

Nonlinear Optimization

Lecture Notes
Spring Semester 2023

Andreas Sommer

(date of build: 29th April 2023)

Preface

These lecture notes have been written for the course Nonlinear Optimization (Spring 2023).

The notes are based on the textbooks listed below and the lecture script *Nonlinear Optimization (Spring 2018)* of Prof. Claudia Schillings (U Mannheim).

Please send corrections (including minor typos) to ansommer@mail.uni-mannheim.de

Literature:

- J. Nocedal and S. J. Wright, Numerical Optimization, Springer-Verlag, Berlin, 2006.
- D. Bertsekas, Nonlinear Programming, Athena Scientific Publisher, Belmont, Massachusetts, 1995.
- A. R. Conn, N. I. M. Gould, P. L. Toint, Trust-Region Methods, SIAM, Philadelphia, 2000.
- J. E. Dennis, R. B. Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations, SIAM Philadelphia, 1996.
- R. Fletcher, Practical Methods of Optimization, Wiley & Sons Publisher, New York, 1980.
- C. Geiger, C. Kanzow, Numerische Verfahren zur Loesung unrestringierter Optimierungsaufgaben, Springer-Verlag, Berlin, 1999.
- F. Jarre, J. Stoer: Optimierung, Springer Verlag.
- C. T. Kelley, Iterative Methods for Optimization, Frontiers in Applied Mathematics, SIAM, Philadelphia, 1999.
- M. Ulbrich, S. Ulbrich, Nichtlineare Optimierung, Birkhäuser, 2012.

Contents

1	Introduction	5
2	Optimality conditions for unconstrained optimization problems	7
3	Convex functions	10
4	Gradient based methods	14
4.1	The method of steepest descent	15
4.2	Armijo step size strategy	16
4.3	Convergence of the method of steepest descent	16
5	General Descent Methods	20
5.1	Admissible Descent Direction	20
5.2	Admissible Step Sizes	21
5.3	Globally convergent descent methods	22
6	Step Size Strategies and Algorithms	23
6.1	Armijo rule	23
6.2	Powell-Wolfe step size strategy	24
7	Newton's Method	26
8	Newton-like methods	32
9	Inexact Newton methods	34
10	Quasi-Newton Methods	36
11	Optimality Conditions for Constrained Optimization Problems	42
12	Penalty Method	48
12.1	The Quadratic Penalty Method	49
13	Sequential Quadratic Programming	51
14	Quadratic Optimization Problems	57

The Greek Alphabet

α , A	alpha	ι , I	iota	ρ , P	rho
β , B	beta	κ , K	kappa	σ , Σ	sigma
γ , Γ	gamma	λ , Λ	lambda	τ , T	tau
δ , Δ	delta	μ , M	my	υ , Υ	upsilon
ε , E	epsilon	ν , N	ny	φ , Φ	phi
ζ , Z	zeta	ξ , Ξ	xi	χ , X	chi
η , H	eta	o , O	omikron	ψ , Ψ	psi
θ , Θ	theta	π , Π	pi	ω , Ω	omega

1. Introduction

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

Let $X \subseteq \mathbb{R}^n$ be an arbitrary set and $f : X \rightarrow \mathbb{R}$ a continuous function. The problem is to find an $\tilde{x} \in X$ such that

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

Using a more compact notation we write

$$\min_{x \in X} f(x) \tag{1.1}$$

or

$$\min f(x) \quad \text{s.t.} \quad x \in X, \tag{1.2}$$

where 's.t.' stands for 'subject to'. If $X = \mathbb{R}^n$, then (1.1) is called **unconstrained**, otherwise **constrained**. In general f is called **objective function** and $X \subseteq \mathbb{R}^n$ **feasible set**.

Definition 1.1 The point $x \in \mathbb{R}^n$ is called **feasible** for (1.1), if $x \in X$.

If $X \neq \mathbb{R}^n$, the feasible set can often be described in the form

$$X = \{x \in \mathbb{R}^n : h(x) = 0, g(x) \leq 0\}$$

with continuous functions $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$. We then write

$$\min f(x) \quad \text{s.t.} \quad h(x) = 0, g(x) \leq 0. \tag{1.3}$$

Constrained optimization problems will be discussed in the second part of the course.

Example 1.2 **Quadratic optimization problems** are important examples of nonlinear problems:

- $f(x) = \frac{1}{2}x^\top Qx + b^\top x + c$ with $Q \in \mathbb{R}^{n \times n}$ (symmetric), $b \in \mathbb{R}^n$, $c \in \mathbb{R}$,
- g, h are linear, i.e. $g(x) = Ax - a$ with $A \in \mathbb{R}^{m \times n}$, $a \in \mathbb{R}^m$
and $h(x) = Dx - d$ with $D \in \mathbb{R}^{p \times n}$, $d \in \mathbb{R}^p$.

△

In this course, we can restrict ourselves to minimization problems, since maximization problems with objective function \tilde{f} and feasible set X are equivalent to minimization problems (1.1) with $f = -\tilde{f}$, i.e.

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

Definition 1.3 The feasible point $\tilde{x} \in X$ is called

- **local minimizer** of (1.1), if there exists $\varepsilon > 0$ such that

$$f(\tilde{x}) \leq f(x) \quad \forall x \in X \cap B_\varepsilon(\tilde{x}),$$

where

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

- **strict local minimizer** of (1.1), if there exists $\varepsilon > 0$ such that

$$f(\tilde{x}) < f(x) \quad \forall x \in (X \cap B_\varepsilon(\tilde{x})) \setminus \{\tilde{x}\}.$$

- **global minimizer** of (1.1), if

$$f(\tilde{x}) \leq f(x) \quad \forall x \in X.$$

- **strict global minimizer** of (1.1), if

$$f(\tilde{x}) < f(x) \quad \forall x \in X \setminus \{\tilde{x}\}.$$

Remark We denote by $\|\cdot\|$ the Euclidean norm $\|\cdot\|_2$. △

The Weierstrass Extreme Value Theorem ensures the existence of solutions:

Theorem 1.4 Let $X \subseteq \mathbb{R}^n$. If the function $f : X \rightarrow \mathbb{R}$ is continuous and there exists $\tilde{x} \in X$, such that the **level set**

$$N_f(\tilde{x}) = \{x \in X : f(x) \leq f(\tilde{x})\}$$

is compact, then there exists a global minimizer of (1.1).

2. Optimality conditions for unconstrained optimization problems

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

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 (First order necessary optimality condition)

Let $X \subseteq \mathbb{R}^n$ be an open set and $f : X \rightarrow \mathbb{R}$ a continuously differentiable function. If $\tilde{x} \in X$ is a local minimizer of f on X , then

$$\nabla f(\tilde{x}) = 0,$$

i.e. \tilde{x} is a **stationary point**.

Proof. We prove the statement by means of contradiction. Let us assume that $\tilde{x} \in X$ is a local minimizer for which $\nabla f(\tilde{x}) \neq 0$. Then there exists $d \in \mathbb{R}^n$ with

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

(for example $d = -\nabla f(\tilde{x})$). By assumption, f is continuously differentiable. Consequently, the directional derivative of f in \tilde{x} in direction d exists

$$f'(\tilde{x}; d) = \lim_{t \rightarrow 0^+} \frac{f(\tilde{x} + td) - f(\tilde{x})}{t} = \nabla f(\tilde{x})^\top d < 0.$$

Due to the continuity of the derivative there exists $\bar{t} > 0$ with $\tilde{x} + td \in X$ and

$$\frac{f(\tilde{x} + td) - f(\tilde{x})}{t} < 0, \quad \forall t \in (0, \bar{t}].$$

Therefore, it holds that

$$f(\tilde{x} + td) - f(\tilde{x}) < 0, \quad \forall t \in (0, \bar{t}].$$

This is a contradiction to the assumption that \tilde{x} is a local minimizer of f in X . \square

Note: 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 we need the following lemma about the continuity of the smallest eigenvalue of a matrix.

