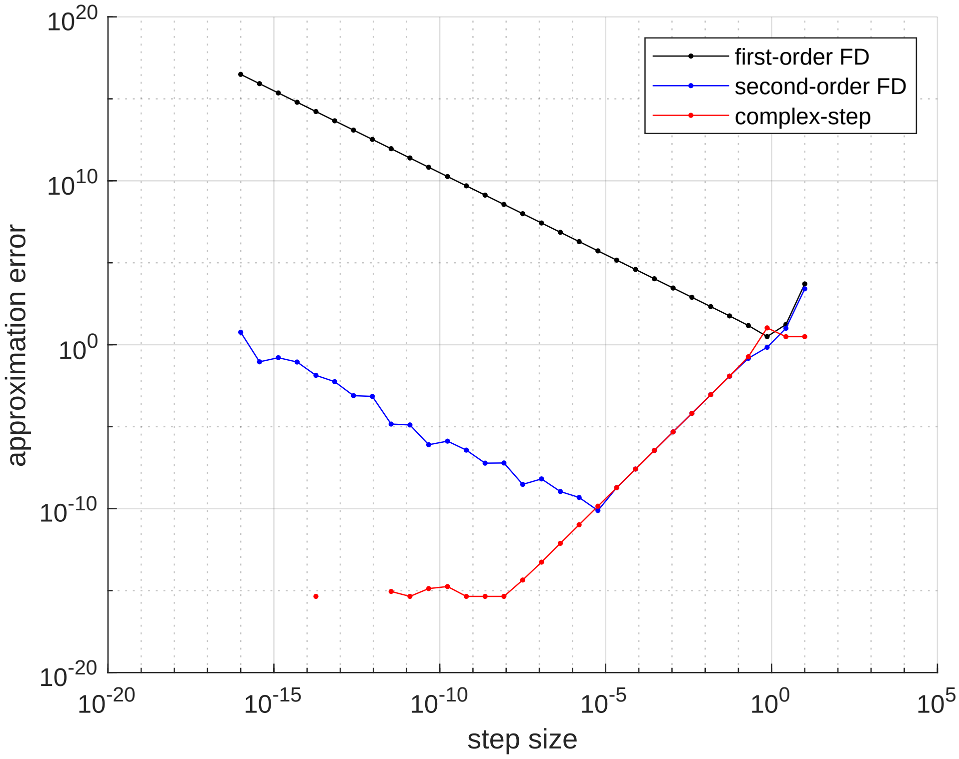


Figure 15.1: Comparison of the approximation errors of the first- and second-order finite differences as well as the complex-step approximation for the function in Example 15.1. We clearly see the limited accuracies due to round-off in the finite difference approximations, while the complex-step method achieves full machine precision.

§ 16 Algorithmic Differentiation

Algorithmic Differentiation (also known as **Automatic Differentiation** or **AD**) is a technique to evaluate derivatives of functions $F\colon \mathbb{R}^n \to \mathbb{R}^m$ that are realized by computer code. The origins of AD reach as far back as the 1950s; see the references in Bischof, Hovland, Norris, 2008; Griewank, Walther,

2008. Algorithmic differentiation is based on the fact that even complicated functions are ultimately compositions of elementary operations such as "+", " · ", "exp", etc. Using a table of the derivatives of these elementary functions and the chain rule, the computer code realizing derivatives (for instance the directional derivative $F'(x)\,\delta x$ for a given direction δx) can be automatically generated. We begin with an example.

Example 16.1 (algorithmic differentiation by hand)**.** *We use the function from Example 15.1. When implemented, e. g., in* Python *syntax, the function may look like this:*

```
y = exp(x) / (cos(x)**3 + sin(x)**3)
```

When this expression is broken down into elementary functions, which is similar to what a compiler would do, we might obtain the following sequence of intermediate results. Next to each elementary operation, we also write the respective derivative.

$$\begin{aligned} e &:= \mathrm{e}^x & \dot e &:= \mathrm{e}^x\,\dot x \\ c &:= \cos(x) & \dot c &:= -\sin(x)\,\dot x \\ c_3 &:= c^3 & \dot c_3 &:= 3\,c^2\,\dot c \\ s &:= \sin(x) & \dot s &:= \cos(x)\,\dot x \\ s_3 &:= s^3 & \dot s_3 &:= 3\,s^2\,\dot s \\ d &:= c_3 + s_3 & \dot d &:= \dot c_3 + \dot s_3 \\ y &:= \frac{e}{d} & \dot y &:= \frac{\dot e}{d} - \frac{e}{d^2}\,\dot d \end{aligned}$$

Here a dot above a quantity means the derivative of that quantity with respect to the input variable x, in a given direction δx in input space. In this example, since x is a scalar, it is reasonable to set $\delta x = 1$, which amounts to $\dot x = 1$.

