Berlin School of Mathematics – Basic Course

Nonlinear Optimization

Part I: Unconstrained and box-constrained problems

Prof. Dr. Michael Hintermüller

Humboldt-University of Berlin
Department of Mathematics
John-von-Neumann Haus, Room 2.426
Rudower Chaussee 25, Berlin-Adlershof
hint@mathematik.hu-berlin.de

Contents

Acknowledgement

These lecture notes grew out of several courses which I held at the Karl-Franzens University of Graz and the University of the Philippines in Manila, respectively.
For the careful typesetting of all the proofs in my original manuscript (and for the tedious task of deciphering my hand-writing) I would like to express my sincere thanks to Mag. Cornelia Kulmer. Mag. Ian Kopacka was invaluable for the input he gave and for tracing typos in an earlier version of the script.
These lecture notes are largely based on the monographs listed in the bibliography.

CHAPTER 1

Introduction

1. Preface and motivation

The following task is known as a finite dimensional *minimization problem* :

$$\text{(1.1)} \qquad \begin{cases} \textit{Let } X \subset \mathbb{R}^n \textit{ an arbitrary set and } f : X \to \mathbb{R} \textit{ a continuous function.} \\ \textit{The problem is to find an } x^* \in X \textit{ such that} \\ \qquad\qquad f(x^*) \leq f(x) \quad \text{for all } x \in X. \end{cases}$$

Using a more compact notation we write:

$$\min \quad f(x) \quad \text{s.t.} \quad x \in X,$$

where "s.t." stands for "subject to", or

$$\text{(1.2)} \qquad \min_{x \in X} \quad f(x).$$

If $X = \mathbb{R}^n$, then (1.1) (resp. (1.2)) is called *unconstrained*, otherwise *constrained*. In general X is called the *feasible set* and f the *objective function*.

REMARK 1.1. Maximizing $f(x)$ for $x \in X$ is equivalent to minimizing $-f(x)$ s.t. $x \in X$. Therefore we can restrict ourselves to problems of the form (1.1) or (1.2), resp.

Being a mathematical model for several problems, e.g., in physics, medicine, economy and engineering science, problem (1.1) is of great importance.

EXAMPLE 1. In many cases one is interested in computing certain parameters by means of observation (measurements) of the corresponding system. Typically, the difference between measured data and data based on the computations should be minimal.
Let M be a mass point with mass m which is fixated on a vertical spring lying on a vertical y-axis. If the spring is relaxed, M is located at the origin (equilibrium position). If M is displaced, the (compressed or expanded) spring applies a restoring force K which tries to replace M in its equilibrium position. For small displacements of y, the force can accurately be modeled by Hooke's law $K = -\hat{k}y$, where $\hat{k}$ is a positive spring constant. If the position of M at time t is denoted by $y(t)$, then (neglecting damping or friction) , according to Newton's law:

$$\text{(1.3)} \qquad m\ddot{y} = -\hat{k}y,$$

which is called the undamped harmonic oscillator equation. In most cases friction- and damping forces behave proportionally to the velocity of M, *i.e.* $-r\dot{y}$ with fixed $r > 0$. Together with (1.3) we obtain

$$m\ddot{y} + r\dot{y} + \hat{k}y = 0.$$

Setting $c := r/m$, $k := \hat{k}/m$ we get

$$\text{(1.4)} \qquad \ddot{y} + c\dot{y} + ky = 0.$$

Let us assume that at time $t = 0$ the displacement is $y(0) = y_0$ and furthermore $\dot{y}(0) = 0$, then the following initial conditions hold true:

$$y(0) = y_0, \quad \dot{y}(0) = 0 \tag{1.5}$$

In the following we will concentrate on the time interval $[0, T]$. Let $\{y^j\}_{j=1}^N$ be measurements of the spring's deviation at time instances $t_j = (j-1)T/(N-1)$. The objective is to determine the spring constant k and the damping factor c with the help of measurements.
Let $x = (c, k)^\top$. To emphasize the dependence of $y(t)$ on x, we also write $y(x; t)$. Following the motivation at the beginning of the example, we try to solve the following unconstrained non-linear minimization problem:

$$\min_{x \in \mathbb{R}^2} \quad f(x) := \frac{1}{2} \sum_{j=1}^{N} |y(x; t_j) - y^j|^2. \tag{1.6}$$

It should be mentioned that y is differentiable with respect to x if $c^2 - 4k \neq 0$ holds true. Problem (1.6) aims at minimizing the sum of the squares of the errors ("nonlinear least squares problem").

A further simple example taken from economy could be as follows:

Example 2. In a company the following fictional relation between the output quantity x and the corresponding total costs is found, using "least squares" estimations:

$$K(x) = K_v(x) + K_f(x).$$

Here $K_v(x)$ denotes the variable costs for output quantity x. Moreover there are fixed costs (lease rental charges, ...) in the amount of $K_f(x) = c$, $c > 0$. Normally one is looking for x^* minimizing the total costs of $K(x)$, i.e.,

$$x^* = \operatorname{argmin}\{K_f(x) + K_v(x) : x \in \mathbb{R}\} = \operatorname{argmin}\{K_v(x) : x \in \mathbb{R}\}, \tag{1.7}$$

which is equivalent to finding an x^* solving

$$\min_{x \in \mathbb{R}} \quad K_v(x).$$

For general problems one cannot expect the set $\operatorname{argmin}\{K_f(x) + K_v(x) : x \in \mathbb{R}\}$ to contain just one element. However, if K_v is uniformly convex (see section 2), then the first equality in (1.7) is justified.

In case that $X \neq \mathbb{R}^n$ is the feasible set, it can often be written in the form

$$X = X_1 \cap X_2 \cap X_3$$

with sets

$$\begin{aligned} X_1 &= \{x \in \mathbb{R}^n : c_i(x) = 0,\ i \in I_1\}, \\ X_2 &= \{x \in \mathbb{R}^n : c_i(x) \leq 0,\ i \in I_2\}, \\ X_3 &= \{x \in \mathbb{R}^n : x_i \in \mathbb{Z},\ i \in I_3\}. \end{aligned}$$

Here $I_1, I_2, I_3 \subset \mathbb{N}$ are finite index sets. The sets X_1, X_2 and X_3 are called *equality-, inequality- and integer constraints.*

Remark 1.2. In example 1 we have

$$X = \big\{x \in \mathbb{R}^2 : x_i \geq 0,\ i \in \{1, 2\}\big\}.$$

If X is a set of discrete points, one refers to (1.1) as a *discrete* (or *combinatorial*) optimization problem; otherwise the optimization problem is called *continuous*.
Occasionally, f is not differentiable, then we call (1.1) a *nondifferentiable* optimization problem (this would also be the case if one of the c_i's in X_1 or X_2 would be non-differentiable, even if f would be differentiable).

REMARK 1.3. For instance, replacing the objective function $f(x)$ in example 1 by

$$g(x) = \sum_{j=1}^{N} |y(x; t_j) - y_j|,$$

we obtain a nondifferentiable optimization problem.

2. Notions of solutions

In the following definition we introduce our basic notions of optimality.

DEFINITION 1.1. *Let $f : X \to \mathbb{R}$ with $X \subset \mathbb{R}^n$. The point $x^* \in X$ is called a*

(i) *(strict) global minimizer of f (on X), if and only if*

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

The optimal objective value $f(x^*)$ is called a (strict) global minimum;

(ii) *(strict) local minimizer of f (on X), if there exists a neighborhood U of x^* such that*

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

The optimal objective value $f(x^*)$ is called a (strict) local minimum.

REMARK 1.4. The point x^* is a ((strict) global, (strict) local) maximizer of f (on X), if and only if x^* is a ((strict) global, (strict) local) minimizer of $-f$ (on X).

In the following the gradient of f in x is denoted by

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

DEFINITION 1.2. *Let $X \subset \mathbb{R}^n$ be an open set and $f : X \to \mathbb{R}$ be a continuously differentiable function. The point $x^* \in X$ is called a stationary point of f, if*

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

holds true.

CHAPTER 2

Optimality conditions

This chapter deals with *necessary* and *sufficient* conditions for characterizing minimizers (under certain differentiability assumptions on f).

1. General case

If f does not possess any structure nor properties apart from differentiability, then we can only make statements about local minimizers, in general.

THEOREM 2.1. *Let $X \subset \mathbb{R}^n$ an open set and $f : X \to \mathbb{R}$ a continuously differentiable function. If $x^* \in X$ is a local minimizer of f (on X), then*

$$\nabla f(x^*) = 0, \tag{2.1}$$

i.e., x^* is a stationary point.

PROOF. We prove the statement by means of contradiction. Let us assume that x^* is a local minimizer for which $\nabla f(x^*) \neq 0$ does not hold true. Then there exists $d \in \mathbb{R}^n$ with

$$\nabla f(x^*)^\top d < 0 \quad \text{(choose for instance } d = -\nabla f(x^*)\text{)}.$$

By assumption, f is continuously differentiable. Consequently, the directional derivative of f at x^* in direction d exists:

$$f'(x^*; d) = \lim_{\alpha \downarrow 0} \frac{f(x^* + \alpha d) - f(x^*)}{\alpha} = \nabla f(x^*)^\top d < 0.$$

Due to the continuity of the derivative there exists $\bar{\alpha} > 0$ satisfying $x^* + \alpha d \in X$ (X is open) and

$$\frac{f(x^* + \alpha d) - f(x^*)}{\alpha} < 0$$

for all $0 < \alpha \leq \bar{\alpha}$. Therefore, it holds that

$$f(x^* + \alpha d) < f(x^*) \quad \forall 0 < \alpha \leq \bar{\alpha},$$

which contradicts the assumption that x^* is a local minimizer of f in X. □

REMARK 2.1.
(1) Since Theorem 2.1 only uses first order derivatives and assumes x^* to be a (local) minimizer, it specifies a *first order necessary condition*.
(2) The condition $\nabla f(x^*) = 0$ is **not** sufficient for a local minimum; consider, e.g., $f(x) = -x^2$ with $x^* = 0$.

As preparation for the next Theorem 2.2 we need the following lemma about the continuity of the smallest eigenvalue of a matrix.

LEMMA 2.1. *Let $\mathcal{S}_n$ be the vector space of symmetrical $n \times n$-matrices. For $A \in \mathcal{S}_n$ let $\lambda(A) \in \mathbb{R}$ be the smallest eigenvalue of A. Then the following estimate holds true:*

$$|\lambda(A) - \lambda(B)| \leq \|A - B\| \quad \textit{for all } A, B \in \mathcal{S}_n.$$

Note that the vector norm and the matrix norm are denoted by the same symbol, *i.e.* $\|\cdot\|$. If f is twice continuously differentiable it follows from Lemma 2.1 and from the continuity of $\nabla^2 f \in \mathbb{R}^{n\times n}$ (the Hessian of f), that $\nabla^2 f(x)$ is positive definite in a neighborhood of x^* if $\nabla^2 f(x^*)$ is positive definite. An analogue statement holds true, if $\nabla^2 f(x^*)$ is negative definite.

THEOREM 2.2. *Let $X \subset \mathbb{R}^n$ be open and $f : X \to \mathbb{R}$ be twice continuously differentiable. If $x^* \in X$ is a local minimizer of f (on X), then $\nabla f(x^*) = 0$ and the Hessian $\nabla^2 f(x^*)$ is positive semi-definite.*

PROOF. The statement that $\nabla f(x^*) = 0$ holds true, follows from Theorem 2.1. Therefore, we only have to consider the positive-semi-definiteness of the Hessian of f at x^*. Again we prove the statement by means of contradiction. Let us assume that x^* is a local minimizer of f, but $\nabla^2 f(x^*)$ is not positive semi-definite. Then there exists $d \in \mathbb{R}^n$ such that

$$d^\top \nabla^2 f(x^*) d < 0. \tag{2.2}$$

Applying Taylor's theorem, we obtain for sufficiently small $\alpha > 0$:

$$f(x^* + \alpha d) = f(x^*) + \alpha \nabla f(x^*)^\top d + \frac{\alpha^2}{2} d^\top \nabla^2 f(\xi(\alpha)) d = f(x^*) + \frac{\alpha^2}{2} d^\top \nabla^2 f(\xi(\alpha)) d, \tag{2.3}$$

where we used $\nabla f(x^*) = 0$ and the existence of $\vartheta = \vartheta(\alpha) \in (0,1)$ with $\xi(\alpha) = x^* + \vartheta \alpha d \in X$. By Lemma 2.1 and (2.2) there exists $\bar{\alpha} > 0$, such that

$$d^\top \nabla^2 f(\xi(\alpha)) d < 0 \quad \forall 0 < \alpha \leq \bar{\alpha}.$$

Now, (2.3) yields

$$f(x^* + \alpha d) < f(x^*) \quad \forall 0 < \alpha \leq \bar{\alpha},$$

which contradicts the assumption that x^* is a local minimizer of f on X. □

REMARK 2.2.
(1) The conditions of Theorem 2.1 and Theorem 2.2 are **not** sufficient for local minimality; consider, e.g., $f(x) = x_1^2 - x_2^4$ with $x^* = (0,0)^\top$.
(2) As Theorem 2.2 involves second order derivatives and assumes x^* to be a (local) minimizer, it defines *second order necessary conditions*.

The subsequent theorem specifies *second order sufficient conditions*: If conditions (a) and (b) of Theorem 2.3 are satisfied at the point x^*, then x^* is a strict local minimizer of f on X.

THEOREM 2.3. *Let $X \subset \mathbb{R}^n$ be open and $f : X \to \mathbb{R}$ twice continuously differentiable. If*
(a) $\nabla f(x^*) = 0$ *and*
(b) $\nabla^2 f(x^*)$ *is positive definite,*
then x^* is a strict *local minimizer of f (on X).*

PROOF. Assumption (b) ensures that $\lambda(\nabla^2 f(x^*)) > 0$, i.e., the smallest eigenvalue of the Hessian of f in x^* is positive. Therefore it holds:

$$d^\top \nabla^2 f(x^*) d \geq \mu d^\top d = \mu \|d\|^2 \quad \forall d \in \mathbb{R}^n,$$

for $0 < \mu \leq \lambda(\nabla^2 f(x^*))$. From Taylor's Theorem we obtain for all d sufficiently close to 0: $x^* + d \in X$ and

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

with $\xi(d) = x^* + \vartheta d$ for $\vartheta = \vartheta(d) \in (0,1)$. Applying (a) and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} f(x^*+d) &= f(x^*) + \frac{1}{2} d^\top \nabla f(x^*) d + \frac{1}{2} d^\top \left(\nabla^2 f(\xi(d)) - \nabla^2 f(x^*) \right) d \\ &\geq f(x^*) + \frac{1}{2} \left(\mu - \|\nabla^2 f(\xi(d)) - \nabla^2 f(x^*)\| \right) \|d\|^2. \end{aligned}$$

Given that $\nabla^2 f$ is continuous, we are able to choose d small enough, such that $\|\nabla^2 f(\xi(d)) - \nabla^2 f(x^*)\| \leq \frac{\mu}{2}$ holds true. Thus,

$$f(x^*+d) \geq f(x^*) + \frac{\mu}{4}\|d\|^2 > f(x^*)$$

for all sufficiently small $d \in \mathbb{R}^n$ with $d \neq 0$. Hence x^* is a strict local minimizer of f on X. □

Remark 2.3. (1) Conditions (a) and (b) in Theorem 2.3 are **not** necessary for the local minimality of x^*; consider $f(x) = x_1^2 + x_2^4$ with $x^* = (0,0)^\top$. To some extent there is a "gap" between necessary and sufficient conditions.

(2) Given (a) of Theorem 2.3 in the case of an indefinite Hessian $\nabla^2 f(x^*)$, we refer to x^* as a *saddle point*.

2. Convex functions

Convex functions are of particular importance for optimization. For a convex function f we are able to show that the first order necessary conditions are also sufficient for local optimality (see Theorem 2.6) . In the following we will introduce procedures that approximate a complicated non-linear minimization problem by a sequence of convex problems. Apart from global properties, these convex problems offer a simple way of computing solutions or approximations.

Definition 2.1. (1) *A set $X \subset \mathbb{R}^n$ is called convex, if for all $x, y \in X$ and all $\lambda \in (0,1)$*

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

i.e. the segment $[x, y]$ lies completely in X.

(2) *Let $X \subset \mathbb{R}^n$ convex. A function $f : X \to \mathbb{R}$ is called*

(i) *(strictly) convex (in X), if for all $x, y \in X$ and for all $\lambda \in (0,1)$ the following holds true:*

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &\leq \lambda f(x) + (1-\lambda) f(y) \\ \big(f(\lambda x + (1-\lambda)y) &< \lambda f(x) + (1-\lambda) f(y),\ x \neq y\big). \end{aligned}$$

Geometrically, the (strict) convexity of f means that the line segment between $f(x)$ and $f(y)$ is located (strictly) above the graph of f.

(ii) *uniformly convex (on X), if there exists $\mu > 0$ with*

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

for all $x, y \in X$ and all $\lambda \in (0,1)$. (In that case, f is sometimes called uniformly convex with module(us) μ)

By definition, every uniformly convex function is also strictly convex and every strictly convex function is also convex. The converse is not true in general!

REMARK 2.4. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a quadratic function, i.e.,

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

with $A \in \mathcal{S}_n$, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then the following statements hold true:

(a) f is convex $\iff$ A positive semi-definite.
(b) f is strictly convex $\iff$ f uniformly convex $\iff$ A is positive definite.

If the (strictly, uniformly) convex function f is continuously differentiable, the following characterizations, referring to the graph of the function and the tangential hyperplane, hold true.

LEMMA 2.2. *Let $X \subset \mathbb{R}^n$ open, convex and $f : X \to \mathbb{R}$ continuously differentiable. Then the following assertions hold true:*

(a) *f is convex (on X) if and only if for all $x, y \in X$ there holds:*

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y). \tag{2.4}$$

(b) *f is strictly convex (on X) if and only if for all $x, y \in X$, with $x \neq y$, there holds:*

$$f(x) > f(y) + \nabla f(y)^\top (x - y). \tag{2.5}$$

(c) *f is uniformly convex (on X) if and only if there exists $\mu > 0$ such that*

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y) + \mu \|x - y\|^2 \tag{2.6}$$

for all $x, y \in X$.

PROOF. We prove (a) and assume that (2.4) is satisfied. If $x, y \in X$ and $\lambda \in [0, 1]$ are arbitrarily chosen and $z := \lambda x + (1 - \lambda)y$, then the following holds true:

$$\begin{aligned} f(x) &\geq f(z) + \nabla f(z)^\top (x - z), \\ f(y) &\geq f(z) + \nabla f(z)^\top (y - z). \end{aligned}$$

If we multiply the first inequality by $\lambda > 0$ and the second by $(1 - \lambda) > 0$, then the addition of the two terms provide the following:

$$\lambda f(x) + (1 - \lambda) f(y) - f(z) \geq \nabla f(z)^\top \left(\lambda x + (1 - \lambda)y - z\right) = 0.$$

Hence

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

which proves the convexity of f. Conversely, for all $x, y \in X$ and $\lambda \in (0, 1)$, the convexity of f provides

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

and

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

Given that f is continuously differentiable and considering the limit $\lambda \downarrow 0$ in the inequality above, yields condition (2.4), i.e.

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

Now, we turn to (b). The verification of the strict convexity of f under (2.5) is analogous to (a). However, the reverse direction cannot be shown by applying the same arguments as in (a), since the application of the limit does not ensure the strict inequality. Hence, we assume

that f is strictly convex. Since a strictly convex function is also convex, we already have the result from (a). For $z := \frac{1}{2}(x+y)$, the inequality of (2.4) yields

$$\nabla f(y)^\top (x-y) = 2\nabla f(y)^\top (z-y) \leq 2\left(f(z) - f(y)\right).$$

If $x \neq y$ holds true then the strict convexity of f implies that $2f(z) < f(x) + f(y)$. Thanks to the relation above, we deduce that

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

which corresponds to (2.5).
Taking into account the quadratic term, the proof of (c) is completely analogous to (a). □

Now we provide a characterization of twice continuously differentiable (strictly, uniformly) convex functions, enabling us to read off the convexity qualities of f from the definiteness of the Hessian of f.

THEOREM 2.4. *Let $X \subset \mathbb{R}^n$ an open, convex set and $f : X \to \mathbb{R}$ twice continuously differentiable. Then the following statements hold true:*

(a) *f is convex (on X) if and only if $\nabla^2 f(x)$ is positive semi-definite for all $x \in X$.*
(b) *If $\nabla^2 f(x)$ is positive definite for all $x \in X$, then f is strictly convex (on X).*
(c) *f is uniformly convex (on X) if and only if $\nabla^2 f(x)$ is uniformly positive definite on X, i.e., if there exists $\mu > 0$ such that*

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

for $x \in X$ and for all $d \in \mathbb{R}^n$.

PROOF. We start with (a) and assume that f is convex (on X). Due to the assumption that f is twice continuously differentiable, the application of Taylor's Theorem yields the following equation:

$$f(y) = f(x) + \nabla f(x)^\top (y-x) + \frac{1}{2}(y-x)^\top \nabla^2 f(x)(y-x) + r(y-x) \tag{2.7}$$

for all $y \in X$ sufficiently close to x. The remainder term has the following property: $r(y-x)/\|y-x\|^2 \to 0$ for $y \to x$. Now we choose $y = x + \alpha d$, where $d \in \mathbb{R}^n$ is arbitrary and $\alpha > 0$ is sufficiently small. Lemma 2.2 (a) yields

$$0 \leq \frac{\alpha^2}{2} d^\top \nabla^2 f(x) d + r(\alpha d).$$

Dividing by $\alpha^2/2$ and considering the limit for $\alpha \downarrow 0$ we obtain

$$0 \leq d^\top \nabla^2 f(x) d.$$

Since $x \in X$ and $d \in \mathbb{R}^n$ were chosen arbitrarily, the statement holds true. Conversely: Given that f is twice continuously differentiable with $\nabla^2 f(x)$ positive semi-definite for all $x \in X$, the subsequent equation follows from Taylor's theorem and the mean value theorem

$$f(y) = f(x) + \nabla f(x)^\top (y-x) + \frac{1}{2}\int_0^1 (y-x)^\top \nabla^2 f(x+\tau(y-x)))(y-x) d\tau. \tag{2.8}$$

The positive semi-definiteness of $\nabla^2 f$ yields

$$f(y) \geq f(x) + \nabla f(x)^\top (y-x)$$

for $y, x \in X$. Then the convexity of f on X follows from Theorem 2.2 (a).

The proof of (b) works analogously to the second part of proof (a). Now let us turn to the verification of (c). We assume that f is uniformly convex. Then, analogously to (a), we obtain (2.7). Theorem 2.2 (c) with $y = x + \alpha d$, where $d \in \mathbb{R}^n$ and $\alpha > 0$ is sufficiently small, provides

$$\mu\alpha^2\|d\|^2 \leq \frac{\alpha^2}{2} d^\top \nabla^2 f(x) d + r(\alpha d).$$

Dividing by α^2 and considering the limit for $\alpha \downarrow 0$ gives

$$\mu\|d\|^2 \leq \frac{1}{2} d^\top \nabla^2 f(x) d$$

for an arbitrary $d \in \mathbb{R}^n$, which proves the assertion. Conversely: Let $\nabla^2 f$ be uniformly positive definite (with modulus $\mu > 0$). Then the assertion follows from relation (2.8),

$$\int_0^1 (y-x)^\top \nabla^2 f(x + \tau(y-x))(y-x) d\tau \geq \mu\|x-y\|^2$$

and Theorem 2.2 (c). □

Note that the statement (b) of Theorem 2.4 can not be reversed in general; consider, e.g., $f(x) = x^4$ in $\mathbb{R}$.
The following lemma deals with the level sets of uniformly convex functions. In the context of general continuously differentiable functions, the statement of Lemma 2.3 is of local importance, *i.e.*, in a neighborhood of the local minimizer x^* of f (in X).

LEMMA 2.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and $x^0 \in \mathbb{R}^n$ arbitrary. Further assume that the level set*

$$L(x^0) := \{x \in \mathbb{R}^n : f(x) \leq f(x^0)\}$$

is convex and that f is uniformly convex in $L(x^0)$. Then the set $L(x^0)$ is compact.

PROOF. First note that due to our construction $L(x^0) \neq \emptyset$ holds true. Let $x \in L(x^0)$. From the uniform convexity of f in $L(x^0)$ we obtain (with $\lambda = \frac{1}{2}$)

$$\frac{\mu}{4}\|x - x^0\|^2 + f(\tfrac{1}{2}(x + x^0)) \leq \frac{1}{2}\left(f(x) + f(x^0)\right),$$

where $\mu > 0$. With the aid of Theorem 2.2 (a) we get the following estimate:

$$\begin{aligned} \frac{\mu}{4}\|x - x^0\|^2 &\leq \frac{1}{2}\left(f(x) - f(x^0)\right) - \left(f(\tfrac{1}{2}(x + x^0)) - f(x^0)\right) \\ &\leq -\left(f(\tfrac{1}{2}(x + x^0)) - f(x^0)\right) \\ &\leq -\frac{1}{2}\nabla f(x^0)^\top (x - x^0) \leq \frac{1}{2}\|\nabla f(x^0)\|\,\|x - x^0\|. \end{aligned}$$

¿From this we infer

$$\|x - x^0\| \leq \frac{2}{\mu}\|\nabla f(x^0)\| \quad \forall x \in L(x^0)$$

or–in other words–the boundedness of $L(x^0)$. Given that f is continuous, we find that $L(x^0)$ is closed. Closedness and boundedness yields the compactness of $L(x^0)$., which ends the proof. □

Now we have gathered all ingredients for proving the following theorem which illustrate why (strictly, uniformly) convex functions are of fundamental importance in optimization.

THEOREM 2.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and $X \subset \mathbb{R}^n$ be convex. Consider the optimization problem*

$$\min \quad f(x) \quad \text{s.t. } x \in X, \tag{2.9}$$

then the following statements hold true:

(a) *If f is convex on X, the solution set of* (2.9) *is convex (possibly empty).*
(b) *If f is strictly convex on X,* (2.9) *has* at most one *solution.*
(c) *If f is uniformly convex (on X) and X is non-empty and closed, then* (2.9) has exactly one *solution.*

PROOF. (a) Let $x^1, x^2 \in X$ be two different solutions of (2.9). Then $f(x^1) = f(x^2) = \min_{x\in X} f(x)$ holds true. For $\lambda \in [0,1]$ the convexity of X ensures that also $\lambda x^1 + (1-\lambda)x^2 \in X$. Moreover we have

$$f(\lambda x^1 + (1-\lambda)x^2) \le \lambda f(x^1) + (1-\lambda) f(x^2) = \min_{x\in X} f(x).$$

Therefore f attains a minimum at $\lambda x^1 + (1-\lambda)x^2$. Given that $\lambda \in [0,1]$ was arbitrary, this proves the convexity of the solution set.
(b) Let us assume that (2.9) has two solutions $x^1 \neq x^2$. Due to the strict convexity of f on X , we find for $\lambda x^1 + (1-\lambda)x^2 \in X$ that

$$f(\lambda x^1 + (1-\lambda)x^2) < \lambda f(x^1) + (1-\lambda) f(x^2) = \min_{x\in X} f(x),$$

holds true, which is a contradiction.
(c) Let $x^0 \in X$ be chosen arbitrarily. Lemma 2.3 ensures the compactness of $L(x^0)$ and also of $X \cap L(x^0)$. The continuous function f attains a minimum on the compact set $X \cap L(x^0)$. Since every uniformly convex function is also strictly convex, (b) holds true. Consequently, the minimizer is unique. □

REMARK 2.5. (1) Even for a strictly convex f, the problem (2.9) need not have a solution; consider for instance $f(x) = \exp(x)$ with $X = \mathbb{R}$.
(2) The requirement for the closeness of X in Lemma 2.5 **cannot** be dismissed! Consider $f(x) = x^2$ on $X = (0, 1]$.

A further immediate consequence is given in the following lemma.

LEMMA 2.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $x^0 \in \mathbb{R}^n$, and let the level set $L(x^0)$ be convex and f be uniformly convex on $L(x^0)$. Further suppose that $x^* \in \mathbb{R}^n$ is the unique global minimizer of f. Then there exists $\mu > 0$ with*

$$\mu \|x - x^*\|^2 \le f(x) - f(x^*) \quad \text{for all } x \in L(x^0).$$

PROOF. The statement follows immediately from Lemma 2.2 (c) and $\nabla f(x^*) = 0$. □

The central result of this section is given in Theorem 2.6. It proves that the necessary condition $\nabla f(x^*) = 0$ is also sufficient for x^* being a global minimizer of the convex function f.

THEOREM 2.6. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable and convex function, and let $x^* \in \mathbb{R}^n$ be a stationary point of f. Then x^* is a* global *minimizer of f in $\mathbb{R}^n$.*

PROOF. Theorem 2.2 (a) yields

$$f(x) - f(x^*) \geq \nabla f(x^*)^\top (x - x^*) = 0 \quad \forall x \in \mathbb{R}^n,$$

where we have used $\nabla f(x^*) = 0$. As an immediate consequence we obtain

$$f(x^*) \leq f(x) \quad \forall x \in \mathbb{R}^n,$$

i.e., x^* is a global minimizer of f. □

CHAPTER 3

General descent methods and step size strategies

In general, only exceptional cases allow the explicit calculation of (local) solutions of the minimization problem

$$\min \quad f(x), \quad x \in \mathbb{R}^n. \tag{3.1}$$

In practice, iterative methods are applied for computing approximate (local) minimizers. After a convergence analysis, these methods are normally represented in algorithmic form and implemented on a computer.
For this reason, we now consider descent methods for finding solutions of problem (3.1), in which $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function. The fundamental idea of the methods in this chapter is as follows:

(i) At a point $x \in \mathbb{R}^n$, one chooses a direction $d \in \mathbb{R}^n$ in which the function value decreases (**descent method**).
(ii) Starting at x, one proceeds along this direction d as long as the function value of f *reduces sufficiently* (**step size strategy**).

These steps will be formalized.

DEFINITION 3.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $x \in \mathbb{R}^n$. The vector $d \in \mathbb{R}^n$ is called a descent direction of f at x, if there exists an $\bar{\alpha} > 0$ such that*

$$f(x + \alpha d) < f(x) \quad \text{for all } \alpha \in (0, \bar{\alpha}].$$

Let us assume that f is continuously differentiable at $x \in \mathbb{R}^n$. Then

$$\nabla f(x)^\top d < 0 \tag{3.2}$$

is sufficient to show that $d \in \mathbb{R}^n$ is a descent direction of f in x. To see this, we define $\varphi(\alpha) := f(x + \alpha d)$. The continuous differentiability of f implies

$$\varphi(\alpha) = \varphi(0) + \alpha\varphi'(0) + r(\alpha), \tag{3.3}$$

where $r(\alpha)/\alpha \to 0$ for $\alpha \to 0^+$. We have

$$\varphi(0) = f(x) \quad \text{and} \quad \varphi'(0) = \nabla f(x)^\top d.$$

Transforming (3.3) and dividing by α yields

$$\frac{\varphi(\alpha) - \varphi(0)}{\alpha} = \nabla f(x)^\top d + \frac{r(\alpha)}{\alpha}.$$

Since $r(\alpha)/\alpha \to 0$ for $\alpha \to 0^+$ and $\nabla f(x)^\top d < 0$ (by assumption (3.2)), the existence of an $\bar{\alpha} > 0$ from Definition 3.1 is proven. Thus, d is a descent direction of f in x.

REMARK 3.1. (1) Condition (3.2) indicates that the angle between d and the negative gradient of f in x is less than 90^o.

(2) The criterion (3.2) is **not** necessary for d to be a descent direction of f at x. Consider, for instance, the case where x is a strict local maximizer. Then all directions $d \in \mathbb{R}^n$ would be descent directions of f in x, but (3.2) does not hold.

Examples for possible descent directions $d = d(x)$ are:

$$\begin{aligned} d &= -\nabla f(x) && (\textit{direction of steepest descent}), \\ d &= -M\nabla f(x) \text{ mit } M \in \mathcal{S}^n \text{ positive definite} && (\textit{gradient-related descent direction}). \end{aligned}$$

1. Globally convergent descent methods

Now let us consider a general descent method. For the time being, we do neither specify the exact choice of the descent direction nor the conditions on the step size along this direction. Below, in Lemma 3.1 we introduce abstract conditions which ensure the convergence (with its meaning still to be made precise) of the subsequent algorithm. In the following paragraphs we then specify methods to determine an appropriate step size, and furthermore we address the choice of the descent direction.

Algorithm 3.1 (General descent method).

input: $f : \mathbb{R}^n \to \mathbb{R}$, *starting point* $x^0 \in \mathbb{R}^n$.
begin
  $k := 1$
  while *convergence criterion is not fulfilled* **do**
  begin
    specify a descent direction d^k *of* f *at* x^k.
    determine a step size $\alpha_k > 0$ *with* $f(x^k + \alpha_k d^k) < f(x^k)$.
    set $x^{k+1} := x^k + \alpha_k d^k$, $k := k + 1$.
  end
end

For the moment, we do not discuss appropriate stopping rules in our theoretical considerations on convergence. Moreover, we assume that an infinite sequence $\{x^k\}$ is generated. But, of course, a practical implementation does not work without a numerically meaningful stopping criterion, and, in fact, we are going to deal with this issue in Chapter 3.3.
Note that the calculation of the new iterate is often called *line search*, as one seeks for the new iterate along the search direction (line). Algorithms aiming at the determination of a suitable step size α_k along the search direction are called *algorithms for the line search* or *step size algorithms* .
The global convergence of Algorithm 3.1 is the topic of the following lemma. Here global convergence refers to the fact that the algorithm converges for an arbitrarily chosen initial value $x^0 \in \mathbb{R}^n$. In this sense, global convergence must **not** be confused with the convergence of the sequence $\{x^k\}$ (or a subsequence) to a global minimizer of f!

Theorem 3.1. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be continuously differentiable and* $\{x^k\}$ *a sequence generated by Algorithm 3.1, such that there exist constants* $\Theta_1 > 0$ *and* $\Theta_2 > 0$ *(independent of* $\{x^k\}$ *and* $\{d^k\}$*) such that*

(a)

$$-\nabla f(x^k)^\top d^k \geq \Theta_1 \|\nabla f(x^k)\| \, \|d^k\| \quad \textit{for all } k \in \mathbb{N} \quad (\textit{angle condition});$$

(b)

$$f(x^k + \alpha_k d^k) \leq f(x^k) - \Theta_2 \left(\frac{\nabla f(x^k)^\top d^k}{\|d^k\|} \right)^2 \quad (\textit{sufficient decrease})$$

with $\alpha_k > 0$ *for all* $k \in \mathbb{N}$.

Then every accumulation point of the sequence $\{x^k\}$ *is a stationary point of* f.

PROOF. According to the assumption, every step size α_k fulfills the sufficient decrease condition (b). Using (a) in (b) results in

$$f(x^{k+1}) \leq f(x^k) - \Theta_1^2 \Theta_2 \|\nabla f(x^k)\|^2. \tag{3.4}$$

The relation (3.4) ensures that the sequence of function values $\{f(x^k)\}$ decreases monotonically. Let x^* be a accumulation point of the sequence $\{x^k\}$. Due to the continuity of f, $\{f(x^k)\}$ converges to $f(x^*)$ for a subsequence of x^k. The monotonicity ensures that $\{f(x^k)\}$ itself converges to $\{f(x^*)\}$, in particular it holds true:

$$f(x^{k+1}) - f(x^k) \to 0 \quad \text{for } k \to \infty.$$

The inequality (3.4) implies that

$$\|\nabla f(x^k)\| \to 0 \quad \text{for } k \to \infty.$$

Thus, every accumulation point x^* of $\{x^k\}$ is a stationary point of f. □

REMARK 3.2. If η_k denotes the angle between d^k and $-\nabla f(x^k)$, then part (a) of Lemma 3.1 means that

$$\cos(\eta_k) = -\frac{\nabla f(x^k)^\top d^k}{\|\nabla f(x^k)\| \, \|d^k\|}$$

is bounded away from 0; or in other words the angle is uniformly smaller than 90^o. A famous example for a descent direction fulfilling the angle condition from Theorem 3.1 is $d^k = -\nabla f(x^k)$, the direction of steepest descent of f at x^k.