Lemma 2.2 Let S_n be the vector space of symmetric matrices in $\mathbb{R}^{n \times n}$. For $A \in S_n$ let $\lambda(A) \in \mathbb{R}$ be the smallest eigenvalue of A . Then the following statement holds true

$$|\lambda(A) - \lambda(B)| \leq \|A - B\| \quad \forall A, B \in S_n.$$

Remark 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 2.2 and from the continuity of $\nabla^2 f \in \mathbb{R}^{n \times n}$ (the Hessian of f), that $\nabla^2 f$ is positive definite in a neighborhood of \tilde{x} if $\nabla^2 f(\tilde{x})$ is positive definite. An analogous statement holds true, if $\nabla^2 f(\tilde{x})$ is negative definite. \triangle

Theorem 2.3 (Second order necessary optimality condition)

Let $X \subseteq \mathbb{R}^n$ be open and $f : X \rightarrow \mathbb{R}$ be twice continuously differentiable. If $\tilde{x} \in X$ is a local minimizer of f (on X), then

- (i) $\nabla f(\tilde{x}) = 0$ and
- (ii) the Hessian $\nabla^2 f(\tilde{x})$ is positive semidefinite, i.e.

$$d^T \nabla^2 f(\tilde{x}) d \geq 0 \quad \forall d \in \mathbb{R}^n.$$

Proof. By Theorem 2.1, the statement that $\nabla f(\tilde{x}) = 0$ holds true. Therefore, we only have to show the positive semidefiniteness of the Hessian of f at x . Again we prove the statement by means of contradiction. Let us assume that \tilde{x} is a local minimizer of f , but $\nabla^2 f(\tilde{x})$ is not positive semidefinite. Then there exists $d \in \mathbb{R}^n$ such that

$$d^T \nabla^2 f(\tilde{x}) d < 0.$$

Applying Taylor's theorem, we obtain for sufficiently small $t > 0$

$$f(\tilde{x} + td) = f(\tilde{x}) + t \nabla f(\tilde{x})^T d + \frac{t^2}{2} d^T \nabla^2 f(\zeta(t)) d = f(\tilde{x}) + \frac{t^2}{2} d^T \nabla^2 f(\zeta(t)) d$$

with $\zeta(t) = \tilde{x} + \nu_t td \in X$ für ein $\nu_t \in (0, 1)$. By Lemma 2.2 there exists $\bar{t} > 0$, such that

$$d^T \nabla^2 f(\zeta(t)) d < 0 \quad \forall t \in (0, \bar{t}].$$

Hence

$$f(\tilde{x} + td) < f(\tilde{x}) \quad \forall t \in (0, \bar{t}]$$

which contradicts the assumption that \tilde{x} is a local minimizer of f on X . \square

Theorem 2.4 (Second order sufficient optimality conditions)

Let $X \subseteq \mathbb{R}^n$ be open and $f : X \rightarrow \mathbb{R}$ twice continuously differentiable. If

- (i) $\nabla f(\tilde{x}) = 0$ and
- (ii) the Hessian $\nabla^2 f(\tilde{x})$ is positive definite, i.e.

$$d^T \nabla^2 f(\tilde{x}) d > 0 \quad \forall d \in \mathbb{R}^n \setminus \{0\}.$$

then \tilde{x} is a strict local minimizer of f on X .

Proof. Since $\nabla^2 f$ is continuous and positive definite at \tilde{x} , there exists an $\varepsilon > 0$, such that

$$\nabla^2 f(x) > 0 \quad \forall x \in B_\varepsilon(\tilde{x}) = \{x \in \mathbb{R}^n : \|x - \tilde{x}\| < \varepsilon\}.$$

For every $d \in \mathbb{R}^n \setminus \{0\}$ sufficiently close to 0 (i.e. d is “small” in the sense $\|d\| < \varepsilon$), we have $\tilde{x} + d \in X$ and

$$f(\tilde{x} + d) = f(\tilde{x}) + \underbrace{\nabla f(\tilde{x})^\top}_{=0} d + \frac{1}{2} \underbrace{d^\top \nabla^2 f(\tilde{x} + td) d}_{\substack{\in B_\varepsilon(\tilde{x}) \\ >0}} > f(\tilde{x})$$

for a $t \in (0, 1)$, i.e. $f(\tilde{x}) < f(\tilde{x} + d) \forall d \in B_\varepsilon(\tilde{x})$, so \tilde{x} is a local minimizer of f . □

Remark Condition (ii) in Theorem 2.4 is not necessary for the local minimality of \tilde{x} . To some extent there is a ‘gap’ between necessary and sufficient conditions. Given (i) of Theorem 2.4 in the case of an indefinite Hessian $\nabla^2 f(\tilde{x})$, we refer to \tilde{x} as a saddle point. △

3. 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. In the following we will introduce procedures that approximate a complicated nonlinear 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 3.1 A set $X \subseteq \mathbb{R}^n$ is called **convex**, if for all $x, y \in X$ and all $\lambda \in (0, 1)$

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

i.e. the line segment $\overline{x, y}$ lies completely in X .

Definition 3.2 Let $X \subseteq \mathbb{R}^n$ be convex. A function $f : X \rightarrow \mathbb{R}$ is called

- **convex**, if

$$f((1 - \lambda)x + \lambda y) \leq (1 - \lambda)f(x) + \lambda f(y) \quad \forall x, y \in X, \quad \forall \lambda \in (0, 1).$$

- **strictly convex**, if

$$f((1 - \lambda)x + \lambda y) < (1 - \lambda)f(x) + \lambda f(y) \quad \forall x, y \in X \text{ mit } x \neq y, \quad \forall \lambda \in (0, 1).$$

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 .

- **uniformly convex**, if there exists $\mu > 0$ with

$$f((1 - \lambda)x + \lambda y) + \mu\lambda(1 - \lambda)\|y - x\|^2 \leq (1 - \lambda)f(x) + \lambda f(y) \quad \forall x, y \in X, \quad \forall \lambda \in (0, 1).$$

Remark 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! \triangle

Theorem 3.3 Let $X \subseteq \mathbb{R}^n$ open, convex and $f : X \rightarrow \mathbb{R}$ continuously differentiable. Then the following assertions hold true:

1. f is convex (on X) if and only if

$$\nabla f(x)^\top (y - x) \leq f(y) - f(x) \quad \forall x, y \in X.$$

2. f is strictly convex (on X) if and only if

$$\nabla f(x)^\top (y - x) < f(y) - f(x) \quad \forall x, y \in X \text{ mit } x \neq y.$$

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

$$\nabla f(x)^\top (y - x) + \mu \|y - x\|^2 \leq f(y) - f(x) \quad \forall x, y \in X.$$

Proof.  Exercise. □

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 3.4 Let $X \subseteq \mathbb{R}^n$ be an open, convex set and $f : X \rightarrow \mathbb{R}$ twice continuously differentiable. Then the following statements hold true:

1. f is convex (on X) if and only if $\nabla^2 f(x)$ is positive semidefinite for all $x \in X$, i.e.

$$d^\top \nabla^2 f(x) d \geq 0 \quad \forall x \in X, \quad \forall d \in \mathbb{R}^n.$$

2. If $\nabla^2 f(x)$ is positive definite for all $x \in X$, i.e.

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

then f is strictly convex (on X).

3. 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

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

Proof.

1) “ \Rightarrow ”: Since f is convex and twice continuously differentiable, we can apply Taylor’s theorem:

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

for all $y \in X$ sufficiently close to x , and for the remainder holds: $r(y - x) / \|y - x\|^2 \rightarrow 0$ for $y \rightarrow x$. Choose $y = x + td$ for an arbitrary $d \in \mathbb{R}^n$ and $t > 0$ sufficiently small. Then, by continuity of eigenvalues (Lemma 2.2), we get

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

Divide by $t^2/2$ and take the limit $t \rightarrow 0$ to obtain $0 \leq d^\top \nabla^2 f(x) d \quad \forall x \in X, \quad d \in \mathbb{R}^n$.

“ \Leftarrow ”: Since f is continuously differentiable and $\nabla^2 f(x)$ positive semi-definite for all $x \in X$, applying Taylor’s theorem and the mean value theorem, we get

$$f(y) = f(x) + \nabla^\top(y-x) + \frac{1}{2} \int_0^1 (y-x)^\top \nabla^2 f(x+t(y-x))(y-x) dt. \quad (*)$$

Using the positive semi-definiteness of f , we get $f(y) \geq f(x) + \nabla f(x)^\top(y-x) \forall x, y \in X$. The convexity of f then follows by Theorem 3.3.

2) “ \Leftarrow ”: Analogously to the second part of 1).

3) “ \Rightarrow ”: Let f be uniformly convex. Similar to the first part of 1), using the continuity of eigenvalues, we get for $y = x + td$ with $d \in \mathbb{R}^n$ and $t > 0$ sufficiently small:

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

Divide by t^2 and take the limit $t \rightarrow 0$ to get: $\mu \|d\|^2 \leq \frac{1}{2} d^\top \nabla^2 f(x) d$ for arbitrary $d \in \mathbb{R}^n$.