While the differentiation process in Example 16.1 was carried out by hand, we were following a simple set of rules, which can be formalized and automated. In general, we can think of the result $y = F(x) \in \mathbb{R}^m$ being obtained from the input x passing through a sequence of elementary operations, which can be arranged in a computational graph.

Definition 16.2 (Directed graph)**.**

(*i*) *A finite* ***directed graph*** *(**digraph**) is a pair $G = (V, E)$ consisting of a finite set V (whose elements are called **vertices**) and a finite set E (whose elements are called **edges** or **arcs**). Every edge is a pair $(u, v) \in E$ with $u, v \in V$.*

(*ii*) *A digraph is called **simple** if no edge is of the form (u, u) (no loops), and if every possible edge (u, v) appears at most once in E.*

(*iii*) *The fact that (u, v) is an edge with tail u and head v will also be denoted by the short-hand notation $u \prec v$ or $v \succ u$. In this case, we also call v a **direct successor** of u and u a **direct predecessor** of v.*

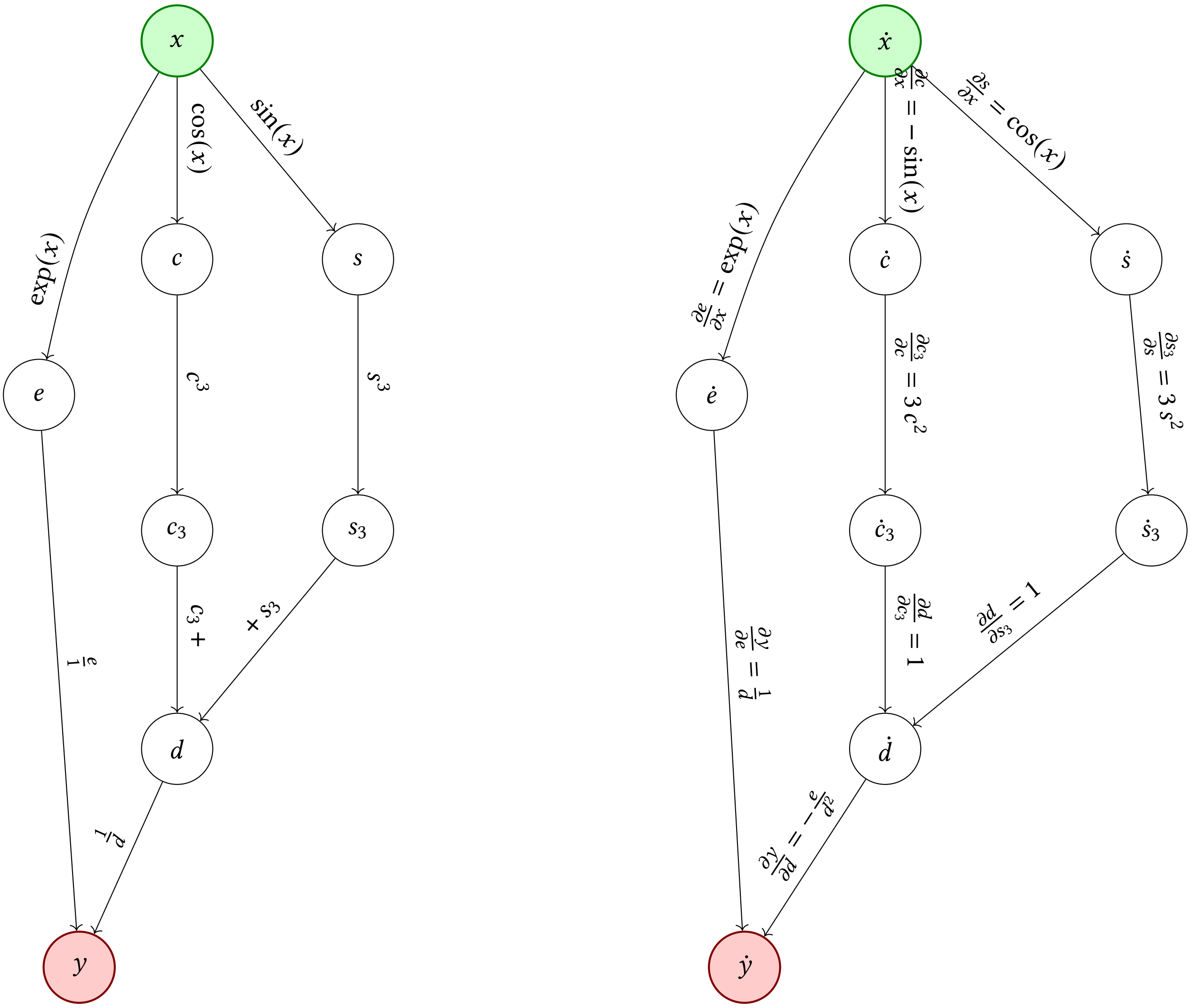

x
$\exp(x)$
$\cos(x)$
$\sin(x)$
c
s
e
c^3
s^3
c_3
s_3
$c_3 +$
$+ s_3$
$\frac{e}{1}$
d
$\frac{1}{d}$
y
$\dot{x}$
$\frac{\partial e}{\partial x} = \exp(x)$
$\frac{\partial c}{\partial x} = -\sin(x)$
$\frac{\partial s}{\partial x} = \cos(x)$
$\dot{c}$
$\dot{s}$
$\dot{e}$
$\frac{\partial c_3}{\partial c} = 3\,c^2$
$\frac{\partial s_3}{\partial s} = 3\,s^2$
$\dot{c}_3$
$\dot{s}_3$
$\frac{\partial y}{\partial e} = \frac{1}{d}$
$\frac{\partial d}{\partial c_3} = 1$
$\frac{\partial d}{\partial s_3} = 1$
$\dot{d}$
$\frac{\partial y}{\partial d} = -\frac{e}{d^2}$
$\dot{y}$

Figure 16.1: Computational graph for a possible realization of the function $F(x) = \frac{\exp(x)}{(\cos x)^3 + (\sin x)^3}$ (left) and the corresponding graph for its derivative (right).

(*iv*) *A tuple* $(v_0, \ldots, v_k)$ *of vertices,* $k \ge 1$*, is said to be a* ***path*** *from* v_0 *to* v_k *if* $(v_i, v_{i+1}) \in E$ *holds for* $i = 0, 1, \ldots, k-1$.

(*v*) *A simple digraph is said to be* ***acyclic*** *if there is no path from any vertex back to itself.*