Additionally assuming that f is uniformly convex on the convex level set $L(x^0)$, it is possible to substitute the angle condition of Lemma 3.1 by the weaker *Zoutendijk-condition*, *i.e.*

$$\sum_{k=0}^{\infty} \delta_k = \infty \quad \text{mit} \quad \delta_k = \left(\frac{\nabla f(x^k)^\top d^k}{\|\nabla f(x^k)\| \, \|d^k\|} \right)^2.$$

The Zoutendijk-conditon ensures that the angle between d^k and $-\nabla f(x^k)$ tends sufficiently slowly to 90^o.

2. Step size strategies and -algorithms

The general descent method (Algorithm 3.1) offers quite some freedom in the choice of the descent direction d^k and the step size $\alpha_k > 0$. The obvious *minimization rule*, *i.e.* $\alpha_k := \alpha_k^{\min}$ with

$$f(x^k + \alpha_k^{\min} d^k) = \min_{\alpha > 0} f(x^k + \alpha d^k),$$

is well-defined, provided that $L(x^0)$ is compact and ∇f Lipschitz-continuous on $L(x^0)$. Certainly, this rule is in general impracticable due to the tremendous effort necessary (at every iteration k there is one exact (!) univariate minimization required). Fortunately, we can abandon the exact univariate minimization without endangering the convergence of the descent

method. In the following, we consider three important representatives of practicable step size strategies, which all resign to an approximate minimization of $f(x^k + \alpha d^k)$ w.r.t. $\alpha > 0$.

2.1. Armijo rule. The subsequent strategy does not fit directly into the framework of Theorem 3.1, because the sufficient decrease condition may be violated. However, it allows to point out essential aspects of an approximate minimization of $f(x^k + \alpha d^k)$. In addition, the Armijo rule is an important element of alternative step size strategies, fulfilling the sufficient decrease condition.
In order to keep the subsequent explanations exemplary, we consider gradient-related descent directions of the form

$$d = d(x) = -M\nabla f(x), \quad M \in \mathcal{S}^n \text{ positive definite.}$$

It has to be mentioned that the analysis can be done in a more general way, i.e. for descent directions in the sense of Definition 3.1.
Let $\sigma \in (0, 1)$ be fixed. The *Armijo rule* is a condition which ensures a sufficient descent in the following sense:

$$f(x + \alpha d) \leq f(x) + \sigma\alpha\nabla f(x)^\top d \tag{3.5}$$

This requirement can be interpreted as a restriction on the step size α. The meaning of (3.5) can be illustrated: The solid line in Figure 1 represents the graph of $f(x + \alpha d)$ for $\alpha \geq 0$.

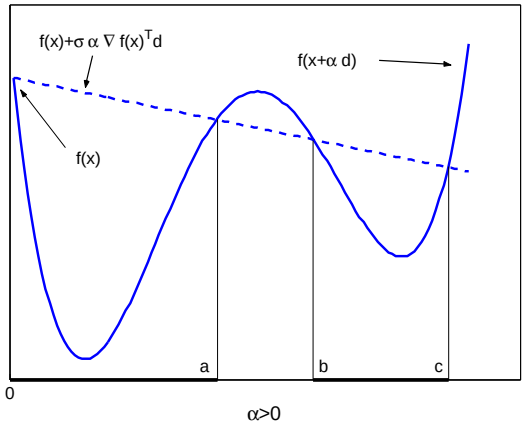

FIGURE 1. Illustration of the Armijo rule

The dashed line represents the half-line $f(x) + \sigma\alpha\nabla f(x)^\top d$. In our example, the condition (3.5) is fulfilled for $\alpha \in [0, a] \cup [b, c]$. Note that due to the requirement $\nabla f(x)^\top d < 0$ and the (Lipschitz-)continuous differentiability of f, the existence of $a > 0$ is ensured. The fact that even $\alpha \in [c, d]$ fulfills the Armijo-condition has accidentally occured in our example.
For the actual calculation of α, one checks (3.5) sequentially for e.g.

$$\alpha = \beta^l, \quad l = 0, 1, 2, \ldots, \tag{3.6}$$

where $\beta \in (0, 1)$ is fixed. One begins with $\alpha^{(0)} = \beta^0 = 1$ and stops the test if (3.5) holds true; otherwise l is incremented (resp. the α-value is decremented) and (3.5) will be checked once again.

In the following algorithm, (3.6) is generalized: If $\alpha^{(l)}$ does not fulfill the Armijo condition (3.5), then $\alpha^{(l+1)}$ is chosen such that,

$$\alpha^{(l+1)} \in [\underline{\nu}\alpha^{(l)}, \overline{\nu}\alpha^{(l)}] \tag{3.7}$$

with $0 < \underline{\nu} \leq \overline{\nu} < 1$.

ALGORITHM 3.2 (Armijo step size strategy).

input: *descent direction* d.
begin
  $l := 0$
  $\alpha^{(0)} := 1$
  while (3.5) *is not fulfilled* **do**
  begin
    determine $\alpha^{(l+1)} \in [\underline{\nu}\alpha^{(l)}, \overline{\nu}\alpha^{(l)}]$.
    set $l := l + 1$.
  end
  $\alpha_k := \alpha^{(l)}$
end

First of all we analyze Algorithms 3.2 and then Algorithm 3.1 with the step size determination according to Algorithm 3.2.

LEMMA 3.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable with Lipschitz-continuous gradient ∇f, where L denotes the Lipschitz-constant. Let $\sigma \in (0, 1)$, $x \in \mathbb{R}^n$ and $M \in \mathcal{S}^n$ be positive definite. Furthermore, let $\lambda_s = \lambda(M^{-1})$ and $\lambda_g \geq \lambda_s$ be the smallest and the largest eigenvalue of M^{-1}, respectively.*
If $\nabla f(x) \neq 0$, then (3.5) is fulfilled for all α with

$$0 < \alpha \leq \frac{2\lambda_s(1-\sigma)}{L\kappa(M^{-1})}, \tag{3.8}$$

with $\kappa(M^{-1}) = \lambda_g/\lambda_s$ the (spectral) condition number of M^{-1}.

PROOF. It holds

$$f(x + \alpha d) - f(x) = \int_0^1 \nabla f(x + \tau\,\alpha d)^\top d\, d\tau.$$

We infer that

$$f(x + \alpha d) = f(x) + \alpha \nabla f(x)^\top d + \alpha \int_0^1 \left(\nabla f(x + \tau\,\alpha d) - \nabla f(x)\right)^\top d\, d\tau.$$

By the Lipschitz-continuity of $\nabla f(x)$, this implies

$$f(x + \alpha d) \leq f(x) + \alpha \nabla f(x)^\top d + \frac{L\alpha^2}{2}\|d\|^2. \tag{3.9}$$

The positive definiteness of M implies $\lambda_g^{-1}\|z\|^2 \leq z^\top M z \leq \lambda_s^{-1}\|z\|^2$ for all $z \in \mathbb{R}^n$. Now,

$$\begin{aligned} \|d\|^2 &= \|M\nabla f(x)\|^2 = \nabla f(x)^\top M^2 \nabla f(x) \leq \lambda_s^{-2}\|\nabla f(x)\|^2 \\ &\leq \lambda_g \lambda_s^{-2} \nabla f(x)^\top M \nabla f(x) = -\lambda_g \lambda_s^{-2} \nabla f(x)^\top d \\ &= -\kappa(M^{-1})\lambda_s^{-1} \nabla f(x)^\top d. \end{aligned}$$

Using inequality (3.9), we obtain

$$f(x+\alpha d) \leq f(x) + \alpha \left(1 - \kappa(M^{-1})\lambda_s^{-1}\frac{L\alpha}{2}\right)\nabla f(x)^\top d.$$

This implies, that (3.5) is fulfilled if

$$\sigma \leq \left(1 - \kappa(M^{-1})\lambda_s^{-1}\frac{L\alpha}{2}\right),$$

which is equivalent to

$$\alpha \leq \frac{2\lambda_s(1-\sigma)}{L\kappa(M^{-1})}.$$

□

Only proving the *finite termination property* of the Armijo step size strategy does not require Lipschitz-continuity of ∇f in general. But without this additional assumption, no upper bound like (3.8) can be expected. We will come back to this point later on.
Assuming that f has the same properties as in Lemma 3.1, we prove that

$$\alpha_k \geq \underline{\alpha} > 0 \quad \text{for all } k \in \mathbb{N}. \tag{3.10}$$

In order to allow for a more general choice of gradient-related directions, let us assume that $\{M^k\}$ is a sequence of symmetric positive definite matrices and that

$$d^k = -M^k \nabla f(x^k).$$

LEMMA 3.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable with Lipschitz-continuous gradient ∇f, where L denotes the Lipschitz-constant. Let $\{x^k\}$ be the iteration sequence generated by Algorithm 3.1 with a step size choice according to Algorithm 3.2. Further let $\{M^k\}$ be a sequence of symmetric positive definite matrices, such that there exists $0 < \underline{\lambda} \leq \overline{\lambda} < +\infty$ with*

$$\underline{\lambda} \leq \lambda_s^{(k)} \leq \lambda_g^{(k)} \leq \overline{\lambda} \quad \textit{for all } k \in \mathbb{N},$$

where $\lambda_s^{(k)}$ and $\lambda_g^{(k)}$ denote the smallest resp. the largest eigenvalue of $(M^k)^{-1}$.
Then the step size α_k fulfills the inequality

$$\alpha_k \geq \underline{\alpha} = \frac{2\underline{\nu}\,\underline{\lambda}(1-\sigma)}{L\overline{\kappa}} \quad \textit{for all } k \in \mathbb{N} \tag{3.11}$$

with $\overline{\kappa} = \overline{\lambda}/\underline{\lambda}$. Furthermore, in every iteration of Algorithm 3.1 there will be at most

$$m \leq \log\left(\frac{2\underline{\lambda}(1-\sigma)}{L\overline{\kappa}}\right) / \log(\overline{\nu}), \quad m \in \mathbb{N} \tag{3.12}$$

step size reductions necessary.

PROOF. Lemma 3.1 proves that Algorithm 3.2 terminates if

$$\alpha \leq \frac{2\lambda_s^k(1-\sigma)}{L\kappa((M^k)^{-1})}$$

or even before that. Given that $\lambda_s^k \geq \underline{\lambda} > 0$ and $\kappa((M^k)^{-1}) = \lambda_g^k/\lambda_s^k \leq \bar{\lambda}/\underline{\lambda} = \bar{\kappa}$ hold true, we have

$$\frac{2\lambda_s^k(1-\sigma)}{L\kappa((M^k)^{-1})} \geq \frac{2\underline{\lambda}(1-\sigma)}{L\bar{\kappa}} > 0 \quad \forall k \in \mathbb{N}.$$

As a result of the step size strategy in Algorithm 3.2, the actual step size cannot be smaller than $\frac{2\underline{\lambda}(1-\sigma)}{L\bar{\kappa}}$ multiplied by the factor $\underline{\nu}$. This proves (3.11).
Since Algorithm 3.2 chooses $\alpha^{(0)} = 1$ and $\alpha^{(l+1)} \leq \bar{\nu}\alpha^{(l)}$, α_k will be found after at most m reductions, where $m \in \mathbb{N}$ fulfills the following relation:

$$\bar{\nu}^m < \frac{2\underline{\lambda}(1-\sigma)}{L\bar{\kappa}} \leq \frac{2\lambda_s^k(1-\sigma)}{L\kappa((M^k)^{-1})}.$$

A simple calculation leads to (3.12). □

Abandoning the assumption that ∇f is Lipschitz-continuous, there might be subsequences $\{\alpha_{k(l)}\}$ such that $\alpha_{k(l)} \to 0$. Then the statement of Lemma 3.2 does not hold true any longer. In this case, let us assume that $x^{k(l)} \to x^*$ with $\nabla f(x^*) \neq 0$. Further we assume that $\alpha_{k(l)} = \beta^{j_k(l)}$. Then the following holds true:

$$\frac{f(x^{k(l)} + \beta^{j_k(l)-1}d^{k(l)}) - f(x^{k(l)})}{\beta^{j_k(l)-1}} > \sigma\nabla f(x^{k(l)})^\top d^{k(l)}$$

and after a transition to a further subsequence, we can observe that

$$0 \geq \nabla f(x^*)^\top d^* \geq \sigma\nabla f(x^*)^\top d^*,$$

which is a contradiction, as $\sigma \in (0,1)$ was assumed. This shows, despite a weakening of assumptions of Lemma 3.2, the convergence of Algorithm 3.1 with a step size choice according to Algorithm 3.2. The indication of an upper resp. lower bound on $\{\alpha_k\}$ is dropped.

THEOREM 3.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let $\{M^k\}$ fulfill the assumption of Lemma 3.2. Then either $\{f(x^k)\}$ is unbounded from below or*

$$\lim_{k\to\infty} \nabla f(x^k) = 0, \tag{3.13}$$

and thus every accumulation point of $\{x^k\}$ is a stationary point of f in $\mathbb{R}^n$. In particular, it holds true that if $\{f(x^k)\}$ is bounded from below and $\lim_{l\to\infty} x^{k(l)} = x^*$, then $\nabla f(x^*) = 0$.

REMARK 3.3.
(1) In general there is no guarantee for the existence of a unique accumulation point.
(2) The following variation of the Armijo rule can be analyzed with Lemma 3.1, Lemma 3.2 and Theorem 3.2: Let $r > 0$ be a scaling factor. Determine

$$\alpha = \max\{r\beta^l : l = 0, 1, 2, \ldots\}, \tag{3.14}$$

such that (3.5) is fulfilled.
(3) The determination of the step size α according to (3.6) or (3.14) is called *backtracking* and fulfills (3.7).
(4) In several cases the restriction of the Armijo rule by *backtracking*, *i.e.* only permitting a reduction of the step size after the initial choice of $\alpha^{(0)}$, is a drawback. In contrast, the following strategy is more flexible: In addition to the Armijo rule (3.5) one tests the subsequent equation

$$f(x + \alpha d) \geq f(x) + \mu\alpha\nabla f(x)^\top d, \quad \text{mit } 0 < \sigma < \frac{1}{2} < \mu < 1. \tag{3.15}$$

The step size strategy (3.5)+(3.15) is called *Armijo-Goldstein rule*. The Armijo condition (3.5) implies that α shall not be too large, whereas the Goldstein condition (3.15) requires that α shall not be too small. Illustrated by Figure 2, the step size

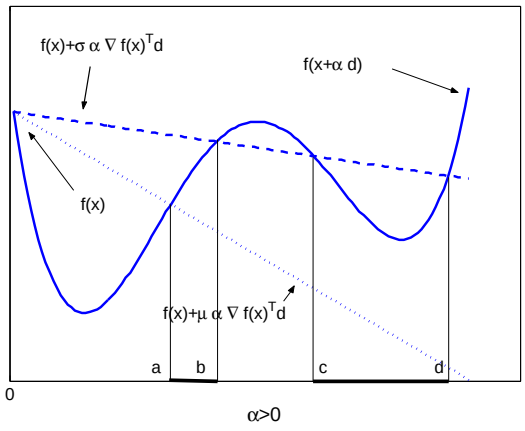

FIGURE 2. Illustration of the Armijo-Goldstein rule

in $[a,b] \cup [c,d]$ fulfills the Armijo-Goldstein rule; compare to Figure 1.

A typical approach to find α, fulfilling the conditions (3.5) and (3.15) simultaneously, is as follows: One sets $\alpha_1^{(0)} := 0, \alpha_2^{(0)} = +\infty, \alpha^{(0)} > 0$[1]. In the l-th iteration we have an $\alpha^{(l)}$ and an interval $[\alpha_1^{(l)}, \alpha_2^{(l)}]$ with the properties that $\alpha_1^{(l)} < \alpha_2^{(l)}$ and $\alpha^{(l)} \in [\alpha_1^{(l)}, \alpha_2^{(l)}]$. If (3.5) is violated for $\alpha = \alpha^{(l)}$, then one sets $\alpha_2^{(l+1)} = \alpha^{(l)}$ and $\alpha_1^{(l+1)} = \alpha_1^{(l)}$ (reduction of the step size). If (3.15) is violated for $\alpha = \alpha^{(l)}$, then one sets $\alpha_1^{(l+1)} = \alpha^{(l)}$ and $\alpha_2^{(l+1)} = \alpha_2^{(l)}$ (enlargement of the step size). The new trial step size $\alpha^{(l+1)}$ will be chosen, s.t.

$$\alpha^{(l+1)} \in [\alpha_1^{(l+1)} + \tau\Delta^{(l+1)}, \alpha_2^{(l+1)} - \tau\Delta^{(l+1)}],$$

where $\Delta^{(l+1)} = \alpha_2^{(l+1)} - \alpha_1^{(l+1)}$ denotes the interval length and $0 < \tau \ll 1$ is fixed (bisection!). It is obvious that this approach only makes sense, if $\alpha_2^{(l+1)} < +\infty$ holds true. If that is not (yet) the case, one chooses $\alpha^{(l+1)} \geq \xi \max(\alpha_1^{(l)}, \epsilon)$, where $\xi > 1$ and $\epsilon > 0$ are fixed. From a numerical point of view, one not only has to make sure that the step size algorithm terminates when meeting the Armijo-Goldstein conditions (3.5) and (3.15), but also when $\Delta^{(l+1)}$ is relatively small.

Finally we note that (3.15) together with (3.5) and *backtracking* does not make sense in general.

Now we discuss another step size choice based on polynomial models of the function $\varphi(\alpha)$. Our discussion is based on quadratic models. However, models of higher order are often applied in practice.

Armijo step size algorithm based on polynomial models. Apart from the simple *backtracking* method, there are strategies which apply polynomial models of $\varphi(\alpha)(= f(x+\alpha d))$.

[1]Superscripts denote the index in iterative methods to determine α resp. α_k in the k-th iteration of the minimization algorithm for f.

In every iteration of Algorithm 3.2 we have the following data at hand:

$$\varphi(0) = f(x), \quad \varphi(\alpha^{(l)}) = f(x + \alpha^{(l)} d), \quad \varphi'(0) = \nabla f(x)^\top d < 0.$$

With the help of this data, we create a *quadratic model of* $\varphi(\alpha)$. The ansatz

$$q(\alpha) = a + b\alpha + c\alpha^2, \quad a, b, c \in \mathbb{R},$$

with the conditions

$$q(0) = \varphi(0), \quad q(\alpha^{(l)}) = \varphi(\alpha^{(l)}), \quad q'(0) = \varphi'(0)$$

leads to

$$q(\alpha) = \varphi(0) + \varphi'(0)\alpha + \frac{1}{(\alpha^{(l)})^2}(\varphi(\alpha^{(l)}) - \varphi(0) - \varphi'(0)\alpha^{(l)})\alpha^2. \tag{3.16}$$

Since the quadratic model (3.16) is only computed when the Armijo rule (3.5) is violated for $\alpha = \alpha^{(l)}$, it follows that

$$\varphi(\alpha^{(l)}) - \varphi(0) - \varphi'(0)\alpha^{(l)} > \varphi(\alpha^{(l)}) - \varphi(0) - \sigma\varphi'(0)\alpha^{(l)} > 0.$$

Thus, a global minimizer $\hat{\alpha}^{\min}$ for $q(\alpha)$ exists:

$$\hat{\alpha}^{\min} = \frac{-\varphi'(0)(\alpha^{(l)})^2}{2(\varphi(\alpha^{(l)}) - \varphi(0) - \varphi'(0)\alpha^{(l)})} > 0.$$

Now, we can choose the next trial step size $\alpha^{(l+1)}$ according to

$$\alpha^{(l+1)} := \begin{cases} \underline{\nu}\alpha^{(l)} & \text{if } \hat{\alpha}^{\min} < \underline{\nu}\alpha^{(l)}, \\ \hat{\alpha}^{\min} & \text{if } \underline{\nu}\alpha^{(l)} \le \hat{\alpha}^{\min} \le \overline{\nu}\alpha^{(l)}, \\ \overline{\nu}\alpha^{(l)} & \text{if } \hat{\alpha}^{\min} > \overline{\nu}\alpha^{(l)}. \end{cases} \tag{3.17}$$

This choice ensures the required property $\alpha^{l+1} \in [\underline{\nu}\alpha^l, \overline{\nu}\alpha^{(l)}]$. In addition, (3.17) ensures that

$$\underline{\nu}^{l+1}\alpha^{(0)} \le \underline{\nu}\alpha^{(l)} \le \alpha^{(l+1)} \le \overline{\nu}\alpha^{(l)} \le \overline{\nu}^{l+1}\alpha^{(0)} \quad \text{for all } l \in \mathbb{N},$$

where $\lim_{l\to\infty}(\underline{\nu}^l, \overline{\nu}^l) = 0$ because of $0 < \underline{\nu} \le \overline{\nu} < 1$.

Remark 3.4. (1) Cubic polynomial models for $\varphi(\alpha)$ can be obtained by using

$$\varphi(0), \quad \varphi(\alpha^{(l)}), \quad \varphi'(0), \quad \varphi'(\alpha^{(l)}) \tag{3.18}$$

or

$$\varphi(0), \quad \varphi(\alpha^{(l)}), \quad \varphi'(0), \quad \varphi(\alpha^{(l-1)}) \quad \text{for } l \ge 1. \tag{3.19}$$

In case the determination of the derivatives of f is expensive, then one prefers (3.19).

(2) As we have already mentioned above, in the case of the Armijo-Goldstein step size strategy, (3.5) and (3.15) together with the bisection idea of Remark 3.3 (4), $\alpha_2^{(l)} = +\infty$ is possible. Methods based on polynomial models have to take this into account when determining coefficients and choosing the step size.

2.2. Wolfe-Powell rule. Let $\sigma \in (0, \frac{1}{2})$ and $\rho \in [\sigma, 1)$ be fixed. The *Wolfe-Powell conditions* are: For $x, d \in \mathbb{R}^n$ with $\nabla f(x)^\top d < 0$ determine a step size $\alpha > 0$ such that

$$f(x + \alpha d) \leq f(x) + \sigma\alpha\nabla f(x)^\top d, \tag{3.20}$$

$$\nabla f(x + \alpha d)^\top d \geq \rho\nabla f(x)^\top d. \tag{3.21}$$

Just like the Armijo-Goldstein rule, the choice of $\sigma < \frac{1}{2}$ enables us to accept the exact minimizer of a quadratic function as a Wolfe-Powell step size. The condition (3.21) ideally implies that the graph of $\varphi(\alpha) = f(x + \alpha d)$ at $\alpha > 0$ does not descend as "steeply" as at $\alpha = 0$. This claim is motivated by the fact that

$$\varphi'(\hat{\alpha}) = \nabla f(x + \hat{\alpha} d)^\top d = 0$$

is satisfied at a (local) minimizer $\hat{\alpha}$ of φ. Similar to condition (3.15) of the Armijo-Goldstein line search, (3.21) prevents α from getting too small. Figure 3 illustrates the conditions (3.20) and (3.21). The first condition yields - as before - a restriction on the extent of the step size. The second condition (3.21) ensures that points with $\nabla f(x + \alpha d)^\top d = 0$ are always located in the acceptable step size set ($[a, b] \cup [c, d]$ in our example).

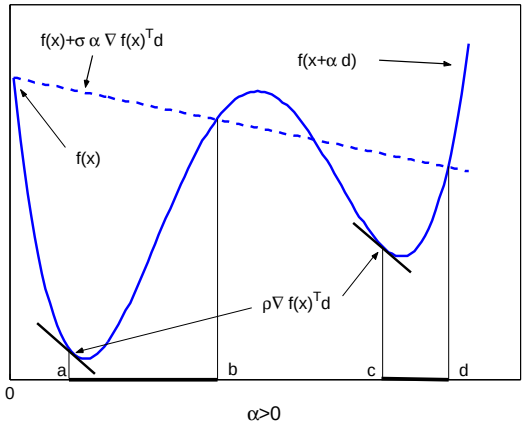

FIGURE 3. Illustration of the Wolfe-Powell rule

Now we want to prove that for given x and d, the set of Wolfe-Powell step sizes is non-empty and that for step sizes which fulfill (3.20) and (3.21), the sufficient decrease condition of Theorem 3.1 is satisfied.

THEOREM 3.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $\sigma \in (0, \frac{1}{2})$, $\rho \in [\sigma, 1)$ and $x^0 \in \mathbb{R}^n$ be fixed. For $x \in L(x^0)$ and $d \in \mathbb{R}^n$ with $\nabla f(x)^\top d < 0$ let*

$$S_{WP}(x, d) := \{\alpha > 0 : (3.20) \text{ and } (3.21) \text{ are fulfilled}\}$$

be the set of Wolfe-Powell step sizes at x in direction d. Then the following statements hold true:

(a) *If f is bounded from below, then $S_{WP}(x, d) \neq \emptyset$, i.e. the Wolfe-Powell step size strategy is well-defined.*

(b) *If, in addition, ∇f is Lipschitz-continuous on $L(x^0)$, then there is a constant $\Theta > 0$ (independent of x and d) with*

$$f(x+\alpha d) \le f(x) - \Theta \left(\frac{\nabla f(x)^\top d}{\|d\|} \right)^2 \quad \text{for all } \alpha \in S_{WP}(x,d).$$

PROOF. Define $\psi(\alpha) := f(x) + \sigma\alpha\nabla f(x)^\top d$. We have to show that there exists a step size $\alpha > 0$, such that

$$\varphi(\alpha) \le \psi(\alpha) \quad \text{and} \quad \varphi'(\alpha) \ge \rho\varphi'(0).$$

Note that $\psi(0) = \varphi(0)$ and

$$\psi'(\alpha) = \sigma\, \nabla f(x)^\top d > \nabla f(x)^\top d$$

hold true. Therefore the graph of φ is located below the graph of ψ for sufficiently small $\alpha > 0$. Let α^* be the smallest step size $\alpha > 0$ with $\varphi(\alpha) = \psi(\alpha)$. The existence of α^* is guaranteed, since f is assumed to be bounded from below and $\psi(\alpha) \to -\infty$ holds true for $\alpha \to \infty$. Obviously,

$$\varphi'(\alpha^*) \ge \psi'(\alpha^*)$$

holds true. We distinguish between two cases:

(1) $\varphi'(\alpha^*) < 0$ holds true. The relation between the derivatives of φ and ψ in $\alpha = \alpha^*$ yields

$$-\varphi'(\alpha^*) \le -\psi'(\alpha^*) = -\sigma\, \nabla f(x)^\top d = -\sigma\, \varphi'(0) \le -\rho\, \varphi'(0),$$

since $0 < \sigma \le \rho$. Due to $\varphi(\alpha^*) = \psi(\alpha^*)$, we have $\alpha^* \in S_{\mathrm{WP}}(x,d)$.

(2) Now let $\varphi'(\alpha^*) \ge 0$. Since $\varphi'(0) < 0$ there exists $\alpha^{**} \in (0, \alpha^*]$ with $\varphi'(\alpha^{**}) = 0$ (by continuity). Since $\alpha^{**} \le \alpha^*$ holds true, the condition $\varphi(\alpha) \le \psi(\alpha)$ is fulfilled for $\alpha = \alpha^{**}$. By definition of α^{**}, $\varphi'(\alpha^{**}) = 0$, thus

$$0 = \varphi'(\alpha^{**}) \ge \rho\, \varphi'(0).$$

Consequently $\alpha = \alpha^{**} \in S_{\mathrm{WP}}(x,d)$.

This proves assertion (a).
Next we establish (b). Let $\alpha \in S_{\mathrm{WP}}(x,d)$. Then $f(x+\alpha d) \le f(x)$ holds true and in particular $x + \alpha d \in L(x^0)$. From the Wolfe-Powell rule, it follows

$$(\rho - 1)\nabla f(x)^\top d \le \left(\nabla f(x+\alpha d) - \nabla f(x)\right)^\top d.$$

Moreover,

$$(\rho - 1)\nabla f(x)^\top d \le \|\nabla f(x+\alpha d) - \nabla f(x)\|\, \|d\| \le L\alpha \|d\|^2,$$

where $L > 0$ denotes the Lipschitz-constant of ∇f on $L(x^0)$. We obtain

$$\alpha \ge \frac{(\rho - 1)\nabla f(x)^\top d}{L\|d\|^2}$$

for the Wolfe-Powell-step size and consequently

$$f(x+\alpha d) \le f(x) + \sigma\alpha\nabla f(x)^\top d \le f(x) - \frac{(1-\rho)\sigma}{L} \left(\frac{\nabla f(x)^\top d}{\|d\|^2} \right)^2.$$

□

If the descent directions d^k in the general descent method, cf. Algorithm 3.1, are chosen such that the angle condition of Theorem 3.1 is fulfilled, then we can easily infer from Theorem 3.3 (and Theorem 3.1) that every accumulation point of the sequence $\{x^k\}$ is a stationary point. It remains to examine the numerical realization of the Wolfe-Powell-rule.
Before specifying the corresponding step size algorithm, we consider the following lemma which is going to be used for the determination of an appropriate starting point for the numerical determination of a Wolfe-Powell step size in the subsequent algorithm.

LEMMA 3.3. *Let $\sigma < \rho$ (cf. Theorem 3.3), $\varphi'(0) < 0$ and $\Phi(\alpha) := \varphi(\alpha) - \varphi(0) - \sigma\alpha\varphi'(0)$. If $[a, b]$ denotes an interval with the properties*

$$\Phi(a) \leq 0, \quad \Phi(b) \geq 0, \quad \Phi'(a) < 0, \tag{3.22}$$

then $[a, b]$ contains a point $\bar{\alpha}$ with

$$\Phi(\bar{\alpha}) < 0, \quad \Phi'(\bar{\alpha}) = 0.$$

$\bar{\alpha}$ is an interior point of an interval I such that for all $\alpha \in I$ there holds:

$$\Phi(\alpha) \leq 0 \quad \textit{and} \quad \varphi'(\alpha) \geq \rho\varphi'(0),$$

i.e. $I \subset S_{WP}(x, d)$.

PROOF. According to the assumption, $\Phi(a) \leq 0$, $\Phi'(a) < 0$ and $\Phi(b) \geq 0$ hold true. Therefore there is at least one point $\xi \in (a, b)$ with $\Phi'(\xi) \geq \epsilon$ for sufficiently small $\epsilon > 0$. If there was no such a point, it would hold that $\Phi'(\alpha) \leq 0$ for all $\alpha \in [a, b]$. Since $\Phi'(a) < 0$ was assumed, then from the continuous differentiability of Φ we would get that $\Phi(b) < 0$, a contradiction. Now let $\hat{\xi}$ be the smallest element ξ in (a, b) satisfying $\Phi'(\xi) \geq \epsilon$. Given that Φ' is continuous, there exists a $\xi_0 \in (a, \hat{\xi})$ with $\Phi'(\xi_0) = 0$ by Bolzano's Root Theorem . Let ξ_0 be the smallest element having this property. Then $\Phi(\xi_0) < 0$ holds true. If that was not the case, there would be an $\xi_1 \in (a, \xi_0)$ with $\Phi(a) = \Phi(\xi_1) \leq 0$ due to the continuity of Φ. Rolle's Theorem however ensures the existence of a $\xi_{00} \in (a, \xi_1)$ with $\Phi'(\xi_{00}) = 0$ which contradicts the choice of ξ_0. To conclude the first part of the proof, we set $\bar{\alpha} = \xi_0$.
For the proof of the second part, note that

$$\begin{aligned}\Phi'(\alpha) =& \varphi'(\alpha) - \sigma\,\varphi'(0) = \nabla f(x + \alpha d)^\top d - \sigma\,\nabla f(x)^\top d \\ =& \nabla f(x + \alpha d)^\top d - \rho\,\nabla f(x)^\top d + (\rho - \sigma)\varphi'(0)\end{aligned}$$

holds true. This implies

$$\nabla f(x + \bar{\alpha} d)^\top d > \nabla f(x + \bar{\alpha} d)^\top d + (\rho - \sigma)\varphi'(0) = \rho\,\nabla f(x)^\top d. \tag{3.23}$$

As $\Phi(\bar{\alpha}) < 0$, there exists a neighborhood $[\bar{\alpha} - r_0, \bar{\alpha} + r_0]$, $r_0 > 0$, s.t. $\Phi(\alpha) \leq 0$ holds true for all $\alpha \in [\bar{\alpha} - r_0, \bar{\alpha} + r_0]$. Due to the continuity of Φ', $\rho > \sigma$ and $\varphi'(0) < 0$, for $0 < \epsilon \leq \frac{1}{2}(\sigma - \rho)\varphi'(0)$ there exists $r_\epsilon > 0$ such that

$$\varphi'(\alpha) = \nabla f(x + \alpha d)^\top d > \nabla f(x + \alpha d)^\top d - \epsilon \geq \rho\,\nabla f(x)^\top d = \rho\,\varphi'(0)$$

for all $\alpha \in [\bar{\alpha} - r_\epsilon, \bar{\alpha} + r_\epsilon]$. Choosing $r = \min(r_0, r_\epsilon) > 0$, it follows that

$$\Phi(\alpha) \leq 0 \quad \text{and} \quad \varphi'(\alpha) \geq \rho\varphi'(0)$$

for all $\alpha \in I := [\bar{\alpha} - r, \bar{\alpha} + r] \subset S_{\mathrm{WP}}(x, d)$. □

Lemma 3.3 is crucial for the following algorithm.

ALGORITHM 3.3 (Wolfe-Powell step size algorithm).

input: *descent direction* $d \in \mathbb{R}^n$.
begin
choose $\alpha^{(0)} > 0$, $\gamma > 1$, $i := 0$
(A.1) **if** $\Phi(\alpha^{(i)}) \geq 0$ **then**
begin
$a := 0$, $b := \alpha^{(i)}$
goto *(B.0)*
end
else
if $\varphi'(\alpha^{(i)}) \geq \rho\varphi'(0)$ **then** $\alpha := \alpha^{(i)}$, *RETURN 1* **end**
if $\varphi'(\alpha^{(i)}) < \rho\varphi'(0)$ **then**
begin
$\alpha^{(i+1)} := \gamma\alpha^{(i)}$, $i := i + 1$
goto *(A.1)*
end
end
end
(B.0) choose $\tau_1, \tau_2 \in (0, \frac{1}{2}]$, $j := 0$, $\alpha_1^{(0)} := a$, $\alpha_2^{(0)} := b$, $\Delta^{(0)} := \alpha_2^{(0)} - \alpha_1^{(0)}$
(B.1) choose $\alpha^{(j)} \in [\alpha_1^{(j)} + \tau_1\Delta^{(j)}, \alpha_2^{(j)} - \tau_2\Delta^{(j)}]$
if $\Phi(\alpha^{(j)}) \geq 0$ **then**
begin
$\alpha_1^{(j+1)} := \alpha_1^{(j)}$, $\alpha_2^{(j+1)} := \alpha^{(j)}$, $\Delta^{(j+1)} := \alpha_2^{(j+1)} - \alpha_1^{(j+1)}$, $j := j + 1$
goto *(B.1)*
end
else
if $\varphi'(\alpha^{(j)}) \geq \rho\varphi'(0)$ **then** $\alpha := \alpha^{(j)}$, *RETURN 2* **end**
if $\varphi'(\alpha^{(j)}) < \rho\varphi'(0)$ **then**
begin
$\alpha_1^{(j+1)} := \alpha^{(j)}$, $\alpha_2^{(j+1)} = \alpha_2^{(j)}$, $\Delta^{(j+1)} := \alpha_2^{(j+1)} - \alpha_1^{(j+1)}$, $j = j + 1$
goto *(B.1)*
end
end
end
end

Concerning the choice of $\alpha^{(0)}$, we remark that from the second iteration of the descent method Algorithm 3.1 on, i.e. $k \geq 2$, $\alpha^{(0)}$ can be chosen as α_{k-1}. If a lower bound $\underline{f}$ of the function values of f on $\mathbb{R}^n$ is known, we can infer immediately

$$0 \geq \Phi(\alpha) = \varphi(\alpha) - \varphi(0) - \sigma\alpha\nabla f(x)^\top d \geq \underline{f} - \varphi(0) - \sigma\alpha\varphi'(0)$$

from $\Phi(\alpha) \leq 0$. Rearranging yields

$$\alpha \leq \frac{\underline{f} - \varphi(0)}{\sigma\varphi'(0)} =: \bar{\alpha}.$$

Thus it is reasonable to choose $\alpha^{(0)} \in (0, \bar{\alpha}]$ in this case. With regard to Newton- or Quasi-Newton methods, the step size $\alpha = 1$ is of particular importance. Hence it is recommandable to choose $\alpha^{(0)} = \min\{1, \bar{\alpha}\}$ for these methods.
Now we demonstrate that Algorithm 3.3 terminates successfully after finitely many steps, under certain conditions.