“ \Leftarrow ”: If $\nabla^2 f$ is uniformly positive definite with modulus $\mu > 0$, we use

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

in (*) and Theorem 3.3 delivers the uniform convexity of f . □

Note that the second statement of Theorem 3.4 cannot be reversed in general; consider, e.g., $f(x) = x^4 \in \mathbb{R}$.

Example 3.5 Let $f : \mathbb{R} \rightarrow \mathbb{R}$.

- The function $f(x) = x$ is convex, but not strictly convex.
- The function $f(x) = \exp(x)$ is strictly convex, but not uniformly.
- The function $f(x) = x^2$ is uniformly convex. △

Remark Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a quadratic function, i.e.,

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

with $Q \in S_n, b \in \mathbb{R}^n, c \in \mathbb{R}$. Then the following statements hold true:

- (i) f is convex if and only if Q is positive semidefinite.
- (ii) f is strictly convex $\Leftrightarrow f$ uniformly convex $\Leftrightarrow Q$ positive definite. △

Theorem 3.6 (Convex optimization problems) Let $f : X \rightarrow \mathbb{R}$ be continuous and convex. Further, let $X \subseteq \mathbb{R}^n$ be convex. Consider the optimization problem

$$\min f(x) \quad \text{s.t. } x \in X, \quad (3.1)$$

then the following statements hold true:

1. Every local minimizer of f on X is a global minimizer of f on X .
2. If f is strictly convex, then (3.1) has at most one solution.
3. If X is open, f continuously differentiable and $\tilde{x} \in X$ a stationary point of f , then \tilde{x} is a global minimizer of f on X .

Proof.  Exercise.

□

4. Gradient based methods

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

$$\min f(x), \quad x \in \mathbb{R}^n. \quad (4.1)$$

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 (4.1), in which $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a continuously differentiable function. The fundamental idea of the methods in this chapter is as follows:

1. 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).
2. Starting at x , one proceeds along this direction d as long as the function value of f reduces sufficiently (step size strategy).

Definition 4.1 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $\bar{t} > 0$ such that

$$f(x + td) < f(x)$$

for all $t \in (0, \bar{t}]$. If $\|d\| = 1$, then d is called a **unit direction**.

Lemma 4.2 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable, $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$ with $\nabla f(x)^\top d < 0$. Then, d is a descent direction of f in x .

Proof. W.l.o.g. $\|d\| = 1$. The continuous differentiability of f implies that for the directional derivative of f in x in direction d , it holds true that

$$f'(x; d) = \lim_{t \rightarrow 0^+} \frac{f(x + td) - f(x)}{t} = \nabla f(x)^\top d < 0.$$

Thus,

$$\frac{f(x + td) - f(x)}{t} < 0$$

for all sufficiently small $t > 0$. Thus, d is a descent direction of f in x . \square

Remark The criterion in the previous lemma 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 the criterion does not hold. \triangle

Algorithm (1) General descent method**Input:** starting point $x^0 \in \mathbb{R}^n$

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: STOP, if $\nabla f(x^k) = 0$
- 3: compute a descent direction $d^k \in \mathbb{R}^n$ with $\nabla f(x^k)^\top d^k < 0$
- 4: compute a step size $\sigma_k > 0$ with $f(x^k + \sigma_k d^k) < f(x^k)$
- 5: set $x^{k+1} = x^k + \sigma_k d^k$
- 6: **end for**

Remark The following stopping rules are common in practice (with $\varepsilon > 0$):

- $\|\nabla f(x^k)\| \leq \varepsilon$,
- $|f(x^k) - f(x^{k-1})| \leq \varepsilon$ (for $k \geq 1$), (not suitable, consider the sequence $f(x^0) = 1$, $f(x^k) = -\sum_{j=1}^k j^{-1}$)
- $\|x^k - x^{k-1}\| \leq \varepsilon$ (für $k \geq 1$) (in combination with other stopping criteria).

△

4.1 The method of steepest descent

The aim is to determine a direction d along which f in x decreases the most.**Definition 4.3** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable, $x \in \mathbb{R}^n$ with $\nabla f(x) \neq 0$ and $\tilde{d} \in \mathbb{R}^n$ the solution of

$$\min_{\|d\|=1} \nabla f(x)^\top d. \quad (4.2)$$

Every vector of the form $\hat{d} = \lambda \tilde{d}$ with $\lambda > 0$ is called **direction of steepest descent** of f in x .**Theorem 4.4** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable, $x \in \mathbb{R}^n$ with $\nabla f(x) \neq 0$. Then, the unique solution of problem (4.2) is given by

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

Proof. By Cauchy-Schwarz, it holds for every $d \in \mathbb{R}^n$ with $\|d\| = 1$

$$\nabla f(x)^\top d \geq -|\nabla f(x)^\top d| \geq -\|\nabla f(x)\| \underbrace{\|d\|}_{=1} = -\|\nabla f(x)\|.$$

The choice $d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$ yields $\nabla f(x)^\top d = -\|\nabla f(x)\|$ and thus solves (4.2) (uniquely). □**Remark** The **steepest descent method** uses the negative gradient in the current iterate as descent direction, i.e.

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

Note that d^k is not a unit direction, but the full negative gradient. △

4.2 Armijo step size strategy

The Armijo step size strategy will be used to determine σ_k in the general descent algorithm.

Algorithm (2) Armijo step size strategy

Input: Iterate $x^k \in \mathbb{R}^n$, descent direction $d^k \in \mathbb{R}^n$, parameter $\beta, \gamma \in (0, 1)$

Output: step size $\sigma_k > 0$

- 1: set $\sigma_k = 1$
 - 2: **while** $f(x^k + \sigma_k d^k) - f(x^k) > \sigma_k \gamma \nabla f(x^k)^\top d^k$ **do**
 - 3: set $\sigma_k = \beta \sigma_k$
 - 4: **end while**
-

Theorem 4.5 Let $X \subseteq \mathbb{R}^n$ be open, $f : X \rightarrow \mathbb{R}$ continuously differentiable and $\gamma \in (0, 1)$. If $d \in \mathbb{R}^n$ is a descent direction in $x \in X$, then there exists $\bar{\sigma} > 0$ such that

$$f(x + \sigma d) - f(x) \leq \sigma \gamma \nabla f(x)^\top d \quad \forall \sigma \in [0, \bar{\sigma}]. \quad (4.3)$$

Proof. For $\sigma = 0$, the relation (4.3) is satisfied, since d is a descent direction. So let $\sigma > 0$ be sufficiently small such that $x + \sigma d \in X$.

If $\nabla f(x)^\top d = 0$, relation (4.3) is satisfied, with the same argument. So, it remains to show (4.3) only for the case $\nabla f(x)^\top d < 0$. In that case, it holds:

$$\underbrace{\frac{f(x + \sigma d) - f(x)}{\sigma}}_{\rightarrow \nabla f(x)^\top d} - \gamma \nabla f(x)^\top d \xrightarrow{\sigma \rightarrow 0} \nabla f(x)^\top d - \gamma \nabla f(x)^\top d = \underbrace{(1 - \gamma)}_{>0} \underbrace{\nabla f(x)^\top d}_{<0} < 0.$$

For $\bar{\sigma} > 0$ sufficiently small, we obtain

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

Multiplying with $\sigma > 0$ gives (4.3). □

The termination of Algorithm (2) after finite number of steps follows then directly from Theorem 4.5, since $\beta^l \in (0, \bar{\sigma}]$ with $\beta \in (0, 1)$ for sufficiently large $l \in \mathbb{N}$.

4.3 Convergence of the method of steepest descent

Theorem 4.6 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable. Then either Algorithm (3) terminates after a finite number of iterations returning a stationary point x^k , or the algorithm generates a sequence of iterates (x^k) such that

1. for all k it holds true that $f(x^{k+1}) < f(x^k)$.
2. every accumulation point of (x^k) is a stationary point of f .

Algorithm (3) Method of the steepest descent**Input:** starting point $x^0 \in \mathbb{R}^n$.