We now generalize the computations in Example 16.1 to a general function $y = F(x)$. We assume that the function evaluation has been represented as a simple, acyclic digraph, which we term a **computational graph** of F. The vertices $v_0, \ldots, v_N$ of the graph correspond to intermediate quantities, which can be scalars or vector-valued quantities. Our convention is that $v_0 = x$ is the input and $v_N = y = F(x)$ is the output. Consequently, v_0 has no direct predecessors and v_N has no direct successors. Every intermediate quantity $v_1, \ldots, v_N$ is going to be a function of some of the other intermediate quantities. We denote this fact in the form

$$v_i = \varphi_i(v_j)_{j \prec i}. \tag{16.1}$$

The functions φ_i are called **elemental functions**. They are typically unary or binary functions, i. e., $\{j \mid j \prec i\}$ has typically zero, one or two entries. (**Quiz 16.1:** How do we usually refer to a quantity v_i that does not depend on any input?) We can assume that the vertices are ordered to that $j \prec i$ implies $j < i$. In other words, no intermediate quantity depends on an intermediate quantity with a higher index.

As illustrated in Example 16.1, we may evaluate a directional derivative $F'(x)\,\delta x$ by propagating, alongside the values of the intermediate quantities $v_0, v_1, \ldots, v_N$, the values of their directional derivatives $\dot{v}_i := \frac{\partial v_i}{\partial x}(x)\,\delta x$ through the computational graph. Due to the chain rule, these values obey

$$\dot{v}_i = \sum_{j \prec i} \frac{\partial \varphi_i}{\partial v_j}(v_j)\,\dot{v}_j \quad \text{for } i = 1, \ldots, N. \tag{16.2}$$

We initialize the process by setting

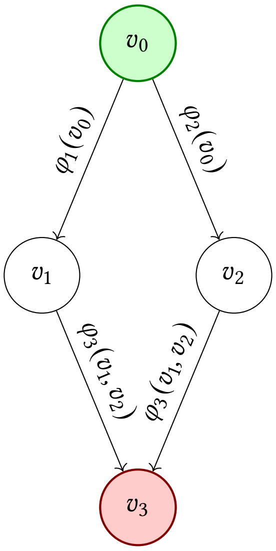
$$\dot{v}_0 := \delta x$$

and then implement (16.2) with an outer loop over the vertices v_i from $i = 1, \ldots, N$ and a (short) inner loop over the few vertices v_j that vertex v_i depends on. Due to $y = v_N$, we will obtain the desired derivative as $\dot{y} = \dot{v}_N = F'(x)\,\delta x$.