THEOREM 3.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and bounded from below. Furthermore let $\sigma \in (0, \frac{1}{2})$ and $\rho \in (\frac{1}{2}, 1)$ be fixed. Then Algorithm 3.3 terminates after finitely many steps at "RETURN 1" or "RETURN 2" with $\alpha \in S_{WP}(x, d)$.*

PROOF. First we consider the case where Algorithm 3.3 terminates at "RETURN 1". Then it is obvious that the Wolfe-Powell conditions are fulfilled.
In the next step we show that the loop in the first part of the algorithm (switch to (A.1)) is finite, i.e. after finitely many steps we continue with (B.0) or we terminate with "RETURN 1". Let us assume that the loop was not finite, resp. the algorithm would return infinitely often to the position (A.1). In this case it holds $\alpha^{(i)} = \gamma\, \alpha^{(0)}$ and $\Phi(\alpha^{(i)}) < 0$ for all $i \in \mathbb{N}$. But the last inequality implies

$$f(x + \alpha^{(i)} d) < f(x) + \sigma\, \alpha^{(i)} \nabla f(x)^\top d \quad \forall i \in \mathbb{N}.$$

By assumption, $\gamma > 1$ and $\varphi'(0) < 0$. Hence we would obtain $f(x + \alpha^{(i)} d) \downarrow -\infty$, which contradicts the boundedness of f from below. Hence the loop (switch to (A.1)) has to terminate with "RETURN 1" in the first part of the algorithm after finitely many steps or it has to jump to position (B.0).
Now, assume one has reached (B.0). In that case, the interval $[a, b]$ has the property of Lemma 3.3 and $\varphi'(a) < \rho\varphi'(0)$. If the algorithm terminates with "RETURN 2", then the Wolfe-Powell conditions are satisfied. It remains to prove that "RETURN 2" can be reached after finitely many trials (switch to (B.1)). First we prove by induction that the interval $[\alpha_1^{(j)}, \alpha_2^{(j)}]$ has, for every $j \in \mathbb{N}$, the property (3.22) (with $a = \alpha_1^{(j)}$ and $b = \alpha_2^{(j)}$) and fulfills $\varphi'(\alpha_1^{(j)}) < \rho\varphi'(0)$.

- $j = 0$. The assertion follows from the conditions which allow to arrive at (B.0).
- $j \to j + 1$. We assume that $[\alpha_1^{(j)}, \alpha_2^{(j)}]$ fulfills (3.22) (with $a = \alpha_1^{(j)}$ and $b = \alpha_2^{(j)}$) and $\varphi'(\alpha_1^{(j)}) < \rho\varphi'(0)$.

  If $\Phi(\alpha^{(j)}) \geq 0$, then setting $\alpha_1^{(j+1)} := \alpha_1^{(j)}$, $\alpha_2^{(j+1)} := \alpha^{(j)}$ gives:

$$\begin{aligned} \Phi(\alpha_1^{(j+1)}) &= \Phi(\alpha_1^{(j)}) \leq 0, \\ \Phi(\alpha_2^{(j+1)}) &= \Phi(\alpha^{(j)}) \geq 0, \\ \Phi'(\alpha_1^{(j+1)}) &= \Phi'(\alpha_1^{(j)}) < 0. \end{aligned}$$

  In case $\Phi(\alpha^{(j)}) < 0$ and $\varphi'(\alpha^{(j)}) < \rho\varphi'(0)$ (otherwise the algorithm will terminate with "RETURN 2"), we have $\alpha_1^{(j+1)} = \alpha^{(j)}$ and $\alpha_2^{(j+1)} = \alpha_2^{(j)}$ and

$$\begin{aligned} \Phi(\alpha_1^{(j+1)}) &= \Phi(\alpha^{(j)}) < 0, \\ \Phi(\alpha_2^{(j+1)}) &= \Phi(\alpha_2^{(j)}) \geq 0, \\ \Phi'(\alpha_1^{(j+1)}) &= \Phi'(\alpha^{(j)}) = \varphi'(\alpha^{(j)}) - \sigma\, \varphi'(0) < 0, \end{aligned}$$

  since $\rho > \sigma > 0$.

In both cases, $[\alpha_1^{(j)}, \alpha_2^{(j)}]$ has the desired property.
Finally we show that the loop (switch to (B.1)) is finite. Let us assume that this was not true. Then the intervals $[\alpha_1^{(j)}, \alpha_2^{(j)}]$ would "shrink" to a point α^*. This results from the fact that

$$0 < \alpha_2^{(j)} - \alpha_1^{(j)} \leq \max\{1-\tau_1, 1-\tau_2\}\,(\alpha_2^{(j+1)} - -\alpha_1^{(j+1)})$$

and $\max\{1-\tau_1, 1-\tau_2\} < 1$ hold true. Lemma 3.3 would yield that for each $j \in \mathbb{N}$ there exists $\hat{\alpha}^{(j)} \in (\alpha_1^{(j)}, \alpha_2^{(j)})$, such that

$$\Phi(\hat{\alpha}^{(j)}) < 0 \quad \text{and} \quad \Phi'(\hat{\alpha}^{(j)}) = 0$$

would be fulfilled. Because of $\hat{\alpha}^{(j)} \to \alpha^*$ for $j \to \infty$, it follows $\Phi'(\alpha^*) = 0$ and also

$$\varphi'(\alpha^*) = \sigma\varphi'(0) > \rho\varphi'(0), \tag{3.24}$$

since by assumption $0 < \sigma < \rho$ and $\varphi'(0) < 0$. On the other hand, $\varphi'(\alpha_1^{(j)}) < \rho\varphi'(0)$ and the continuity of φ' would imply $\varphi'(\alpha^*) \leq \rho\varphi'(0)$. This, however, would contradict (3.24). Hence, the loop (switch to (B.1)) has to terminate after finitely many iterations with "RETURN 2". □

The freedom of choice w.r.t. $\alpha^{(j)}$ in (B.1) can again be used to apply quadratic or cubic polynomial models.

2.3. Strong Wolfe-Powell rule. Let $\sigma \in (0, \frac{1}{2})$ and $\rho \in [\sigma, 1)$ fixed. The *Strong Wolfe-Powell rule* requires: For $x, d \in \mathbb{R}^n$ with $\nabla f(x)^\top d < 0$ determine a step size $\alpha > 0$ with

$$f(x+\alpha d) \leq f(x) + \sigma\alpha\nabla f(x)^\top d, \tag{3.25}$$

$$|\nabla f(x+\alpha d)^\top d| \leq -\rho\nabla f(x)^\top d. \tag{3.26}$$

In comparison to the Wolfe-Powell rule, condition (3.26) requires not only that the graph of $\varphi(\alpha)(= f(x+\alpha d))$ in $\alpha > 0$ does not decrease as steeply as in $\alpha = 0$, but also that the graph does not increase too steeply. A step size, which fulfills (3.26) for a very small ρ (and thus also for a very small σ), is near to a stationary point of $\varphi(\cdot)$.
For the *set of Strong Wolfe-Powell step sizes* in x in direction d, *i.e.*

$$S_{\text{SWP}}(x,d) := \{\alpha > 0 : (3.25) + (3.26) \text{ are fulfilled}\}$$

an analogue statement to Theorem 3.3 holds true. Furthermore we can prove an analogous result to Lemma 3.3, in which the third condition in (3.22) has to be modified. The corresponding step size algorithm is structured similarly to Algorithm 3.3.

3. Practical aspects

The algorithms in section 2 of this chapter are idealized. In *numerical practice* it has to be taken into account that the exactness in the evaluation of functions and derivatives is machine- and problem-dependent. If these "inaccuracies" are not taken into account in the conditions of Paragraphs 2.1–2.3, *dead loops* are likely to occur. In the best case, error bounds $\epsilon(\alpha), \epsilon(0) \geq 0$ and $\hat{\epsilon}(\alpha), \hat{\epsilon}(0) \geq 0$ are known for the function values $\varphi(\alpha), \varphi(0)$ and derivatives $\varphi'(\alpha), \varphi'(0)$. Then it would be possible to modify resp. attenuate the condition $\varphi(\alpha) \leq \varphi(0) + \sigma\alpha\varphi'(0)$ in the following way:

$$\varphi(\alpha) \leq \varphi(0) + \sigma\alpha\big(\varphi'(0) + \hat{\epsilon}(0)\big) + \epsilon(\alpha) + \epsilon(0).$$

Further, $\varphi'(\alpha) \geq \rho\varphi'(0)$ could be implemented in the form

$$\varphi'(\alpha) \geq \rho\big(\varphi'(0) - \hat{\epsilon}(0)\big) - \hat{\epsilon}(\alpha).$$

In most cases, such error bounds are not available. Then a possible approach contains the application of error bounds of the following form

$$\epsilon(\alpha) := \epsilon\big(1 + |\varphi(\alpha)|\big)$$

(or also $\epsilon(\alpha) := \epsilon|\varphi(\alpha)|$ for sufficiently large $|\varphi(\alpha)|$), where $\epsilon \geq \epsilon_M$. The value $\epsilon_M > 0$ corresponds to *the machine precision* (or the relative accuracy in computations). If the analytic form of the derivative is implemented, then one may choose $\hat{\epsilon}(\alpha) = \hat{\epsilon}(0) = 0$ (and ϵ slightly enlarged).

Furthermore the step size algorithm has to be terminated whenever the interval $[\alpha_1^{(j)}, \alpha_2^{(j)}]$ gets "too small", i.e., when $\alpha_2^{(j)} - \alpha_1^{(j)} > 0$ becomes small. A practicable criterion uses a tolerance level of the form

$$\epsilon_\Delta := \epsilon(1 + \alpha_2^{(j)}) \quad \text{for } \alpha_2^{(j)} < +\infty$$

(or $\epsilon_\Delta := \epsilon\alpha_2^{(j)}$).

Since Theorem 3.4 assumes the boundedness of f from below, a lower bound to φ should be employed in the step size strategy, terminating the algorithm when ϕ drops below this bound.

CHAPTER 4

Rate of convergence

For the realization of a numerical method to solve

$$\min \quad f(x), \quad x \in \mathbb{R}^n,$$

not only the convergence of iterates x^k to a solution (or probably only a stationary point) is of importance, but also "how fast" this convergence takes place.

1. Q-convergence and R-convergence

First we discuss classical approaches to characterize the rate of convergence.

DEFINITION 4.1. (a) *For the sequence $\{x^k\} \subset \mathbb{R}^n$ with limit $x^* \in \mathbb{R}^n$ and $p \in [1, +\infty)$ we refer to*

$$Q_p\{x^k\} := \begin{cases} \limsup_{k\to\infty} \frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|^p}, & \text{if } x^k \neq x^* \ \forall k \geq k_0, \\ 0, & \text{if } x^k = x^* \ \forall k \geq k_0, \\ +\infty, & \text{otherwise}, \end{cases}$$

as the quotient-convergence factor (Q-factor) of $\{x^k\}$.

(b) *We refer to*

$$O_Q\{x^k\} := \inf\{p \in [1, +\infty) : Q_p\{x^k\} = +\infty\}$$

as the Q-convergence order (Q-order) of the sequence $\{x^k\}$.

We summarize some important properties in the following remark.

REMARK 4.1. (1) The Q-factor depends on the applied norm, but the Q-order does not.

(2) There always exists a value $p_0 \in [1, +\infty)$ such that

$$Q_p\{x^k\} = \begin{cases} 0 & \text{for } p \in [1, p_0), \\ +\infty & \text{for } p \in (p_0, +\infty). \end{cases}$$

(3) The Q-orders 1 and 2 are of particular importance. The following notions have become popular:

$Q_1\{x^k\} = 0$	Q-superlinear convergence
$0 < Q_1\{x^k\} < 1$	Q-linear convergence
$Q_2\{x^k\} = 0$	Q-superquadratic convergence
$0 < Q_2\{x^k\} < +\infty$	Q-quadratic convergence

The implementation of the criterion

$$\|x^k - x^*\| \leq \epsilon \tag{4.1}$$

for Q-superlinearly convergent iteration sequences requires the knowledge of x^*, which, however, is unrealistic! The following result allows to replace (4.1) by the practical criterion

$$\|x^{k+1} - x^k\| \leq \epsilon. \tag{4.2}$$

THEOREM 4.1. *Any sequence $\{x^k\} \subset \mathbb{R}^n$ with $\lim_k x^k = x^*$ satisfies*

$$\left|1 - \frac{\|x^{k+1} - x^k\|}{\|x^k - x^*\|}\right| \leq \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \quad \text{for } x^k \neq x^*.$$

If $\{x^k\}$ converges Q-superlinearly to x^* and $x^k \neq x^*$ for $k \geq k_0$ then

$$\lim_{k\to\infty} \frac{\|x^{k+1} - x^k\|}{\|x^k - x^*\|} = 1.$$

PROOF. Using the triangle inequality we obtain

$$\left|\|x^* - x^k\| - \|x^{k+1} - x^*\|\right| \leq \|x^{k+1} - x^k\| \leq \|x^* - x^k\| + \|x^{k+1} - x^*\|.$$

Dividing by both sides by $\|x^k - x^*\|$, the assertions follow immediately. □

In addition to Q-convergence, the notion of R-convergence plays a decisive role.

DEFINITION 4.2. (a) *For a sequence $\{x^k\} \subset \mathbb{R}^n$ with limit $x^* \in \mathbb{R}^n$ and $p \in [1, +\infty)$ we refer to*

$$R_p\{x^k\} := \begin{cases} \limsup_{k\to\infty} \|x^k - x^*\|^{1/k}, & \text{if } p = 1, \\ \limsup_{k\to\infty} \|x^k - x^*\|^{1/p^k}, & \text{if } p > 1 \end{cases}$$

as the root-convergence factor (R-factor) of $\{x^k\}$.

(b) *We refer to*

$$O_R\{x^k\} := \inf\{p \in [1, +\infty) : R_p\{x^k\} = 1\}$$

as the R-convergence order (R-order) of the sequence $\{x^k\}$.

REMARK 4.2. (1) In contrast to the Q-factor the R-factor is independent of the applied norm due to the norm equivalence in $\mathbb{R}^n$. To see this, let $\|\cdot\|_a$ and $\|\cdot\|_b$ be two norms in $\mathbb{R}^n$, necessarily satisfying $c_1\|x\|_b \leq \|x\|_a \leq c_2\|x\|_b$ for all $x \in \mathbb{R}^n$, where c_1, c_2 are positive constants. Furthermore let $\{\gamma^k\}$, $\gamma_k > 0$ for all $k \in \mathbb{N}$, be a zero sequence. Then

$$\begin{aligned} \limsup_{k\to\infty} \|x^k - x^*\|_a^{\gamma_k} &\leq \lim_{k\to\infty} c_2^{\gamma_k} \limsup_{k\to\infty} \|x^k - x^*\|_b^{\gamma_k} \\ &= \limsup_{k\to\infty} \|x^k - x^*\|_b^{\gamma_k} \leq \lim_{k\to\infty} c_1^{-\gamma_k} \limsup_{k\to\infty} \|x^k - x^*\|_a^{\gamma_k} \\ &= \limsup_{k\to\infty} \|x^k - x^*\|_a^{\gamma_k}. \end{aligned}$$

(2) There always exists a $p_0 \in [1, +\infty)$ such that

$$R_p\{x^k\} = \begin{cases} 0 \text{ for } p \in [1, p_0), \\ 1 \text{ for } p \in (p_0, +\infty). \end{cases}$$

(3) Between Q- and R-convergence resp. the Q- and R-factor, the following relations hold true:

$$O_Q\{x^k\} \leq O_R\{x^k\} \quad \text{and} \quad R_1\{x^k\} \leq Q_1\{x^k\}.$$

It is often convenient to use the *Landau symbols* o and $\mathcal{O}$ for describing the convergence behavior.

DEFINITION 4.3. *Let $f, g : \mathbb{R}^n \to \mathbb{R}^m$ and $x^* \in \mathbb{R}^n$. We write*

(a) *$f(x) = \mathcal{O}(g(x))$ for $x \to x^*$ if and only if there exist a uniform constant $\lambda > 0$ and a neighborhood U of x^* such that for all $x \in U \setminus \{x^*\}$ the following relation holds true:*

$$\|f(x)\| \leq \lambda \|g(x)\|.$$

(b) *$f(x) = o(g(x))$ for $x \to x^*$ if and only if for all $\epsilon > 0$ there exists a neighborhood U of x^* such that for all $x \in U \setminus \{x^*\}$ we have*

$$\|f(x)\| \leq \epsilon \|g(x)\|.$$

REMARK 4.3. If $\lim_k x^k = x^*$, then $\{x^k\}$ converges to x^* (at least)

(1) Q-superlinearly, if $\|x^{k+1} - x^*\| = o(\|x^k - x^*\|)$;
(2) Q-quadratically, if $\|x^{k+1} - x^*\| = \mathcal{O}(\|x^k - x^*\|^2)$.

2. Characterizations

The aim is to specify an alternative characterization for (Q-)superlinear and (Q-)quadratic convergence of a sequence $\{x^k\}$. For that purpose we need some auxiliary results which will also be applied in the subsequent chapters.

LEMMA 4.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $\{x^k\} \subset \mathbb{R}^n$ with $\lim_{k\to\infty} x^k = x^* \in \mathbb{R}^n$. Then the following assertions holds true:*

(a) *If f is twice continuously differentiable, then*

$$\|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^k)(x^k - x^*)\| = o(\|x^k - x^*\|).$$

(b) *If, in addition, $\nabla^2 f$ is locally Lipschitz-continuous, then*

$$\|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^k)(x^k - x^*)\| = \mathcal{O}(\|x^k - x^*\|^2).$$

PROOF. (a)The triangle inequality yields

$$\begin{aligned}\|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^k)(x^k - x^*)\| \leq \\ \leq \|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^*)(x^k - x^*)\| + \|\nabla^2 f(x^*) - \nabla^2 f(x^k)\| \cdot \|x^k - x^*\|.\end{aligned}$$

Since $f \in \mathcal{C}^2$ by assumption, we have

$$\|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^*)(x^k - x^*)\| = o(\|x^k - x^*\|)$$

as well as

$$\|\nabla^2 f(x^k) - \nabla^2 f(x^*)\| \longrightarrow 0 \text{ for } k \to \infty.$$

and thus assertion (a).

(b) The Mean Value Theorem yields

$$\begin{aligned}\|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^k)(x^k - x^*)\| = \\ &= \| \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))(x^k - x^*) d\tau - \nabla^2 f(x^k)(x^k - x^*)\| \\ &\leq \int_0^1 \|\nabla^2 f(x^* + \tau(x^k - x^*)) - \nabla^2 f(x^k)\| \cdot \|x^k - x^*\| d\tau \\ &\leq L\|x^k - x^*\| \int_0^1 \|(\tau - 1)(x^k - x^*)\| d\tau \\ &= \frac{L}{2}\|x^k - x^*\|^2 = \mathcal{O}(\|x^k - x^*\|^2).\end{aligned}$$

□

The following lemma ensures that for sufficiently small $\epsilon > 0$,

$$\|\nabla f(x^k)\| \leq \epsilon \tag{4.3}$$

represents a reasonable (and frequently applied) stopping criterion.

LEMMA 4.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable, $\{x^k\} \subset \mathbb{R}^n$ with $\lim_k x^k = x^*$, $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ nonsingular. Then there exist an index $k_0 \in \mathbb{N}$ and a constant $\beta > 0$ such that*

$$\|\nabla f(x^k)\| \geq \beta \|x^k - x^*\| \quad \textit{for all } k \geq k_0.$$

PROOF. Since $f \in \mathcal{C}^2$ we get

$$\|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^*)(x^k - x^*)\| = o\left(\|x^k - x^*\|\right).$$

Hence, for every $\varepsilon > 0$ there exists an index $k_0(\varepsilon) \in \mathbb{N}$ with

$$\|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^*)(x^k - x^*)\| \leq \varepsilon \|x^k - x^*\| \; \forall k \geq k_0(\varepsilon).$$

We assume, with out loss of generality, that $\epsilon < 1/\|\nabla^2 f(x^*)^{-1}\|$. Consequently $\forall k \geq k_0(\varepsilon)$ it holds that

$$\begin{aligned}\|\nabla f(x^k)\| =& \|\nabla f(x^k) - \nabla f(x^*) - \nabla^2(x^*)(x^k - x^*) + \nabla^2 f(x^*)(x^k - x^*)\| \\ \geq& \|\nabla^2 f(x^*)(x^k - x^*)\| - \|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^*)(x^k - x^*)\| \\ \geq& \frac{1}{\|\nabla^2 f(x^*)^{-1}\|}\|x^k - x^*\| - \varepsilon\|x^k - x^*\| \\ =& \beta\|x^k - x^*\| \;\text{ with }\; \beta = \frac{1}{\|\nabla^2 f((x^*)^{-1})\|} - \varepsilon.\end{aligned}$$

Here we used

$$\|x^k - x^*\| = \|\nabla^2 f(x^*)^{-1}\nabla^2 f(x^*)(x^k - x^*)\| \leq \|\nabla^2 f(x^*)^{-1}\| \cdot \|\nabla^2 f(x^*)(x^k - x^*)\|.$$

□

In numerical practice (4.3) is usually implemented as a *relative criterion*. This aspect will be addressed later.
Now we can characterize superlinear convergence of a sequence $\{x^k\}$ as follows.

THEOREM 4.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable, $\{x^k\} \subset \mathbb{R}^n$ with $\lim_k x^k = x^* \in \mathbb{R}^n$, $x^k \neq x^*$ for all $k \in \mathbb{N}$ and $\nabla^2 f(x^*)$ nonsingular. Then the following assertions are equivalent:*

(a) *$\{x^k\}$ converges superlinearly to x^* and $\nabla f(x^*) = 0$.*
(b) *$\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$.*
(c) *$\|\nabla f(x^k) + \nabla^2 f(x^*)(x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$.*

PROOF. (c) $\Longrightarrow$ (a): The Mean Value Theorem yields:

$$\begin{aligned}\|\nabla f(x^{k+1})\| =& \|\nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^*)(x^{k+1} - x^k) + \nabla f(x^k) + \nabla^2 f(x^*)(x^{k+1} - x^k)\| \\ =& \| \int_0^1 (\nabla^2 f(x^k + \tau(x^{k+1} - x^k) - \nabla^2 f(x^*))(x^{k+1} - x^k) d\tau \\ &+ \nabla f(x^k) + \nabla^2 f(x^*)(x^{k+1} - x^k)\| \\ &\leq \int_0^1 \|\nabla^2 f(x^k + \tau(x^{k+1} - x^k) - \nabla^2 f(x^*)\| \, d\tau \cdot \|x^{k+1} - x^k\| \\ &+ \|\nabla f(x^k) + \nabla^2 f(x^*)(x^{k+1} - x^k)\|.\end{aligned}$$

By assumption (c), the continuity of $\nabla^2 f(\cdot)$ and $\lim_{k\to\infty} x^k = x^*$ we infer the existence of a zero sequence $(\varepsilon_k) \subset \mathbb{R}$ with

$$\|\nabla f(x^{k+1})\| \leq \varepsilon_k \|x^{k+1} - x^k\|. \tag{4.4}$$

Hence $\nabla f(x^{k+1}) \to 0$ and consequently $\nabla f(x^*) = 0$. Lemma 4.3 ensures the existence of $\beta \geq 0$ with

$$\|\nabla f(x^{k+1})\| \geq \beta \|x^{k+1} - x^*\|$$

for all sufficiently large $k \in \mathbb{N}$. We infer (using (4.2)) that

$$\beta \|x^{k+1} - x^*\| \leq \varepsilon_k \|x^{k+1} - x^k\| \leq \varepsilon_k (\|x^{k+1} - x^*\| + \|x^k - x^*\|).$$

Now we obtain for sufficiently large $k \in \mathbb{N}$

$$\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \leq \frac{\varepsilon_k}{\beta - \varepsilon_k}$$

and thus superlinear convergence of $\{x^k\}$ to x^*.
(a)$\Longrightarrow$(c): By assumption f is twice continuously differentiable $\Longrightarrow \nabla f$ is locally Lipschitz-continuous. Since $x^k \longrightarrow x^*$, there exists a constant $L > 0$ with

$$\|\nabla f(x^{k+1}) - \nabla f(x^*)\| \leq L \|x^{k+1} - x^*\| \qquad \forall k \in \mathbb{N} \text{ sufficiently large.}$$

Thus $\nabla f(x^*) = 0$ implies

$$\|\nabla f(x^{k+1})\| = \|\nabla f(x^{k+1}) - \nabla f(x^*)\| \leq \frac{L\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \cdot \frac{\|x^k - x^*\|}{\|x^{k+1} - x^k\|} \cdot \|x^{k+1} - x^k\|$$

Since, by assumption, $\{x^k\}$ converges superlinearly to x^* there exists a zero sequence $\{\varepsilon_k\} \subset \mathbb{R}_+$ with

$$\|\nabla f(x^{k+1})\| \leq \varepsilon_k \|x^{k+1} - x^k\| \text{ (use Theorem 4.1)}$$

Moreover it holds that

$$\begin{aligned}\|\nabla f(x^k) + \nabla^2 f(x^*)(x^{k+1} - x^*)\| \leq \\ \leq &\|\nabla f(x^{k+1})\| + \int_0^1 \|\nabla^2 f(x^k + \tau(x^{k+1} - x^k)) - \nabla^2 f(x^*)\| d\tau \cdot \|x^{k+1} - x^k\| \\ \leq &\left(\varepsilon_k + \int_0^1 \|\nabla^2 f(x^k + \tau(x^{k+1} - x^k)) - \nabla^2 f(x^*)\| d\tau\right) \|x^{k+1} - x^k\|\end{aligned}$$

Since x^k converges to x^* we have $x^k + \tau(x^{k+1} - x^k) \longrightarrow x^*$ uniformly in $\tau \in [0, 1]$. Together with the continuity of $\nabla^2 f(\cdot)$, this yields

$$\int_0^1 \|\nabla^2 f(x^k + \tau(x^{k+1} - x^k)) - \nabla^2 f(x^*)\| d\tau \overset{x^k \to x^*}{\longrightarrow} 0,$$

which implies (c). The equivalence (b) $\Longleftrightarrow$ (c) is easy to verify. □

A simple, but very important consequence of Theorem 4.2 is associated with gradient-related methods. For this purpose let $\{H^k\} \subset \mathbb{R}^{n \times n}$ be a sequence of nonsingular matrices and let the sequence $\{x^k\}$ be defined by

$$x^{k+1} := x^k - (H^k)^{-1} \nabla f(x^k), \quad k = 0, 1, 2, \ldots.$$

Suppose $\{x^k\}$ converges to x^* and $\nabla^2 f(x^*)$ is nonsingular, then the following assertions are equivalent:

(a) $\{x^k\}$ converges superlinearly to x^* and $\nabla f(x^*) = 0$.
(b) $\|(\nabla^2 f(x^k) - H^k)(x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$.
(c) $\|(\nabla^2 f(x^*) - H^k)(x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$.

Remark 4.4. The assertion of Theorem 4.2 even holds true if "superlinear" and "$o(\|x^{k+1} - x^k\|)$ are replaced by "quadratically" and "$\mathcal{O}(\|x^{k+1} - x^k\|^2)$".

CHAPTER 5

Gradient based methods

Concerning the general descent method (cf. Algorithm 3.1), we still have to make a reasonable choice of the descent direction d^k.

1. The method of steepest descent

An obvious way to choose d (also with respect to the angle condition from Theorem 3.1 (a)) is by solving

$$\min \quad \nabla f(x)^\top d \quad \text{s.t.} \quad \|d\| = 1. \tag{5.1}$$

The aim is to determine a direction d along which f in x decreases the most (*steepest descent*). Obviously, the solution of (5.1) satisfies

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

The choice

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

yields $\nabla f(x)^\top d = -\|\nabla f(x)\|$ and thus solves (5.1). Applying the Wolfe-Powell step size strategy, it follows immediately from Theorem 3.3 and Theorem 3.1, that every accumulation point of the sequence $\{x^k\}$ is a stationary point of f. An analogous statement holds true for the strict Wolfe-Powell rule. As the Armijo condition does not necessarily satisfy assertion (b) from Theorem 3.1, we would like to indicate the proof. Thereby we assume (for simplicity) a "backtracking" strategy. To begin with, we consider the following lemma.

LEMMA 5.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, $x, d \in \mathbb{R}^n$, $\{x^k\}$, $\{d^k\} \subset \mathbb{R}^n$ with $\lim_k x^k = x$ and $\lim_k d^k = d$ as well as $\{\alpha_k\} \subset \mathbb{R}_{++}$ with $\lim_k \alpha_k = 0$. Then it holds that*

$$\lim_{k\to\infty} \frac{f(x^k + \alpha_k d^k) - f(x^k)}{\alpha_k} = \nabla f(x)^\top d.$$

PROOF. Due to the mean-value theorem, for all k there exists a vector $\xi^k \in \mathbb{R}^n$ on the line segment joining x^k and $x^k + \alpha_k d^k$ with

$$f(x^k + \alpha_k d^k) - f(x^k) = \alpha_k \nabla f(\xi^k)^\top d^k.$$

Since $\xi^k \to x$ as $\alpha_k \to 0$ and $f \in C^1$, it follows

$$\nabla f(\xi^k)^\top d^k \longrightarrow \nabla f(x)^\top d.$$

□

This lemma is useful for the following convergence theorem.

THEOREM 5.1. *If $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, then every accumulation point of a sequence $\{x^k\}$ generated by Algorithm 3.1 with Armijo step size strategy and $d^k = -\nabla f(x^k)/\|\nabla f(x^k)\|$ is a stationary point of f.*

PROOF. Let $x^* \in \mathbb{R}^n$ be an accumulation point of $\{x^k\}$, and let $\{x^{k(l)}\}$ be a subsequence converging to x^*. Assume $\nabla f(x^*) \neq 0$. Since $\{f(x^k)\}$ is monotonically decreasing and the subsequence $\{f(x^{k(l)})\}$ converges to $f(x^*)$, the entire sequence $\{f(x^k)\}$ converges to $f(x^*)$. Thus,

$$f(x^{k+1}) - f(x^k) \to 0 \text{ für } k \to \infty.$$

¿From $\nabla f(x^{k(l)}) \to \nabla f(x^*) \neq 0$ it follows that $\alpha_{k(l)} \to 0$, where $\alpha_{k(l)} = \beta^{m_{k(l)}}$, $m_{k(l)} \in \mathbb{N}$ denoting the uniquely determined exponent from the Armijo rule. Consequently it holds that

$$f(x^{k(l)} + \beta^{m_{k(l)}-1} d^{k(l)}) > f(x^{k(l)}) + \sigma \beta^{m_{k(l)}-1} \nabla f(x^{k(l)})^\top d^{k(l)}$$

for sufficiently large l. This implies

$$\frac{f(x^{k(l)} + \beta^{m_{k(l)}-1} d^{k(l)})}{\beta^{m_{k(l)}-1}} > \sigma \nabla f(x^{k(l)})^\top d^{k(l)}.$$

For $l \to \infty$, it follows from $\beta^{m_{k(l)}-1} \to 0$ and Lemma 5.1 that

$$-\|\nabla f(x^*)\|^2 \geq -\sigma \|\nabla f(x^*)\|^2.$$

Since $\nabla f(x^*) \neq 0$ and $\sigma \in (0,1)$ this yields a contradiction. □

The rate of convergence of the steepest descent method can be very slow. We want to illustrate this with the help of quadratic functions: Let

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

with $Q \in \mathcal{S}^n$ positive definite, $c \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$. Let x^k, d^k be given. We have

$$\begin{aligned}\varphi(\alpha) :=& \frac{1}{2}(x^k + \alpha d^k)^\top Q (x^k + \alpha d^k) + c^\top (x^k + \alpha d^k) + \gamma \\ =& \frac{\alpha^2}{2} (d^k)^\top Q d^k + \alpha ((d^k)^\top Q x^k + c^\top d^k) + \frac{1}{2}(x^k)^\top Q x^k + c^\top x^k + \gamma.\end{aligned}$$

The quadratic nature of $\varphi(\alpha) = f(x^k + \alpha d^k)$ allows to explicitly determine the minimizing step size α_k, i.e.

$$f(x^k + \alpha_k d^k) = \min\{f(x^k + \alpha d^k) : \alpha \geq 0\}.$$

The step size α_k fulfills $\varphi'(\alpha_k) = 0$ with

$$\alpha_k = -\frac{(Qx^k + c)^\top d^k}{(d^k)^\top Q d^k} = -\frac{\nabla f(x^k)^\top d^k}{(d^k)^\top Q d^k} = \frac{\nabla f(x^k)^\top \nabla f(x^k)}{\nabla f(x^k)^\top Q \nabla f(x^k)},$$

where we use $d^k = -\nabla f(x^k)$ (without scaling by $\|\nabla f(x^k)\|^{-1}$). It can be shown that

$$f(x^k) - f(x^{k+1}) \leq \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^2 (f(x^k) - f(x^*)), \tag{5.2}$$

where $\lambda_{\max} \geq \lambda_{\min} > 0$ are the largest resp. the smallest eigenvalues of Q and x^* is the global minimizer of f. Let $\kappa = \lambda_{\max}/\lambda_{\min}$ be the spectral condition number of the matrix Q. Then

(5.2) is equivalent to

$$f(x^k) - f(x^{k+1}) \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^2 (f(x^k) - f(x^*)).$$

Furthermore, $\lambda_{\min} x^\top x \leq x^\top Q x \leq \lambda_{\max} x^\top x$ implies

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

Evidently the rate of convergence is very slow if κ is very large (*zig-zagging effect*).

2. Gradient-related methods

A possible remedy for the slow convergence of the method of steepest descent consists in choosing

$$d^k = -H^{-1} \nabla f(x^k),$$

where $H \in \mathcal{S}^n$ is positive definite. Additionally, the matrix H should be chosen such that

$$0 < \frac{\lambda_{\max}(H^{-1}Q)}{\lambda_{\min}(H^{-1}Q)} < \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}$$

and $Hd^k = -\nabla f(x^k)$ is simple to solve.

Definition 5.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and $\{x^k\} \subset \mathbb{R}^n$. A sequence $\{d^k\} \subset \mathbb{R}^n$ is called gradient-related w.r.t. f and $\{x^k\}$, if for every subsequence $\{x^{k(l)}\}$ converging to a nonstationary point of f, there exists $c > 0$ and $\epsilon > 0$ such that*

(a) $\|d^{k(l)}\| \leq c$ *for all* $l \in \mathbb{N}$,
(b) $\nabla f(x^{k(l)})^\top d^{k(l)} \leq -\epsilon$ *for all sufficiently large* $l \in \mathbb{N}$.

Remark 5.1.

(1) Let $\{H^k\} \subset \mathcal{S}^n$ be a sequence of positive definite matrices, which fulfill

$$c_1 \|x\|^2 \leq x^\top H^k x \leq c_2 \|x\|^2 \quad \text{for all } x \in \mathbb{R}^n,\ k \in \mathbb{N},$$

with constants $c_2 \geq c_1 > 0$. Then $\{d^k\}$ given by

$$H^k d^k = -\nabla f(x^k) \quad \text{for all } k \in \mathbb{N}$$

is gradient-related.

(2) For Algorithm 3.1 with gradient-related search directions and Armijo step size strategy an analogous statement to Theorem 5.1.