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: STOP, if $\nabla f(x^k) = 0$.
- 3: compute $d^k = -\nabla f(x^k)$.
- 4: compute $\sigma_k > 0$ with the Armijo step size strategy (Algorithm (2)).
- 5: set $x^{k+1} = x^k + \sigma_k d^k$.
- 6: **end for**

Proof. If the algorithm terminates after a finite number of steps, there is nothing to prove. We consider the case that Algorithm (3) computes a sequence $(x^k) \subseteq \mathbb{R}^n$ and $(\sigma_k) \subseteq (0, 1]$ where $\nabla f(x^k) \neq 0$. The application of the Armijo step size strategy (and Theorem 4.5) implies

$$f(x^{k+1}) - f(x^k) = f(x^k + \sigma_k d^k) - f(x^k) \leq \sigma_k \gamma \nabla f(x^k)^\top d^k = -\sigma_k \gamma \|\nabla f(x^k)\|^2 < 0,$$

thus the first statement holds true.

Let \tilde{x} be an accumulation point of (x^k) and $(x^k)_K$ the subsequence with limit \tilde{x} .

The monotonicity of the function values $(f(x^k))$ implies convergence to a limit $f^* \in \mathbb{R} \cup \{-\infty\}$. In particular, we have

$$(f(x^k))_K \rightarrow f^*.$$

The continuity of f implies (since $(x^k)_K \rightarrow \tilde{x}$)

$$(f(x^k))_K \rightarrow f(\tilde{x})$$

and thus

$$f(x^k) \rightarrow f^* = f(\tilde{x}).$$

The application of the Armijo step size strategy gives

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

Hence,

$$\sigma_k \|\nabla f(x^k)\|^2 \rightarrow 0. \quad (4.4)$$

We will now show by contradiction that $\nabla f(\tilde{x}) = 0$. Let us therefore assume that $\nabla f(\tilde{x}) \neq 0$. By continuity of ∇f and $(x^k)_K \rightarrow \tilde{x}$, there exists $l \in K$ such that

$$\|\nabla f(x^k)\| \geq \frac{\|\nabla f(\tilde{x})\|}{2} > 0 \quad \forall k \in K, k \geq l.$$

Due to (4.4) it follows that

$$(\sigma_k)_K \rightarrow 0.$$

Thus, there exists $l' \in K$ such that

$$\sigma_k \leq \beta \quad \forall k \in K, k \geq l'.$$

For all these indices, the stopping criterion of the Armijo step size rule has been violated at least once, in particular for $t_k = \beta^{-1} \sigma_k$. Therefore, we obtain

$$f(x^k + t_k d^k) - f(x^k) > \gamma t_k \nabla f(x^k)^\top d^k = -\gamma t_k \|\nabla f(x^k)\|^2 \quad \forall k \in K, k \geq l'.$$

Division by t_k and applying the mean value theorem leads to

$$-\gamma \underbrace{\|\nabla f(x^k)\|^2}_{\rightarrow \|\nabla f(\tilde{x})\|^2} < \frac{f(x^k + t_k d^k) - f(x^k)}{t_k} = \underbrace{\nabla f(x^k + \tau_k d^k)^\top}_{\rightarrow \nabla f(\tilde{x})^\top} \underbrace{d^k}_{\rightarrow -\nabla f(\tilde{x})} \quad \forall k \in K, k \geq l'$$

with $\tau_k \in [0, t_k]$. Since $(\sigma_k)_K$ is a null sequence, $(t_k)_K$ is a null sequence and we obtain in the limit $k \rightarrow \infty$

$$-\gamma \|\nabla f(\tilde{x})\|^2 \leq -\|\nabla f(\tilde{x})\|^2.$$

Since $\gamma \in (0, 1)$, this is a contradiction to $\nabla f(\tilde{x}) \neq 0$ and the second statement follows. \square

Remark Under the assumption that the level set $N_f(x^0)$ is compact, existence of an accumulation point of (x^k) follows. \triangle

For the realization of a numerical method to solve the minimization problem not only the convergence of iterates to a solution (or probably only a stationary point) is of importance, but also 'how fast' this convergence takes place. We will focus on quadratic functions to investigate the rate of convergence of the method of steepest descent:

- The function f is quadratic and strictly convex, i.e. $f(x) = \frac{1}{2}x^\top Qx + b^\top x + c$ with symmetric and positive definite matrix Q .
- The **exact step size** strategy (i.e. exact line search) will be used.

Definition 4.7 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable, $x \in \mathbb{R}^n$, $d \in \mathbb{R}^n$ descent direction of f in x , and the level set $N_f(x)$ be compact. The step size $\sigma_E = \sigma_E(x, d)$ with

$$\sigma_E = \arg \min_{\sigma \geq 0} f(x + \sigma d)$$

is called **exact step size**.

Example 4.8 Let $f(x) = \frac{1}{2}x^\top Qx + b^\top x$ be a quadratic function with Q symmetric, positive definite. The exact step size for a descent direction d is given by

$$\sigma_E = -\frac{\nabla f(x)^\top d}{d^\top Q d}. \quad \text{✎ exercise} \quad \triangle$$

Theorem 4.9 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be strictly convex and quadratic, i.e. $f(x) = \frac{1}{2}x^\top Qx + b^\top x + c$ with positive definite and symmetric $Q \in \mathbb{R}^{n \times n}$. Let (x^k) and (σ_k) be the iterates and step sizes, respectively, of the steepest descent algorithm (3) with exact line search.

Then, the following statements hold true:

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

$$\|x^k - \tilde{x}\| \leq \sqrt{\frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}} \left(\frac{\lambda_{\max}(Q) - \lambda_{\min}(Q)}{\lambda_{\max}(Q) + \lambda_{\min}(Q)} \right)^k \|x^0 - \tilde{x}\|,$$

where $\tilde{x} = -Q^{-1}b$ is the global minimizer of f and $\lambda_{\max}(Q)$, $\lambda_{\min}(Q)$ denote the largest and smallest eigenvalue of Q .

Remark

- $\kappa := \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)}$ is called the **spectral condition number** of Q .
- If κ is large, the contraction rate $\frac{\kappa-1}{\kappa+1}$ becomes very close to 1, i.e. slow convergence.
- As a remedy, one may choose instead of the steepest descent direction a scaled version, i.e. choose

$$d^k := -M^{-1}\nabla f(x^k)$$

for a positive definite matrix M . This matrix should be chosen to ensure that

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

and such that $M \cdot d^k = -\nabla f(x^k)$ is easy to solve.

△

Theorem 4.10 (*Zig-zagging theorem*)

Let $(x^k)_k$ be the sequence generated by the steepest descent method with exact line search. Then, for all $k = 0, 1, 2, \dots$, it holds

$$(x^{k+1} - x^k) \perp (x^{k+2} - x^{k+1})$$

Proof. The iterates are $x^{k+1} = x^k - \sigma_k \nabla f(x^k)$ and $x^{k+2} = x^{k+1} - \sigma_{k+1} \nabla f(x^{k+1})$, and we have

$$\langle x^{k+1} - x^k, x^{k+2} - x^{k+1} \rangle = \sigma_k \cdot \sigma_{k+1} \cdot \langle \nabla f(x^k), \nabla f(x^{k+1}) \rangle$$

Using the exact line search, we get

$$\sigma_k = \arg \min_{\sigma \geq 0} \underbrace{f(x^k - \sigma \cdot \nabla f(x^k))}_{=: \phi_k(\sigma)}$$

and thus (by application of the chain rule):

$$\begin{aligned} 0 &= \frac{d}{d\sigma} \phi_k(\sigma_k) \\ &= \langle -\nabla(f(x^k)), \nabla f(x^k - \sigma_k \nabla f(x^k)) \rangle \\ &= \langle -\nabla(f(x^k)), \nabla(f(x^{k+1})) \rangle \quad . \end{aligned}$$

□

5. General Descent Methods

Due to the generally slow convergence of steepest descent method, we return to the choice of search direction and step size strategies in the general gradient method. Without specifying the exact choice of the descent direction nor the conditions on the step size along this direction, we introduce abstract conditions which ensure convergence.