Remark 16.3 (on the forward mode of algorithmic differentiation).

(i) *The process of propagating directional derivatives through the computational graph simultaneously with the quantities needed for the evaluation of $F(x)$ is known as the* ***forward mode*** *of algorithmic differentiation. The quantities $\dot{v}_i$ are also known as* ***tangents****.*

(ii) *Mathematically speaking, the forward mode computes $F'(x)\,\delta x$, i. e., a Jacobian-vector product, by evaluating the components of the chain rule from the right. For example, in the graph*



we have (omitting arguments for readability)

$$\dot{v}_3 = \frac{\partial \varphi_3}{\partial v_1}\,\dot{v}_1 + \frac{\partial \varphi_3}{\partial v_2}\,\dot{v}_2 = \frac{\partial \varphi_3}{\partial v_1}\Big(\frac{\partial \varphi_1}{\partial v_0}\,\dot{v}_0\Big) + \frac{\partial \varphi_3}{\partial v_2}\Big(\frac{\partial \varphi_2}{\partial v_0}\,\dot{v}_0\Big). \tag{16.3}$$

(iii) *It is easily possible to evaluate several directional derivatives of $F'(x)\,\delta x$ at once. All we have to do is make δx a matrix with one column for each direction in the input space. We will then propagate all directional derivatives at once. This is known as the* ***vector forward mode****, and δx is called the* ***seed matrix****. For example, when $\delta x = \mathrm{Id}$, we will obtain the full Jacobian $\dot{y} = F'(x)\,\delta x = F'(x)$ instead of a single Jacobian-vector product.*

(iv) *A computational graph is an idealized model of a function that is realized as computer code. For instance, when variables in a code are overwritten, an acyclic computational graph representation requires new variable names to be generated for disambiguation. In addition, we often have constructs which make the conversion into a* static *computational graph impossible, e. g., loops with a variable number of iterations. Nonetheless, the forward mode of AD works the same way, even when the computational graph is developing only at run-time and when it depends on the input x.*

(v) *The smallest computational entity in a computational graph are the elemental functions φ_i. What precisely qualifies as an elemental function depends on the programming language and the libraries in use. What is important from the AD tool point of view is that the derivatives of an elemental function w.r.t. its input parameters are available. For instance, when the computational graph contains an elemental function which realizes $\varphi(A, b) = A^{-1}b$, i. e., the solution of a linear system, then we may need to provide the AD tool with functions that realize the partial directional derivatives*

$$\frac{\partial \varphi}{\partial A}(A, b)\,\delta A = -A^{-1}\,\delta A\,A^{-1}b \quad \textit{and} \quad \frac{\partial \varphi}{\partial b}(A, b)\,\delta b = A^{-1}\delta b.$$

*(**Quiz 16.2:** Can you confirm these formulas for the partial directional derivatives of φ?)*

The forward mode is not the only way to obtain derivatives of a function $y = F(x)$ using algorithmic differentiation. The **reverse mode** starts from the output $y = v_N$ and works its way backwards through the computational graph to evaluate the derivatives

$$\overline{v}_i := \delta y\,\frac{\partial y}{\partial v_i}(v_i, \ldots, v_N)$$

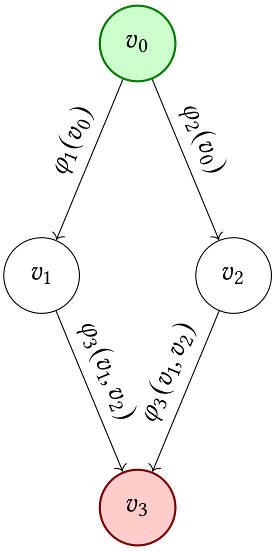
of y w.r.t. each of the intermediate quantities v_i. The quantities $\overline{v}_i$ are also known as **adjoints**. That is, the quantities propagated through the graph are the derivatives of the linear combination $\delta y\, y$ of the components of the output y w.r.t. the intermediate quantities v_i. Due to the chain rule, these values obey

$$\overline{v}_j = \sum_{i \succ j} \overline{v}_i\,\frac{\partial \varphi_i}{\partial v_j} \quad \text{for } j = 1, \ldots, N. \tag{16.4}$$

We initialize the process by setting

$$\overline{v}_N := \delta y.$$

For our simple example graph



we obtain (omitting arguments for readability)

$$\begin{aligned} \overline{y} &= \overline{v}_3 := \delta y \\ \overline{v}_2 &= \overline{v}_3 \, \frac{\partial \varphi_3}{\partial v_2} \\ \overline{v}_1 &= \overline{v}_3 \, \frac{\partial \varphi_3}{\partial v_1} \\ \overline{v}_0 &= \overline{v}_2 \, \frac{\partial \varphi_2}{\partial v_0} + \overline{v}_1 \, \frac{\partial \varphi_1}{\partial v_0}. \end{aligned}$$