(3) Sometimes the choice

$$H^k = \operatorname{diag}(h_{ii}^k) \quad \text{with} \quad h_{ii}^k = \frac{\partial^2 f(x^k)}{\partial x_i^2}$$

results in a significant improvement of the rate of convergence.

CHAPTER 6

Conjugate gradient method

We first derive the conjugate gradient method for minimizing strictly convex quadratic functions. Then we transfer the technique to minimization problems of general nonlinear functions. In this context we consider the Fletcher-Reeves and the Polak-Ribière variants of the conjugate gradient (CG) method. The two versions differ in the update strategy of a scalar which has an impact on the determination of the search direction and the line search algorithm. While the original Polak-Ribière method requires an impractical step size strategy in order to be analyzed successfully, we will briefly elaborate on a modified Polak-Ribière variant of the conjugate gradient method, which is based on an implementable step size strategy.

1. Quadratic minimization problems

As we have already seen in section 1 of chapter 5, the method of steepest descent may be very slow even if an exact line search is performed. Still utilizing gradient information only, in this section our goal is to devise a strategy which overcomes (to some extent) this difficulty of the steepest descent method and which is computationally efficient. In fact, often one is confronted with minimizing

$$f(x) = \frac{1}{2}x^\top Ax - b^\top x + c \quad \text{with } A \in \mathcal{S}^n \text{ pos. def.},$$

where A is a very large matrice often possessing special structure. The last aspect frequently allows an efficient computation of the matrix-vector product Ax: For instance if A is a dense matrix being the product of sparse matrices, i.e.,

$$A = Q_l \cdot Q_{l-1} \cdots Q_1, \quad l \in \mathbb{N},$$

then it is easy to calculate the product Ax iteratively according to

$$y^{i+1} = Q_i y^i, \quad y^0 = x$$

with $y^l = Ax$, exploiting the structure of Q_i. In the subsequent chapter 8 we will study a strategy which approximates the Hessian $\nabla^2 f$ by means of special difference methods. In every iteration k of the resulting method (Quasi-Newton method) one determines the search direction d^k as the solution of $H^k d = -\nabla f(x^k)$, where H^k is an approximation of $\nabla^2 f(x^k)$. Then a possible choice of H^{k+1} yielding a positive definite symmetric approximation of $\nabla^2 f(x^k)$ is given by

$$H^{k+1} = H^k + \frac{y^k (y^k)^\top}{(s^k)^\top y^k} - \frac{H^k s^k (s^k)^\top H^k}{(s^k)^\top H^k s^k},$$

where $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$ and $s^k = x^{k+1} - x^k$. If indeed one performs the multiplications on the right hand side, then one usually obtains a dense matrix H^{k+1}, even if $\nabla^2 f(x^{k+1})$ is sparse. In case of very large Hessians this would directly lead to storage problems. Alternatively one could only store the vectors (s^k, y^k) in every iteration k and determine the

matrix-vector product $H^{k+1}d$ iteratively as a vector-vector product according to the update formula. In this way one can avoid the storage problems mentioned above.
The conjugate gradient method introduced in this section, when applied to quadratic functions with positive definite Hessians A, terminates after finitely many steps and requires only matrix-vector products involving A. Hence it is not necessary to store A as an array.
The necessary and sufficient condition for the minimizer x^* of

$$\frac{1}{2}x^\top Ax - b^\top x + c \quad \text{with } A \in \mathcal{S}^n \text{ pos. def.}$$

is

$$Ax^* = b.$$

For a given initial value x^0 we now specify a method which iteratively constructs a solution to

$$Ax = b. \tag{6.1}$$

The following result motivates our strategy.

LEMMA 6.1. *Let $f(x) = \frac{1}{2}x^\top Ax - b^\top x + c$ with $A \in \mathcal{S}^n$ positive definite, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$ and $x^0 \in \mathbb{R}^n$. Further let $d^0, d^1, \ldots, d^{n-1} \in \mathbb{R}^n$ be different from the null vector with*

$$(d^i)^\top Ad^j = 0, \quad \forall i, j = 0, 1, \ldots, n-1,\ i \neq j. \tag{6.2}$$

Then the method of successive one-dimensional minimization along the directions $d^0, d^1, \ldots, d^{n-1}$, i.e. the computation of the sequence $\{x^k\}$ according to

$$x^{k+1} = x^k + \alpha_k d^k$$

with

$$f(x^k + \alpha_k d^k) = \min_{\alpha \in \mathbb{R}} f(x^k + \alpha d^k), \quad k = 0, 1, \ldots, n-1 \tag{6.3}$$

yields, after at most n steps, the minimizer $x^n = x^*$ of f. Furthermore, for $k = 0, \ldots, n-1$ with $g^k := Ax^k - b = \nabla f(x^k)$ it holds that

$$\alpha_k = -\frac{(g^k)^\top d^k}{(d^k)^\top Ad^k} \tag{6.4}$$

and

$$(g^{k+1})^\top d^j = 0, \quad j = 0, 1, \ldots, k. \tag{6.5}$$

PROOF. We have

$$\begin{aligned} f(x^k + \alpha d^k) &= \frac{1}{2}\alpha^2 d^{kT} Ad^k + \alpha(x^{kT} Ad^k - b^\top d^k) + c + \frac{1}{2}(x^k)^\top Ax^k - b^\top x^k \\ &= \frac{1}{2}\alpha^2 (d^k)^\top Ad^k + \alpha(g^k)d^k + c + \frac{1}{2}(x^k)^\top Ax^k - b^\top x^k. \end{aligned}$$

From (6.3) it follows that the minimizer α_k fulfills

$$\alpha_k (d^k)^\top Ad^k + (g^k)^\top d^k = 0.$$

This proves (6.4). Moreover, we obtain

$$\begin{aligned} 0 &= \alpha_k (d^k)^\top Ad^k + (g^k)^\top d^k \\ &= (\alpha_k (d^k)^\top A + (x^k)^\top A - b^\top) d^k \\ &= (A(x^k + \alpha_k d^k) - b)^\top d^k = (Ax^{k+1} - b)^\top d^k. \end{aligned}$$

We infer

$$(g^{k+1})^\top d^k = 0 \qquad \text{for } k = 0, 1, ..., n-1. \tag{6.6}$$

The application of (6.2) for $i \neq j$ implies

$$(g^{i+1} - g^i)^\top d^j = (Ax^{i+1} - Ax^i)^\top d^j = \alpha_j (d^i)^\top A d^j = 0.$$

Together with (6.6) this yields

$$(g^{k+1})^\top d^j = (g^{j+1})^\top d^j + \Sigma_{i=j+1}^{k} (g^{i+1} - g^i)^\top d^j = 0$$

for $j = 0, ..., k$. This proves (6.5). Due to (6.2), the vectors $d^0, \ldots, d^{n-1}$ are pairwise orthogonal w.r.t. the scalar product

$$< u, v >_A := u^\top A v.$$

Consequently these vectors are also linearly independent. From (6.5), it follows immediately that $g^n = 0$ or equivalently $x^n = x^*$, the solution of the problem (6.1). □

REMARK 6.1. Vectors $d^0, d^1, \ldots, d^{n-1}$ with the property (6.2) are called *A-conjugated* or *A-orthogonal*.

We continue by defining a strategy to determine A-conjugate directions $d^0, d^1, \ldots, d^{n-1}$: We begin with

$$d^0 = -\nabla f(x^0) = -g^0.$$

Assume we already know $l + 1$ vectors $d^0, \ldots, d^l$ with

$$(d^i)^\top A d^j = 0 \quad \text{for } i, j = 0, \ldots, l \text{ mit } i \neq j. \tag{6.7}$$

According to Lemma 6.1, (6.4) and (6.5) hold true for $k = 0, \ldots, l$. We suppose that $g^{l+1} \neq 0$, otherwise we have already found the solution. We make the ansatz

$$d^{l+1} := -g^{l+1} + \sum_{i=0}^{l} \beta_i^l d^i. \tag{6.8}$$

(because of (6.5) g^{l+1} is linearly independent of $d^0, \ldots, d^l$). Our goal is to have

$$(d^{l+1})^\top A d^j = 0 \quad j = 0, \ldots, l.$$

It holds that

$$\begin{aligned}
(d^{l+1})^\top A d^j &= \left(-g^{l+1} + \sum_{i=0}^{l} \beta_i^l d^i \right)^\top A d^j \\
&= -(g^{l+1})^\top A d^j + \sum_{i=0}^{l} \beta_i^l (d^i)^\top A d^j \\
&= -(g^{l+1})^\top A d^j + \beta_j^l (d^j)^\top A d^j,
\end{aligned}$$

since by (6.7), $(d^i)^\top A d^j = 0$ for $i, j = 0, \ldots, l$ with $i \neq j$. This implies

$$\beta_j^l = \frac{(g^{l+1})^\top A d^j}{(d^j)^\top A d^j} \quad \text{for } j = 0, \ldots, l. \tag{6.9}$$

We study further properties. In fact, multiplying (6.8) by $(g^{l+1})^\top$, we obtain

$$(g^{l+1})^\top \left(-g^{l+1} + \sum_{i=0}^{l} \beta_i^l d^i \right) = -\|g^{l+1}\|^2 + \sum_{i=0}^{l} \beta_i^l (g^{l+1})^\top d^i = -\|g^{l+1}\|^2 < 0.$$

Obviously d^{l+1} is a descent direction for f at x^{l+1}. Then, by (6.4),

$$\alpha_{l+1} = -\frac{(g^{l+1})^\top d^{l+1}}{(d^{l+1})^\top A d^{l+1}} > 0.$$

As a result of the construction of d^k, $k = 0, \ldots, l$, it holds that

$$(g^k)^\top d^k = -\|g^k\|^2 < 0 \text{ and } \alpha_k > 0 \text{ for } k = 0, \ldots, l.$$

A further orthogonality relation can be obtained in the following way:

$$(g^{l+1})^\top g^j = (g^{l+1})^\top \left(\sum_{i=0}^{j-1} \beta_i^{j-1} d^i - d^j \right) \tag{6.10}$$

$$= \sum_{i=0}^{j-1} \beta_i^{j-1} (g^{l+1})^\top d^i - (g^{l+1})^\top d^j = 0. \tag{6.11}$$

Thus, for the left hand side in (6.9) we get

$$(g^{l+1})^\top A d^j = \frac{1}{\alpha_j} (g^{l+1})^\top (g^{j+1} - g^j) \overset{(6.11)}{=} 0.$$

because of $g^{j+1} - g^j = Ax^{j+1} - Ax^j = \alpha_j A d^j$ for $j = 0, \ldots, l-1$. Hence,

$$\beta_j^l = 0 \quad \text{for } j = 0, \ldots, l-1$$

and

$$\begin{aligned} \beta_l^l &= \frac{(g^{l+1})^\top A d^l}{(d^l)^\top A d^l} = \frac{1}{\alpha_l} \frac{(g^{l+1})^\top (g^{l+1} - g^l)}{(d^l)^\top A d^l} \\ &= \frac{\|g^{l+1}\|^2}{(d^l)^\top (g^{l+1} - g^l)} = \frac{\|g^{l+1}\|^2}{(-d^l)^\top g^l} = \frac{\|g^{l+1}\|^2}{\|g^l\|^2} =: \beta_l. \end{aligned}$$

Consequently (6.8) is reduced to

$$d^{l+1} = -g^{l+1} + \beta_l d^l.$$

Due to $g^k = Ax^k - b$ we also have:

$$g^{k+1} - g^k = Ax^{k+1} - Ax^k = \alpha_k A d^k.$$

Therefore g^k can be updated at each step without requiring a further matrix-vector product. The product Ad^k was already necessary to determine α_k. To spare the evaluation of the scalar product, one can rearrange (6.4) by means of $(g^k)^\top d^k = -\|g^k\|^2$ to

$$\alpha_k = \frac{\|g^k\|^2}{(d^k)^\top A d^k}.$$

The CG-algorithm is as follows:

ALGORITHM 6.1 (CG-algorithm for quadratic functions).

input: $x^0 \in \mathbb{R}^n$
begin
set $g^0 := Ax^0 - b$, $d^0 := -g^0$, $k := 0$; *choose* $\epsilon \geq 0$.
while $\|g^k\| > \epsilon$
begin
set

$$\begin{aligned} \alpha_k &:= \frac{\|g^k\|^2}{(d^k)^\top A d^k} \\ x^{k+1} &:= x^k + \alpha_k d^k \\ g^{k+1} &:= g^k + \alpha_k A d^k \\ \beta_k &:= \frac{\|g^{k+1}\|^2}{\|g^k\|^2} \\ d^{k+1} &:= -g^{k+1} + \beta_k d^k \\ k &:= k+1 \end{aligned}$$

end
end
end

REMARK 6.2. (1) The main complexity of Algorithm 6.1 consists of the calculation of Ad^k. Given that the product is needed twice, one should store it as $z^k := Ad^k$.

(2) Because of (6.11) β_k can be calculated according to

$$\beta_k = \frac{(g^{k+1} - g^k)^\top g^{k+1}}{\|g^k\|^2}. \tag{6.12}$$

In this case, however, an additional scalar product is required. But from a numerical point of view (6.12) is often more appropriate. The reason is that the directions d^k quickly lose their A-conjugacy due to numerical errors. Consequently the descent property of the direction d^k might get lost and α_k might become very small. Approximately evaluating Ad^k might even imply a negative α_k. Therefore it is recommended to choose

$$\begin{aligned} \beta_k &= \max\{0, \frac{(g^{k+1} - g^k)^\top g^{k+1}}{\|g^k\|^2}\}, \\ \alpha_k &= \max\{0, \frac{\|g^k\|^2}{(d^k)^\top A d^k}\}. \end{aligned}$$

If $\alpha_k \approx 0$ or $\beta_k \approx 0$ due to error influences, then $x^{k+1} = x^k + \alpha_k d^k$ and $g^{k+1} = g^k + \alpha_k A d^k$ yield

$$x^{k+1} \approx x^k \quad \text{and} \quad g^{k+1} \approx g^k.$$

The choice of the next direction d^{k+1} is dominated by $-g^{k+1}$. Thus

$$d^{k+1} \approx -g^{k+1}$$

basically corresponds to the direction of steepest descent in x^{k+1}. In this sense a kind of *automatical restart* is carried out.

(3) Even if (6.12) is applied, it is recommended to execute a restart from time to time. The CG-method, which is theoretically a *direct method*, i.e. terminating after finitely many steps with the exact solution, can be numerically regarded as an *iterative method*.

(4) As mentioned before, the CG-method finds the exact solution after at most n steps. Moreover it can be shown: If A possesses m ($\leq n$) different eigenvalues, then the CG-method terminates after m steps with the exact solution. In addition, the method terminates after m steps if b can be represented as a linear combination of at most m eigenvectors of A and if $x^0 = 0$ is used.

Let $\kappa = \lambda_{\max}(A)/\lambda\min(A)$, then the iteration sequence $\{x^k\}$ of the CG-method satisfies

$$\|x^k - x^*\| \leq 2\sqrt{\kappa}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \|x^0 - x^*\|.$$

Obviously, the closer κ approaches 1, the faster the method converges. This leads to the concept of *preconditioning*.
For an efficient preconditioning, our aim is to find a transformation of A, such that as many eigenvalues (of the transformed matrix) as possible are 1 (or cluster near 1). For this purpose, let $W \in \mathcal{S}^n$ be positive definite. The solution of $Ax = b$ can be found by solving the system

$$W^{-1/2}AW^{-1/2}y = W^{-1/2}b$$

and using $x = W^{-1/2}y$. The matrix $W^{-1/2}AW^{-1/2} =: R$ has the same eigenvalues as $W^{-1}A$, since $W^{-1/2}RW^{1/2} = W^{-1}A$. The quest of determining W such that as many eigenvalues as possible are 1 and all others are close to 1 corresponds to the fact that the condition number of $W^{-1}A$ is preferably small. The matrix W is called *preconditioning matrix* or simply *preconditioner*. We note that in practice one only employs W, but not $W^{1/2}$.

Algorithm 6.2 (Preconditioned CG-algorithm for quadratic functions).

input: $x^0 \in \mathbb{R}^n$, $W \in \mathcal{S}^n$ *positive definite.*
begin
set $g^0 := Ax^0 - b$, $d^0 := -W^{-1}g^0$, $k := 0$; *choose* $\epsilon \geq 0$.
while $\|g^k\| > \epsilon$
begin
set

$$\begin{aligned}
\alpha_k &:= \frac{(g^k)^\top W^{-1} g^k}{(d^k)^\top A d^k} \\
x^{k+1} &:= x^k + \alpha_k d^k \\
g^{k+1} &:= g^k + \alpha_k A d^k \\
\beta_k &:= \frac{(g^{k+1})^\top W^{-1} g^{k+1}}{(g^k)^\top W^{-1} g^k} \\
d^{k+1} &:= -W^{-1}g^{k+1} + \beta_k d^k \\
k &:= k+1
\end{aligned}$$

end
end
end

Naturally, in numerical practice one does not evaluate W^{-1}. Merely

$$Wd = g$$

is solved. This demands a certain efficiency when solving the system and often requires a compromise between "$\kappa(W^{-1}A)$ preferably close to 1" and simple solvability of $Wd = g$.

2. Nonlinear functions

2.1. Fletcher-Reeves method. The elementary structure of Algorithm 6.1 is the motivation for the following variant of the conjugate gradient method for the minimization of continuously differentiable but not necessarily quadratic functions.

ALGORITHM 6.3 (Fletcher-Reeves method).

input: $x^0 \in \mathbb{R}^n$, $0 < \sigma < \rho < \frac{1}{2}$.
begin
set $d^0 := -\nabla f(x^0)$, $k := 0$*; choose* $\epsilon \geq 0$.
while $\|\nabla f(x^k)\| > \epsilon$
begin
determine α_k *s.t. the strong Wolfe-Powell rule is satisfied, i.e.*

$$\begin{aligned} &f(x^k + \alpha_k d^k) \leq f(x^k) + \sigma\alpha_k \nabla f(x^k)^\top d^k \ \textit{and} \\ &|\nabla f(x^k + \alpha_k d^k)^\top d^k| \leq -\rho \nabla f(x^k)^\top d^k. \end{aligned}$$

set

$$\begin{aligned} x^{k+1} &:= x^k + \alpha_k d^k \\ \beta_k^{FR} &:= \frac{\|\nabla f(x^{k+1})\|^2}{\|\nabla f(x^k)\|^2} \\ d^{k+1} &:= -\nabla f(x^{k+1}) + \beta_k^{FR} d^k \\ k &:= k + 1 \end{aligned}$$

end
end
end

Note that we now require $\rho \in (\sigma, \frac{1}{2})$ which is more restrictive than the condition $\rho \in [\sigma, 1)$ introduced in chapter 2.3. The current restriction is due to the convergence analysis of the method. Firstly it can be shown that Algorithm 6.3 is well-defined for a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ which is bounded from below. Moreover we have the following convergence property.

THEOREM 6.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable, bounded from below and ∇f Lipschitz continuous on $L(x^0) := \{x \in \mathbb{R}^n : f(x) \le f(x^0)\}$. Then it holds that*

$$\liminf_{k\to\infty} \|\nabla f(x^k)\| = 0.$$

If $L(x^0)$ is convex, f twice continuously differentiable and uniformly convex on $L(x^0)$, then the sequence $\{x^k\}$ generated by the Fletcher-Reeves method (Algorithm 6.3) converges to the unique global minimizer of f.

2.2. Polak-Ribière method and modifications. In numerical practice it is often observed that variants of the following nonlinear CG-method (*Polak-Ribière method*) have better convergence behavior.

ALGORITHM 6.4 (Polak-Ribière method).

input: $x^0 \in \mathbb{R}^n$.
begin
set $d^0 := -\nabla f(x^0)$, $k := 0$; *choose* $\epsilon \ge 0$.
while $\|\nabla f(x^k)\| > \epsilon$
begin
determine α_k *such that*

$$\alpha_k = \min\{\alpha > 0 | \nabla f(x^k + \alpha d^k)^\top d^k = 0\}. \tag{6.13}$$

set

$$\begin{aligned} x^{k+1} &:= x^k + \alpha_k d^k \\ \beta_k^{PR} &:= \frac{(\nabla f(x^{k+1}) - \nabla f(x^k))^\top \nabla f(x^{k+1})}{\|\nabla f(x^k)\|^2} \\ d^{k+1} &:= -\nabla f(x^{k+1}) + \beta_k^{PR} d^k \\ k &:= k+1 \end{aligned}$$

end
end
end

The Fletcher-Reeves and the Polak-Ribière-method differ in the strategy for determining α_k and for the choice of β_k:

- The step size choice (6.13) in the Polak-Ribière method is impractical, but necessary for the convergence analysis. However in numerical practice the strong Wolfe-Powell rule (here with small ρ), which is also applied in the Fletcher-Reeves algorithm, yields satisfying results. There also exist so-called *modified Polak-Ribière methods*, which work with an implementable step size strategy. In one instance, one computes α_k such that $x^{k+1} = x^k + \alpha_k d^k$ and $d^{k+1} = -\nabla f(x^{k+1}) + \beta_k^{PR} d^k$ satisfy the following conditions:

$$f(x^{k+1}) \le f(x^k) - \sigma \alpha_k^2 \|d^k\|^2 \text{ and} \tag{6.14}$$

$$-\delta_2 \|\nabla f(x^{k+1})\|^2 \le \nabla f(x^{k+1})^\top d^{k+1} \le -\delta_1 \|\nabla f(x^{k+1})\|^2, \tag{6.15}$$

where $\alpha_k = \max\{\rho_k\beta^\ell : \ell = 0, 1, 2, \ldots\}$ with $\rho_k := |\nabla f(x^k)^\top d^k|/\|d^k\|^2$ s.t. (6.14) and (6.15) are satisfied. The parameter restrictions are: $\beta \in (0,1)$, $\sigma \in (0,1)$ and $0 < \delta_1 < 1 < \delta_2$.

Under stronger assumptions compared to the ones for the Fletcher-Reeves method, it can be shown that a sequence $\{x^k\}$ generated by the modified Polak-Ribière method satisfies:

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

Provided that the level set $L(x^0)$ is convex and f uniformly convex on $L(x^0)$, then it holds again that the sequence $\{x^k\}$ converges to the uniquely determined global minimizer of f.

- Concerning the choice of β_k^{FR} resp. β_k^{PR} note that in the case where d^k yields hardly any progress in the minimization of f, owing to bad descent- and conjugation properties, the Polak-Ribière method tends to perform better than the Fletcher-Reeves method. This can be seen in the following way: In the just mentioned situation we can expect x^{k+1} to be close to x^k, because α_k is very small. Hence $\nabla f(x^{k+1})$ will also be close to $\nabla f(x^k)$, even if $\|\nabla f(x^{k+1})\|$ is still relatively large. In such cases $0 \leq |\beta_k^{PR}| \ll \beta_k^{FR}$ can be expected. If β_k^{PR} is close to 0, then it holds that $d^{k+1} \approx -\nabla f(x^{k+1})$ and the method almost carries out a restart.

CHAPTER 7

Newton's method

From now on we assume that f and its local minimizers x^* fulfill the following conditions:

(A)
1. f is twice continuously differentiable with $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq \gamma \|x - y\|$ in a neighborhood of x^* with
2. $\nabla f(x^*) = 0$,
3. $\nabla^2 f(x^*)$ positive definite.

To simplify notations we will frequently denote the current iteration point by x_a and the new iterate by x_+.
We consider the following quadratic model of f near x_a:

$$m_a(x) = f(x_a) + \nabla f(x_a)^\top (x - x_a) + \frac{1}{2}(x - x_a)^\top \nabla^2 f(x_a)(x - x_a).$$

If $\nabla^2 f(x_a)$ is positive definite, then we define x_+ as the minimizer of $m_a(x)$:

$$0 = m_a'(x_+) = \nabla f(x_a) + \nabla^2 f(x_a)(x_+ - x_a).$$

Rearranging yields the *iteration rule of Newton's method, i.e.*

$$x_+ = x_a - (\nabla^2 f(x_a))^{-1} \nabla f(x_a).$$

Naturally the inverse $(\nabla^2 f(x_a))^{-1}$ is not computed. Merely

$$\nabla^2 f(x_a) d = -\nabla f(x_a)$$

is solved, and we set $x_+ = x_a + d$. If x_a is far away from a local minimizer, then $\nabla^2 f(x_a)$ might possess negative eigenvalues. Then x_+ can be a local maximizer or a saddle point of m_a. In order to cope with this, we have to introduce certain modifications. But for the time being we assume that x_a is sufficiently close to a local minimizer.
In what follows we will often make use of the following result:

LEMMA 7.1. *Let* (A) *be satisfied. Then there exists $\delta > 0$ such that for all $x \in B(\delta) := \{y : \|y - x^*\| < \delta\}$ it holds that*

$$\begin{aligned}
&\|\nabla^2 f(x)\| \leq 2\|\nabla^2 f(x^*)\|, \\
&\|(\nabla^2 f(x))^{-1}\| \leq 2\|(\nabla^2 f(x^*))^{-1}\|, \\
&\frac{\|x - x^*\|}{2\|(\nabla^2 f(x^*))^{-1}\|} \leq \|\nabla f(x)\| \leq 2\|\nabla^2 f(x^*)\| \, \|x - x^*\|.
\end{aligned}$$

This enables us to prove *local convergence* of Newton's method.

THEOREM 7.1. *Let* (A) *be satisfied. Then there exists constants $K > 0$ and $\delta > 0$ (independent of x_a and x_+) such that, for $x_a \in B(\delta)1$, the Newton step*

$$x_+ = x_a - (\nabla^2 f(x_a))^{-1} \nabla f(x_a)$$

satisfies the following estimate:

$$\|x_+ - x^*\| \leq K\|x_a - x^*\|^2.$$

PROOF. Let $\delta > 0$ be chosen sufficiently small such that the assertion of Lemma 7.1 holds true. Then it holds that

$$\begin{aligned} x_+ - x^* &= x_a - x^* - (\nabla^2 f(x_a))^{-1}\nabla f(x_a) \\ &= (\nabla^2 f(x_a))^{-1}(\nabla^2 f(x_a)(x_a - x^*) - \nabla f(x_a)) \\ &= (\nabla^2 f(x_a))^{-1}(\nabla^2 f(x_a)(x_a - x^*) - \nabla f(x^*) \\ &\quad - \int_0^1 \nabla^2 f(x^* + t(x_a - x^*))(x_a - x^*)dt) \\ &= (\nabla^2 f(x_a))^{-1} \int_0^1 \left(\nabla^2 f(x_a) - \nabla^2 f(x^* + t(x_a - x^*))\right)(x_a - x^*)dt \end{aligned}$$

Thus, we have

$$\begin{aligned} \|x_+ - x^*\| &\leq \|(\nabla^2 f(x_a))^{-1}\| \cdot \int_0^1 \|\nabla^2 f(x_a) - \nabla^2 f(x^* + t(x_a - x^*))\|dt\|x_a - x^*\| \\ &\leq 2\|(\nabla^2 f(x^*))^{-1}\| \cdot \gamma \int_0^1 \|x_a - x^* - t(x_a - x^*)\|dt \cdot \|x_a - x^*\| \\ &= 2\gamma\|(\nabla^2 f(x^*))^{-1}\| \cdot \|x_a - x^*\|^2 \cdot \int_0^1 (1-t)dt \\ &= 2\gamma\|(\nabla^2 f(x^*))^{-1}\| \cdot \frac{1}{2} \cdot \|x_a - x^*\|^2 \\ &= \gamma\|(\nabla^2 f(x^*))^{-1}\| \cdot \|x_a - x^*\|^2. \end{aligned}$$

Setting $K := \gamma\|(\nabla^2 f(x^*))^{-1}\|$ proves the assertion. □

Theorem 7.1 immediately implies the local Q-quadratic convergence of Newton's method.

THEOREM 7.2. *Let* (A) *be satisfied. Then there exists $\delta > 0$ such that Newton's method*

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1}\nabla f(x^k)$$

for $x^0 \in B(\delta)$ converges Q-quadratically to x^*.

PROOF. Let $\delta > 0$ be sufficiently small such that the assertion of Theorem 7.1 holds true. If necessary reduce δ to guarantee $K\delta =: \mu < 1$. For $k \geq 0$ and $x^k \in B(\delta)$, Theorem 7.1 yields

$$\|x^{k+1} - x^*\| \leq K\|x^k - x^*\|^2 \leq \mu\|x^k - x^*\| < \|x^k - x^*\| < \delta. \tag{7.1}$$

Thus $x^{k+1} \in B(\mu\delta) \subset B(\delta)$. As $x^0 \in B(\delta)$, this implies that Newton's method is well-defined and $\{x^k\} \subset B(\delta)$. Now, (7.1) yields the Q-quadratic convergence of $x^k \to x^*$. □

A canonical stopping criterion for Newton's method (as well as for the gradient based methods of section 5 and the CG-method of section 1) consists of a *relative* and an *absolute error bound*. Let $\tau_r \in (0,1)$ be a desired reduction in the gradient norm and τ_a, with $1 \gg \tau_a > 0$, an absolute error bound, then the algorithm terminates as soon as

$$\|\nabla f(x^k)\| \leq \tau_r\|\nabla f(x^0)\| + \tau_a$$

holds true.

1. Inaccuracies in function, gradient and Hessian evaluation

We discuss the error influence by means of an one-dimensional problem. For that purpose assume that f can be approximately evaluated, *i.e.*

$$\tilde{f}(\cdot) = f(\cdot) + \tilde{\epsilon}_f(\cdot) \text{ with errors } \tilde{\epsilon}_f(\cdot) \geq 0 \text{ and } |\tilde{\epsilon}_f(\cdot)| \leq \epsilon_f.$$

Determining the derivatives numerically e.g. by forward differences

$$D_h^+ f(x) = \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h}$$

results in

$$\begin{aligned}
\|D_h^+ f(x) - f'(x)\| &= \|\frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} - f'(x)\| \\
&= \|\frac{f(x+h) + \tilde{\epsilon}_f(x+h) - f(x) - \tilde{\epsilon}_f(x)}{h} - f'(x)\| \\
&\leq \|\frac{f(x+h) - f(x)}{h} - f'(x)\| + \frac{2\epsilon_f}{h} \\
&= \frac{1}{2}\|f''(\xi)h\| + \frac{2\epsilon_f}{h} = \mathcal{O}(h + \frac{\epsilon_f}{h}).
\end{aligned}$$

Here ξ lies on the line segment joining x and $x+h$. The minimizer h^* of the error function $\text{err}_+(h) = h + \frac{\epsilon_f}{h}$ fulfills

$$\text{err}'_+(h^*) = 1 - \frac{\epsilon_f}{(h^*)^2} = 0.$$

This implies

$$h^* = \sqrt{\epsilon_f} \quad \text{and} \quad \text{err}_+(h^*) = 2\sqrt{\epsilon_f}.$$

For the error in the gradient we obtain

$$\epsilon_g = \mathcal{O}(h^*) = \mathcal{O}(\sqrt{\epsilon_f}).$$

Applying once again forward differences to calculate the Hessian, we obtain that the error ϵ_H is of order

$$\epsilon_H = \mathcal{O}(\sqrt{\epsilon_g}) = \mathcal{O}(\epsilon_f^{1/4}).$$

This implies that Hessian matrices computed by two numerical differentiations are relatively inaccurate. Even if we assume that $\epsilon_f \approx 10^{-16}$ (order of machine accuracy!) the error in the approximation of the Hessian is $\epsilon_H \approx 10^{-4}$!
The application of central (or symmetric) differences yields better results, *i.e.*

$$D_h^0 f(x) = \frac{\tilde{f}(x+h) - \tilde{f}(x-h)}{2h}.$$

It holds that

$$\begin{aligned}
\|D_h^0 f(x) - f'(x)\| &= \|\frac{f(x+h) + \tilde{\epsilon}_f(x+h) - f(x-h) - \tilde{\epsilon}_f(x-h)}{2h} - f'(x)\| \\
&\leq \|\frac{f(x+h) - f(x-h)}{2h} - f'(x)\| + \frac{\epsilon_f}{h} \\
&= \frac{1}{12}(\|f'''(\xi_1)\| + \|f'''(\xi_2)\|)h^2 + \frac{\epsilon_f}{h} = \mathcal{O}(h^2 + \frac{\epsilon_f}{h}).
\end{aligned}$$

For the estimates above one uses third order Taylor expansions and intermediate values ξ_1, ξ_2. The minimizer h^* of $\mathrm{err}_0(h) = h^2 + \frac{\epsilon_f}{h}$ fulfills

$$\mathrm{err}_0'(h^*) = 2h^* - \frac{\epsilon_f}{(h^*)^2} = 0,$$

yielding

$$h^* = \epsilon_f^{1/3} \quad \text{and} \quad \mathrm{err}_0(h^*) = \mathcal{O}(\epsilon_f^{2/3}).$$

Thus the gradient error is of the order

$$\epsilon_g = \mathcal{O}(\epsilon_f^{2/3}).$$

For the approximation of the Hessian we obtain

$$\epsilon_H = \mathcal{O}(\epsilon_g^{2/3}) = \mathcal{O}(\epsilon_f^{4/9}),$$

which is significantly better than in the case of forward differences.
Naturally, one can only expect convergence of an iterative scheme if $\epsilon_g \to 0$ during the iteration. This is illustrated by the following result. Errors denoted by $\epsilon_f(\cdot)$, $\epsilon_g(\cdot)$ and $\epsilon_H(\cdot)$ have to be understood scalar-, vector- and matrix-valued. Inequalities like $\epsilon_H(\cdot) < \epsilon$ are to be read elementwise.