5.1 Admissible Descent Direction

Definition 5.1 A subsequence $(d^k)_K$ of the sequence of descent directions (d^k) generated by algorithm (1) is called **admissible**, if

$$\nabla f(x^k)^T d^k < 0 \quad \forall k \geq 0, \quad (5.1)$$

$$\left(\frac{\nabla f(x^k)^T d^k}{\|d^k\|} \right)_K \rightarrow 0 \quad \implies \quad (\nabla f(x^k))_K \rightarrow 0. \quad (5.2)$$

The first condition (5.1) ensures that all vectors d^k are descent directions. The second condition (5.2) becomes more intuitive when realizing that the expression $\left(\frac{\nabla f(x^k)^T d^k}{\|d^k\|} \right)_K$ is the slope of f at x^k in direction d^k , i.e. if the slopes along the search direction d^k become smaller and smaller then the steepest possible slopes $\|\nabla f(x^k)\|$ have to become smaller and smaller. This can be guaranteed by bounding the angle between the search direction d^k and the negative gradient $-\nabla f(x^k)$, since

$$\frac{|\nabla f(x^k)^T d^k|}{\|d^k\|} = \frac{-\nabla f(x^k)^T d^k}{\underbrace{\|d^k\| \|\nabla f(x^k)\|}_{\cos \angle(-\nabla f(x^k), d^k)}} \|\nabla f(x^k)\|.$$

The **angle condition**

$$\cos \angle(-\nabla f(x^k), d^k) = \frac{-\nabla f(x^k)^T d^k}{\|d^k\| \|\nabla f(x^k)\|} \geq \eta \quad (5.3)$$

for all $k \in K$ with fixed $\eta \in (0, 1)$ (independent of k) implies (5.2).

Another sufficient condition for (5.2) is the **generalized angle condition**

$$\|\nabla f(x^k)\| \leq \Phi \left(\frac{-\nabla f(x^k)^T d^k}{\|d^k\|} \right) \quad (5.4)$$

with a suitable function $\Phi : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ with $\Phi(0) = 0$, where Φ is assumed to be continuous at 0.

Theorem 5.2 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable and $(d^k)_K$ be the subsequence of the descent directions (d^k) generated by (1). Then, the following statement holds true:

$$\begin{aligned} & d^k \text{ satisfies for all } k \in K \text{ the angle condition (5.3)} \\ \Rightarrow & d^k \text{ satisfies for all } k \in K \text{ the generalized angle condition (5.4)} \\ \Rightarrow & (d^k)_K \text{ satisfies the condition (5.2),} \end{aligned}$$

where η and Φ are independent of $k \in K$.

Proof. „(5.3) \Rightarrow (5.4)“: We set $\Phi : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ with $\Phi(t) = \frac{t}{\eta}$. Thus, with (5.3), it holds true that

$$\|\nabla f(x^k)\| \leq \frac{1}{\eta} \frac{-\nabla f(x^k)^\top d^k}{\|d^k\|} = \Phi\left(\frac{-\nabla f(x^k)^\top d^k}{\|d^k\|}\right).$$

„(5.4) \Rightarrow (5.2)“: Since Φ is assumed to be continuous at 0 with $\Phi(0) = 0$, we obtain

$$\left(\frac{\nabla f(x^k)^\top d^k}{\|d^k\|}\right)_K \rightarrow 0 \quad \Longrightarrow \quad \|\nabla f(x^k)\| \leq \Phi\left(\frac{-\nabla f(x^k)^\top d^k}{\|d^k\|}\right) \rightarrow \Phi(0) = 0 \quad (K \ni k \rightarrow \infty).$$


Hence $(\nabla f(x^k))_K \rightarrow 0$ and the assertion follows. \square

Example 5.3 Newton-like methods use descent directions d^k given by the solution of the following linear system

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

If M_k is symmetric and positive definite with

$$0 < \mu_1 \leq \lambda_{\min}(M_k) \leq \lambda_{\max}(M_k) \leq \mu_2 < \infty$$

for all $k \in \mathbb{N}$, then every subsequence $(d^k)_K$ of search directions is admissible.  Exercise. \triangle

5.2 Admissible Step Sizes

Definition 5.4 The subsequence $(\sigma_k)_K$ of the step sizes (σ_k) generated by Algorithm (1) is called **admissible**, if

$$f(x^k + \sigma_k d^k) \leq f(x^k) \quad \forall k \geq 0, \quad (5.5)$$

and

$$f(x^k + \sigma_k d^k) - f(x^k) \rightarrow 0 \quad \Longrightarrow \quad \left(\frac{\nabla f(x^k)^\top d^k}{\|d^k\|}\right)_K \rightarrow 0 \quad (5.6)$$

hold true.


Admissible step sizes are given e.g. by **efficient** step sizes.

Definition 5.5 Let d^k be a descent direction of f at x^k . The step size $\sigma_k > 0$ is called **efficient**, if

$$f(x^k + \sigma_k d^k) \leq f(x^k) - \theta \left(\frac{\nabla f(x^k)^T d^k}{\|d^k\|} \right)^2$$

with $\theta > 0$.

Theorem 5.6 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable. The sequences (x^k) , (d^k) and (σ_k) are generated by Algorithm (1) and (5.5) is assumed to hold true. If $(\sigma_k)_K$ is a subsequence such that the step sizes σ_k with $k \in K$ are efficient, then the subsequence of step sizes $(\sigma_k)_K$ will be admissible.

Proof.  Exercise □

5.3 Globally convergent descent methods

The global convergence of general descent methods with admissible descent directions and admissible step sizes is the topic of the following theorem. 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 $(x^k)_k$ (or a subsequence) to a global minimizer of f !

Theorem 5.7 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable. Assume that Algorithm (1) generates an infinite sequence of iterates (x^k) , (d^k) and (σ_k) . Let \tilde{x} be an accumulation point of (x^k) and $(x^k)_K$ be a subsequence with limit \tilde{x} , such that $(d^k)_K$ and the step sizes $(\sigma_k)_K$ are admissible. Then, \tilde{x} is a stationary point of f .

Proof. As in the proof of Theorem 4.6, the monotonicity of the sequence $(f(x^k))$ implies

$$\lim_{k \rightarrow \infty} f(x^k) = \lim_{K \ni k \rightarrow \infty} f(x^k) = f(\tilde{x}).$$

Thus, it holds: $f(\tilde{x}) - f(x^0) = \lim_{k \rightarrow \infty} f(x^k) - f(x^0) = \sum_{k=0}^{\infty} (f(x^{k+1}) - f(x^k))$

and therefore (due to the convergence of the series on the right hand side), it holds true that

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

The admissibility of the step sizes $(\sigma_k)_K$ gives

$$\left(\frac{\nabla f(x^k)^T d^k}{\|d^k\|} \right)_K \rightarrow 0,$$

and the admissibility of the descent directions $(d^k)_K$ leads to $(\nabla f(x^k))_K \rightarrow 0$.

Due to the continuity of the gradient ∇f , it holds true that

$$\nabla f(\tilde{x}) = \lim_{K \ni k \rightarrow \infty} \nabla f(x^k) = 0.$$

□

6. Step Size Strategies and Algorithms

The general descent method offers quite some freedom in the choice of the descent direction d_k and the step size $\sigma_k > 0$. The exact minimization rule, i.e. $\sigma_k = \sigma_k^E$ with


$$f(x^k + \sigma_k^E d^k) = \min_{\sigma > 0} f(x^k + \sigma d^k),$$

is well-defined, provided that the level set $N_f(x^0)$ is compact and ∇f Lipschitz continuous on $N_f(x^0)$. 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 two important representatives of practicable step size strategies, the Armijo rule and the Powell-Wolfe strategy.

6.1 Armijo rule

In Algorithm (3) we already introduced the Armijo rule (Algorithm (2)). This strategy does not fit directly into the framework of the previous chapter, since this rule does not generate admissible step sizes in general.

Example 6.1 Let $f(x) = \frac{x^2}{8}$, $x^0 = 1$ und $d^k = -2^{-k} \nabla f(x^k)$. The general descent algorithm (1) with Armijo rule generates a sequence of iterates $(x^k)_k$ which does not converge to the global minimum at 0.

 Exercise

△

However, under additional assumptions on the search directions, admissibility of the step sizes generated by the Armijo rule can be shown:

Theorem 6.2 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable and the subsequence $(x^k)_K$ be bounded. Further, denote by $\phi : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ a strictly increasing function, such that the search directions generated by the general descent algorithm satisfy the following condition:

$$\|d^k\| \geq \phi \left(\frac{-\nabla f(x^k)^T d^k}{\|d^k\|} \right) \quad \forall k \in K. \quad (6.1)$$

Then, the step sizes $(\sigma_k)_K$ generated by the Armijo rule are admissible.

6.2 Powell-Wolfe step size strategy

The Powell-Wolfe rule requires in addition to the Armijo condition

$$f(x^k + \sigma_k d^k) - f(x^k) \leq \sigma_k \gamma \nabla f(x^k)^T d^k \quad (6.2)$$

that the following condition

$$\nabla f(x^k + \sigma_k d^k)^T d^k \geq \eta \nabla f(x^k)^T d^k \quad (6.3)$$

holds true with $0 < \gamma < 1/2$ and $\gamma < \eta < 1$.