This amounts to the formula

$$\delta y \, F'(x) = \left(\delta y \, \frac{\partial \varphi_3}{\partial v_2}\right) \frac{\partial \varphi_2}{\partial v_0} + \left(\delta y \, \frac{\partial \varphi_3}{\partial v_1}\right) \frac{\partial \varphi_1}{\partial v_0}. \tag{16.5}$$

We observe that the reverse mode computes $\delta y \, F'(x)$, i. e., a vector-Jacobian product, by evaluating the components of the chain rule from the left. Compare this to the forward mode representation

$$\dot{v}_3 = \frac{\partial \varphi_3}{\partial v_2} \left(\frac{\partial \varphi_2}{\partial v_0} \, \dot{v}_0\right) + \frac{\partial \varphi_3}{\partial v_1} \left(\frac{\partial \varphi_1}{\partial v_0} \, \dot{v}_0\right), \tag{16.3}$$

where the components of the chain rule are evaluated from the right.

We now work out a slightly more complex example for the reverse mode by hand by revisiting Example 16.1.

Example 16.4 (algorithmic differentiation by hand: reverse mode).

$$
\begin{aligned}
e &:= \mathrm{e}^x \\
c &:= \cos(x) \\
c_3 &:= c^3 \\
s &:= \sin(x) \\
s_3 &:= s^3 \\
d &:= c_3 + s_3 \\
y &:= \frac{e}{d} \\
\overline{y} &:= \delta y \\
&\textit{initialize } \overline{d}, \overline{s}_3, \overline{s}, \overline{c}_3, \overline{c} \textit{ and } \overline{e} \textit{ to zero} \\
\overline{e} &:= \overline{e} + \delta y \, \frac{\partial y}{\partial e} = \overline{e} + \overline{y} \, \frac{1}{d} \\
\overline{d} &:= \overline{d} + \delta y \, \frac{\partial y}{\partial d} = \overline{d} - \overline{y} \, \frac{e}{d^2} \\
\overline{c}_3 &:= \overline{c}_3 + \delta y \, \frac{\partial y}{\partial d} \, \frac{\partial d}{\partial c_3} = \overline{c}_3 + \overline{d} \, 1 \\
\overline{s}_3 &:= \overline{s}_3 + \delta y \, \frac{\partial y}{\partial d} \, \frac{\partial d}{\partial s_3} = \overline{s}_3 + \overline{d} \, 1 \\
\overline{s} &:= \overline{s} + \delta y \, \frac{\partial y}{\partial s_3} \, \frac{\partial s_3}{\partial s} = \overline{s} + \overline{s}_3 \, 3 \, s^2 \\
\overline{x} &:= \overline{x} + \delta y \, \frac{\partial y}{\partial s} \, \frac{\partial s}{\partial x} = \overline{x} + \overline{s} \, \cos(x) \\
\overline{c} &:= \overline{c} + \delta y \, \frac{\partial y}{\partial c_3} \, \frac{\partial c_3}{\partial c} = \overline{c} + \overline{c}_3 \, 3 \, c^2 \\
\overline{x} &:= \overline{x} + \delta y \, \frac{\partial y}{\partial c} \, \frac{\partial c}{\partial x} = \overline{x} + \overline{c} \, (-\sin(x)) \\
\overline{x} &:= \overline{x} + \delta y \, \frac{\partial y}{\partial e} \, \frac{\partial e}{\partial x} = \overline{x} + \overline{e} \, \mathrm{e}^x
\end{aligned}
$$

Remark 16.5 (on the reverse mode of algorithmic differentiation).

(*i*) *The reverse mode propagates the derivatives in reverse order through the computional graph, compared to the flow of data during the evaluation of $F(x)$ itself. Therefore, all intermediate quantities from the forward pass need to be stored! For example, near the end of the backward pass, we require access to the intermediate quantity c from near the beginning of the forward pass to evaluate $\overline{c} := \overline{c} + \overline{c}_3 \, 3 \, c^2$.*

(*ii*) *The chain rule*

$$
\overline{v}_j = \sum_{i \succ j} \overline{v}_i \, \frac{\partial \varphi_i}{\partial v_j} \quad \textit{for } j = 1, \ldots, N \tag{16.4}
$$