THEOREM 7.3. *Let* (A) *be satisfied. Then there exist constants* $\bar{K} > 0$, $\delta > 0$ *and* $\epsilon > 0$ *such that for* $x_a \in B(\delta)$ *and* $\epsilon_H(x_a) < \epsilon$ *it holds that*

$$x_+ = x_a - \left(\nabla^2 f(x_a) + \epsilon_H(x_a)\right)^{-1} \left(\nabla f(x_a) + \epsilon_g(x_a)\right)$$

is well-defined, i.e. $\left(\nabla^2 f(x_a) + \epsilon_H(x_a)\right)^{-1}$ *is nonsingular and satisfies*

$$\|x_+ - x^*\| \leq \bar{K} \left(\|x_a - x^*\|^2 + \|\epsilon_H(x_a)\|\|x_a - x^*\| + \|\epsilon_g(x_a)\|\right).$$

PROOF. Let δ be chosen such that Lemma 7.1 and Theorem 7.1 hold true.
Define

$$x_+^N = x_a - (\nabla^2 f(x_a))^{-1} \nabla f(x_a)$$

and note that

$$\begin{aligned} x_+ = x_+^N + &((\nabla^2 f(x_a))^{-1} - (\nabla^2 f(x_a) + \epsilon_H(x_a))^{-1})\nabla f(x_a) \\ &- (\nabla^2 f(x_a) + \epsilon_H(x_a))^{-1}\epsilon_g(x_a). \end{aligned}$$

Lemma 7.1 and Theorem 7.1 imply

$$\begin{aligned} \|x_+ - x^*\| \leq \| x_+^N - x^*\| &+ \|((\nabla^2 f(x_a))^{-1} - (\nabla^2 f(x_a) + \epsilon_H(x_a))^{-1})\nabla f(x_a) \\ &+ \|(\nabla^2 f(x_a) + \epsilon_H(x_a))^{-1}\| \cdot \|\epsilon_g(x_a)\| \\ \leq\ & K\|x_a - x^*\|^2 + \|(\nabla^2 f(x_a))^{-1} - (\nabla^2 f(x_a) + \epsilon_H(x_a))^{-1}\| \cdot \|\nabla f(x_a)\| \\ &+ \|(\nabla^2 f(x_a) + \epsilon_H(x_a))^{-1}\| \cdot \|\epsilon_g(x_a)\| \qquad (7.2) \\ \leq\ & K\|x_a - x^*\|^2 + 2\|(\nabla^2 f(x_a))^{-1} - (\nabla^2 f(x_a) + \epsilon_H(x_a))^{-1}\| \cdot \|\nabla^2 f(x^*)\| \cdot \\ &\cdot \|x_+^N - x_a\| + \|(\nabla^2 f(x_a) + \epsilon_H(x_a))^{-1}\| \cdot \|\epsilon_g(x_a)\| \\ \leq\ & \tilde{K}\|x_a - x^*\|^2 + 2 \cdot \|(\nabla^2 f(x_a))^{-1} - (\nabla^2 f(x_a) + \epsilon_H(x_a))^{-1}\| \cdot \|\nabla^2 f(x^*)\| \\ &\cdot \|x_a - x^*\| + \|(\nabla^2 f(x_a) + \epsilon_H(x_a))^{-1}\| \cdot \|\epsilon_g(x_a)\|. \end{aligned}$$

The last inequality holds due to

$$\|x_+^N - x_a\| \leq \|x_+^N - x^*\| + \|x_a - x^*\| \leq K\|x_a - x^*\|^2 + \|x_a - x^*\|.$$

For $\|\epsilon_H(x_a)\| \leq \|(\nabla^2 f(x^*))^{-1}\|^{-1}/4$ Lemma 7.1 yields

$$\|\epsilon_H(x_a)\| \leq \|(\nabla^2 f(x_a))^{-1}\|^{-1}/2.$$

Setting $B = \nabla^2 f(x_a) + \epsilon_H(x_a)$ and $A = (\nabla^2 f(x_a))^{-1}$, one obtains:

$$\begin{aligned} \|I - BA\| =& \|I - (\nabla^2 f(x_a) + \epsilon_H(x_a))(\nabla^2 f(x_a))^{-1}\| \\ &\leq \|\epsilon_H(x_a)\| \cdot \|(\nabla^2 f(x_a))^{-1}\| \leq \frac{1}{2} \end{aligned} \tag{7.3}$$

The Banach Lemma implies that $B = \nabla^2 f(x_a) + \epsilon_H(x_a)$ is nonsingular and additionally:

$$\begin{aligned} \|(\nabla^2 f(x_a) + \epsilon_H(x_a))^{-1}\| &\leq \frac{\|(\nabla^2 f(x_a)^{-1}\|}{1 - \|\epsilon_H(x_a)(\nabla^2 f(x_a)^{-1}\|} \\ &\leq \frac{\|(\nabla^2 f(x_a))^{-1}\|}{\frac{1}{2}} = 2\|(\nabla^2 f(x_a))^{-1}\| \\ &\leq 4\|(\nabla^2 f(x^*))^{-1}\|. \end{aligned} \tag{7.4}$$

Thus we have by (7.4), (7.3) and Lemma 7.1:

$$\begin{aligned} \|(\nabla^2 f(x_a))^{-1} - (\nabla^2 f(x_a) + \epsilon_H(x_a))^{-1}\| \leq \\ \underbrace{\|(\nabla^2 f(x_a) + \epsilon_H(x_a))^{-1}\|}_{\underset{(7.4)}{\leq} 4\cdot\|(\nabla^2 f(x^*))^{-1}\|} \cdot \underbrace{\|I - (\nabla^2 f(x_a) + \epsilon_H(x_a))(\nabla^2 f(x_a))^{-1}\|}_{\underset{(7.3)}{\leq} \|\epsilon_H(x_a)\| \underbrace{\|(\nabla^2 f(x_a))^{-1}\|}_{\leq 2\cdot\|(\nabla^2 f(x^*))^{-1}\|}} \\ \leq 8\|(\nabla^2 f(x^*))^{-1}\|^2 \cdot \|\epsilon_H(x_a)\| \end{aligned}$$

From (7.1) we infer that

$$\begin{aligned} \|x_+ - x^*\| \leq \overline{K}\ \|x_a - x^*\|^2 + 16\|(\nabla^2 f(x^*)^{-1}\|^2\|\nabla^2 f(x^*)\| \cdot \|\epsilon_H(x_a)\| \cdot \|x_a - x^*\| \\ + 4\|(\nabla^2 f(x^*))^{-1}\| \cdot \|\epsilon_g(x_a)\|. \end{aligned}$$

□

The interpretation of Theorem 7.3 is as follows: The error ϵ_g of the gradient evaluation influences the accuracy of Newton's method. The error ϵ_H of the Hessian evaluation reduces the rate of convergence. In addition Theorem 7.3 gives a hint on how the individual errors should behave in order to get superlinear convergence.
Now we want to discuss some variants of Newton's method. The evaluation and factorization of the Hessian of f can be very expensive. If x^0 is sufficiently close to x^*, then the following iteration rule reduces this effort considerably:

$$x^{k+1} = x^k - (\nabla^2 f(x^0))^{-1}\nabla f(x^k), \quad k = 0, 1, \ldots. \tag{7.5}$$

Here we have $\epsilon_H(x^k) = \nabla^2 f(x^0) - \nabla^2 f(x^k)$ and

$$\|\epsilon_H(x^k)\| \leq \|\nabla^2 f(x^0) - \nabla^2 f(x^k)\| \leq \gamma\|x^0 - x^k\| \leq \gamma(\|x^0 - x^*\| + \|x^k - x^*\|). \tag{7.6}$$

The convergence of method (7.5) follows from Theorem 7.3 with $\epsilon_g = 0$ and $\epsilon_H = \mathcal{O}(\|x^0 - x^*\|)$.

THEOREM 7.4. *Let* (A) *be satisfied. Then there exist $K > 0$ and $\delta > 0$ such that for $x^0 \in B(\delta)$ it holds that the sequence $\{x^k\}$ generated by* (7.5) *converges q-linearly to x^* and satisfies*

$$\|x^{k+1} - x^*\| \leq K\|x^0 - x^*\| \, \|x^k - x^*\|.$$

PROOF. Let δ be chosen such that Theorem 7.3 holds true. Then (7.6) implies

$$\begin{aligned}\|x^{k+1} - x^*\| &\leq \overline{K}(\|x^k - x^*\|^2 + \gamma(\|x^0 - x^*\| + \|x^k - x^*\|)\|x^k - x^*\| \\ &= \overline{K}(\|x^k - x^*\|(1+\gamma) + \gamma\|x^0 - x^*\|)\|x^k - x^*\| \\ &\leq \overline{K}(1+2\gamma)\delta\|x^k - x^*\|.\end{aligned}$$

Decreasing δ until $\overline{K}(1+2\gamma)\delta = \gamma < 1$ yields the assertion. □

The *Sharmanskii-method* is a generalization of (7.5). In this variant, the Hessian will be updated at each $(m+1)$-st iteration by the Hessian at the current iterate. For $m = 0$ one obtains Newton's method and for $m = \infty$ iteration (7.5). We state the following result.

THEOREM 7.5. *Let* (A) *be satisfied and $m \geq 0$ be given. Then there exist constants $K \geq 0$ and $\delta > 0$ such that the Sharmanskii-method converges q-superlinearly to x^* for all $x^0 \in B(\delta)$.*

Appropriate stopping criteria. We have already mentioned that the stopping criterion

$$\|\nabla f(x^k)\| \leq \tau_r\|\nabla f(x^0)\| + \tau_a \tag{7.7}$$

with $\tau_r \in (0,1)$ and $1 \gg \tau_a > 0$ is adequate whereas testing the difference between two consecutive functions values (as sole stopping criterion) is not reasonable. Consider e.g.

$$f(x^k) = -\sum_{j=1}^{k} j^{-1}.$$

Then it holds: $f(x^k) \to -\infty$ for $k \to \infty$ and $f(x^{k+1}) - f(x^k) \to 0$.
Very often one is not only interested in a sufficiently small gradient norm, but also in the proximity of the current iterates to a stationary point (or a local minimizer). It turns out that when designing a corresponding stopping criterion we have to take special care of the rate of convergence of the method. In view of Lemma 4.2 we state the following result about the relation between *relative errors* (in x) and *relative gradients*.

LEMMA 7.2. *Let* (A) *be satisfied. Let $\delta > 0$ be chosen such that Lemma 7.1 is satisfied for all $x \in B(\delta)$. Then it holds for all $x \in B(\delta)$ that*

$$\frac{\|x - x^*\|}{4\|x^0 - x^*\|\kappa(\nabla^2 f(x^*))} \leq \frac{\|\nabla f(x)\|}{\|\nabla f(x^0)\|} \leq \frac{4\|x - x^*\|\kappa(\nabla^2 f(x^*))}{\|x^0 - x^*\|} \quad \text{with } x^0 \in B(\delta).$$

Lemma 7.2 demonstrates that the relative error in x, *i.e.* $\|x - x^*\|/\|x^0 - x^*\|$ corresponds –except for a constant factor– to the relative gradient *i.e.* $\|\nabla f(x)\|/\|\nabla f(x^0)\|$.
In case of Newton's method (*quadratically convergent*) the length of the search direction d can be considered as a sufficiently exact error estimator for the order of $\|x_+ - x^*\|$, since

$$\|x_a - x^*\| = \|d\| + \mathcal{O}(\|x_a - x^*\|^2). \tag{7.8}$$

In case one aspires to terminate the algorithm if $\|x_+ - x^*\|$ is of the same order as the the bound $\tau_s > 0$, then it should be checked whether

$$\|d\| = \mathcal{O}(\sqrt{\tau_s}).$$

Moerover relation (7.8) yields

$$\|x_+ - x^*\| = \mathcal{O}(\|x_a - x^*\|^2) = \mathcal{O}(\tau_s).$$

In case of *superlinearly convergent* methods it holds that

$$\|x_a - x^*\| = \|d\| + o(\|x_a - x^*\|).$$

Here $\|x_a - x^*\| \leq \tau_s$ in general only implies $\|x_+ - x^*\| < \tau_s$.
In *q-linear convergent* methods one should be careful when applying $\|d\| \leq \tau_s$. One has

$$\frac{\big|\|x_a - x^*\| - \|d\|\big|}{\|x_a - x^*\|} \leq \frac{\|x_a - x^* + d\|}{\|x_a - x^*\|} \leq \frac{\|x_+ - x^*\|}{\|x_a - x^*\|}.$$

A termination due to the smallness of $\|d\|$ is only allowed in case of very fast linearly convergent methods. Suppose we have a good estimate ρ for the Q-factor, *i.e.*

$$\|x_+ - x^*\| \leq \rho\|x_a - x^*\|,$$

then it follows that

$$(1-\rho)\|x_a - x^*\| \leq \|x_a - x^*\| - \|x_+ - x^*\| \leq \|x_a - x_+\| = \|d\|.$$

Thus it holds that

$$\|x_+ - x^*\| \leq \rho\|x_a - x^*\| \leq \frac{\rho}{1-\rho}\|d\|.$$

In this case one applies the stopping criterion

$$\|d\| \leq \frac{(1-\rho)}{\rho}\tau_s.$$

yielding $\|x_+ - x^*\| \leq \tau_s$.
In fact one frequently implements more than one stopping criterion. It is conceivable to combine (7.7) with an aforementioned reasonable criterion for the smallness of $\|d^k\| = \|x^{k+1} - x^k\|$, *i.e.*

$$\|x^{k+1} - x^k\| \leq \tau_x(1 + \|x^k\|) =: \tau_s \quad \text{with } \tau_x \in (0,1).$$

Additionally, one can test the smallness of the difference of consecutive function values (but this criterion should never be applied alone):

$$|f(x^{k+1}) - f(x^k)| \leq \tau_f(1 + |f(x^k)|) \quad \text{with } \tau_f \in (0,1).$$

A typical choice of τ_f is given by $\tau_f = \tau_x^2$.

2. Nonlinear least-squares problems

A *nonlinear least-squares problem* is a minimization task with an objective function of the form

$$f(x) = \frac{1}{2}\sum_{i=1}^{M} \|r_i(x)\|^2 = \frac{1}{2}R(x)^\top R(x),$$

where the vector $R = (r_1, r_2, \ldots, r_M)^\top$ is called the *residuum*. Problems of this kind typically appear in data fitting (regression). Thereby M represents the number of observations (data) and n is the number of parameters which have to be determined. The problem is called *overdetermined*, if $M > n$, and *underdetermined* for $M < n$. If $M = n$, then the problem reduces to solving a nonlinear equation.

If x^* is a local minimizer of f and $f(x^*) = 0$, then the problem min $f(x)$ is called a *null-residuum problem*. In case $f(x^*)$ is small, i.e. the data-fitting is good, then one refers to a *problem with small residuum*.
Let $R' \in \mathbb{R}^{M\times n}$ be the Jacobian of R, then it holds that

$$\nabla f(x) = R'(x)^\top R(x) \in \mathbb{R}^n.$$

The necessary condition for a local minimizer x^* is given by

$$0 = \nabla f(x^*) = R'(x^*)^\top R(x^*). \tag{7.9}$$

For an underdetermined problem with rank $(R'(x^*)) = M$, (7.9) does imply $R(x^*) = 0$. However for $n > M$ this is not the case. The Hessian of f is given by

$$\nabla^2 f(x) = R'(x)^\top R'(x) + \sum_{j=1}^{M} r_i(x)\nabla^2 r_i(x).$$

Observe that for the computation of $\nabla^2 f(x)$, the M Hessians $\nabla^2 r_i(x)$ have to be evaluated.

2.1. Gauss-Newton iteration. Let us assume that min $f(x)$ is a null-residuum problem. Then it holds that

$$\nabla^2 f(x^*) = R'(x^*)^\top R'(x^*), \text{ as } r_i(x^*) = 0\ \forall i.$$

This suggests to use $R'(x)^\top R'(x)$ as an approximation of the Hessian of f, which converges to $\nabla^2 f(x^*)$ for $x \to x^*$. In case of small residuals $r_i(x^*)$, $R'(x)^\top R'(x)$ typically represents a good Hessian approximation at x near x^*.
With the help of this Hessian approximation, we construct the following quadratic model:

$$m_a(x) = f(x_a) + R(x_a)^\top R'(x_a)(x - x_a) + \frac{1}{2}(x - x_a)^\top R'(x_a)^\top R'(x_a)(x - x_a).$$

Assuming that $R'(x_a)^\top R'(x_a)$ has full rank, there exists a unique minimizer x_+ of $m_a(x)$ which satisfies

$$0 = R'(x_a)^\top R(x_a) + R'(x_a)^\top R'(x_a)(x_+ - x_a).$$

In the following we will consider over- and underdetermined problems separately. In any case we make the following assumption:

ASSUMPTION 7.1. *The point x^* is a local minimizer of* min $\|R(x)\|^2$, *$R'(x)$ is Lipschitz continuous at x^*, and $R'(x^*)^\top R'(x^*)$ has full rank. The last assumption means that*

- *$R'(x^*)$ is nonsingular for $M = n$;*
- *$R'(x^*)$ has a full column rank for $M > n$;*
- *$R'(x^*)$ has a full row rank for $M < n$.*

2.2. Overdetermined problems. The Gauss-Newton method is given by the iteration rule

$$x^{k+1} = x^k - \left(R'(x^k)^\top R'(x^k)\right)^{-1} R'(x^k)^\top R(x^k), \quad x^0 \in \mathbb{R}^n \text{ given.}$$

We have the following result.

THEOREM 7.6. *Let $M > n$ and Assumption 7.1 satisfied. Then there exist $K > 0$ and $\delta > 0$ such that the Gauss-Newton-step*

$$x_+ = x_a - \left(R'(x_a)^\top R'(x_a)\right)^{-1} R'(x_a)^\top R(x_a) \tag{7.10}$$

fulfills the following estimate for $x_a \in B(\delta)$:

$$\|x_+ - x^*\| \leq K(\|x_a - x^*\|^2 + \|R(x^*)\| \, \|x_a - x^*\|).$$

PROOF. Let δ be chosen such that $\|x - x^*\| < \delta$ implies $\operatorname{rank}(R'(x)^\top R'(x)) = n$. Moreover let γ be the Lipschitz constant of R near x^* . From (7.10) we infer

$$\begin{aligned} x_+ - x^* &= x_a - x^* - (R'(x_a)^\top R'(x_a))^{-1} R'(x_a)^\top R(x_a) \\ &= (R'(x_a)^\top R'(x_a))^{-1} R'(x_a)^\top (R'(x_a)(x_a - x^*) - R(x_a)). \end{aligned}$$

It holds that

$$R'(x_a)(x_a - x^*) - R(x_a) = R'(x_a)(x_a - x^*) - R(x^*) + R(x^*) - R(x_a)$$

as well as

$$\begin{aligned} &\|R(x^*) - R(x_a) - R'(x_a)(x^* - x_a)\| = \\ &\quad = \|R(x_a) + \int_0^1 R'(x_a + \tau(x^* - x_a))(x^* - x_a) d\tau - R'(x_a)(x^* - x_a) - R(x_a)\| \\ &\quad \leq \int_0^1 \|R'(x_a + \tau(x^* - x_a)) - R'(x_a)\| d\tau \|x^* - x_a\| \leq \frac{\gamma}{2} \|x_a - x^*\|^2. \end{aligned}$$

The first order necessary conditions yield $R'(x^*)^\top R(x^*) = 0$ and thus

$$-R'(x_a)^\top R(x^*) = (R'(x^*) - R'(x_a))^\top R(x^*).$$

This gives

$$\begin{aligned} \|x_+ - x^*\| &\leq \|(R'(x_a)^\top R'(x_a)^{-1}\| \cdot \|R'(x_a)^\top (R(x^*) - [R(x^*) - R(x_a) \\ &\qquad - R'(x_a)(x^* - x_a)])\| \\ &\leq \|(R'(x_a)^\top R'(x_a)^{-1}\| \cdot [\|(R'(x^*) - R'(x_a)^\top R(x^*)\| \\ &\qquad + \|R'(x_a)\| \cdot \frac{\gamma}{2} \|x^* - x_a\|^2] \end{aligned} \tag{7.11}$$

$$\leq \|(R'(x_a)^\top R'(x_a))^{-1}\| \cdot \gamma \|x^* - x_a\| \cdot [\|R(x^*)\| + \frac{\|R'(x_a)\|}{2} \|x^* - x_a\|].$$

The choice

$$K = \gamma \max_{x \in B(\delta)} \|(R'(x)^\top R'(x)^{-1}\| \cdot (1 + \frac{\|R'(x)\|}{2})$$

proves the assertion. □

Theorem 7.6 shows that the local rate of convergence is Q-quadratic in case of $R(x^*) = 0$. In addition we observe that for $R(x^*) \neq 0$ not even linear convergence follows immediately. The proof shows that linear convergence requires $K\|R'(x^*)\| < 1$.
A more subtle estimate in the proof of Theorem 7.6 can be obtained by using

$$\begin{aligned} R'(x_a)^\top R(x^*) &= \left(R'(x^*) + R''(x^*)(x_a - x^*) + \mathcal{O}(\|x_a - x^*\|^2)\right)^\top R(x^*) \\ &= (x_a - x^*)^\top R''(x^*)^\top R(x^*) + \mathcal{O}(\|x_a - x^*\|^2). \end{aligned}$$

We have tacitly introduced the tensor R'' and applied $R'(x^*)^\top R(x^*) = 0$. We obtain the estimate

$$\|(R'(x^*) - R'(x_a))^\top R(x^*)\| \leq \|\nabla^2 f(x^*) - R'(x^*)^\top R'(x^*)\| \, \|R(x^*)\| + \mathcal{O}(\|x_a - x^*\|^2).$$

Thus we have seen that the Gauss-Newton method converges even for problems with large residuum, provided that $R''(x^*)$ is sufficiently small.

2.3. Underdetermined problems. At first we consider the following underdetermined linear least-squares problem

$$\min \|Ax - b\|^2, \quad A \in \mathbb{R}^{M\times n},\ M < n.$$

It can be demonstrated that there is no unique minimizer, but a unique *minimizer with minimal norm*. This special solution can be expressed with the help of the *singular value decomposition* of A, which is given by given by

$$A = U\Sigma V^\top$$

with $\Sigma = \operatorname{diag}(\sigma_i) \in \mathbb{R}^{M\times n}$ a diagonal matrix whose diagonal entries are called singular values. It holds that $\sigma_i \geq 0$ and $\sigma_i = 0$ for $i > M$. The columns of $U \in \mathbb{R}^{M\times M}$ and $V \in \mathbb{R}^{n\times n}$ are called left and right singular vectors. The matrices U and V are orthogonal.
The solution with minimal norm is given by

$$x = A^\dagger b,$$

where $A^\dagger = V\Sigma^\dagger U^\top$, $\sigma^\dagger = \operatorname{diag}(\sigma_i^\dagger)$ and

$$\sigma_i^\dagger = \begin{cases} \sigma_i^{-1} & \text{for } \sigma_i \neq 0, \\ 0 & \text{for } \sigma_i = 0. \end{cases}$$

The matrix $A^\dagger$ is called the *Moore-Penrose inverse of A*. If A is a nonsingular quadratic matrix, then it holds that $A^\dagger = A^{-1}$. The singular value decomposition also exists for $M > n$, and—if A has full column rank—one obtains $A^\dagger = (A^\top A)^{-1}A^\top$. In addition it holds that $A^\dagger A$ is a projection onto the image of $A^\dagger$ and $AA^\dagger$ is a projection onto the image of A, *i.e.*

$$A^\dagger AA^\dagger = A^\dagger,\ (A^\dagger A)^\top = A^\dagger A \text{ and } AA^\dagger A = A,\ (AA^\dagger)^\top = AA^\top.$$

The solution with minimal norm of

$$\min \frac{1}{2}\|R(x_a) + R'(x_a)(x - x_a)\|^2$$

in case of underdetermined problems is

$$x_+ = x_a - R'(x_a)^\dagger R(x_a),$$

which corresponds to the Gauss-Newton iteration for the associated nonlinear least-squares problem. In the linear case, *i.e.* $R(x) = Ax - b$, it follows

$$x_+ = x_a - A^\dagger(Ax_a - b) = (I - A^\dagger A)x_a + A^\dagger b.$$

Let $e_a = x_a - A^\dagger b$ and $e_+ = x_+ - A^\dagger b$, then $A^\dagger AA^\dagger b = A^\dagger b$ implies

$$e_+ = (I - A^\dagger A)e_a.$$

This does not ensure that $x_+ = A^\dagger b$ is the solution with minimal norm, but it does imply that x_+ solves the problem and that the method terminates after one step. Let $\mathcal{Z} = \{x : R(x) = 0\}$.

THEOREM 7.7. *Let $M < n$ and the Assumption 7.1 be fulfilled for $z^* \in \mathcal{Z}$. Then there exists $\delta > 0$ such that the Gauss-Newton iteration*

$$x^{k+1} = x^k - R'(x^k)^\dagger R(x^k)$$

is well-defined for $\|x^0 - x^*\| \leq \delta$ and converges R-quadratically to $z^* \in \mathcal{Z}$.

3. Inexact Newton methods

Inexact Newton methods use an approximate Newton step $\tilde{d}$, which satisfies

$$\|\nabla^2 f(x_a)\tilde{d} + \nabla f(x_a)\| \leq \eta_a \|\nabla f(x_a)\|. \tag{7.12}$$

We refer to $\tilde{d}$ as *an inexact Newton step*. In our context we consider Newton methods with iterative solvers for

$$\nabla^2 f(x_a)\tilde{d} = -\nabla f(x_a). \tag{7.13}$$

In particular we know that $\nabla^2 f(x_a)$ is positive definite for x_a near x^*. Therefore the CG method of chapter 1 is appropriate for the iterative solution of (7.13). The resulting overall algorithm is called the *Newton-CG method*.

THEOREM 7.8. *Let* (A) *be fulfilled. Then there exist constants $K_I \geq 0$, $\delta > 0$ such that for $x_a \in B(\delta)$ with $\tilde{d}$ from* (7.12) *and $x_+ = x_a + \tilde{d}$ it holds that*

$$\|x_+ - x^*\| \leq K_I(\|x_a - x^*\| + \eta_a)\|x_a - x^*\|.$$

PROOF. Let δ be chosen such that Lemma 7.1 and Theorem 7.1 hold true. Let $r = -\nabla^2 f(x_a)\tilde{d} - \nabla f(x_a)$. Then one obtains

$$\tilde{d} + \nabla^2 f(x_a)^{-1}\nabla f(x_a) = -(\nabla^2 f(x_a)^{-1} r$$

as well as the equation

$$x_+ - x^* = x_a - x^* + \tilde{d} = x_a - x^* - (\nabla^2 f(x_a))^{-1}\nabla f(x_a) - (\nabla^2 f(x_a))^{-1} r. \tag{7.14}$$

From (7.7) and Lemma 7.1 it follows that

$$\begin{aligned}
\|\tilde{d} + (\nabla^2 f(x_a))^{-1}\nabla f(x_a)\| &= \|(\nabla^2 f(x_a))^{-1}(\nabla^2 f(x_a)\tilde{d} + \nabla f(x_a))\| \\
&\leq \|(\nabla^2 f(x_a))^{-1}\| \cdot \|\nabla^2 f(x_a)\tilde{d} + \nabla f(x_a)\| \\
&\leq \|(\nabla^2 f(x_a))^{-1}\| \cdot \eta_a \|\nabla f(x_a)\| \leq 4 \cdot \|\underbrace{(\nabla^2 f(x^*))^{-1}\| \cdot \|\nabla^2 f(x^*)\|\eta_a}_{=K(\nabla^2 f(x^*))\cdot\|x_a - x^*\|} \\
&= 4K(\nabla^2 f(x^*))\eta_a \cdot \|x_a - x^*\|.
\end{aligned}$$

Theorem 7.1 and (7.14) yield

$$\begin{aligned}
\|x_+ - x^*\| &\leq \|x_a - x^* - (\nabla^2 f(x_a))^{-1}\nabla f(x_a)\| + \|\tilde{d} + (\nabla^2 f(x_a))^{-1}\nabla f(x_a)\| \\
&\leq K \cdot \|x_a - x^*\|^2 + 4K(\nabla^2 f(x^*)) \cdot \eta_a \|x_a - x^*\|.
\end{aligned}$$

Setting

$$K_I = K + 4K(\nabla^2 f(x^*))$$

proves the assertion. □

Theorem 7.8 also contains a rule on how to control η_a in order to achieve fast convergence.

THEOREM 7.9. *Let* (A) *be satisfied. Then there exist $\delta > 0$ and $\bar{\eta} > 0$ such that the inexact Newton iteration $x^{k+1} = x^k + \tilde{d}^k$ with*

$$\|\nabla^2 f(x^k)\tilde{d}^k + \nabla f(x^k)\| \leq \eta_k \|\nabla f(x^k)\|$$

converges Q-linearly to x^* for $x^0 \in B(\delta)$ and $\{\eta_k\} \subset [0, \bar{\eta}]$. Furthermore it holds that

- *if $\eta_k \to 0$, then the rate of convergence is Q-superlinear;*

- *if $\eta_k \leq K_\eta \|\nabla f(x^k)\|^p$ for $K_\eta > 0$, then the rate of convergence is Q-superlinear with Q-order $1+p$.*

3.1. Implementation of the Newton-CG method. As already mentioned, in the Newton-CG method, the Newton-direction

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

is determined with the help of the CG-method. In addition we assume that $D_h^2 f(x;d)$ is a sufficiently exact and disposable approximation of the Hessian-vector product $\nabla^2 f(x)d$. The quantity h can be interpreted for example as the step size of a difference-approximation of the second derivative of f in the direction d. We now specify a variant of the preconditioned CG-method, which terminates with an error message, if $\nabla^2 f(x)$ is singular (w.r.t. d), *i.e.* $d^\top \nabla^2 f(x) d = 0$; or if d turns out to be a direction of negative curvature *i.e.* $d^\top \nabla^2 f(x) d < 0$. Later we will see that the case of a negative curvature can also lead to meaningful search directions.

ALGORITHM 7.1.
input: $W \in \mathcal{S}^n$ *positive definite,* $\eta \in \mathbb{R}_0^+$, $x \in \mathbb{R}^n$.
begin
set $d^0 := 0$, $r^0 := \nabla f(x)$, $p^0 := -W^{-1} r^0$, $l := 0$.
while $\|r^l\| > \eta \|\nabla f(x)\|$
begin
$w^l := D_h^2 f(x; p^l)$
if $(p^l)^\top w^l = 0$ **then** *RETURN("indefiniteness")*
if $(p^l)^\top w^l < 0$ **then** *RETURN("negative curvature")*
set

$$\begin{aligned}
\alpha_l &:= \frac{(r^l)^\top W^{-1} r^l}{(p^l)^\top w^l} \\
d^{l+1} &:= d^l + \alpha_l p^l \\
r^{l+1} &:= r^l + \alpha_l w^l \\
\beta_{l+1} &:= \frac{(r^{l+1})^\top W^{-1} r^{l+1}}{(r^l)^\top W^{-1} r^l} \\
p^{l+1} &:= -W^{-1} r^{l+1} + \beta_{l+1} p^l \\
l &:= l+1
\end{aligned}$$

end
end

In the implementation of the Newton-CG Algorithm, the preconditioner W and the error bound η from Theorem 7.9 will be adjusted at each Newton iteration. This idea is implemented in the following Newton-CG method:

ALGORITHM 7.2 (Newton-CG method).
input: $x^0 \in \mathbb{R}^n$.
begin

$r_0 := \|\nabla f(x^0)\|$, $k := 0$.
while $\|\nabla f(x^k)\| > \tau_r r_0 + \tau_a$
begin
choose $\eta_k, W^k \in \mathcal{S}^n$ *positive definite*
calculate d^k *with Algorithm 7.1 with input* W^k, η_k, x^k.
if *"indefiniteness"* **then** *STOP with error message.*
set $x^{k+1} = x^k + d^k$.
if $f(x^{k+1}) \geq f(x^k)$ **then** *STOP with error message.*
end
end

REMARK 7.1. (1) In Algorithm 7.2 we have made use of a rather simple stopping criterion. Naturally, combined criteria (see the discussion in section 1 of this chapter) can be applied as well.
(2) The algorithm terminates whenever "indefiniteness" occurs in the CG iteration. Possible modifications in case that d^k is a direction of negative curvature will be discussed later on.
(3) We also stop the Newton iteration as soon as the full step $x^{k+1} = x^k + d^k$ does not contribute to a decrease in f.

4. Global convergence

Regarding Newton methods we have only considered local convergence (results) so far. Until now, we have always assumed that x^0 is sufficiently close to a local solution x^*. Now we will introduce *globalization* approaches which allow for a relaxation of the choice of the starting point.
If one ensures that $\nabla^2 f(x^k)$ or a corresponding Hessian approximation satisfies

$$c_1\|d\|^2 \leq d^\top \nabla^2 f(x^k) d \leq c_2\|d\|^2 \quad \forall k \in \mathbb{N}\, \forall d \in \mathbb{R}^n,$$

$(0 < c_1 \leq c_2)$, then d^k defined as the solution of $\nabla^2 f(x^k)d = -\nabla f(x^k)$ is a gradient-related search direction. Inserting this into the general descent method 3.1 yields (cf. chapter 5.2) the global convergence of the global Newton method which means that the method converges to a stationary point regardless of the choice of the starting point x^0. Furthermore it can be shown that in the vicinity of x^* , $\alpha_k = 1$ will be accepted as the step size if one initiates the step size algorithm with $\alpha^{(0)} = 1$.
Here, we focus on another way to globalize Newton's method. Given a current approximation of a solution, due to its strategy of confining the next iterate to a sufficiently small neighborhood of the current iterate this strategy is called *trust-region method* (or trust region globalization).

4.1. Trust-Region method. A major drawback of the general descent method 3.1 with one of the step size strategies of chapter 3.2 is the necessity of ensuring that $\{\nabla^2 f(x^k)\} \subset \mathcal{S}^n$ is positive definite. Trust region methods deal with this problem in a suitable way and solve it algorithmically. Roughly speaking, these methods realize a smooth transition from the method of steepest descent to Newton's method. In this way the global convergence property of steepest descent is combined with the fast local convergence of Newton's method (Theorem 7.2).

The idea can be described as follows: Let $m_a(x)$ be a quadratic model of f in a neighborhood of x_a, which is given by

$$m_a(x) = f(x_a) + \nabla f(x_a)^\top (x - x_a) + \frac{1}{2}(x - x_a)^\top \nabla^2 f(x_a)(x - x_a),$$

and let Δ be the radius of a ball about x_a where we "trust" the model m_a to represent f well. The quantity Δ is called the *trust region radius*, and one refers to

$$\mathcal{T}(\Delta) = \{x : \|x - x_a\| \leq \Delta\}$$

as the *trust region*.
Given x_a, the next iterate x_+ is chosen as an approximate minimizer of m_a in $\mathcal{T}(\Delta)$. The associated *trust region subproblem* is defined as

$$\min\ m_a(x_a + d) \quad \text{s.t.} \quad \|d\| \leq \Delta. \tag{7.15}$$

We denote the solution of (7.15) by d_v (trial step) and the associated trial solution as $x_v = x_a + d_v$. Then we have to decide whether the step is acceptable or whether the trust region radius needs to be changed. Usually, both options are checked simultaneously. For the former one verifies whether the quadratic model is a good approximation of f in $\mathcal{T}(\Delta)$. For this purpose we define

$$ared = f(x_a) - f(x_v) \qquad \text{(actual reduction)}$$

and also

$$pred = m_a(x_a) - m_a(x_v) \qquad \text{(predicted reduction)}.$$

Note that (with $H_a = \nabla^2 f(x_a)$)

$$\begin{aligned} pred &= m_a(x_a) - m_a(x_v) = -\nabla f(x_a)^\top (x_v - x_a) - \frac{1}{2}(x_v - x_a)^\top H_a (x_v - x_a) \\ &= -\nabla f(x_a)^\top d_v - \frac{1}{2} d_v^\top H_a d_v. \end{aligned}$$

In the following algorithm we need the parameters

$$\mu_0 \leq \underline{\mu} \leq \overline{\mu}$$

to decide whether we reject the trial step ($ared/pred < \mu_0$) and/or reduce Δ ($ared/pred < \underline{\mu}$), whether we increase Δ ($ared/pred > \overline{\mu}$) or leave the trust region radius unchanged. The reduction resp. the increase of Δ are realized by multiplication by $0 < \underline{\omega} < 1 < \overline{\omega}$. Further let $C > 1$ be fixed.

Algorithm 7.3.
input: $x_a \in \mathbb{R}^n$, $x_v \in \mathbb{R}^n$, $\Delta \in \mathbb{R}_+$.
begin
  $z^0 := x_a$, $z_v^0 := x_v$, $\hat{\Delta}^{(0)} := \Delta$, $l := 0$.
  while $z^l = x_a$
  begin
    $ared^{(l)} := f(x_a) - f(z_v^l)$, $d_v^l := z_v^l - x_a$, $pred^{(l)} := -\nabla f(x_a)^\top d_v^l - \frac{1}{2}(d_v^l)^\top H_a d_v^l$
    if $ared^{(l)}/pred^{(l)} < \mu_0$ **then**
      $z^{l+1} := x_a$, $\hat{\Delta}^{(l+1)} := \underline{\omega}\hat{\Delta}^{(l)}$
      if $l \geq 1$ **and** $\hat{\Delta}^{(l)} > \hat{\Delta}^{(l-1)}$
        $z^{l+1} := z_v^{l-1}$, $\hat{\Delta}^{(l+1)} := \hat{\Delta}^{(l-1)}$
      else

compute the solution d_v^{l+1} *of the trust region sub-problem with radius* $\hat{\Delta}^{(l+1)}$
$z_v^{l+1} := x_a + d_v^{l+1}$
end
elseif $\mu_0 \leq ared^{(l)}/pred^{(l)} \leq \underline{\mu}$ **then**
$z^{l+1} := z_v^l$, $\hat{\Delta}^{(l+1)} := \underline{\omega}\hat{\Delta}^{(l)}$
elseif $\underline{\mu} \leq ared^{(l)}/pred^{(l)} \leq \overline{\mu}$ **then**
$z^{l+1} := z_v^l$
elseif $\overline{\mu} \leq ared^{(l)}/pred^{(l)}$
if $\|d_v^l\| = \hat{\Delta}^{(l)} \leq C\|\nabla f(x_a)\|$ **then**
$z^{l+1} := x_a$, $\hat{\Delta}^{(l+1)} := \overline{\omega}\hat{\Delta}^{(l)}$
compute the solution d_v^{l+1} *of the trust region sub-problem with radius* $\hat{\Delta}^{(l+1)}$
$z_v^{l+1} := x_a + d_v^{l+1}$
else
$z^{l+1} = z_v^l$
end
end
$l := l + 1$
end
$x_+ := z^l$, $\Delta_+ = \hat{\Delta}^{(l)}$
end

In Algorithm 7.3 we require

$$\hat{\Delta}^{(l)} \leq C\|\nabla f(x_a)\|,$$

which bounds the trust region radius from above. The *while*-loop in Algorithm 7.3 is comparable to the loops of the step size algorithms and should terminate after finitely many iterations. Algorithm 7.3 now fits into a general trust region paradigm.