The second condition ideally implies that the graph of $f(x + \sigma d)$ at $\sigma > 0$ does not descend as 'steeply' as at $\sigma = 0$.

The following theorem gives a sufficient condition to ensure the existence of step sizes σ satisfying (6.2) and (6.3):

Theorem 6.3 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable, $d \in \mathbb{R}^n$ descent direction of f at $x \in \mathbb{R}^n$ and f be bounded from below in direction d , i.e.

$$\inf_{t \geq 0} f(x + td) > -\infty.$$

Then, there exists $\sigma > 0$ for given $\gamma \in (0, \frac{1}{2})$ and $\eta \in (\gamma, 1)$ such that

$$f(x + \sigma d) - f(x) \leq \sigma \gamma \nabla f(x)^T d \quad (6.4)$$

and

$$\nabla f(x + \sigma d)^T d \geq \eta \nabla f(x)^T d. \quad (6.5)$$

Under the assumptions of Theorem 6.3, Algorithm (4) terminates after a finite number of iterations and returns a step size $\sigma > 0$ satisfying (6.4) and (6.5):

Theorem 6.4 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable, $d \in \mathbb{R}^n$ descent direction of f at $x \in \mathbb{R}^n$ and f be bounded from below in direction d , i.e.

$$\inf_{t \geq 0} f(x + td) > -\infty.$$

For given $\gamma \in (0, \frac{1}{2})$ and $\eta \in (\gamma, 1)$, Algorithm (4) terminates after a finite number of iterations. The generated step size $\sigma > 0$ satisfies (6.4) and (6.5).

The admissibility of the step sizes follows:

Theorem 6.5 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable and $x^0 \in \mathbb{R}^n$ a starting point, such that the level set $N_f(x^0)$ is compact. Algorithm (1) employs the Powell-Wolfe rule (Algorithm (4)). Then, the algorithm is well-defined and every subsequence of step sizes is admissible.

Algorithm (4) Powell-Wolfe Step Size Strategy

Input: Iterate $x \in \mathbb{R}^n$, descent direction $d \in \mathbb{R}^n$, parameters $\gamma \in (0, \frac{1}{2})$, $\eta \in (\gamma, 1)$, $\sigma_- := 1$.**Output:** Step size $\sigma > 0$ fulfilling the Wolfe conditions.

- 1: **if** σ_- satisfies the Armijo condition (6.4) **then**
 - 2: **if** σ_- satisfies the curvature condition (6.5) **then**
 - 3: STOP with σ_-
 - 4: **end if**
 - 5: Determine the smallest $\sigma_+ \in \{2^1, 2^2, 2^3, \dots\}$, such that $\sigma = \sigma_+$ does not satisfy the sufficient descent condition (Armijo rule (6.4))
 - 6: Set $\sigma_- = \frac{\sigma_+}{2}$
 - 7: **else**
 - 8: Determine the largest $\sigma_- \in \{2^{-1}, 2^{-2}, 2^{-3}, \dots\}$, such that $\sigma = \sigma_-$ satisfies the sufficient descent condition (Armijo rule (6.4))
 - 9: Set $\sigma_+ = 2\sigma_-$
 - 10: **end if**
 - 11: **while** σ_- does not satisfy the curvature condition (6.5) **do**
 - 12: Set $\sigma = \frac{\sigma_- + \sigma_+}{2}$
 - 13: **if** σ satisfies the Armijo rule (6.4) **then**
 - 14: Set $\sigma_- = \sigma$
 - 15: **else**
 - 16: Set $\sigma_+ = \sigma$
 - 17: **end if**
 - 18: **end while**
 - 19: STOP with step size σ_-
-

7. Newton's Method

Newton's method, with all its variations, is the most important method in unconstrained optimization. Compared to gradient methods, we will achieve a faster convergence by using second order information, the Hessian $\nabla^2 f$, in addition to the first order derivatives. The basic idea is to use Newton's method as an algorithm for the solution of the system of first order necessary conditions $\nabla f(x) = 0$.

Newton's method is a method for the solution of nonlinear systems of the form

$$F(x) = 0$$

with $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$. If the Jacobian matrix of F exists and is continuous, then Taylor's theorem leads to:

$$F(x^k + d) = F(x^k) + F'(x^k)d + o(\|d\|).$$

Hence, given a point x^k , we can determine x^{k+1} setting d^k such that

$$x^{k+1} = x^k + d^k \quad \text{with} \quad F'(x^k)d^k = -F(x^k).$$

With $F(x) := \nabla f(x)$ and thus $F'(x) = \nabla^2 f(x)$, we obtain a Newton method for minimizing f by solving the first order necessary optimality condition $\nabla f(x) = 0$, as in Algorithm (5).

Algorithm (5) Local Newton's Method (for optimization problems)

Input: starting point $x^0 \in \mathbb{R}^n$.

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: STOP, if $\nabla f(x^k) = 0$.
- 3: Compute the Newton step $d^k \in \mathbb{R}^n$ by solving the linear system


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

- 4: Set $x^{k+1} = x^k + d^k$.

- 5: **end for**

Remark An equivalent way to introduce Newton's method for unconstrained optimization is to compute the next iterate x^{k+1} by minimizing a quadratic approximation of f , i.e. we consider the following quadratic model of f near x^k :

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

If $\nabla^2 f(x^k)$ is positive definite, there exists a unique minimizer of the quadratic model yielding the next iterate x^{k+1} .  Exercise.

Note that the Hessian $\nabla^2 f(x)$ is positive definite in a neighborhood $B_\varepsilon(\tilde{x})$ of the local minimizer provided that second order sufficient conditions are satisfied.

△

Definition 7.1 (Convergence rate) The sequence $(x^k) \subset \mathbb{R}^n$

- **converges linearly** with rate $0 < \gamma < 1$ to $\tilde{x} \in \mathbb{R}^n$, iff there exists $l \geq 0$ such that:

$$\|x^{k+1} - \tilde{x}\| \leq \gamma \|x^k - \tilde{x}\| \quad \forall k \geq l.$$

- **converges superlinearly** to $\tilde{x} \in \mathbb{R}^n$, if $x^k \rightarrow \tilde{x}$ and

$$\|x^{k+1} - \tilde{x}\| = o(\|x^k - \tilde{x}\|) \quad \text{for } k \rightarrow \infty,$$

i.e. there exists a null sequence $(\epsilon_k) \subset \mathbb{R}_+$ with

$$\|x^{k+1} - \tilde{x}\| \leq \epsilon_k \|x^k - \tilde{x}\| \quad \text{for all } k \rightarrow \infty$$

- **converges quadratically** to $\tilde{x} \in \mathbb{R}^n$, if $x^k \rightarrow \tilde{x}$ and

$$\|x^{k+1} - \tilde{x}\| = \mathcal{O}(\|x^k - \tilde{x}\|^2) \quad \text{for } k \rightarrow \infty.$$

The above condition is equivalent to the existence of a constant $C > 0$ such that

$$\|x^{k+1} - \tilde{x}\| \leq C \|x^k - \tilde{x}\|^2 \quad \forall k \geq 0.$$

Remark Note that the superlinear and quadratic convergence are independent of the chosen norm, i.e. if $(x^k) \subset \mathbb{R}^n$ converges superlinearly to \tilde{x} according to Definition 7.1, then it holds true that

$$\|x^{k+1} - \tilde{x}\|_a = \eta_k \|x^k - \tilde{x}\|_a \quad \text{for all } k \rightarrow \infty$$

for an arbitrary norm $\|\cdot\|_a$ and null sequence $(\eta_k) \subset \mathbb{R}_+$. Analogously,

$$\|x^{k+1} - \tilde{x}\|_a \leq C_a \|x^k - \tilde{x}\|_a^2 \quad \forall k \geq 0.$$

converges quadratically to \tilde{x} , where C_a depends on the norm.

However, the linear convergence depends on the applied norm. For every linearly convergent sequence $(x^k) \subset \mathbb{R}^n$, it holds true that

$$\|x^{k+1} - \tilde{x}\|_a \leq \gamma_a \|x^k - \tilde{x}\|_a \quad \forall k \geq l$$

with a constant γ_a depending on the applied norm, but the constant is in general not necessarily smaller than 1. \triangle

The convergence proof of Newton's method is based on some auxiliary results, which we establish in the following lemmas 7.2 and 7.3.

Lemmas 7.4 and 7.5 give characterizations of superlinear and quadratic convergence and will be also be used in the convergence proof of the local Newton method.