*is implemented with the outer loop over the vertices v_i from $i = N, \ldots, 1$, and the (short) inner loop over the few vertices v_j that vertex v_i depends on. (**Quiz 16.3:** Why can't the outer loop be over j and the inner loop over i?) This loop ordering means that at the time vertex v_i is visited in the*

	forward mode $F(x)$ plus $F'(x)\,\delta x$	reverse mode $F(x)$ plus $\delta y\, F'(x)$
computational effort (time)	$\le (1 + 1.5\,p)\,\text{effort}(F(x))$	$\le (1.5 + 2.5\,q)\,\text{effort}(F(x))$
memory requirement	$\le (1 + p)\,\text{memory}(F(x))$	$\le (1 + q)\,\text{memory}(F(x))$

Table 16.1: Computation effort and memory requirements for the (vector) forward and (vector) reverse mode of AD, where p denotes the number of columns in δx, and q denotes the number of rows in δy. See Griewank, Walther, 2008, Chapter 4 for details.

*backward pass, the values $\overline{v}_j$ of all vertices v_j that vertex v_i depends on are updated. At a later stage, the value $\overline{v}_j$ may be updated again. (**Quiz 16.4:** When does it happen that a value $\overline{v}_j$ is updated multiple times?) In Example 16.4, this happens only for vertex x, whose adjoint $\overline{x}$ is touched three times, while all other adjoints such as $\overline{e}$, $\overline{d}$ etc. are touched only once during the backward pass.*

(*iii*) *The **vector reverse mode** evaluates several vector-Jacobian products at once. All we have to do is make δy a matrix with one row for each direction in the output space. Again, δy is called the **seed matrix** for the reverse mode. When $\delta y = \mathrm{Id}$, we will obtain the full Jacobian $\overline{y} = \delta y\, F'(x) = F'(x)$ instead of a single vector-Jacobian product.*

(*iv*) *We emphasize again that the intermediate values v_i as well as their tangents $\dot{v}_i$ (forward mode) or adjoints $\overline{v}_i$ (reverse mode) need not necessarily be scalars (as they were in our Examples 16.1 and 16.4), but they can be vectors, matrices, or any other data type representing numerical values.*

Let us briefly discuss the numerical effort incurred by the forward and the reverse mode of AD to see in which case which variant is to be preferred. Table 16.1 summarizes the results in simplified form. More details can be found in Griewank, Walther, 2008, Chapter 4.

It can be shown that the evaluation of $F(x)$ plus a single Jacobian-vector product $F'(x)\,\delta x$ has an effort which is a small multiple of the effort to evaluate $F(x)$ alone. The effort grows linearly with every additional Jacobian-vector product, i. e., every column of the seed matrix δx. Therefore, an evaluation of the full Jacobian of a function $F\colon \mathbb{R}^n \to \mathbb{R}^m$ via the forward mode has a time and memory complexity proportional to the dimension n of the input, relative to the complexity of evaluating $F(x)$ alone.

Similarly, the evaluation of $F(x)$ plus a single vector-Jacobian product $\delta y\, F'(x)$ also has an effort which is a small multiple of the effort to evaluate $F(x)$ alone. This time, however, the effort grows linearly with every additional row of the seed matrix δy. Consequently, an evaluation of the full Jacobian of a function $F\colon \mathbb{R}^n \to \mathbb{R}^m$ via the reverse mode has a time and memory complexity proportional to the dimension m of the output, relative to the complexity of evaluating $F(x)$ alone.

For instance, the derivative of a scalar-valued function $f\colon \mathbb{R}^n \to \mathbb{R}$ ($m = 1$!) should clearly be obtained using the reverse mode! By the way, the famous **backpropagation** in the training of neural networks is nothing but the reverse mode of AD, applied to the scalar-valued loss (objective) function.

Algorithmic differentiation can be implemented via

(1) **source transformation**, where the AD tool takes code realizing a function $F\colon \mathbb{R}^n \to \mathbb{R}^m$ as input and returns augmented code to evaluate $F(x)$ plus $F'(x)\,\delta x$ (forward mode[2]), or $F(x)$ plus $\delta y\, F'(x)$ (reverse mode[3]) as output,

(2) or via **operator overloading**, where the AD tool overloads all algorithmic operations of the programming language and augments them with derivative functionality.

There is much more to be said about algorithmic differentiation, its efficient implementation, floating point error analysis, higher-order derivatives etc. Suffice it to mention that numerous stand-alone AD tools exist for a range of programming languages, while other AD tools are included in libraries devoted, for instance, to optimization and machine learning. A good source is `https://autodiff.org`.

End of Week 14

[2] The code that implements the forward mode derivative is also known as **tangent code** or **tangent model**.
[3] The code that implements the reverse mode derivative is also known as **adjoint code** or **adjoint model**.

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.