ALGORITHM 7.4 (Trust region framework).
input: $x^0 \in \mathbb{R}^n$, $\Delta_0 \in \mathbb{R}_+$
begin
$k := 0$, $r_0 := \|\nabla f(x^0)\|$
while $\|\nabla f(x^k)\| > \tau_r r_0 + \tau_a$
begin
compute an approximation H^k *of the Hessian* $\nabla^2 f(x^k)$
compute d_v^k *as the solution of*

$$\min \quad f(x^k) + \nabla f(x^k)^\top d + \frac{1}{2} d^\top H^k d \quad s.t. \quad \|d\| \leq \Delta_k$$

compute (x^{k+1}, Δ_{k+1}) *by Algorithm 7.3 with input* x^k, $x_v^k := x^k + d_v^k$, Δ_k
$k := k + 1$
end
end

4.2. Global convergence of the trust region algorithm. Theoretically, the trust region subproblem can be solved exactly. It turns out that even a relatively inaccurate solution of the trust region subproblem suffices to prove global and locally superlinear convergence. For the proof we invoke the following assumption.

ASSUMPTION 7.2.

(1) *There exists $\sigma > 0$ such that*

$$pred = f(x_a) - m_a(x_v) \geq \sigma \|\nabla f(x_a)\| \min\{\|d_v\|, \|\nabla f(x_a)\|\}. \tag{7.16}$$

(2) *There exists $M > 0$ such that*

$$\|d_v\| \geq \frac{\|\nabla f(x_a)\|}{M} \quad \text{or} \quad \|d_v\| = \Delta_a.$$

We obtain the following global convergence result.

THEOREM 7.10. *Let ∇f be Lipschitz continuous with modulus L. Let $\{x^k\}$ be the sequence generated by Algorithm 7.4, and further assume that the solutions of the trust region subproblems fulfill Assumption 7.2. Moreover suppose that the matrices $\{H^k\}$ are bounded. Then either f is bounded from below or $\nabla f(x^k) = 0$ for a finite k, or*

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

PROOF. Assume that $\nabla f(x^k) \neq 0 \ \forall k$ and f is bounded from below; otherwise the assertion is immediate. We show that in case the step is accepted (and, hence, the radius is not further enlarged), there exists $M_T \in (0, 1)$ such that

$$\|d_v^k\| \geq M_T \|\nabla f(x^k)\|. \tag{7.17}$$

Assume for the moment that (7.17) holds true. Since d_v^k is accepted, Algorithm 7.3 and Assumption 7.2 yield

$$ared_k \geq \mu_0 \, pred_k \geq \mu_0 \sigma \|\nabla f(x^k)\| \min\{\|d_v^k\|, \|\nabla f(x^k)\|\}.$$

Applying (7.17), we obtain

$$ared_k \geq \mu_0 \sigma \|\nabla f(x^k)\|^2 \min\{M_T, 1\} = \mu_0 \sigma M_T \|\nabla f(x^k)\|^2. \tag{7.18}$$

Since $\{f(x^k)\}$ is monotonically decreasing and f is bounded from below, it follows that $\lim_{k\to\infty} ared_k = 0$. Thus, (7.18) implies $\lim_{k\to\infty} \|\nabla f(x^k)\| = 0$.
It remains to prove (7.17). First note that for $\|d_v^k\| < \Delta_k$ from Assumption 7.2(2), we get

$$\|d_v^k\| \geq \frac{\|\nabla f(x^k)\|}{M}.$$

The case

$$\|d_v^k\| = \Delta_k \text{ and } \|d_v^k\| < \|\nabla f(x^k)\| \tag{7.19}$$

remains. In fact, if (7.19) does not hold true, then (7.17) follows from $M_T = \min\{1, \frac{1}{M}\}$. Provided that (7.19) is satisfied and d_v^k is accepted we show that

$$\|d_v^k\| = \Delta_k \geq \frac{2\sigma \min\{1 - \overline{\mu}, (1 - \mu_0)\overline{\omega}^{-2}\}}{M + L} \|\nabla f(x^k)\|. \tag{7.20}$$

Then the assertion follows with

$$M_T = \min\left\{1, \frac{1}{M}, \frac{2\sigma \min\{1 - \overline{\mu}, (1 - \mu_0)\overline{\omega}^{-2}\}}{M + L}\right\}.$$

Let M of Assumption 7.2 be chosen sufficiently large such that

$$\|H^k\| \le M \quad \forall k \in \mathbb{N}. \tag{7.21}$$

We prove (7.20) by showing that the trust region radius is enlarged and the step associated with the larger radius is accepted if (7.19) is fulfilled and (7.20) does not hold true. For this purpose, let d_v^k be a trial step such that $\|d_v^k\| < \|\nabla f(x^k)\|$ and

$$\|d_v^k\| < \frac{2\sigma \min\{1-\overline{\mu}, (1-\mu_0)\overline{\omega}^{-2}\}}{M+L}\|\nabla f(x^k)\|. \tag{7.22}$$

The Lipschitz continuity of ∇f and (7.21) yield

$$\begin{aligned}
ared_k &= f(x^k) - f(x^k + d_v^k) = -\nabla f(x^k)^\top d_v^k - \int_0^1 (\nabla f(x^k + \tau d_v^k) - \nabla f(x^k))^\top d_v^k d\tau \\
&= -\nabla f(x^k)^\top d_v^k - \frac{1}{2}(d_v^k)^\top H^k d_v^k + \frac{1}{2}(d_v^k)^\top H^k d_v^k - \int_0^1 (\nabla f(x^k + \tau d_v^k) - \nabla f(x^k))^\top d_v^k d\tau \\
&= pred_k + \frac{1}{2}(d_v^k)^\top H^k d_v^k - \int_0^1 (\nabla f(x^k + \tau d_v^k) - \nabla f(x^k))^\top d_v^k d\tau \\
&\ge pred_k - \frac{M}{2}\|d_v^k\|^2 - L\|d_v^k\|^2 \int_0^1 \tau d\tau \\
&= pred_k - \frac{1}{2}(M+L)\|d_v^k\|^2.
\end{aligned}$$

Assumption 7.2(1) implies

$$\frac{ared_k}{pred_k} \ge 1 - \frac{(M+L)\|d_v^k\|^2}{2\,pred_k} \ge 1 - \frac{(M+L)\|d_v^k\|^2}{2\sigma\|\nabla f(x^k)\| \min\{\|d_v^k\|, \|\nabla f(x^k)\|\}}. \tag{7.23}$$

As $\|d_v^k\| < \|\nabla f(x^k)\|$ due to (7.19), it holds that

$$\min\{\|d_v^k\|, \|\nabla f(x^k)\|\} = \|d_v^k\|,$$

and thus we obtain, cf. (7.22),

$$\frac{ared_k}{pred_k} \ge 1 - \frac{(M+L)\|d_v^k\|}{2\sigma\|\nabla f(x^k)\|} > 1 - \min\{1-\bar{\mu}, (1-\mu_0)\bar{w}^2\} \ge \bar{\mu}.$$

Thus, an enlargement step is carried out by setting $\Delta_k^+ = \bar{w}\Delta_n$ and replacing d_v^k by $dv^{k,+}$, the minimizer of the quadratic model with radius Δ_k^+. Then, (7.23) is still fulfilled and it follows that

$$\|d_v^{k,+}\| \le \bar{w}\|d_v^k\| < \bar{w}\|\nabla f(x^k)\|.$$

Consequently,

$$\min\{\|\nabla f(x^k)\|, \|d_v^{k,+}\|\} > \frac{\|d_v^{k,+}\|}{\bar{w}} \qquad (\bar{w} > 1).$$

Thus,

$$\begin{aligned}
\frac{ared_k^+}{pred_k^+} &\ge 1 - \frac{(M+L)\|d_v^{k,+}\|^2}{2\sigma\|\nabla f(x^k)\| \min\{\|\nabla f(x^k)\|, \|d_v^{k,+}\|\}} \\
&\ge 1 - \frac{(M+L)\bar{w}\|d_v^{k,+}\|}{2\|\nabla f(x^k)\|\sigma} \ge 1 - \frac{(M+L)\bar{w}^2\|d_v^k\|}{2\|\nabla f(x^k)\|\sigma} \ge \mu_0
\end{aligned}$$

owing to (7.22). Hence, the enlargement of the radius produces an acceptable step which would be taken instead of d_v^k. Thus, (7.20) has to hold true. □

Next we study the computation of the trial step d_v resp. the trial points $x_v = x_a + d_v$. It suffices to compute approximate solutions of the trust region subproblem (7.15), such that Assumption 7.2 is satisfied. To this end, a simple idea is based on fixing the direction to the steepest descent direction under the trust region constraint. Let x_a be the current iterate and Δ_a be the current trust region radius. Then the trial point $x_v := x_v(\alpha)$ is defined as the minimizer α_a of

$$\min_{t\geq 0} \Psi_a(\alpha) := m_a(x_a - \alpha\nabla f(x_a)) \quad \text{s.t.} \quad x_v(\alpha) := x_a - \alpha\nabla f(x_a) \in \mathcal{T}(\Delta_a).$$

It holds that

$$\begin{aligned} \Psi_a(\alpha) &= m_a(x_a - \alpha\nabla f(x_a)) = f(x_a) - \alpha\|\nabla f(x_a)\|^2 + \frac{\alpha^2}{2}\nabla f(x_a)^\top H_a \nabla f(x_a), \\ \Psi_a'(\alpha) &= -\|\nabla f(x_a)\|^2 + \alpha\nabla f(x_a)^\top H_a \nabla f(x_a). \end{aligned}$$

For determining α_a we have to distinguish between the following cases:

(1) $\nabla f(x_a)^\top H_a \nabla f(x_a) \leq 0$. Obviously the Trust-Region constraint becomes active. It holds that

$$\|x_v(\alpha_a) - x_a\| = \alpha_a\|\nabla f(x_a)\| = \Delta_a,$$

which yields $\alpha_a = \frac{\Delta_a}{\|\nabla f(x_a)\|}$.

(2) $\nabla f(x_a)^\top H_a \nabla f(x_a) > 0$. In this case, we have

$$m_a'(x_a - \hat{\alpha}_a\nabla f(x_a)) = 0 \quad \Rightarrow \quad \hat{\alpha}_a = \frac{\|\nabla f(x_a)\|^2}{\nabla f(x_a)^\top H_a \nabla f(x_a)}.$$

If $\|x_v(\hat{\alpha}_a) - x_a\| \leq \Delta_a$ is fulfilled, then we accept $\hat{\alpha}_a$ as α_a; otherwise the trust region constraint becomes active and analogously to (1) it follows that $\alpha_a = \frac{\Delta_a}{\|\nabla f(x_a)\|}$.

To summarize, we have

$$\alpha_a := \begin{cases} \frac{\Delta_a}{\|\nabla f(x_a)\|} & \text{if } \nabla f(x_a)^\top H_a \nabla f(x_a) \leq 0, \\ \min\{\frac{\Delta_a}{\|\nabla f(x_a)\|}, \frac{\|\nabla f(x_a)\|^2}{\nabla f(x_a)^\top H_a \nabla f(x_a)}\} & \text{if } \nabla f(x_a)^\top H_a \nabla f(x_a) > 0. \end{cases} \tag{7.24}$$

The minimizer of the quadratic model m_a in the direction of the negative gradient is called the *Cauchy point* and will be denoted by x_a^{CP} [1]. The Cauchy point has the following properties which will prove to be useful for the global convergence result.

(1) $\nabla f(x_a)^\top H_a \nabla f(x_a) \leq 0$. Then it follows that

$$\begin{aligned} f(x_a) - m_a(x_a^{CP}) &= \alpha_a\|\nabla f(x_a)\|^2 - \frac{\alpha_a^2}{2}\nabla f(x_a)^\top H_a \nabla f(x_a) \\ &\geq \Delta_a\|\nabla f(x_a)\| = \|d_v\| \cdot \|\nabla f(x_a)\|. \end{aligned}$$

(2) $\nabla f(x_a)^\top H_a \nabla f(x_a) > 0$. Depending on where the minimum in (7.24) is attained, we have the following situations:

[1] When using the iteration index k, then the Cauchy point will be written as x_{CP}^k.

(i) $\alpha_a = \frac{\Delta_a}{\|\nabla f(x_a)\|}$. Then it holds that

$$\alpha_a \leq \hat{\alpha}_a = \frac{\|\nabla f(x_a)\|^2}{\nabla f(x_a)^\top H_a \nabla f(x_a)}$$

and thus

$$\alpha_a \nabla f(x_a)^\top H_a \nabla f(x_a) \leq \|\nabla f(x_a)\|^2.$$

This implies

$$\begin{aligned} f(x_a) - m_a(x_a^{CP}) &= \Delta_a \|\nabla f(x_a)\| - \frac{\alpha_a^2}{2} \nabla f(x_a)^\top H_a \nabla f(x_a) \\ &\geq \Delta_a \|\nabla f(x_a)\| - \frac{\alpha_a}{2} \|\nabla f(x_a)\|^2 \\ &= \Delta_a \|\nabla f(x_a)\| - \frac{\Delta_a}{2} \|\nabla f(x_a)\| = \frac{\Delta_a}{2} \|\nabla f(x_a)\| \\ &= \frac{1}{2} \|d_v\| \cdot \|\nabla f(x_a)\|. \end{aligned}$$

(ii) The second case occurs, if $\alpha_a = \hat{\alpha}_a = \frac{\|\nabla f(x_a)\|^2}{\nabla f(x_a)^\top H_a \nabla f(x_a)}$. Then we have

$$\begin{aligned} f(x_a) - m_a(x_a^{CP}) &= \frac{\|\nabla f(x_a)\|^4}{\nabla f(x_a)^\top H_a \nabla f(x_a)} - \frac{1}{2} \frac{\|\nabla f(x_a)\|^4}{(\nabla f(x_a)^\top H_a \nabla f(x_a))^2} \nabla f(x_a)^\top H_a \nabla f(x_a) \\ &= \frac{\|\nabla f(x_a)\|}{2} \frac{\|\nabla f(x_a)\|^3}{\nabla f(x_a)^\top H_a \nabla f(x_a)} = \frac{\|\nabla f(x_a)\|}{2} \hat{\alpha}_a \|\nabla f(x_a)\| \\ &= \frac{\|\nabla f(x_a)\|}{2} \cdot \|d_v\|. \end{aligned}$$

Cases (1) and (2) yield

$$f(x_a) - m_a(x_a^{CP}) \geq \frac{1}{2} \|\nabla f(x_a)\| \min\{\|d_v\|, \|\nabla f(x_a)\|\}. \tag{7.25}$$

Therefore the Cauchy point x_a^{CP} fulfills condition (1) of Assumption 7.2. The second condition of Assumption 7.2 is satisfied too, because in case of $\|d_v\| < \Delta_a$ the definition of x_a^{CP} implies

$$d_v = x_v(\alpha_a) - x_a = x_v(\hat{\alpha}_a) - x_a = -\hat{\alpha}_a \nabla f(x_a) = -\frac{\|\nabla f(x_a)\|^2}{\nabla f(x_a)^\top H_a \nabla f(x_a)} \nabla f(x_a).$$

If $\|H_a\| \leq M$, then

$$\|d_v\| \geq \frac{\|\nabla f(x_a)\|}{M}.$$

The second case, *i.e.* $\|d_v\| = \Delta_a$, is trivial. Therefore the global convergence result of Theorem 7.10 immediately yields the next result.

THEOREM 7.11. *Let ∇f be Lipschitz continuous with modulus L. Let $\{x^k\}$ be generated by Algorithm 7.4 with $x_v^k = x_{CP}^k$ and (7.24). Furthermore let the sequence of matrices $\{H^k\}$ be bounded. Then either $f(x^k)$ is bounded from below or $\nabla f(x^k) = 0$ for a finite k or*

$$\lim_{k \to \infty} \nabla f(x^k) = 0.$$

REMARK 7.2. Weakening the assumptions of Theorem 7.11, one can still show $\liminf_{k \to \infty} \|\nabla f(x^k)\| = 0$.

4.2.1. *Superlinear convergence.* The idea of always fixing the direction $-\nabla f(x_a)$ for the approximate solution of the trust region subproblem and determining the step size α_a such that $x_v(\alpha_a) - x_a \in \mathcal{T}(\Delta_a)$ often leads to a very slow (only Q-linear) rate of convergence (comparable to the method of steepest descent). For this reason, we discuss a technique which locally realizes the transition to the Newton direction. For this purpose, at x_a we define the *Newton point*

$$x_a^N = x_a - H_a^{-1}\nabla f(x_a).$$

If $H_a \in \mathcal{S}^n$ is positive definite, then x_a^N is the global minimizer of the usual quadratic model of f at x_a. In case that H_a possesses directions of negative curvature, then the quadratic model has no finite minimizer. The Newton point, however, remains meaningful.
Now we consider a special approximate solution of the trust region subproblem, which finally yields a locally superlinearly convergent method. This is achieved by minimizing m_a along a piecewise linear path $\mathcal{P} \in \mathcal{T}(\Delta)$, which is called the *dogleg path*. Its classical variant makes use of three points: x_a, x_a^N and $\hat{x}_a^{CP}$, the global minimizer of the quadratic model in the direction of steepest descent. It holds that $\hat{x}_a^{CP}$ only exists, if $\nabla f(x_a)^\top H_a \nabla f(x_a) > 0$ is satisfied. Whenever $\hat{x}_a^{CP}$ exists and fulfills

$$(x_a^N - \hat{x}_a^{CP})^\top (\hat{x}_a^{CP} - x_a) \geq 0, \tag{7.26}$$

then we define x_a^N as the last node on the path. If $H_a \in \mathcal{S}^n$ is positive definite, then it can be shown that (7.26) is fulfilled. In case (7.26) is violated, then x_a^N is not used. A closer examination of (7.26) yields

$$\begin{aligned} 0 \leq & (x_a^N - \hat{x}_a^{CP})^\top (\hat{x}_a^{CP} - x_a) = (x_a^N - x_a + x_a - \hat{x}_a^{CP})^\top (\hat{x}_a^{CP} - x_a) \\ = & (x_a^N - x_a)^\top (\hat{x}_a^{CP} - x_a) - \|\hat{x}_a^{CP} - x_a\|^2. \end{aligned}$$

Assuming that $\hat{x}_a^{CP} \neq x_a$, this immediately implies

$$0 < \|\hat{x}_a^{CP} - x_a\|^2 \leq (x_a^N - x_a)^\top (\hat{x}_a^{CP} - x_a) = -\hat{\alpha}_a \nabla f(x_a)^\top (x_a^N - x_a).$$

Since $\hat{\alpha}_a > 0$, we obtain

$$\nabla f(x_a)^\top (x_a^N - x_a) > 0.$$

The classical trial solution $x^D(\Delta_a)$ is computed according to

$$x^D(\Delta_a) = \begin{cases} x_a^{CP} & \text{if } \|x_a - x_a^{CP}\| = \Delta_a \text{ or } \hat{x}_a^{CP} \text{ exists} \\ & \text{and (7.26) is not fulfilled,} \\ x_a^N & \text{if } \|x_a - x_a^{CP}\| < \|x_a - x_a^N\| \leq \Delta_a \\ & \text{and (7.26) holds true,} \\ y^D(\Delta_a) & \text{else.} \end{cases} \tag{D}$$

Here, $y^D(\Delta_a)$ is the uniquely determined point between x_a^{CP} and x_a^N, which fulfills $\|y^D(\Delta_a) - x_a\| = \Delta_a$.
Typical properties of the "dogleg"-method are the following ones:

- There do not exist two points on $\mathcal{P}$, which have the same distance to x_a. Thus $\mathcal{P}$ can be parameterized by $x_a(s)$ with $s = \|x_a(s) - x_a\|$.
- $m_a(s)$ is a monotonically decreasing function of s.

LEMMA 7.3. *Let x_a, H_a and Δ_a be given, where H_a is nonsingular,*

$$d_a^N = -H_a^{-1}\nabla f(x_a) \text{ and } x_a^N = x_a + d_a^N.$$

We assume that $\nabla f(x_a)^\top H_a \nabla f(x_a) > 0$ *and*

$$\hat{d}_a^{CP} = \hat{x}_a^{CP} - x_a = -\frac{\|\nabla f(x_a)\|^2}{\nabla f(x_a)^\top H_a \nabla f(x_a)} \nabla f(x_a) \neq d_a^N.$$

Let $\mathcal{P}$ *be the piecewise linear path of* x_a *via* $\hat{x}_a^{CP}$ *to* x_a^N*. If*

$$(d_a^N - \hat{d}_a^{CP})^\top \hat{d}_a^{CP} \geq 0, \tag{7.27}$$

then, for arbitrary $\delta \leq \|d_a^N\|$*, there exists a unique point* $x(\delta) \in \mathcal{P}$ *such that* $\|x(\delta) - x_a\| = \delta$*.*

Proof. On the segment from x_a to $\hat{x}_a^{CP}$ the assertion holds trivially. Consequently the segment from $\hat{x}_a^{CP}$ to x_a^N remains to be discussed. We have to show that

$$\phi(\lambda) = \frac{1}{2}\|(1-\lambda)\hat{d}_a^{CP} + \lambda d_a^N\|^2$$

increases strictly monotonically for $\lambda \in (0,1)$. We start by assuming that (7.27) holds with strict inequality. Then we have

$$\|d_a^N\| \cdot \|\hat{d}_a^{CP}\| \geq (d_a^N)^\top \hat{d}_a^{CP} \underset{(7.27)}{>} \|\hat{d}_a^{CP}\|^2$$

and therefore $\|d_a^N\| > \|\hat{d}_a^{CP}\|$. Thus, we obtain with (7.27)

$$\begin{aligned}\phi(\lambda) =& (d_a^N - \hat{d}_a^{CP})^\top ((1-\lambda)\hat{d}_a^{CP} + \lambda d_a^N) = -(1-\lambda)\|\hat{d}_a^{CP}\|^2 \\ &+ (1-\lambda)(d_a^N)^\top \hat{d}_a^{CP} - \lambda(\hat{d}_a^{CP})^\top d_a^N + \lambda\|d_a^N\|^2 \\ &\overset{(7.27)}{>} \lambda(\|d_a^N\|^2 - (\hat{d}_a^{CP})^\top d_a^N) \geq \lambda(\|d_a^N\| - \|\hat{d}_a^{CP}\|)\|\hat{d}_a^{CP}\| > 0.\end{aligned}$$

Hence, ϕ is strictly monotonically increasing.
If (7.27) holds with equality, then $d_a^N \neq \hat{d}_a^{CP}$ by assumption. Rearranging terms, we find that

$$\phi'(\lambda) = \lambda\|\hat{d}_a^{CP} - d_a^N\|^2 + (d_a^N - \hat{d}_a^{CP})^\top \hat{d}_a^{CP} = \lambda\|\hat{d}_a^{CP} - d_a^N\|^2 > 0$$

for $\lambda > 0$. Hence, ϕ is strictly monotonically increasing, which concludes the proof. □

Now we show that the quadratic model decreases monotonically along the "dogleg"-path.

Lemma 7.4. *Let the assumptions of Lemma 7.3 be satisfied. Then the local quadratic model*

$$m_a(x) = f(x_a) + \nabla f(x_a)^\top (x - x_a) + \frac{1}{2}(x - x_a)^\top H_a(x_a)(x - x_a)$$

is monotonically decreasing along $\mathcal{P}$*.*

Proof. Let $\hat{x}_a^{CP}$ $(\neq x_a)$ be the minimizer of m_a in the direction $-\nabla f(x_a)$. Thus, m_a is strictly monotonically decreasing from x_a to $\hat{x}_a^{CP}$. Hence, only the segment from $\hat{x}_a^{CP}$ to x_a^N remains to be discussed. For this purpose, let

$$\begin{aligned}\psi(\lambda) = m_a(x_a + (1-\lambda)\hat{d}_a^{CP} + \lambda d_a^N) = f(x_a) + \nabla f(x_a)^\top ((1-\lambda)\hat{d}_a^{CP} + \lambda d_a^N) \\ + \frac{1}{2}((1-\lambda)\hat{d}_a^{CP} + \lambda d_a^N)^\top H_a((1-\lambda)\hat{d}_a^{CP} + \lambda d_a^N).\end{aligned}$$

The relations $H_a d_a^N = -\nabla f(x_a)$ and $\hat{d}_a^{CP} = -\hat{\alpha}_a \nabla f(x_a)$ imply

$$\begin{aligned}
\psi(\lambda) &= f(x_a) - \hat{\alpha}_a(1-\lambda)\|\nabla f(x_a)\|^2 - \lambda \nabla f(x_a)^\top H_a^{-1} \nabla f(x_a) + \frac{1}{2}(1-\lambda)^2 \hat{\alpha}_a^2 \nabla f(x_a)^\top H_a \nabla f(x_a) \\
&\quad + (-\hat{\alpha}_a)(1-\lambda)\nabla f(x_a)^\top H_a \lambda(-H_a^{-1}\nabla f(x_a)) + \frac{1}{2}\lambda^2 \nabla f(x_a)^\top H_a^{-1} H_a H_a^{-1} \nabla f(x_a) \\
&= f(x_a) - \hat{\alpha}_a(1-\lambda)\|\nabla f(x_a)\|^2 - \lambda \nabla f(x_a)^\top H_a^{-1} \nabla f(x_a) + \frac{1}{2}(1-\lambda)^2 \hat{\alpha}_a^2 \nabla f(x_a)^\top H_a \nabla f(x_a) \\
&\quad + \hat{\alpha}_a(\lambda - \lambda^2)\|\nabla f(x_a)\|^2 + \frac{1}{2}\lambda^2 \nabla f(x_a)^\top H_a^{-1} \nabla f(x_a) \\
&= f(x_a) - \hat{\alpha}_a(1-\lambda)^2\|\nabla f(x_a)\|^2 + \frac{1}{2}(1-\lambda)^2 \hat{\alpha}_a^2 \nabla f(x_a)^\top H_a \nabla f(x_a) + \lambda(1-\frac{\lambda}{2})\nabla f(x_a)^\top d_a^N
\end{aligned}$$

Using $\hat{\alpha}_a = \frac{\|\nabla f(x_a)\|^2}{\nabla f(x_a)^\top H_a \nabla f(x_a)}$, we obtain

$$\begin{aligned}
\psi'(\lambda) &= 2\hat{\alpha}_a(1-\lambda)\|\nabla f(x_a)\|^2 - (1-\lambda)\hat{\alpha}_a^2 \nabla f(x_a)^\top H_a \nabla f(x_a) + (1-\lambda)\nabla f(x_a)^\top d_a^N \\
&= 2\hat{\alpha}_a(1-\lambda)\|\nabla f(x_a)\|^2 - (1-\lambda)\hat{\alpha}_a\|\nabla f(x_a)\|^2 + (1-\lambda)\nabla f(x_a)^\top(-H_a^{-1}\nabla f(x_a)) \\
&= (1-\lambda)\nabla f(x_a)^\top(\hat{\alpha}_a \nabla f(x_a) - H_a^{-1}\nabla f(x_a)) \\
&= \frac{1-\lambda}{\hat{\alpha}_a}(x_a - \hat{x}_a^{CP})^\top(x_a - H_a^{-1}\nabla f(x_a) - (x_a - \hat{\alpha}_a \nabla f(x_a))) \\
&= \frac{1-\lambda}{\hat{\alpha}_a}(x_a - \hat{x}_a^{CP})^\top(x_a^N - \hat{x}_a^{CP}) \leq 0.
\end{aligned}$$

□

We have shown that the trust region subproblem possesses a unique solution. Now we can prove the following global convergence theorem.

THEOREM 7.12. *Let ∇f be Lipschitz continuous with modulus L. Let $\{x^k\}$ be generated by Algorithm 7.4, where the solutions of the trust region subproblem are given by* (D). *Furthermore we assume that the sequence of matrices $\{H^k\}$ is bounded. Then either $\{f(x^k)\}$ is bounded from below or $\nabla f(x^k) = 0$ for a finite k or*

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

PROOF. We have to prove that the conditions of Assumption 7.2 are satisfied.
Condition 2: Let $\|H^k\| \leq \Delta$. In case $\|d^k\| \leq \Delta$, the definition of x^D yields (7.26) and $x_v = x_N^k$. Further we have

$$\|d^k\| = \|x^k - x_N^k\| = \|(H^k)^{-1}\nabla f(x^k)\| \geq \frac{\|\nabla f(x^k)\|}{M}.$$

Condition 1: We distinguish the different cases for the determination of x^D. If $x^D = x_a^{CP}$, then either $\|d_a^{CP}\| = \Delta_a$ or (7.26) does not hold true. First we consider the case $\nabla f(x_a)^\top H_a \nabla f(x_a) \leq 0$. Then $\|d_a^{CP}\| = \Delta_a$ and $\alpha_a = \frac{\Delta_a}{\|\nabla f(x_a)\|}$. Therefore it holds that

$$\begin{aligned}
pred_k &= \alpha_a\|\nabla f(x_a)\|^2 - \frac{\alpha_a^2}{2}\nabla f(x)^\top H_a \nabla f(x_a) \\
&= \Delta_a\|\nabla f(x_a)\| - \Delta_a^2 \frac{\nabla f(x_a)^\top H_a \nabla f(x_a)}{2\|\nabla f(x_a)\|^2} \geq \Delta_a\|\nabla f(x_a)\| = \|d_v\| \cdot \|\nabla f(x_a)\|
\end{aligned}$$

Condition 1 holds true for $\sigma = 1$. The following cases can be checked similarly:

$$\nabla f(x_a)^\top H_a \nabla f(x_a) > 0 \quad \wedge \quad \begin{cases} \|d_a^{CP}\| = \Delta_a \\ \|d_a^{CP}\| < \Delta_a \end{cases} \Rightarrow \sigma = \frac{1}{2}.$$

Finally, there is the case, where (7.26) holds true whereas $x^D \neq x_a^{CP}$. In this situation we have

$$pred_k \geq m_a(x_a) - m_a(x_a^{CP}) \geq \frac{\|\nabla f(x_a)\|^2}{M}.$$

Now we can apply Theorem 7.10 to obtain the assertion. □

Finally, we establish fast local convergence.

THEOREM 7.13. *Let ∇f be Lipschitz continuous with modulus L. Let $\{x^k\}$ be generated by Algorithm 7.4, where the solutions of the trust region subproblems are given by* (D)*. Further assume that $H^k = \nabla^2 f(x^k)$ and $\{H^k\}$ is bounded, f is bounded from below and x^* is a minimizer of f satisying* (A)*. If $\lim_k x^k = x^*$, then x^k converges Q-quadratically to x^*.*

PROOF. Since x^* is the limit of $\{x^k\}$, there exists a $\delta > 0$ such that for sufficiently large k it holds that

$$\|x^k - x^*\| \leq \delta, \|H^k\| \leq 2\|\nabla^2 f(x^*)\|, \|(H^k)^{-1}\| \leq 2\|(\nabla^2 f(x^*))^{-1}\|.$$

Furthermore, let δ be chosen such that the assertions of Theorem 7.1 hold true. If $H^k \in \mathcal{S}^k$ is positive definite, then $(H^k)^{-1} \in \mathcal{S}^k$ is positive definite as well and (7.26) is fulfilled. Hence, the "dogleg" path starts at x_a and runs through x_a^{CP} to x_a^N. For sufficiently small ρ, the it holds that

$$\|(H^k)^{-1}\nabla f(x^k)\| \leq 2\|x^k - x^*\| \leq 2\rho.$$

From this we infer (see proof of Theorem 7.14):

$$pred_k \geq \frac{1}{2}\|d_v^k\| \cdot \|\nabla f(x^k)\|.$$

Moreover, we have

$$\begin{aligned} ared_k &= -\nabla f(x^k)^\top d_v^k - \int_0^1 (\nabla f(x^k + \tau d_v^k) - \nabla f x^k)^\top d_v^k d\tau \\ &= pred_k + \frac{1}{2}(d_v^k)^\top \nabla^2 f(x^k) d_v^k - \int_0^1 (\nabla f(x^k + \tau d_v^k) - \nabla f x^k)^\top d_v^k d\tau \\ &= pred_k + \mathcal{O}\left(\|d_v^k\| \cdot \|\nabla f(x^k)\|\rho\right) \end{aligned}$$

Hence, $\frac{ared_k}{pred_k} = 1 - \mathcal{O}(\rho)$. For sufficiently small ρ, the trust region radius will be enalrged until x_a^N is located in the trust region, and x_a^N will be accepted. □

REMARK 7.3. Some trust region methods realize the inexact Newton idea. The local rate of convergence can be derived analogously to Theorem 7.9.

CHAPTER 8

Quasi-Newton methods

Unlike Newton's method, quasi-Newton methods do not make use of second order derivatives of f. Rather, they approximate the second order derivatives iteratively with the help of first order derivatives. Here, we consider quasi-Newton methods, which essentially work like Newton methods with line search, but $\nabla^2 f(x^k)$ is approximated by a positive definite matrix H^k. At each iteration, H^k is updated in an appropriate way. The general algorithmic structure is as follows:

(1) Set $d^k = -(H^k)^{-1}\nabla f(x^k)$.
(2) Determine $x^{k+1} = x^k + \alpha_k d^k$ by a step size strategy.
(3) Use x^k, x^{k+1} and H^k, to update H^k to H^{k+1}.

For the initial matrix H^0 choose a (symmetric) positive definite matrix. A standard choice is given by $H^0 = I$, but sometimes better scaling might be necessary. The benefits of quasi-Newton methods are (amongst others):

- Only first order derivatives are required.
- H^k is always positive definite such that d^k is a descent direction for f at x^k and our line search framework is applicable.
- Some variants require only $\mathcal{O}(n^2)$ multiplications per iteration (instead of $\mathcal{O}(n^3)$ like Newton's method).

The last point is related to quasi-Newton variants, which approximate $(\nabla^2 f(x^k))^{-1}$ directly, thus sparing the cost of solving the linear system to determine d^k.
As we will see soon, positive definiteness is not guaranteed for all quasi-Newton methods. In the case where $\{H^k\}$ consists only of positive definite matrices, then one speaks of a *"variable metric"-method*.