Lemma 7.2 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $(x^k) \subset \mathbb{R}^n$ be a sequence converging to $\tilde{x} \in \mathbb{R}^n$. Then, the following statements hold true:

1. If f is twice continuously differentiable, then

$$\|\nabla f(x^k) - \nabla f(\tilde{x}) - \nabla^2 f(x^k)(x^k - \tilde{x})\| = o(\|x^k - \tilde{x}\|).$$

2. If f is twice continuously differentiable and $\nabla^2 f$ is locally Lipschitz continuous, then

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

Lemma 7.3 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable, $\tilde{x} \in \mathbb{R}^n$ and $\nabla^2 f(\tilde{x})$ invertible. Then, there exists $\delta > 0$ such that $\nabla^2 f(x)$ is invertible for all $x \in B_\delta(\tilde{x})$. Further, there exists $c > 0$, such that

$$\|\nabla^2 f(x)^{-1}\| \leq c \quad \forall x \in B_\delta(\tilde{x}).$$

Lemma 7.4 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable, $(x^k) \subset \mathbb{R}^n$ be a convergent sequence with limit $\tilde{x} \in \mathbb{R}^n$, $x^k \neq \tilde{x}$ for all $k \in \mathbb{N}$ and $\nabla^2 f(\tilde{x})$ invertible. Then, the following statements are equivalent:

1. $x^k \rightarrow \tilde{x}$ superlinear und $\nabla f(\tilde{x}) = 0$.
2. $\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$.
3. $\|\nabla f(x^k) + \nabla^2 f(\tilde{x})(x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$.

Lemma 7.5 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable and $\nabla^2 f$ be locally Lipschitz continuous, $(x^k) \subset \mathbb{R}^n$ be a convergent sequence with limit $\tilde{x} \in \mathbb{R}^n$, $x^k \neq \tilde{x}$ for all $k \in \mathbb{N}$ and $\nabla^2 f(\tilde{x})$ invertible. Then, the following statements are equivalent:

1. $x^k \rightarrow \tilde{x}$ quadratically and $\nabla f(\tilde{x}) = 0$.
2. $\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\| = O(\|x^{k+1} - x^k\|^2)$.
3. $\|\nabla f(x^k) + \nabla^2 f(\tilde{x})(x^{k+1} - x^k)\| = O(\|x^{k+1} - x^k\|^2)$.

With these auxiliary results, we may now show the convergence of the (local/full-step) Newton method.

Theorem 7.6 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable and $\tilde{x} \in \mathbb{R}^n$ be a stationary point of f and $\nabla^2 f(\tilde{x})$ invertible. Then, there exists $\delta > 0$ such that for all $x \in B_\delta(\tilde{x})$ it holds true that

1. the (local) Newton's method (5) is well defined and generates a convergent sequence (x^k) with limit \tilde{x} .
2. the convergence rate is superlinear.
3. If $\nabla^2 f$ is locally Lipschitz continuous, the convergence rate will be quadratic.

Proof. By Lemma 7.3, there exists $\delta_1 > 0$ such that $\nabla^2 f(x)$ is invertible for all $x \in B_{\delta_1}(\tilde{x})$ with

$$\|\nabla^2 f(x)^{-1}\| \leq c \quad \forall x \in B_{\delta_1}(\tilde{x}).$$

for a constant $c > 0$. Furthermore, by Lemma 7.2, there exists $\delta_2 > 0$ with

$$\|\nabla f(x^k) - \nabla f(\tilde{x}) - \nabla^2 f(x^k)(x^k - \tilde{x})\| \leq \frac{1}{2c} \|x^k - \tilde{x}\|.$$

for all $x \in B_{\delta_2}(\tilde{x})$. We set $\delta = \min\{\delta_1, \delta_2\}$ and choose $x^0 \in B_\delta(\tilde{x})$. Then, x^1 is well defined and we have

$$\begin{aligned} \|x^1 - \tilde{x}\| &= \|x^0 - \tilde{x} - \nabla^2 f(x^0)^{-1} \nabla f(x^0)\| \\ &\leq \|\nabla^2 f(x^0)^{-1}\| \|\nabla f(x^0) - \nabla f(\tilde{x}) - \nabla^2 f(x^0)(x^0 - \tilde{x})\| \\ &\leq c \frac{1}{2c} \|x^0 - \tilde{x}\| = \frac{1}{2} \|x^0 - \tilde{x}\|. \end{aligned}$$

Thus, $x^1 \in B_\delta(\tilde{x})$ and by induction, it follows that


$$\|x^k - \tilde{x}\| \leq \left(\frac{1}{2}\right)^k \|x^0 - \tilde{x}\|$$

for all $k \in \mathbb{N}$. Hence, the sequence of iterates (x^k) is well defined and converges to \tilde{x} . Lemma 7.4, Lemma 7.5 and

$$\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k) = 0$$

imply the statements on the convergence rates. □

Theorem 7.6 states a local convergence results. The following example illustrates that Algorithm (5) does not converge for arbitrary starting points x^0 in general:

Example 7.7 We consider the function $f(x) = \sqrt{x^2 + 1}$. The second order sufficient conditions are satisfied at $\tilde{x} = 0$. However, Algorithm (5) does not converge for starting points x^0 with $|x^0| \geq 1$.  Exercise. △

To design a globally convergent method based on the convergence results for general descent methods (Theorem 5.7), we introduce a step size strategy and ensure that step sizes and descent directions are both admissible.

Algorithm (6) Globalization of Newton's method

Input: Starting point $x^0 \in \mathbb{R}^n$, parameters $\beta \in (0, 1)$, $\gamma \in (0, 1)$, $\alpha_1, \alpha_2 > 0$ and $p > 0$.

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: STOP, if $\nabla f(x^k) = 0$
- 3: Compute Newton direction $d_N^k \in \mathbb{R}^n$ by solving the linear system

$$\nabla^2 f(x^k) d_N^k = -\nabla f(x^k) \quad (*)$$
- 4: **if** $d_N^k \neq 0$ is a solution of $(*)$ that satisfies the generalized angle condition

$$-\nabla f(x^k)^T d_N^k \geq \min\{\alpha_1, \alpha_2 \|d_N^k\|^p\} \|d_N^k\|^2 \quad (7.1)$$
- then**
- 5: $d^k := d_N^k$
- 6: **else**
- 7: $d^k := -\nabla f(x^k)$ *(fallback if $(*)$ not solvable or Newton-step is unsatisfying)*
- 8: **end if**
- 9: Determine the step size $\sigma_k > 0$ using the Armijo rule (Algorithm (2))
- 10: $x^{k+1} := x^k + \sigma_k d^k$
- 11: **end for**

Theorem 7.8 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable. Then, either Algorithm (6) terminates after a finite number of iterations with $\nabla f(x^k) = 0$ or every accumulation point of the sequence of iterates (x^k) generated by the algorithm is a stationary point of f .

Proof. Let $K_G = \{k \geq 0 : d^k = -\nabla f(x^k)\}$ and $K_N = \{k \geq 0 : d^k \neq -\nabla f(x^k)\}$.

To apply the convergence results for general descent methods (Theorem 5.7) and to ensure the well posedness of the Armijo rule, we first show that the directions generated by Algorithm (6) are descent directions. By definition of K_G , it follows directly that

$$\frac{-\nabla f(x^k)^T d^k}{\|d^k\|} = \|\nabla f(x^k)\| > 0, \quad (7.2)$$

for all $k \in K_G$. By (7.1), we obtain

$$\frac{-\nabla f(x^k)^T d^k}{\|d^k\|} \geq \min\{\alpha_1, \alpha_2 \|d^k\|^p\} \|d^k\| > 0. \quad (7.3)$$

for all $k \in K_N$. Hence, the generated directions d^k are descent directions for all $k \in \mathbb{N}_0$. In particular, the Armijo rule is well defined.

If the algorithm terminates after a finite number of steps, there is nothing to prove. Therefore, we focus on the case that the algorithm generates a sequence of iterates (x^k) with accumulation \tilde{x} . We denote by $(x^k)_K$ a subsequence converging to \tilde{x} , i.e. $(x^k)_K \rightarrow \tilde{x}$.