Alt, W. (2002). *Nichtlineare Optimierung*. Vieweg Studium: Aufbaukurs Mathematik. Eine Einführung in Theorie, Verfahren und Anwendungen. [An introduction to theory, procedures and applications]. Friedrich Vieweg & Sohn, Braunschweig. DOI: 10.1007/978-3-322-84904-5.

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.

Bertsekas, D. P. (1999). *Nonlinear Programming*. Belmont: Athena Scientific.

Bischof, C. H.; P. D. Hovland; B. Norris (2008). “On the implementation of automatic differentiation tools”. *Higher-Order and Symbolic Computation* 21.3, pp. 311–331. DOI: 10.1007/s10990-008-9034-4.

Blum, E.; W. Oettli (1972). “Direct proof of the existence theorem for quadratic programming”. *Operations Research* 20, pp. 165–167. DOI: 10.1287/opre.20.1.165.

Burke, J. V. (2014). *Nonlinear optimization*. Lecture notes, Math 408, University of Washington, Seattle, WA. URL: https://sites.math.washington.edu/~burke/crs/408/notes/Math408_Spring2014/math408text.pdf.

Cartan, H. (1971). *Differential Calculus*. Translated from the French. Hermann, Paris; Houghton Mifflin Co., Boston, Massachusetts.

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.

Conn, A. R.; N. I. M. Gould; P. L. Toint (2000). *Trust-Region Methods*. Philadelphia: SIAM. DOI: 10.1137/1.9780898719857.

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.

Dennis Jr., J. E.; J. J. Moré (1974). “A characterization of superlinear convergence and its application to quasi-Newton methods”. *Mathematics of Computation* 28, pp. 549–560. DOI: 10.1090/s0025-5718-1974-0343581-1.

Dontchev, A. L.; R. T. Rockafellar (1996). “Characterizations of strong regularity for variational inequalities over polyhedral convex sets”. *SIAM Journal on Optimization* 6.4, pp. 1087–1105. DOI: 10.1137/s1052623495284029.

Elman, H. C.; D. J. Silvester; A. J. Wathen (2014). *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*. 2nd ed. Numerical Mathematics and Scientific Computation. Oxford University Press. DOI: 10.1093/acprof:oso/9780199678792.001.0001.

Fischer, A. (1992). "A special Newton-type optimization method". *Optimization. A Journal of Mathematical Programming and Operations Research* 24.3-4, pp. 269–284. DOI: 10.1080/02331939208843795.

Flegel, M. L.; C. Kanzow (2005). "Abadie-type constraint qualification for mathematical programs with equilibrium constraints". *Journal of Optimization Theory and Applications* 124.3, pp. 595–614. DOI: 10.1007/s10957-004-1176-x.

Fletcher, R.; S. Leyffer (2002). "Nonlinear programming without a penalty function". *Mathematical Programming* 91, pp. 239–269. DOI: 10.1007/s101070100244.

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.

Frank, M.; P. Wolfe (1956). "An algorithm for quadratic programming". *Naval Research Logistics Quarterly* 3, pp. 95–110. DOI: 10.1002/nav.3800030109.

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

Geiger, C.; C. Kanzow (2002). *Theorie und Numerik restringierter Optimierungsaufgaben*. New York: Springer. DOI: 10.1007/978-3-642-56004-0.

Gilbert, J. C.; J. Nocedal (1992). "Global convergence properties of conjugate gradient methods for optimization". *SIAM Journal on Optimization* 2.1, pp. 21–42. DOI: 10.1137/0802003.

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.

Gould, N.; M. Hribar; J. Nocedal (2001). "On the solution of equality constrained quadratic problems arising in optimization". *SIAM Journal on Scientific Computing* 23.4, pp. 1375–1394. DOI: 10.1137/s1064827598345667.

Griewank, A.; A. Walther (2008). *Evaluating Derivatives*. 2nd ed. Principles and techniques of algorithmic differentiation. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM). DOI: 10.1137/1.9780898717761.

Guignard, M. (1969). "Generalized Kuhn-Tucker conditions for mathematical programming in a Banach space". *SIAM Journal on Control and Optimization* 7.2, pp. 232–241. DOI: 10.1137/0307016.

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

Hestenes, M. R.; E. Stiefel (1952). "Methods of conjugate gradients for solving linear systems". *Journal of Research of the National Bureau of Standards* 49, 409–436 (1953). DOI: 10.6028/jres.049.044.

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

Izmailov, A. F.; A. S. Kurennoy; M. V. Solodov (2012). "The Josephy–Newton method for semismooth generalized equations and semismooth SQP for optimization". *Set-Valued and Variational Analysis* 21.1, pp. 17–45. DOI: 10.1007/s11228-012-0218-z.