1. Update rules

In this section we will discuss several update rules for the Hessian approximation H. Let

$$\begin{aligned} s_a &= \alpha_a d_a \quad (= -\alpha_a (H^k)^{-1}\nabla f(x^k)), \\ y_a &= \nabla f(x_+) - \nabla f(x_a) \quad \text{with } x_+ = x_a + s_a. \end{aligned}$$

It holds that

$$\begin{aligned} y_a &= \nabla f(x_+) - \nabla f(x_a) = \nabla f(x_a) + \nabla^2 f(x_a)(x_+ - x_a) + \mathcal{O}(\|x_+ - x_a\|) - \nabla f(x_a) \\ &= \nabla^2 f(x_a) s_a + \mathcal{O}(\|s_a\|). \end{aligned}$$

Therefore we require

$$H_+ s_a = y_a. \tag{8.1}$$

This condition is the called *quasi-Newton condition* (or secant condition). A simple ansatz for H_+ in (8.1) is

$$H_+ = H_a + \alpha u u^\top, \quad \alpha \in \mathbb{R}, u \in \mathbb{R}^n,$$

which is referred to as the *symmetric rank-1-update*. Inserting this update into (8.1) yields

$$H_a s_a + \alpha u (u^\top s_a) = y_a.$$

Thus, u is proportional to $y_a - H_a s_a$. We set $u = y_a - H_a s_a$ (as the length can be adjusted by α) which implies $\alpha u^\top s_a = 1$. This results in the *symmetric rank-1-formula*

$$H_+ = H_a + \frac{(y_a - H_a s_a)(y_a - H_a s_a)^\top}{(y_a - H_a s_a)^\top s_a}. \tag{8.2}$$

Unfortunately this formula has a few drawbacks. In particular the positive definiteness gets lost (even if H^0 is chosen positive definite) and numerical problems appear whenever $y_a - H_a s_a \approx 0$ resp. $(y_a - H_a s_a)^\top s_a \approx 0$.
The non-symmetric ansatz

$$H_+ = H_a + \alpha u v^\top, \quad \alpha \in \mathbb{R}, u, v \in \mathbb{R}^n,$$

with $v := s_a$ inserted into (8.1) yields

$$H_a s_a + \alpha u (s_a^\top s_a) = y_a.$$

Hence, u is proportional to $y_a - H_a s_a$, which immediately implies $\alpha(s_a^\top s_a) = 1$. We obtain *the non-symmetric rank-1-formula*

$$H_+ = H_a + \frac{(y_a - H_a s_a) s_a^\top}{s_a^\top s_a}. \tag{8.3}$$

The fact that positive definiteness cannot be guaranteed and the absence of symmetry represent crucial disadvantages of (8.3).
More flexible update formulae can be derived by applying *symmetric rank-2-updates*, i.e.

$$H_+ = H_a + \alpha u u^\top + \beta v v^\top, \quad \alpha, \beta \in \mathbb{R}, u, v \in \mathbb{R}^n.$$

Inserting this into (8.1) yields

$$H_a s_a + \alpha u u^\top s_a + \beta v v^\top s_a = y_a. \tag{8.4}$$

The vectors u and v are no longer uniquely determined. In view of (8.4), it is adequate to choose

$$u = y_a \quad \text{and} \quad v = H_a s_a.$$

Then we obtain

$$\alpha y_a y_a^\top s_a + \beta (H_a s_a)(H_a s_a)^\top s_a = y_a - H_a s_a,$$

which implies

$$\alpha(y_a^\top s_a) = 1 \quad \text{and} \quad \beta(s_a^\top H_a s_a) = -1.$$

Thus,

$$\alpha = \frac{1}{y_a^\top s_a} \quad \text{and} \quad \beta = -\frac{1}{s_a^\top H_a s_a}$$

and finally

$$H_+ = H_a + \frac{y_a y_a^\top}{y_a^\top s_a} - \frac{(H_a s_a)(H_a s_a)^\top}{s_a^\top H_a s_a}. \tag{8.5}$$

The update rule (8.5) is called the *BFGS-formula* (named after Broyden-Fletcher-Goldfarb-Shanno).
One may also approximate $\nabla^2 f(x^k)^{-1}$ by B^k. In this case the quasi-Newton condition reads

$$B_+ y_a = s_a.$$

Applying a symmetric rank-2-update analogous to (8.4) with $u = s_a$ and $v = B_a y_a$, one obtains

$$B_+ = B_a + \frac{s_a s_a^\top}{s_a^\top y_a} - \frac{(B_a y_a)(B_a y_a)^\top}{y_a^\top B_a y_a}. \tag{8.6}$$

This formula is called the *DFP-formula* (after Davidon-Fletcher-Powell). Owing to the relations $B \leftrightarrow H$ and $y \leftrightarrow s$, (8.5) and (8.6) are considered as dual to each other. Numerical practice shows that the BFGS-method is often superior to the DFP-method. In the following, we therefore will focus on the update according to (8.5).
First of all, note that if $H^0 \in \mathcal{S}^n$ also $H^k \in \mathcal{S}^n$ due to the structure of (8.5). The positive definiteness is considered in the following lemma.

LEMMA 8.1. *Let $H_a \in \mathcal{S}^n$ be positive definite, $y_a^\top s_a > 0$ and H_+ determined according to (8.5). Then $H_+ \in \mathcal{S}^n$ is positive definite.*

PROOF. Positive definiteness of H_a and $y_a^\top s_a \neq 0$ yield for all $z \neq 0$:

$$\begin{aligned} z^\top H_+ z =& z^\top H_a z + \frac{z^\top y_a y_a^\top z}{y_a^\top s_a} - \frac{z^\top H_a s_a \cdot (H_a s_a)^\top z}{s_a^\top H_a s_a} \\ =& \frac{(z^\top y_a)^2}{y_a^\top s_a} + z^\top H_a z - \frac{(z^\top H_a s_a)^2}{s_a^\top H_a s_a}. \end{aligned}$$

Since $H_a \in \mathcal{S}^n$ is positive definite, there exists $H_a^{\frac{1}{2}}$ with $H_a = H_a^{\frac{1}{2}} \cdot H_a^{\frac{1}{2}}$.
Thus, the following holds true:

$$|(z^\top H_a s_a)| = |(z^\top H_a^{\frac{1}{2}} \cdot H_a^{\frac{1}{2}} s_a)| \leq \|H_a^{\frac{1}{2}} z\| \cdot \|H_a^{\frac{1}{2}} s_a\|$$

and also

$$(z^\top H_a s_a)^2 \leq \|H_a^{\frac{1}{2}} z\|^2 \cdot \|H_a^{\frac{1}{2}} s_a\|^2 = (z^\top H_a z) \cdot (s_a^\top H_a s_a).$$

Equality only holds, if $z = 0$ or $s_a = 0$ (but neither is relevant in our situation), or $z = \kappa s_a, \kappa \in \mathbb{R}$. The latter implies either $\frac{(z^\top H_a s_a)^2}{s_a^\top H_a s_a} < z^\top H_a z$ or $z = \kappa s_a$ and $\frac{(z^\top y_a)^2}{y_a^\top s_a} > 0$. Altogether we obtain: $z^\top H_+ z > 0$. As z was arbitrarily chosen, the assertion is proven. □

The condition $y_a^\top s_a > 0$ is realistic. For quadratic problems with positive definite Hessian G, it holds that

$$y_a^\top s_a = (\nabla f(x_+) - \nabla f(x_a))^\top (x_+ - x_a) = (x_+ - x_a)^\top G (x_+ - x_a) > 0.$$

For general problems, $y_a^\top s_a > 0$ is ensured by the Wolfe-Powell step size strategy.

2. Local convergence theory

Before discussing local convergence properties, we demonstrate that Newton's method resp. the BFGS-method are invariant under affine transformations. For this purpose, let A be a nonsingular $n \times n$-matrix, $b \in \mathbb{R}^n$ and $y = Ax + b$ resp. $x = A^{-1}(y - b)$. The chain rule implies

$$\frac{\partial}{\partial x_i} = \sum_k \frac{\partial y_k}{\partial x_i} \frac{\partial}{\partial y_k} = \sum_k A_{ki} \frac{\partial}{\partial y_k}.$$

It follows "$\nabla_x = A^\top \nabla_y$" and

$$\begin{aligned} \nabla_x f &= A^\top \nabla_y f, \\ \nabla_x^2 f &= A^\top \nabla_y^2 f A. \end{aligned}$$

Let H_a^x and H_a^y denote the current BFGS-matrices corresponding to the differentiation w.r.t. x resp. y. Suppose that

$$(H_a^x)^{-1} = A^{-1}(H_a^y)^{-1}A^{-\top}, \quad y_a = Ax_a + b,$$

and that in both cases the same step size α_a is chosen. Then the BFGS -method is *invariant under affine transformations* $T(x) = Ax + b$. To see this, consider

$$x_+ = x_a - \alpha_a (H_a^x)^{-1} \nabla_x f(x_a)$$

and w.r.t. y,

$$\begin{aligned} y_+ =& y_a - \alpha_a (H_a^y)^{-1} \nabla_y f(y_a) \\ =& Ax_a + b - \alpha_a A (H_a^x)^{-1} A^\top A^{-\top} \nabla_x f(x_a) \\ =& A(x_a - \alpha_a (H_a^x)^{-1} \nabla_x f(x_a)) + b = Ax_+ + b. \end{aligned}$$

An analogous argument can be used for Newton's method. As a consequence, from now on we may suppose that $\nabla^2 f(x^*) = I$ holds true (this is ensured by the transformation $T(x) = (\nabla^2 f(x^*))^{-1/2}$).
We next specify the central convergence theorem. Its proof is based on some auxiliary results, which we establish in the remainder of this section.

THEOREM 8.1. *Let* (A) *be satisfied. Then there exists* $\delta > 0$ *such that for*

$$\|x^0 - x^*\| \le \delta \quad \textit{and} \quad \|H^0 - \nabla^2 f(x^*)\| \le \delta$$

the BFGS-method is well-defined and converges q-superlinearly to x^*.

As announced, for the proof of the above result we need several auxiliary results. For this, the error in the approximation of the inverse of the Hessian is denoted by

$$F = H^{-1} - \nabla^2 f(x^*)^{-1} = H^{-1} - I.$$

We state without proof the following lemma.

LEMMA 8.2. *Let* (A) *be satisfied. If* $H_a \in \mathcal{S}^n$ *is positive definite and*

$$x_+ = x_a - H_a^{-1} \nabla f(x_a),$$

then there exists δ_0 *such that for*

$$0 < \|x_a - x^*\| \le \delta_0 \quad \textit{and} \quad \|F_a\| \le \delta_0,$$

it holds that

$$y_a^\top s_a > 0.$$

Furthermore we have that the BFGS-update H_+ of H_a satisfies

$$F_+ = H_+^{-1} - I = (I - w_a w_a^\top) F_a (I - w_a w_a^\top) + D_a$$

with $w_a = \frac{s_a}{\|s_a\|}$, $D_a \in \mathbb{R}^{n\times n}$ and $\|D_a\| \le K_D \|s_a\|$ with $K_D > 0$.

Basically, Lemma 8.2 indicates that the approximation is close to the exact Hessian, if the initial values are "good" enough. This property is elementary for proving local superlinear convergence.

COROLLARY 8.1. *Under the assumptions of Lemma 8.2, it holds that*

$$\|F_+\| \le \|F_a\| + K_D\|s_a\| \le \|F_a\| + K_D(\|x_a - x^*\| + \|x_+ - x^*\|).$$

The second inequality in the assertion of Corollary 8.1 follows directly from $s_a = x_+ - x_a$. The first inequality can be obtained by expanding the representation of F_+ from Lemma 8.2 and estimating the subsequent expression using $\|D_a\| \le K_D\|s_a\|$.
At this point, we can prove local q-linear convergence.

THEOREM 8.2. *Let* (A) *be satisfied and $\sigma \in (0,1)$ be given. Then there exists δ_l such that for*

$$\|x^0 - x^*\| \le \delta_l \quad \text{and} \quad \|(H^0)^{-1} - \nabla^2 f(x^*)^{-1}\| \le \delta_l \tag{8.7}$$

the BFGS-iteration is well-defined and converges q-linearly to x^*. The q-factor is bounded by σ.

Note that in general δ_l is directly proportional to σ.

PROOF. For sufficiently small $\hat\delta$ and

$$\|x_a - x^*\| \le \hat\delta \ \text{ and } \ \|F_a\| = \|H_a^{-1} - I\| \le \hat\delta, \tag{8.8}$$

(A) yields

$$\|x_+ - x^*\| \le \|F_a\|\,\|x_a - x^*\| + o(\|x_a - x^*\|) \le \hat\delta\|x_a - x^*\| + o(\|x_a - x^*\|).$$

Let $\hat\delta$ be small enough such that

$$\|x_+ - x^*\| \le \sigma\|x_a - x^*\| < \|x_a - x^*\| \le \hat\delta.$$

Choose δ_l such that (8.8) holds true for the entire iteration, provided that the initial value satisfies (8.7). We choose

$$\delta_l = \frac{\delta^*}{2}\left(1 + \frac{K_D(1+\sigma)}{1-\sigma}\right)^{-1} < \frac{\delta^*}{2} \tag{8.9}$$

with K_D from Lemma 8.2. In case that $\|I - H^0\| < \delta_l$ with $\delta_l < \frac{1}{2}$, we infer

$$\begin{aligned}\|F^0\| =& \|(H^0)^{-1} - I\| \le \|(H^0)^{-1}\|\,\|I - H^0\| = \|(I - (I - H^0))^{-1}\|\,\|I - H^0\| \\ \le& \frac{1}{1 - \|I - H^0\|}\,\|I - H^0\| \le \frac{\delta_l}{1 - \delta_l} \le 2\delta_l \le \delta^*.\end{aligned}$$

Corollary 8.1 yields

$$\|F^1\| \le \|F^0\| + K_D(1+\sigma)\|x^0 - x^*\|.$$

It remains to prove that (8.7) and (8.9) imply

$$\|F^k\| < \delta^* \quad \forall k.$$

For this purpose we proceed inductively. Let $\|F^k\| < \delta^*$ and $\|x^{j+1} - x^*\| \leq \sigma \|x^j - x^*\| \; \forall j \leq n$. Then Corollary 8.1 implies

$$\begin{aligned}
\|F^{k+1}\| &\leq \|F^k\| + K_D(\|x^k - x^*\| + \|x^{k+1} - x^*\|) \\
&\leq \|F^k\| + K_D(1+\sigma)\|x^k - x^*\| \\
&\leq \|F^k\| + K_D(1+\sigma)\sigma^k\|x^0 - x^*\| \\
&\leq \|F^k\| + K_D(1+\sigma)\sigma^k\delta_l \\
&\leq \|F^0\| + \delta_l K_D(1+\sigma)\sum_{j=0}^{k}\sigma^j \\
&\leq \delta_l(1 + \frac{K_D(1+\sigma)}{1-\sigma}) < \delta^*.
\end{aligned}$$

□

We derive now some useful relations. Assumption (A) also ensures $\nabla f(x_a) \neq 0$ for x_a ($x_a \neq x^*$) sufficiently close to x^*. Moreover, it holds that

$$\nabla f(x_a) = \int_0^1 \nabla^2 f(x^* + \tau(x_a - x^*))(x_a - x^*)d\tau = (I + R_1)(x_a - x^*) \tag{8.10}$$

with

$$R_1 = \int_0^1 (\nabla^2 f(x^* + \tau(x_a - x^*)) - I)d\tau.$$

Thus we obtain $\|R_1\| \leq \frac{\gamma}{2}\|x_a - x^*\|$ as well as

$$s_a = -H_a^{-1}\nabla f(x_a) = -(I + F_a)(I + R_1)(x_a - x^*). \tag{8.11}$$

In case $\|F_a\| \leq \delta_0$ und $\|x_a - x^*\| \leq \delta_0$ (cf. Lemma 8.2), we have

$$\|x_a - x^*\|(1-\delta_0)(1 - \frac{\gamma\delta_0}{2}) \leq \|s_a\| \leq \|x_a - x^*\|(1+\delta_0)(1 + \frac{\gamma\delta_0}{2})$$

and consequently

$$0 < \frac{1}{2}\|x_a - x^*\| \leq \|s_a\| \leq 2\|x_a - x^*\| \tag{8.12}$$

for $\delta_0 \leq \min(\frac{1}{4}, \frac{1}{2\gamma})$. Moreover,

$$\begin{aligned}
y_a =& \nabla f(x_+) - \nabla f(x_a) = \int_0^1 \nabla^2 f(x_a + \tau s_a)s_a d\tau \\
=& s_a + \int_0^1 (\nabla^2 f(x_a + \tau s_a) - I)s_a d\tau = s_a + R_2 s_a,
\end{aligned} \tag{8.13}$$

where

$$R_2 = \int_0^1 (\nabla^2 f(x_a + \tau s_a) - I)d\tau.$$

Using $I = \nabla^2 f(x^*)$ and the Lipschitz continuity of $\nabla^2 f$, we infer

$$\begin{aligned}\|R_2\| \leq \int_0^1 \gamma \|x_a + \tau s_a - x^*\| d\tau \leq \gamma \|x_a - x^*\| + \frac{\gamma}{2}\|s_a\| \\ \leq 2\gamma \|s_a\| + \frac{\gamma}{2}\|s_a\| = \frac{5\gamma}{2}\|s_a\|. \end{aligned} \tag{8.14}$$

In the proof of the following theorem, the *Dennis-Moré condition* is applied. It represents a necessary and sufficient condition for superlinear convergence of quasi-Newton methods. The Dennis-Moré condition reads

$$\lim_{k\to\infty} \frac{\|F^k s^k\|}{\|s^k\|} = 0. \tag{8.15}$$

THEOREM 8.3. *Suppose the standard assumption* (A) *is fulfilled. Let* $\{H^k\}_{k\in\mathbb{N}}$ *be a sequence of nonsingular matrices with* $\|H^k\| \leq M$ *for all* $k \in \mathbb{N}$. *Further let* x^0 *be given and* $\{x^k\}_{k=1}^{\infty}$ *defined by*

$$x^{k+1} = x^k - (H^k)^{-1}\nabla f(x^k).$$

If x^k *converges q-linearly to* x^*, $x^k \neq x^*$ *for all* $k \in \mathbb{N}$, *and* (8.15) *is met, then* x^k *converges q-superlinearly to* x^*.

PROOF. The equations (8.13) and (8.14) yield

$$F^k s^k \overset{(8.13)}{=} \left((H^k)^{-1} - I\right)(y^k - R_2 s^k) \overset{(8.14)}{=} F^k y^k + \mathcal{O}\left(\|s^k\|^2\right).$$

It holds that $x^k \to x^*$ and, hence, $s^k \to 0$. The Dennis-Moré condition (8.15) can be rewritten as

$$\lim_{k\to\infty} \frac{\|F^k y^k\|}{\|s^k\|} = 0. \tag{8.16}$$

Now let σ denote the q-factor of $\{x^k\}$. Then it holds that

$$(1-\sigma)\|x_a - x^*\| \leq \|s_a\| \leq (1+\sigma)\|x_a - x^*\|,$$

as $\|s_a\| = \|x_+ - x_a\|$ and

$$\|x_+ - x_a + x^* - x^*\| \leq \|x_+ - x^*\| + \|x_a - x^*\| \leq \sigma\|x_a - x^*\| + \|x_a - x^*\|,$$

resp.

$$\|x_+ - x_a\| \geq (1-\sigma)\|x_a - x^*\|.$$

Therefore (8.16) is equivalent to

$$\lim_{k\to\infty} \frac{\|F^k y^k\|}{\|x^k - x^*\|} = 0.$$

Since $(H^k)^{-1}\nabla f(x^k) = -s^k$ and $s^k = y^k + \mathcal{O}(\|s^k\|^2)$ (owing to (8.13)), it follows that

$$\begin{aligned} F^k y^k &= \left((H^k)^{-1} - I\right)(\nabla f(x^{k+1}) - \nabla f(x^k)) \\ &= (H^k)^{-1}\nabla f(x^{k+1}) + s^k - y^k = (H^k)^{-1}\nabla f(x^{k+1}) + \mathcal{O}(\|s^k\|^2) \\ &= (H^k)^{-1}(x^{k+1} - x^*) + \mathcal{O}(\|x^k - x^*\|^2 + \|s^k\|^2) = (H^k)^{-1}(x^{k+1} - x^*) + \mathcal{O}(\|x^k - x^*\|^2). \end{aligned}$$

Thus we have:

$$\begin{aligned}\frac{\|F^k y^k\|}{\|x^k - x^*\|} &= \frac{\|(H^k)^{-1}(x^{k+1} - x^*)\|}{\|x^k - x^*\|} + \mathcal{O}(\|x^k - x^*\|) \\ &\geq M^{-1}\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} + \mathcal{O}(\|x^k - x^*\|) \to 0,\end{aligned}$$

which yields the q-superlinear convergence of x^k to x^*. □

Now we can state the proof of Theorem 8.1.

Proof of Theorem 8.1. Assume (8.7) of Theorem 8.2 is fulfilled with δ_l such that (8.2) holds true for $\sigma \in (0,1)$. It follows immediately that

$$\sum_{k=0}^{\infty} \|s^k\| < \infty. \tag{8.17}$$

Let $\|A\|_F^2 = \Sigma_{i,j=1}^N (A)_{ij}^2 = trace(A^\top A)$ be the Frobenius norm of the matrix A. For $v \in \mathbb{R}^n$ with $\|v\| \leq 1$, it holds that

$$\|A(I - vv^\top)\|_F^2 \leq \|A\|_F^2 - \|Av\|^2,$$

as

$$\begin{aligned}\|A(I - vv^\top)\|_F^2 &= trace((I - vv^\top)^\top A^\top A(I - vv^\top)) \\ &= trace((A^\top A - vv^\top A^\top A)(I - vv^\top)) \\ &= trace((A^\top A) - vv^\top A^\top A - A^\top Avv^\top + vv^\top A^\top Avv^\top),\end{aligned}$$

$trace(vv^\top A^\top A) = \|Av\|^2$ and $trace(vv^\top A^\top Avv^\top) \leq trace(vv^\top) trace(A^\top Avv^\top) = \|v\|^2\|Av\|^2 \leq \|Av\|^2$. Moreover, $\|(I - vv^\top)A\|_F^2 \leq \|A\|_F^2$. Thus, Lemma 8.2 implies

$$\|F^{k+1}\|_F^2 \leq \|F^k\|_F^2 - \|F^k w^k\|^2 + \mathcal{O}(\|s^k\|) = (1 - \Theta_k^2)\|F^k\|_F^2 + \mathcal{O}(\|s^k\|)$$

with

$$w^k = \frac{s^k}{\|s^k\|},\ \Theta_k = \begin{cases} \frac{\|F^k w^k\|}{\|F^k\|_F} & \text{if} \quad F^k \neq 0. \\ 1 & \text{if} \quad F^k = 0. \end{cases}$$

With (8.17), we have for $k \geq 0$:

$$\begin{aligned}\sum_{k=0}^{\infty} \Theta_k^2 \|F^k\|_F^2 &\leq \sum_{k=0}^{\infty} (\|F^k\|_F^2 - \|F^{k+1}\|_F^2) + \mathcal{O}(1) \\ &= \|F^0\|_F^2 - \|F^{k+1}\|_F^2 + \mathcal{O}(1) < \infty.\end{aligned}$$

Hence, $\Theta_k \|F^k\|_F \to 0$ and finally

$$\Theta_k \|F^k\|_F = \begin{cases} \|F^k w^k\| & \text{if} \quad F^k \neq 0. \\ 0 & \text{if} \quad F^k = 0. \end{cases}$$

Moreover,

$$\|F^k w^k\| = \frac{\|F^k s^k\|}{\|s^k\|}.$$

Thus, the Dennis-Moré condition is met. □

3. Global convergence

Under the assumption that there exist some constants $0 < c_1 < c_2 < +\infty$ with

$$c_1\|x\|^2 \leq x^\top H^k x \leq c_2\|x\|^2 \quad \forall x \in \mathbb{R}^n \forall k \in \mathbb{N},$$

$d^k = -(H^k)^{-1}\nabla f(x^k)$ is a gradient-related search direction. If the Armijo step size strategy is applied, a statement analogous to Theorem 5.1 holds true. Note that for a local minimizer x^* one cannot expect in general to have $x^k \to x^*$ for x^0 close to x^*. In view of the local theory, the situation where x^0 is situated close to x^*, but H^0 is not close to $\nabla^2 f(x^*)$ is not better than the case where x^0 is not sufficiently close to x^*.

THEOREM 8.4. *Let $D := \{x : f(x) \leq f(x^*)\}$ be convex, f twice continuously differentiable in D and the spectrum $\sigma(\nabla^2 f(x)) \subset [c_1, c_2]$ for all $x \in D$. If $H^0 \in \mathcal{S}^n$ is positive definite, then the BFGS-method with Armijo step size strategy converges q-superlinearly to x^*.*

Similar statements hold true for the (strict) Wolfe-Powell step size strategy.

4. Numerical aspects

4.1. Memory-efficient updating. The BFGS-method always uses the current Hessian approximation to calculate the new approximation with the aid of the rank-2-update. In general, one expects that the performance of the approximation is improving in the course of the iteration. Indeed, one can show that for quadratic problems, adequate initialization and an exact step size strategy, the exact Hessian is perfectly approximated after at most n (for $\nabla^2 f(x^*) \in \mathbb{R}^{n\times n}$) steps. Nonetheless, for general problems the situation is more complicated and due to numerical reasons, a reset of H^k to a well-scaled, positive definite matrix is occasionally implemented. However, for high-dimensional problems storing the BFGS-matrix is unesireable because of memory restrictions. The so-called *limited memory-BFGS method* takes this issue into account. In this method, from the point of view of the current iterate x^k, the preceding pairs $\{(y^l, s^l)\}$, $k - m \leq l \leq k$, are stored and the BFGS-formula is realized iteratively at each new iteration. Here, $m \in \mathbb{N}$ is a fixed number.
We will now specify a strategy, which approximates the inverse Hessian without occupying to much memory space. Let $s_a = x_+ - x_a$, $y_a = \nabla f(x_+) - \nabla f(x_a)$ and H_+ be computed from H_a by the BFGS-formula. Then, for positive definite $H_a \in \mathcal{S}^n$ and $y_a^\top s_a \neq 0$ it holds that H_+ is nonsingular and

$$H_+^{-1} = \left(I - \frac{s_a y_a^\top}{y_a^\top s_a}\right) H_a^{-1} \left(I - \frac{y_a s_a^\top}{y_a^\top s_a}\right) + \frac{s_a s_a^\top}{y_a^\top s_a}. \tag{8.18}$$

Rearranging yields

$$H_+^{-1} = H_a^{-1} + \beta_0 s_a s_a^\top + \gamma_0 \left((H_a^{-1} y_a) s_a^\top + s_a (H_a^{-1} y_a)^\top\right)$$

with coefficients

$$\beta_0 = \frac{y_a^\top s_a + y_a^\top H_a^{-1} y_a}{(y_a^\top s_a)^2} \quad \text{and} \quad \gamma_0 = -\frac{1}{y_a^\top s_a}.$$

From the relation

$$H_a^{-1} y_a = H_a^{-1}\nabla f(x_+) - H_a^{-1}\nabla f(x_a) = H_a^{-1}\nabla f(x_+) + \frac{s_a}{\alpha_a},$$

it follows that

$$H_+^{-1} = H_a^{-1} + \beta_1 s_a s_a^\top + \gamma_0 \left(s_a (H_a^{-1}\nabla f(x_+))^\top + (H_a^{-1}\nabla f(x_+))^\top s_a\right) \tag{8.19}$$

with $\beta_1 = \beta_0 + \frac{2\gamma_0}{\alpha_a}$. For the new direction d_+ this yields

$$\begin{aligned} d_+ =& - H_+^{-1}\nabla f(x_+) = -\left(I - \frac{s_a y_a^\top}{y_a^\top s_a}\right) H_a^{-1} \left(I - \frac{y_a s_a^\top}{y_a^\top s_a}\right) \nabla f(x_+) - \frac{s_a s_a^\top \nabla f(x_a+)}{y_a^\top s_a} \\ =& A_a s_a + C_a H_a^{-1} \nabla f(x_+) \end{aligned}$$

with

$$\begin{aligned} A_a &= \frac{y_a^\top}{y_a^\top s_a} H_a^{-1} \left(I - \frac{y_a s_a^\top}{y_a^\top s_a}\right) \nabla f(x_+) + \left(\frac{1}{\alpha_a} - 1\right) \frac{s_a^\top \nabla f(x_+)}{y_a^\top s_a}, \\ C_a &= -1 + \frac{s_a^\top \nabla f(x_+)}{y_a^\top s_a}. \end{aligned}$$

Thus, we can determine d_+ and consequently α_+, s_+ just by means of $H_a^{-1}\nabla f(x_+)$ (and s_a). Obviously H_+ is not required. Furthermore, we do not need to store the vectors $\{y_a\}$: As $C_a \neq 0$, it holds that

$$-H_a^{-1}\nabla f(x_+) = -\frac{s_+}{C_a \alpha_+} + \frac{A_a s_a}{C_a}.$$

Inserting into (8.1) yields

$$H_+^{-1} = H_a^{-1} + \beta_a s_a s_a^\top + \gamma_a (s_a s_+^\top + s_+ s_a^\top)$$

with coefficients

$$\beta_a = \beta_1 - 2\gamma_0 \frac{A_a}{C_a} \quad \text{and} \quad \gamma_a = \frac{\gamma_0}{C_a \alpha_+}.$$

Finally that leads to

$$(H^{k+1})^{-1} = (H^0)^{-1} + \sum_{l=0}^{k} \left(\beta_l s^l (s^l)^\top + \gamma_l \left(s^l (s^{l+1})^\top s^{l+1} (s^l)^\top\right)\right).$$

4.2. Positive definiteness. In Lemma 8.1 we have already seen that $y_a^\top s_a > 0$ has to be assumed, in order to guarantee the positive definiteness of $H_+ \in \mathcal{S}^n$. A simple strategy which ensures this property, consists of the following choice:

$$H_+ = \begin{cases} H_+^{BFGS} & \text{if } y_a^\top s_a > 0, \\ R & \text{if } y_a^\top s_a \leq 0 \end{cases}$$

with $R \in \mathcal{S}^n$ positive definite. Often one employs $R = I$.

5. Further Quasi-Newton formulae

Now we will indicate two other interesting updating rules for the Hessian.
The direct *DFP-formula* is

$$H_+ = H_a + \frac{(y_a - H_a s_a) y_a^\top + y_a (y_a - H_a s_a)^\top}{y_a^\top s_a} - \frac{\left((y_a - H_a s_a)^\top y_a\right) y_a y_a^\top}{(y_a^\top s_a)^2}.$$

This updating strategy can be analyzed similarly to the BFGS-method. However, many numerical observations favor the standard BFGS-formula.

A formula which is applied especially in connection with trust region methods is the *PSB-formula (Powell-symmetric-Broyden)*. This formula preserves symmetry, however the positive definiteness generally gets lost:

$$H_+ = H_a + \frac{(y_a - H_a s_a)s_a^\top + s_a(y_a - H_a s_a)^\top}{s_a^\top s_a} - \frac{(s_a^\top y_a - H_a s_a)s_a s_a^\top}{(y_a^\top s_a)^2}.$$

CHAPTER 9

Box-constrained problems

Let $X = \{x \in \mathbb{R} : L_i \leq x_i \leq U_i\}$ with $-\infty < L_i < U_i < +\infty$ for $i = 1, \ldots, n$. We consider the following problem: Find a local minimizer $x^* \in X$ of f, i.e.

$$f(x^*) \leq f(x) \quad \forall x \in X \cap \{z : \|z - x^*\| \leq \epsilon\}.$$

for an $\epsilon > 0$. Given that X is compact [1], there always exists a solution to

$$\min f(x) \quad \text{s.t.} \quad x \in X. \tag{9.1}$$

In case that $x_i = L_i$ or $x_i = U_i$, the index i is called *active*; otherwise i is called *inactive*. We collect active indices in the *active set* $\mathcal{A}(x)$ and inactive indices in the *inactive set* $\mathcal{I}(x)$.

1. Necessary conditions

The first and second order necessary conditions in the scalar case $X \subset \mathbb{R}$ are as follows.

THEOREM 9.1. *Let f be twice continuously differentiable on $[a, b]$, $-\infty < a < b < +\infty$. Let x^* be a local minimizer of f in $[a, b]$. Then it holds that*

$$f'(x^*)(x - x^*) \geq 0 \quad \forall x \in [a, b]$$

and

$$f''(x^*)(x^* - a)(b - x^*) \geq 0.$$

This theorem may serve as a basis for the corresponding conditions in the multidimensional case.
The point $x^* \in X$ is called a stationary point for (9.1), if

$$\nabla f(x^*)^\top (x - x^*) \geq 0 \quad \forall x \in X. \tag{9.2}$$

At the same time, (9.2) represents the first order necessary condition.
In the following we use the term "solution" for "local minimizer".

THEOREM 9.2. *Let f be twice continuously differentiable in X and x^* a solution of (9.1). Then x^* is a stationary point for (9.1).*

PROOF. Let $y \in X$. Since X is convex, we have $z(t) = x^* + t(y - x^*) \in X$ for $t \in [0, 1]$. Consider

$$\phi(t) = f(z(t)).$$

It holds that $\phi(t)$ has a minimum in $t = 0$.
Now Theorem 9.1 implies

$$0 \leq \phi'(t)_{|t=0} = \nabla f(z(t))^\top_{|t=0} \cdot (y - x^*) = \nabla f(x^*)^\top (y - x^*),$$

[1] Of course, this is due to our assumption $-\infty < L_i < U_i < +\infty$. However, if $L_i = -\infty$ or $U_i = +\infty$, then we had to ensure–like in the unconstrained case–the existence of a solution, for instance by assuming certain convexity properties of f.

which completes the proof. □

In order to gain information on second order derivatives, we consider the following situation in $\mathbb{R}^2$. Let $X = [0,1]^2$. If $x^* \in (0,1)^2$, then it follows analogously to the unconstrained case, that $\nabla^2 f(x^*)$ is positive-semidefinite. Assume $x^* = (0, x_2^*)$ is a solution with $0 < x_2^* < 1$. Consider $\phi(t) = f(0,t)$ for $t \in [0,1]$. It necessarily holds:

$$0 \leq \phi''(x_2^*) = \frac{\partial^2 f}{\partial x_2^2}(0, x_2^*).$$

However we cannot make a statement about $\frac{\partial^2 f}{\partial x_1^2}$.
We will now introduce the *reduced Hessian of f*. Let f be twice continuously differentiable, then the reduced Hessian $\nabla_R^2 f(x)$ of f is defined as

$$(\nabla_R^2 f(x))_{ij} = \begin{cases} \delta_{ij} & \text{for } i \in \mathcal{A}(x) \text{ or } j \in \mathcal{A}(x), \\ (\nabla^2 f(x))_{ij} & \text{else.} \end{cases}$$

Here, δ_{ij} denotes the Kronecker-delta. Now we are able to formulate the second order necessary condition.

THEOREM 9.3. *Let f be twice continuously differentiable in X and x^* a solution to* (9.1). *Then $\nabla_R^2 f(x^*)$ is positive-semidefinite.*

PROOF. W.l.o.g. we consider the following partition of x^* :

$$x^* = (x^*_{\mathcal{A}(x^*)}, x^*_{\mathcal{I}(x^*)}),$$

with $x^*_{\mathcal{A}(x^*)} \in \mathbb{R}^{|\mathcal{A}(x^*)|}$ consisting of the components $i \in \mathcal{A}(x^*)$; $x^*_{\mathcal{I}(x^*)}$ analogously.
Thus it holds with $\phi(\xi) := f(x^*_{\mathcal{A}(x^*)}, \xi)$, $\xi \in \mathbb{R}^{|\mathcal{I}(x^*)|}$ that

$$\nabla_R^2 f(x^*) = \begin{pmatrix} I & 0 \\ 0 & \nabla^2 \phi(x^*_{\mathcal{I}(x^*)}) \end{pmatrix}.$$

Now assume that $\nabla^2 \phi(x^*(\mathcal{I}(x^*))$ has a negative eigenvalue λ^*. Let u^* denote a corresponding eigenvector. Then it holds with $z := x^*_{\mathcal{I}(x^*)} + tu^*$ that

$$\begin{aligned} \phi(z) &= \phi(x^*_{\mathcal{I}(x^*)}) + t\nabla\phi(x^*_{\mathcal{I}(x^*)})^\top u^* + \frac{t^2}{2} u^{*\top} \nabla^2 \phi(x^*_{\mathcal{I}(x^*)}) u^* + \mathcal{O}(t^3) \\ &= \phi(x^*_{\mathcal{I}(x^*)}) + \frac{t^2}{2}\lambda^* \|u^*\|^2 + \mathcal{O}(t^3) < \phi(x^*_{\mathcal{I}(x^*)}) \end{aligned}$$

for t sufficiently small. Note that the latter represents a contradiction to the optimality of x^*. Thus, $\nabla^2 \phi(x^*_{\mathcal{I}(x^*)})$ is positive-semidefinite. □

Let $P : \mathbb{R}^n \to X$ denote the projection onto X, which is given by

$$P(x)_i = \begin{cases} L_i & \text{if } x_i < L_i, \\ x_i & \text{if } L_i \leq x_i \leq U_i, \\ U_i & \text{if } x_i > U_i. \end{cases}$$

Furthermore define

$$x(\alpha) = P(x - \alpha \nabla f(x)). \tag{9.3}$$

Obviously it holds for $\mathcal{A}(x^*) = \emptyset$ that $x(\alpha) = P(x^*) = x^*$. Also, we have

$$\|x(\alpha) - x + \alpha \nabla f(x^*)\| \leq \|y - x + \alpha \nabla f(x)\| \quad \forall y \in X.$$

Thus

$$\psi(\lambda) = \frac{1}{2}\|(1-\lambda)x(\alpha) + \lambda y - x + \alpha\nabla f(x)\|^2$$

has a local minimizer in $\lambda = 0$, i.e.:

$$\begin{aligned} 0 \leq \psi'(0) &= ((1-\lambda)x(\alpha) + \lambda y - x + \alpha\nabla f(x))^\top_{|\lambda=0}(y - x(\alpha)) \\ &= (x(\alpha) - x + \alpha\nabla f(x))^\top (y - x(\alpha)). \end{aligned} \tag{9.4}$$

Furthermore, it holds for $y = x$ that

$$\|x(\alpha) - x\|^2 \leq \alpha\nabla f(x)^\top (x - x(\alpha)) \quad \forall\alpha \geq 0. \tag{9.5}$$

With the help of these auxiliary results, we can prove the following variant of the first order necessary conditions.

THEOREM 9.4. *Let f be continuously differentiable on X. A point $x^* \in X$ is stationary for* (9.1) *if and only if*

$$x^* = P(x^* - \alpha\nabla f(x^*)) \quad \forall\alpha \geq 0.$$

PROOF. Let x^* be a stationary point and $x^*(\alpha) = P(x^* - \alpha\nabla f(x^*)) \in X$. (9.5) implies:

$$0 \leq \|x^*(\alpha) - x^*\|^2 \leq \alpha\nabla f(x^*)^\top (x^* - x^*(\alpha)) \; \forall\alpha \geq 0. \tag{9.6}$$

As x^* is stationary, it holds that

$$\nabla f(x^*)^\top (x^* - x^*(\alpha)) \leq 0,$$

whereby, using (9.6), it follows, that $x^* = x^*(\alpha)$ $\forall\alpha \geq 0$. Now let $x^* = x^*(\alpha)$ for all $\alpha \geq 0$. This implies

$$x^* = P(x^* - \alpha\nabla f(x^*)) \;\; \forall\alpha > 0.$$

For $i \in \mathcal{A}(x^*)$ it can be inferred that

$$\begin{aligned} x_i^* = L_i &\Rightarrow \nabla f(x^*)_i \geq 0 \Rightarrow \nabla f(x^*)_i(\xi - x_i^*) \geq 0 \; \forall\xi : L_i \leq \xi \leq U_i, \\ x_i^* = U_i &\Rightarrow \nabla f(x^*)_i \leq 0 \Rightarrow \nabla f(x^*)_i(\xi - x_i^*) \geq 0 \; \forall\xi : L_i \leq \xi \leq U_i. \end{aligned}$$

For $i \in L(x^*)$ it follows: $\nabla f(x^*)_i = 0 \Rightarrow \nabla f(x^*)_i(\xi - x_i^*) = 0 \; \forall\xi : L_i \leq \xi \leq U_i$. Finally we obtain

$$\nabla f(x^*)^\top (x - x^*) \geq 0 \qquad \forall x \in X,$$

thus x^* is a stationary point. □

2. Sufficient conditions

When formulating the sufficient condition, we make use of the notion of a non-degenerate stationary point.

DEFINITION 9.1. *A point $x^* \in X$ is a non-degenerate stationary point for* (9.1)*, if x^* is a stationary point and*

$$\nabla f(x^*)_i \neq 0 \quad \forall i \in \mathcal{A}(x^*).$$

If x^* is a local minimizer of (9.1)*, then x^* is called non-degenerate local minimizer.*

A non-degenerate stationary point also fulfills

$$(x_i = L_i \quad \dot\vee \quad x_i = U_i) \quad \wedge \quad \nabla f(x)_i \neq 0$$

or

$$L_i < x_i < U_i \quad \wedge \quad \nabla f(x)_i = 0.$$

One also refers to this situation as *strict complementarity.*
Let $\mathcal{M}$ be an index set, then we define

$$(P_{\mathcal{M}} x)_i = \begin{cases} x_i & \text{for } i \in \mathcal{M}, \\ 0 & \text{else.} \end{cases}$$

By means of this projection, we obtain the following useful relation.

LEMMA 9.1. *Let x^* be a non-degenerate stationary point. Further let $\mathcal{A}(x^*) \neq \emptyset$. Then there exists a constant $\gamma > 0$, s.t.*

$$\nabla f(x^*)^\top (x - x^*) = \nabla f(x^*)^\top P_{\mathcal{A}(x^*)}(x - x^*) \geq \gamma \|P_{\mathcal{A}(x^*)}(x - x^*)\| \quad \forall x \in X.$$

PROOF. For $i \in \mathcal{A}(x^*)$, non-degeneracy and stationarity of x^* imply that there exists a $\gamma > 0$ such that either

$$x_i^* = L_i \text{ and } \nabla f(x^*)_i \geq \gamma$$

or

$$x_i^* = U_i \text{ and } \nabla f(x^*)_i \leq -\gamma.$$

For $x \in X$ it follows for all $i \in \mathcal{A}(x^*)$ that

$$(\nabla f(x^*))_i (x - x^*)_i \geq \gamma |(x - x^*)_i|.$$

Since $\|\cdot\|_1 \geq \|\cdot\|_2$ and $(\nabla f(x^*))_i = 0$ for $i \in \mathcal{I}(x^*)$, it holds that

$$\sum_i (\nabla f(x^*))_i (x - x^*)_i = \nabla f(x^*)^\top (x - x^*) = \nabla f(x^*)^\top P_{\mathcal{A}(x^*)}(x - x^*) \geq \gamma \|P_{\mathcal{A}(x^*)}(x - x^*)\|.$$

□

Now we can state the sufficient conditions.

THEOREM 9.5. *Let $x^* \in X$ be a non-degenerate stationary point for (9.1). Let f be twice continuously differentiable in a neighborhood of x^*. If the reduced Hessian $\nabla_R^2 f(x^*)$ is positive definite, then x^* is a local minimizer of (9.1).*

PROOF. Let $x \in X$. Define $\phi(\alpha) = f(x^* + \alpha(x - x^*))$. If we can show that either

$$\phi'(0) > 0$$

or

$$\phi'(0) = 0 \text{ and } \phi''(0) > 0,$$

then x^* is a non-degenerate local minimizer of f s.t. $x \in X$. It holds that

$$\phi'(0) = \nabla f(x^*)^\top (x - x^*) = \nabla f(x^*)^\top (P_{\mathcal{A}(x^*)}(x - x^*) + P_{\mathcal{I}(x^*)}(x - x^*)).$$

Since x^* is a stationary point, it follows that $\nabla f(x^*)^\top P_{\mathcal{I}(x^*)}(x - x^*) = 0$. In case $P_{\mathcal{A}(x^*)}(x - x^*) \neq 0$, non-degeneracy of x^* implies

$$\nabla f(x^*)^\top P_{\mathcal{A}(x^*)}(x - x^*) > 0,$$

and thus $\phi'(0) > 0$.

In case $P_{\mathcal{A}(x^*)}(x - x^*) = 0$, then we can deduce

$$\phi''(0) = (x - x^*)^\top P_{\mathcal{I}(x^*)} \nabla^2 f(x^*) P_{\mathcal{I}}(x^*)(x - x^*) = (x - x^*)^\top \nabla_R^2 f(x^*)(x - x^*).$$

Thus, $\phi'(0) = 0$ and $\phi''(0) > 0$. □

3. Projected gradient method

The projected gradient method is a natural extension of steepest descent to box-constrained problems. Hence, it shows similar advantages and disadvantages.
Let x_a denote the current iterate. In the projected gradient approach, the new iterate x_+ is given by

$$x_+ = P(x_a - \alpha \nabla f(x_a)).$$

Here, α is a step size computed e.g. by means of an Armijo step size strategy. For the application of step size strategies, the expected descent has to be specified. Obviously, the corresponding quantities from the unconstrained case are no longer adequate. In the case of an Armijo step size strategy, we now apply the following condition:

$$f(x(\alpha)) - f(x) \leq -\frac{\sigma}{\alpha} \|x - x(\alpha)\|^2, \quad 0 < \sigma < 1, \tag{9.7}$$

with $x(\alpha) = P(x - \alpha \nabla f(x))$.

Algorithm 9.1 (Projected gradient method).
input: $x^0 \in \mathbb{R}^n$, $0 < \sigma < 1$, $0 < \beta < 1$.
begin
  $k := 0$
  while "stopping criterion not fulfilled"
  begin
    find $m \in \mathbb{N}$ *as small as possible such that* (9.7) *is met for* $\alpha_k = \beta^m$.
    $x^{k+1} = x^k(\alpha_k)$
  end
end

The stopping criterion will be specified later. In any case, one should fix an upper bound $k_{\max}$ for the maximum number of iterations and terminate the algorithm if this bound is exceeded. Now we want to focus on the stopping criterion. Evidently, $\|\nabla f(x^k)\| \leq \tau_r \|\nabla f(x^0)\| + \tau_a$ is in general not appropriate. We start by studying the active and inactive sets of adjacent points.

Lemma 9.2. *Let f be twice continuously differentiable on X, and let x^* be a non-degenerate stationary point for* (9.1). *Let $\alpha \in (0, 1]$. Then it holds for x sufficiently close to x^* that*

(1) $\mathcal{A}(x) \subset \mathcal{A}(x^*)$ *and* $x_i = x_i^*$ $\forall i \in \mathcal{A}(x)$.
(2) $\mathcal{A}(x(\alpha)) = \mathcal{A}(x^*)$ *and* $x(\alpha)_i = x_i^*$ $\forall i \in \mathcal{A}(x^*)$.

Proof. Let

$$\delta_1 = \min_{i \in \mathcal{I}(x^*)} \{(U_i - x_i^*), (x_i^* - L_i)\}.$$

If $i \in \mathcal{I}(x^*)$ and $\|x - x^*\| < \delta_1$, then $L_i < x_i < U_i$. Moreover, $\mathcal{I}(x^*) \subset \mathcal{I}(x)$ and thus $\mathcal{A}(x) \subset \mathcal{A}(x^*)$, which proves (1).

Let $\mathcal{A}(\alpha)$ and $\mathcal{I}(\alpha)$ be the active resp. inactive indices for $x(\alpha)$. Let $i \in \mathcal{A}(x^*) \neq \emptyset$. According to Lemma 9.1 and the continuity of ∇f there exists a constant $\delta_2 > 0$ with

$$\|x - x^*\| < \delta_2 \;\Rightarrow\; (\nabla f(x^* + (x - x^*))_i (x - x^*)_i \geq \frac{\sigma}{2}(x - x^*)_i.$$

For

$$\delta_3 < \min(\frac{\sigma}{2}, \delta_2) \wedge \|x - x^*\| < \delta_3$$

it follows: $i \in \mathcal{A}(\alpha) \wedge x(\alpha)_i = x_i^*$. Hence, $\mathcal{A}(x^*) \subset \mathcal{A}(\alpha)$. On the other hand, the definition of P implies

$$\|P(x) - P(y)\| \leq \|x - y\| \; \forall x, y \in \mathbb{R}^n.$$

By continuity of $\nabla^2 f$, ∇f is Lipschitz continuous on X. Let L denote the corresponding Lipschitz constant. We have

$$x^* = x^*(\alpha) = P(x^* - \alpha \nabla f(x^*))$$

and therefore

$$\begin{aligned} \|x^* - x(\alpha)\| &= \|P(x^* - \alpha \nabla f(x^*)) - P(x - \alpha \nabla f(x))\| \\ &\leq \|x^* - x\| + \alpha \|\nabla f(x^*) - \nabla f(x)\| \leq (1 + L\alpha)\|x - x^*\|. \end{aligned} \tag{9.8}$$

If $i \in \mathcal{A}(\alpha) \cap \mathcal{I}(x^*)$, then it holds:

$$\|x^* - x(\alpha)\| \geq \delta_1 = \min_{i \in \mathcal{I}(x^*)} \{(U_i - x_i^*), (x_i^* - L_i)\}. \tag{9.9}$$

If now $\|x - x^*\| < \delta_4 := \min\{\delta_3, \frac{\delta_1}{1+L}\}$, then (9.8) implies, that (9.9) cannot be satisfied. □

Now we can prove the equivalence of $\|x - x^*\|$ und $\|x - x(1)\|$, which will lead to a suitable stopping criterion.

THEOREM 9.6. *Let f be twice continuously differentiable on X, and x^* a non-degenerate stationary point for (9.1). Further assume that the second order necessary condition holds at x^*. Then there exist $\delta > 0$ and $K > 0$ such that for $\|x - x^*\| \leq \delta$ and $\mathcal{A}(x) = \mathcal{A}(x^*)$ it holds that*

$$K^{-1}\|x - x^*\| \leq \|x - x(1)\| \leq K\|x - x^*\|. \tag{9.10}$$

PROOF. We have

$$\begin{aligned} \|x - x(1)\| &= \|x - x^* - (x(1) - x^*(1))\| \\ &\leq \|x - x^*\| + \|P(x - \nabla f(x)) - P(x^* - \nabla f(x^*))\| \\ &\leq 2\|x - x^*\| + \|\nabla f(x) - \nabla f(x^*)\| \leq (2 + L)\|x - x^*\|. \end{aligned}$$

This implies the inequality on the right hand side of (9.10). Choose δ_1 such that $\|x - x^*\| < \delta_1$ implies that Lemma 9.2 holds for $\alpha = 1$. One has

$$(x - x(1))_i = \begin{cases} \nabla f(x)_i & i \in \mathcal{I}(x^*), \\ (x - x^*)_i & i \in \mathcal{A}(x^*). \end{cases}$$

It remains to consider $i \in \mathcal{I}(x^*)$. The sufficient conditions yield the existence of a $\mu > 0$ with

$$u^\top P_{\mathcal{I}(x^*)} \nabla^2 f(x^*) P_{\mathcal{I}(x^*)} u \geq \mu \|P_{\mathcal{I}(x^*)} u\|^2 \quad \forall u \in \mathbb{R}^n.$$

Thus, there exists another constant δ_2 such that for $\|x - x^*\| < \delta_2$:

$$u^\top P_{\mathcal{I}(x^*)} \nabla^2 f(x) P_{\mathcal{I}(x^*)} u \geq \frac{\mu}{2} \|P_{\mathcal{I}(x^*)} u\|^2 \quad \forall u \in \mathbb{R}^n.$$

Since $x - x^* = P_{\mathcal{I}(x^*)}(x - x^*)$, we conclude:

$$\begin{aligned}\|P_{\mathcal{I}(x^*)}(x - x(1))\|^2 &= \int_0^1 (x - x^*)^\top P_{\mathcal{I}(x^*)} \nabla^2 f\left(x^* + t(x - x^*)\right)(x - x^*) dt \\ &= \int_0^1 (x - x^*)^\top P_{\mathcal{I}(x^*)} \nabla^2 f(x^* + t(x - x^*)) P_{\mathcal{I}(x^*)}(x - x^*) dt \\ &\geq \frac{\mu}{2} \|P_{\mathcal{I}(x^*)}(x - x^*)\|^2.\end{aligned}$$

This implies $\|x - x(1)\| \geq \min(1, \frac{\sqrt{2\mu}}{2})\|x - x^*\|$. By choosing

$$K = \max(2 + L, 1, \frac{\sqrt{2\mu}}{2}),$$

the assertion is proven. □

At the end, we obtain the following stopping criterion:

$$\|x^k - x^k(1)\| \leq \tau_r \|x^0 - x^0(1)\| + \tau_a.$$

Convergence analysis.

At first we demonstrate that the Armijo step sizes are bounded away from 0.

THEOREM 9.7. *Let ∇f be Lipschitz continuous with modulus L. Let $x \in X$. Then,* (9.7) *is satisfied for all α with*

$$0 < \alpha \leq \frac{2(1 - \sigma)}{L}.$$

PROOF. Let $y := x - x(\alpha)$. Then it holds that

$$f(x - y) - f(x) = f(x(\alpha)) - f(x) = -\int_0^1 \nabla f(x - \tau y)^\top y d\tau.$$

By definition of y,

$$f(x(\alpha)) = f(x) + \nabla f(x)^\top (x(\alpha) - x) - \int_0^1 (\nabla f(x - \tau y) - \nabla f(x))^\top y d\tau.$$

Thus:

$$\alpha(f(x) - f(x(\alpha))) = \alpha \nabla f(x)^\top (x - x(\alpha)) + \alpha \int_0^1 (\nabla f(x - \tau y) - \nabla f(x)^\top y d\tau.$$

It holds that

$$\| \int_0^1 (\nabla f(x - \tau y) - \nabla f(x))^\top y d\tau \| \leq \frac{L}{2} \|x - x(\alpha)\|^2,$$

which implies

$$\alpha(f(x) - f(x(\alpha))) \geq \alpha \nabla f(x)^\top (x - x(\alpha)) - \frac{\alpha L}{2} \|x - x(\alpha)\|^2.$$

Using (9.5), we infer

$$\alpha(f(x) - f(x(\alpha))) \geq (1 - \frac{\alpha L}{2}) \|x - x(\alpha)\|^2.$$

Now,

$$f(x(\alpha)) - f(x) \leq (\frac{L}{2} - \frac{1}{\alpha}) \|x - x(\alpha)\|^2,$$

where

$$\frac{L}{2} - \frac{1}{\alpha} \leq -\frac{\sigma}{\alpha} \quad \Leftrightarrow \quad \alpha \leq \frac{2(1-\sigma)}{L}.$$

□

According to this, the step size strategy terminates successfully, if

$$\beta^m \leq \frac{2(1-\sigma)}{L} < \beta^{m-1}.$$

Furthermore, we observe that

$$\underline{\alpha} = \frac{2\beta(1-\sigma)}{L} > 0$$

is a (uniform) lower bound on the step sizes α_k.
For the projected gradient method, the following convergence result can now be shown:

THEOREM 9.8. *Let ∇f be Lipschitz-continuous with modulus L. Let $\{x^k\}$ be generated by algorithm 9.1. Then every accumulation point of $\{x^k\}$ is a stationary point for* (9.1).

PROOF. Owing to the Armijo step size strategy, $\{f(x^k)\}$ is monotonically decreasing. In addition, $\{f(x^k)\}$ is bounded from below on X. Hence there exists a limit point $f^* \in \mathbb{R}$. Conditions (9.7) and (9.10) imply

$$\|x^k - x^{k+1}\|^2 \leq \frac{\alpha}{\sigma}\left(f(x^k) - f(x^{k+1})\right) \leq \frac{1}{\sigma}\left(f(x^k) - f(x^{k+1})\right) \overset{k\to\infty}{\longrightarrow} 0.$$

For all $y \in X$ it holds that

$$\begin{aligned}\nabla f(x^k)^\top (x^k - y) &= \nabla f(x^k)^\top (x^{k+1} - y) + \nabla f(x^k)^\top (x^k - x^{k+1}) \\ &\overset{(9.4)}{\leq} \frac{1}{\alpha^k}(x^k - x^{k+1})^\top (x^{k+1} - y) + \nabla f(x^k)^\top (x^k - x^{k+1})\end{aligned}$$

and

$$\begin{aligned}\nabla f(x^k)^\top (x^k - y) &\leq \|x^k - x^{k+1}\| \left(\frac{1}{\alpha^k}\|x^{k+1} - y\| + \|\nabla f(x^k)\|\right) \\ &\leq \|x^k - x^{k+1}\| \left(\frac{1}{\underline{\alpha}}\|x^{k+1} - y\| + \|\nabla f(x^k)\|\right).\end{aligned} \tag{9.11}$$

Let $\{x^{k(l)}\}$ be a subsequence converging to x^*, then (9.11) implies

$$\nabla f(x^*)^\top (x^* - y) \leq 0 \quad \forall y \in X.$$

□

The projected gradient method has the interesting property of identifying the active set after finitely many steps. At this point, non-degeneracy of the local minimizer represents an essential assumption.

THEOREM 9.9. *Let ∇f be Lipschitz continuous. If $\{x^k\}$ converges to a non-degenerate local minimizer of* (9.1)*, then there exists an index $k_0 \in \mathbb{N}$ such that $\mathcal{A}(x^k) = \mathcal{A}(x^*)$ for all $k \geq k_0$.*

PROOF. Choose $\underline{\alpha}$ sufficiently small such that Lemma 9.2 holds. By choosing k_0 ensuring

$$\|x^k - x^*\| < \delta_4 \quad \forall k \geq k_0 - 1,$$

where δ_4 denotes the constant from the proof of Lemma 9.2, the assertion is proven. □

4. Superlinearly convergent methods

The theory presented in the preceding section cannot be extended to iterations of the form

$$x_+ = P(x_a - \alpha H_a^{-1}\nabla f(x_a))$$

with positive definite $H_a \in \mathcal{S}^n$. This can be shown by means of a rather simple counter example. For example, it can easily happen that $x_+ = x_a$ for all $\alpha \geq 0$, although x_a is not a local minimizer.

A possible remedy consists of the introduction of the *ϵ-active set*

$$\mathcal{A}^\epsilon(x) = \{i : U_i - \epsilon \leq x_i \ \vee \ x_i \leq L_i + \epsilon\},$$

with $0 \leq \epsilon < \min\{\frac{1}{2}(U_i - L_i) : i = 1, \ldots, n\} =: \bar{\epsilon}$. By $\mathcal{I}^\epsilon(x)$ we denote the complement of $\mathcal{A}^\epsilon(x)$. The magnitude ϵ may well be varied depending on x_a. Then we write ϵ_a.

As a model for the reduced Hessian, we use

$$\begin{aligned} R(x_a, \epsilon_a, H_a) &= P_{\mathcal{A}^{\epsilon_a}(x_a)} I P_{\mathcal{A}^{\epsilon_a}(x_a)} + P_{\mathcal{I}^{\epsilon_a}(x_a)} H_a P_{\mathcal{I}^{\epsilon_a}(x_a)} \\ &= \begin{cases} \delta_{ij} & \text{if } i \in \mathcal{A}^{\epsilon_a}(x_a) \text{ or } j \in \mathcal{A}^{\epsilon_a}(x_a), \\ (H_a)_{ij} & \text{else.} \end{cases} \end{aligned}$$

It holds: $\nabla^2_R f(x_a) = R(x_a, 0, H_a)$. For $0 \leq \epsilon < \bar{\epsilon}$ and positive definite $H_a \in \mathcal{S}^n$, we define

$$x^{H,\epsilon}(\alpha) = P(x - \alpha R(x, \epsilon, H_a)^{-1}\nabla f(x_a)).$$

In view of the step size strategy, the following lemma proves very useful.

LEMMA 9.3. *Let $x \in X$, $0 \leq \epsilon < \bar{\epsilon}$ and $H_a \in \mathcal{S}^n$ positive definite. Further let ∇f Lipschitz continuous on X with modulus L. Then there exists $\alpha^{H,\epsilon} > 0$ such that*

$$f(x^{H,\epsilon}(\alpha)) - f(x) \leq -\sigma \nabla f(x)^\top (x - x^{H,\epsilon}(\alpha)) \quad \forall \alpha \in [0, \alpha^{H,\epsilon}]. \tag{9.12}$$

PROOF. It holds that

$$\nabla f(x)^\top (x - x^{H,\epsilon}(\alpha)) = \left(P_{\mathcal{A}^\epsilon(x)} \nabla f(x)\right)^\top (x - x^{H,\epsilon}(\alpha)) + \left(P_{\mathcal{I}^\epsilon(x)} \nabla f(x)\right)^\top (x - x^{H,\epsilon}(\alpha)).$$

We have $(x^{H,\epsilon}(\alpha))_i = (x(\alpha))_i$ for all $i \in \mathcal{A}^\epsilon(x)$. Consider α with

$$\alpha < \hat{\alpha}_1 = \frac{\bar{\epsilon}}{\max\limits_{x \in X} \|\nabla f(x)\|_\infty}. \tag{9.13}$$

Note that $\mathcal{A}(x) \subset \mathcal{A}^\epsilon(x)$. Thus: $\mathcal{A}^\epsilon(x) = \mathcal{A}(x) \cup (\mathcal{I}(x) \cap \mathcal{A}^\epsilon(x))$. For $i \in \mathcal{A}(x)$, (9.13) implies either

$$(x - x(\alpha))_i = \alpha \nabla f(x)_i$$

or

$$(x - x(\alpha))_i = 0.$$

In both cases it holds that

$$(x - x(\alpha))_i \nabla f(x)_i \geq 0.$$

For $i \in \mathcal{I}(x) \cap \mathcal{A}^\epsilon(x)$ and $(x - x(\alpha))_i \neq \alpha \nabla f(x)_i$, we have: $i \in \mathcal{A}(x(\alpha))$ and consequently $(x - x(\alpha))_i \nabla f(x)_i \geq 0$. Altogether, we obtain

$$(P_{\mathcal{A}^\epsilon(x)} \nabla f(x))^\top (x - x(\alpha)) \geq 0. \tag{9.14}$$

Now consider α with

$$\alpha \leq \overline{\alpha}_2 = \frac{\epsilon}{\max\limits_{x \in X} \|R(x, \epsilon, H_a)^{-1} \nabla f(x)\|_\infty}.$$

Then i is inactive for $x^{H,\epsilon}(\alpha)$ and $x(\alpha)$. Thus it follows:

$$\begin{aligned}\left(P_{\mathcal{I}^\epsilon(x)}\nabla f(x)\right)^\top (x - x^{H,\epsilon}(\alpha)) &= \alpha \left(P_{\mathcal{I}^\epsilon(x)}\nabla f(x)\right)^\top H_a^{-1} P_{\mathcal{I}^\epsilon(x)}\nabla f(x) \\ &\geq \frac{1}{\lambda_{\min}\alpha}\|P_{\mathcal{I}^\epsilon(x)}(x - x(\alpha))\|^2 \\ &= \frac{1}{\lambda_{\min}} \left(P_{\mathcal{I}^\epsilon(x)}\nabla f(x)\right)^\top (x - x(\alpha)),\end{aligned}$$

where $\lambda_{\min} > 0$ denotes the smallest eigenvalue of H. Moreover,

$$\begin{aligned}\nabla f(x)^\top (x - x^{H,\epsilon}(\alpha)) &= \left(P_{\mathcal{A}^\epsilon(x)}\nabla f(x)\right)^\top (x - x^{H,\epsilon}(\alpha)) + \left(P_{\mathcal{I}^\epsilon(x)}\nabla f(x)\right)^\top (x - x^{H,\epsilon}(\alpha)) \\ &\geq (P_{\mathcal{A}^\epsilon(x)}\nabla f(x))^\top (x - x(\alpha)) + \frac{1}{\lambda_{\min}}(P_{\mathcal{I}^\epsilon(x)}\nabla f(x))^\top (x - x(\alpha)) \\ &\geq \min(1, \frac{1}{\lambda_{\min}})\nabla f(x)^\top (x - x(\alpha)) \\ &\overset{(9.5)}{\geq} \frac{\min(1, \frac{1}{\lambda_{\min}})}{\alpha}\|x - x(\alpha)\|^2.\end{aligned}$$

Now,

$$f(x^{H,\epsilon}(\alpha)) - f(x) \leq -\nabla f(x)^\top (x - x^{H,\epsilon}(\alpha)) + L\|x - x^{H,\epsilon}(\alpha)\|^2,$$

and finally

$$f(x^{H,\epsilon}(\alpha)) - f(x) \leq -\left(1 - L\ \alpha\ \max(1, \lambda_{min})\right) \nabla f(x)^\top (x - x^{H,\epsilon}(\alpha)).$$

Thus (9.12) holds for

$$\alpha \leq \overline{\alpha}_3 = \frac{1 - \sigma}{L \max(1, \lambda_{\min})}.$$

Choose $\alpha^{H,\epsilon} = \min(\bar{\alpha_1}, \bar{\alpha_2}, \bar{\alpha_3})$. □

The method which realizes these ideas is referred to as *scaled projected gradient method.*

Algorithm 9.2 (Scaled projected gradient method).
input: $x^0 \in \mathbb{R}^n$, $0 < \sigma < 1$, $0 < \beta < 1$.
begin
$k := 0$
while $\|x^k - x^k(1)\| > \tau_r\|x^0 - x^0(1)\| + \tau_a$
begin
determine $\epsilon_k, H^k \in \mathcal{S}^n$ *positive definite*
solve $R(x^k, \epsilon_k, H^k)d = -\nabla f(x^k)$
find $m \in \mathbb{N}$ *preferably small such that* (9.7) *holds for* $\alpha_k = \beta^m$.
$x^{k+1} = x^k(\alpha_k)$
end
end

We state the following global convergence result.

THEOREM 9.10. *Let ∇f be Lipschitz continuous with modulus L. Let $\{H^k\} \subset \mathcal{S}^n$ uniformly positive definite and bounded. Further we suppose the existence of $\underline{\epsilon}, \overline{\epsilon} > 0$ such that $\underline{\epsilon} \leq \epsilon_k \leq \overline{\epsilon}$ for all k. Then it holds:*

$$\lim_{k\to\infty} \|x^k - x^k(1)\| = 0,$$

i.e. every accumulation point of $\{x^k\}$ is a stationary point for (9.1).

Furthermore we can deduce that for subsequences $\{x^{k(l)}\}$ with $x^{k(l)} \to x^*$, it holds that $x^* = x^*(1)$. If x^* is non-degenerate, then we have $\mathcal{A}(x^k) = \mathcal{A}(x^*)$ for all sufficiently large k.

4.1. Projected Newton method. If x^0 is located sufficently close to a non-degenerate local minimizer of (9.1) and $H^k = \nabla^2_R f(x^k)$, then one refers to the iteration rule

$$x^{k+1} = P(x^k - \alpha(H^k)^{-1}\nabla f(x^k))$$

as the *projected Newton method*. Since x^0 is sufficiently close to x^*, $\alpha = 1$ is always accepted. If ϵ_k is chosen according to

$$\epsilon_k = \min(\|x^k - x^k(1)\|, \bar{\epsilon}),$$

then the projected Newton method converges locally Q-quadratically to x^*.

THEOREM 9.11. *Let x^* be a non-degenerate local minimizer of* (9.1). *If x^0 is sufficiently close to x^*, $\mathcal{A}(x^0) = \mathcal{A}(x^*)$ and $\epsilon_k = \min(\|x^k - x^k(1)\|, \bar{\epsilon})$, then the projected Newton method converges Q-quadratically to x^*.*

PROOF. By assumption, we have: $\mathcal{A}(x_a) = \mathcal{A}(x_+) = \mathcal{A}(x^*)$, hence

$$P_{\mathcal{A}(x_a)}(x_a - x^*) = P_{\mathcal{A}(x_+)}(x_+ - x^*) = 0.$$

Let $\delta^* = \min\limits_{i\in\mathcal{I}(x^*)} (|x_i - U_i|, |x_i - L_i|) > 0$. Consider $x \in \mathbb{R}^n$ with $\|x - x^*\| \leq \frac{\delta^*}{K}$, with the constant K from Theorem 9.5. Theorem 9.5 implies $\epsilon_a < \delta^*$ and $\|x_a - x^*\| \leq \delta^*$. For $i \in \mathcal{A}^{\epsilon_a}(x_a)$, it holds that $i \in \mathcal{A}(x_a) = \mathcal{A}(x^*)$ and thus

$$\mathcal{A}^{\epsilon_a}(x_a) = \mathcal{A}(x_a) = \mathcal{A}(x^*).$$

Consequently,

$$R(x_a, \epsilon_a, \nabla^2_R f(x_a)) = \nabla^2_R f(x_a).$$

For $\|x_a - x^*\|$ sufficiently small, we obtain:

$$x_+ = P(x_a - (\nabla^2_R f(x_a))^{-1}\nabla f(x_a)).$$

We have

$$\nabla f(x_a) = \nabla f(x^*) + \nabla^2 f(x_a)(x_a - x^*) + E_1$$

with

$$E_1 = \int_0^1 \left(\nabla^2 f(x^* + t(x_a - x^*)) - \nabla^2 f(x_a)\right)(x_a - x^*)dt.$$

Hence $\|E_1\| \leq K_1\|x_a - x^*\|^2$ for $K_1 > 0$. The necessary condition yields

$$P_{\mathcal{I}(x)}\nabla f(x^*) = P_{\mathcal{I}(x^*)}\nabla f(x^*) = 0$$

for x sufficiently close to x^*. Since $\mathcal{I}(x_a) = \mathcal{I}(x^*)$, it follows that

$$x_a - x^* = P_{\mathcal{I}(x_a)}(x_a - x^*) \ \wedge \ P_{\mathcal{A}(x_a)}(x_a - x^*) = 0.$$

Thus:

$$\begin{aligned} P_{\mathcal{I}(x_a)}\nabla f(x_a) &= P_{\mathcal{I}(x_a)}\nabla^2 f(x_a)P_{\mathcal{I}(x_a)}(x_a - x^*) + P_{\mathcal{I}(x_a)}E_1 \\ &= P_{\mathcal{A}(x_a)}(x_a - x^*) + P_{\mathcal{I}(x_a)}\nabla^2 f(x_a)P_{\mathcal{I}(x_a)}(x_a - x^*) + P_{\mathcal{I}(x_a)}E_1 \\ &= \nabla^2_R f(x_a)(x_a - x^*) + P_{\mathcal{I}(x_a)}E_1. \end{aligned}$$

By definition of $\nabla^2_R f$, we have

$$P_{\mathcal{I}(x_a)}(\nabla^2_R f(x_a))^{-1}\nabla f(x_a) = x_a - x^* + E_2$$

with $\|E_2\| \leq K_2\|x_a - x^*\|^2$, $K_2 > 0$. As $P_{\mathcal{I}(x_a)}(P(w)) = P(P_{\mathcal{I}(x_a)}w)$ for all $w \in \mathbb{R}^n$, it holds

$$\begin{aligned} P_{\mathcal{I}(x_a)}x_+ &= P_{\mathcal{I}(x_a)}P(x_a - (\nabla^2_R f(x_a))^{-1}\nabla f(x_a)) \\ &= P(P_{\mathcal{I}(x_a)}(x_a - (\nabla^2_R f(x_a))^{-1}\nabla f(x_a))) \\ &= P(x^* - E_2). \end{aligned}$$

Thus $\|x_+ - x^*\| \leq K_2\|x_a - x^*\|^2$. □

REMARK 9.1. It should be mentioned that there exist projected variants of the BFGS method. In order to take account of the box-constraints, the update formula has to be slightly modified by means of the projections $P_{\mathcal{I}(x)}$ and $P_{\mathcal{A}(x)}$. Moreover, under the assumptions of Theorem 9.11 and a sufficiently good initial approximation of the reduced Hessian, local superlinear convergence of the projected BFGS method can be proven.

Bibliography

[1] D. Bertsekas, *Nonlinear Programming*, Athena Scientific Publisher, Belmont, Massachusetts, 1995.

[2] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, C. Sagastizábal, *Optimisation Numérique*, Mathématiques & Applications 27, Springer-Verlag, Berlin, 1997.

[3] A. R. Conn, N. I. M. Gould, P. L. Toint, *Trust-Region Methods*, SIAM, Philadelphia, 2000.

[4] J. E. Dennis, R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM Philadelphia, 1996.

[5] R. Fletcher, *Practical Methods of Optimization I + II*, Wiley & Sons Publisher, New York, 1980.

[6] C. Geiger, C. Kanzow, *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*, Springer-Verlag, Berlin, 1999.

[7] P. E. Gill, W. Murray, M. Wright, *Practical Optimization*, Academic Press, San Diego, 1981.

[8] F. Jarre, J. Stoer, *Optimierung*, Springer-Verlag, Berlin, 2004.

[9] C. T. Kelley, *Iterative Methods for Optimization*, Frontiers in Applied Mathematics, SIAM, Philadelphia, 1999.

[10] P. Spellucci, *Numerische Verfahren der nichtlinearen Optimierung*, Birkhäuser-Verlag, Basel, 1993.