John, F. (1948). "Extremum problems with inequalities as subsidiary conditions". *Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948*. Interscience Publishers, Inc., New York, NY, pp. 187–204. DOI: 10.1007/978-1-4612-5412-6_25.

Josephy, N. H. (1979a). *Newton's method for generalized equations*. MRC Technical Summary Report 1965. University of Wisconsin–Madison. URL: https://apps.dtic.mil/sti/citations/ADA077096.

Josephy, N. H. (1979b). "Newton's Method for Generalized Equations and the PIES Energy Model". PhD thesis. University of Wisconsin–Madison.

Karush, W. (1939). “Minima of Functions of Several Variables with Inequalities as Side Constraints”. M.Sc. Thesis. Department of Mathematics, University of Chicago.

Kosmol, P. (1989). *Methoden zur numerischen Behandlung nichtlinearer Gleichungen und Optimierungsaufgaben*. Teubner Studienbücher Mathematik. [Teubner Mathematical Textbooks]. B. G. Teubner, Stuttgart. DOI: 10.1007/978-3-663-12239-5.

Kuhn, H. W.; A. W. Tucker (1951). “Nonlinear programming”. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*. Berkeley and Los Angeles: University of California Press, pp. 481–492.

Lyness, J. N.; C. B. Moler (1967). “Numerical differentiation of analytic functions”. *SIAM Journal on Numerical Analysis* 4.2, pp. 202–210. DOI: 10.1137/0704019.

Maratos, N. (1978). “Exact Penalty Function Algorithms for Finite Dimensional and Control Optimization Problems”. PhD thesis. Imperial College, London. URL: https://hdl.handle.net/10044/1/7283.

Martínez, J. M. (1994). “Local minimizers of quadratic functions on Euclidean balls and spheres”. *SIAM Journal on Optimization* 4.1, pp. 159–176. DOI: 10.1137/0804009.

Mifflin, R. (1977). “Semismooth and semiconvex functions in constrained optimization”. *SIAM Journal on Control and Optimization* 15.6, pp. 959–972. DOI: 10.1137/0315061.

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.

Omojokun, E. O. (1989). “Trust Region Algorithms for Optimization with Nonlinear Equality and Inequality Constraints”. PhD thesis. University of Colorado, Boulder.

Pang, J.-S.; L. Q. Qi (1993). “Nonsmooth equations: motivation and algorithms”. *SIAM Journal on Optimization* 3.3, pp. 443–465. DOI: 10.1137/0803021.

Powell, M. J. D. (1978). “A fast algorithm for nonlinearly constrained optimization calculations”. *Numerical Analysis (Proceedings of the Biennial Conference held at Dundee, June 28–July 1, 1977)*. Vol. 630. Lecture Notes in Mathematics. Springer, Berlin, pp. 144–157. DOI: 10.1007/BFb0067703.

Powell, M. J. D. (1986). “Convergence properties of algorithms for nonlinear optimization”. *SIAM Review* 28.4, pp. 487–500. DOI: 10.1137/1028154.

Qi, L. Q. (1993). “Convergence analysis of some algorithms for solving nonsmooth equations”. *Mathematics of Operations Research* 18.1, pp. 227–244. DOI: 10.1287/moor.18.1.227.

Qi, L.; J. Sun (1993). “A nonsmooth version of Newton’s method”. *Mathematical Programming* 58.1–3, pp. 353–367. DOI: 10.1007/bf01581275.

Robinson, S. (1980). “Strongly regular generalized equations”. *Mathematics of Operations Research* 5.1, pp. 43–62. DOI: 10.1287/moor.5.1.43.

Squire, W.; G. Trapp (1998). “Using complex variables to estimate derivatives of real functions”. *SIAM Review* 40.1, pp. 110–112. DOI: 10.1137/s003614459631241x.

Steihaug, T. (1983). “The conjugate gradient method and trust regions in large scale optimization”. *SIAM Journal on Numerical Analysis* 20, pp. 626–637. DOI: 10.1137/0720042.

Toint, P. (1981). “Towards an efficient sparsity exploiting newton method for minimization”. *Sparse Matrices and Their Uses*. Ed. by I. S. Duff. Based on the Proceedings of the IMA Numerical Analysis Group Conference, organised by the Institute of Mathematics and Its Applications and held at the University of Reading, 9th–11th July, 1980. London: Academic Press, pp. 57–88.

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

Ziemer, W. P. (1989). *Weakly Differentiable Functions.* Vol. 120. Graduate Texts in Mathematics. Sobolev spaces and functions of bounded variation. New York: Springer-Verlag. DOI: 10.1007/978-1-4612-1015-3.