We first prove the admissibility of the subsequence of descent directions $(d^k)_K$: Since $(x^k)_K$ is bounded and $\nabla^2 f$ is continuous by assumption, there exists $C > 0$, such that $\|\nabla^2 f(x^k)\| \leq C$ for all $k \in K$. Thus, we obtain

$$\|\nabla f(x^k)\| = \|\nabla^2 f(x^k) d^k\| \leq C \|d^k\| \quad \forall k \in K \cap K_N. \quad (7.4)$$

To show the admissibility, let

$$\left(\frac{\nabla f(x^k)^T d^k}{\|d^k\|} \right)_K \rightarrow 0.$$

By (7.3), it follows that

$$\|d^k\| \xrightarrow{K \cap K_N \ni k \rightarrow \infty} 0$$

and (7.4) implies

$$\|\nabla f(x^k)\| \xrightarrow{K \cap K_N \ni k \rightarrow \infty} 0.$$

For the indices $k \in K \cap K_G$, it follows by (7.2) that

$$\|\nabla f(x^k)\| = \frac{-\nabla f(x^k)^T d^k}{\|d^k\|} \xrightarrow{K \cap K_G \ni k \rightarrow \infty} 0.$$

Thus,

$$\left(\nabla f(x^k)\right)_K \rightarrow 0$$

implies the admissibility of $(d^k)_K$.

It remains to show that the step sizes $(\sigma_k)_K$ are also admissible: For all $k \in K \cap K_G$, it holds true that

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

and for $k \in K \cap K_N$, it follows that

$$\|d^k\| \stackrel{(7.4)}{\geq} \frac{1}{C} \|\nabla f(x^k)\| \geq \frac{1}{C} \frac{-\nabla f(x^k)^T d^k}{\|d^k\|}.$$

The function $\phi : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ with

$$\phi(t) = \min\left\{t, \frac{1}{C}t\right\}$$

is continuous, strictly monotonically increasing from $\phi(0) = 0$, and we obtain for all $k \in K$

$$\|d^k\| \geq \phi\left(\frac{-\nabla f(x^k)^T d^k}{\|d^k\|}\right).$$

Thus, Theorem 6.2 gives the admissibility of the subsequence of step sizes $(\sigma_k)_K$ and the assertion follows. □

Theorem 7.9 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable. Further, let (x^k) be a sequence of iterates generated by Algorithm (6) with accumulation point \tilde{x} such that the Hessian is positive definite at \tilde{x} . Then, the following statements hold true:

1. The point \tilde{x} is a strict local minimizer of f .
2. The sequence (x^k) converges to \tilde{x} .
3. If $\gamma \in (0, \frac{1}{2})$, then there exists $l \geq 0$, such that the algorithm performs Newton's method with step size 1 for all $k \geq l$. In particular, Algorithm (6) converges superlinearly. If $\nabla^2 f$ is Lipschitz continuous in a neighborhood of \tilde{x} , then the convergence rate will be quadratic.

8. Newton-like methods

In this chapter we discuss some variants of Newton's method. The evaluation and factorization of the Hessian of f can be very expensive. On the other hand, far from a local minimum, the Hessian may be singular or the Newton direction may not be a direction of descent because the Hessian $\nabla^2 f$ is not positive definite. The basic idea of Newton-like methods is to replace the Hessian $\nabla^2 f(x^k)$ by an approximation M_k and then to solve the linear system

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

in order to compute the descent direction.

Algorithm (7) Newton-like method

Input: Starting point $x^0 \in \mathbb{R}^n$.

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: STOP, if $\nabla f(x^k) = 0$
- 3: choose an invertible matrix $M_k \in \mathbb{R}^{n \times n}$
- 4: compute the direction $d^k \in \mathbb{R}^n$ by solving the linear system

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

- 5: $x^{k+1} := x^k + d^k$
 - 6: **end for**
-

The convergence analysis of Algorithm (7) is based on Lemma 7.4:

Theorem 8.1 (Dennis-Moré condition) Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable, and let (x^k) be the sequence generated by Algorithm (7). Assume that (x^k) converges to \tilde{x} and $\nabla^2 f(\tilde{x})$ is invertible. Then, the following statements are equivalent:

1. (x^k) converges superlinearly to \tilde{x} and it holds true that $\nabla f(\tilde{x}) = 0$.
2. $\|(M_k - \nabla^2 f(\tilde{x}))(x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$.
3. $\|(M_k - \nabla^2 f(x^k))(x^{k+1} - x^k)\| = o(\|x^{k+1} - x^k\|)$.

Proof. In Algorithm (7), the directions are computed by

$$M_k d^k = -\nabla f(x^k) \quad \text{and} \quad d^k = x^{k+1} - x^k,$$

thus

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

Substituting in (2) gives

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

and in (3) leads to

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

The statements (2) und (3) are therefore equivalent to the corresponding statements in Lemma 7.4. Hence, the assertions follow by Lemma 7.4. \square

We will apply Theorem 8.1 in the following two chapters to prove superlinear convergence of Quasi-Newton methods and inexact Newton methods. The following example illustrates the application of the Dennis-Moré condition to prove superlinear convergence:

Example 8.2 Let (x^k) and (M_k) be sequences generated by Algorithm (7) such that

$$x^k \rightarrow \tilde{x} \quad \text{and} \quad M_k \rightarrow \nabla^2 f(\tilde{x}).$$

Then, it holds true that

$$\|(M_k - \nabla^2 f(\tilde{x}))(x^{k+1} - x^k)\| \leq \underbrace{\|M_k - \nabla^2 f(\tilde{x})\|}_{\rightarrow 0} \|x^{k+1} - x^k\| = o(\|x^{k+1} - x^k\|).$$

Thus, condition (2) of Theorem 8.1 is satisfied. If $\nabla^2 f(\tilde{x})$ is invertible, then (x^k) will superlinearly converge to \tilde{x} . \triangle

Analogously to Algorithm (6), Newton-like methods can be globalized as follows.

Algorithm (8) Globalization of Newton-like methods (for optimization problems)

Input: Starting point $x^0 \in \mathbb{R}^n$, parameters $\beta \in (0, 1)$, $\gamma \in (0, 1)$, $\alpha_1, \alpha_2 > 0$ and $p > 0$.

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: STOP, if $\nabla f(x^k) = 0$
- 3: Choose an invertible, symmetric matrix $M_k \in \mathbb{R}^{n \times n}$
- 4: Compute Newton direction $d_N^k \in \mathbb{R}^n$ by solving the linear system

$$M_k d_N^k = -\nabla f(x^k) \quad (*)$$

- 5: **if** $d_N^k \neq 0$ is a solution of $(*)$ that satisfies the generalized angle condition

$$-\nabla f(x^k)^T d_N^k \geq \min\{\alpha_1, \alpha_2 \|d_N^k\|^p\} \|d_N^k\|^2$$

then

- 6: $d^k := d_N^k$
 - 7: **else**
 - 8: $d^k := -\nabla f(x^k)$ *(fallback if $(*)$ not solvable or Newton-like-step is unsatisfying)*
 - 9: **end if**
 - 10: Compute the step size $\sigma_k > 0$ by the Armijo rule (Algorithm (2)).
 - 11: $x^{k+1} := x^k + \sigma_k d^k$.
 - 12: **end for**
-

The convergence result (Theorem 7.8) can be straightforwardly generalized to Algorithm (8) provided that the sequence $(\|M_k\|)$ is bounded.

9. Inexact Newton methods

Newton's method requires the solution of a linear system in each iteration

$$\nabla^2 f(x^k)d^k = -\nabla f(x^k). \quad (9.1)$$

For very large scale problems, the arising linear systems can only be solved by iterative methods (e.g. CG). Then Newton's iteration appears as outer iteration. The question of interest will be to control the accuracy of the inner iteration such that the convergence speed of Newton's method is preserved. The resulting algorithm (local variant) is summarized in Algorithm (9). Globalization of Algorithm (9) for optimization problems is completely analogous to the globalization of Newton's method (Algorithm (6)).

Algorithm (9) Inexact Newton method

Input: Starting point $x^0 \in \mathbb{R}^n$.

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: STOP if $\nabla f(x^k) = 0$.
- 3: compute the direction $d^k \in \mathbb{R}^n$ by approximately solving the linear system

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

- 4: set $x^{k+1} = x^k + d^k$
 - 5: **end for**
-

We assume that the error in the solution of the linear system (9.1) is bounded by

$$\|\nabla f(x^k) + \nabla^2 f(x^k)d^k\| \leq \eta_k \|\nabla f(x^k)\| \quad (9.2)$$

for sufficiently small $\eta_k > 0$. In this setting, the following convergence result can be shown:

Theorem 9.1 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable, $\tilde{x} \in \mathbb{R}^n$ be a stationary point of f and $\nabla^2 f(\tilde{x})$ be invertible. Then, there exists $\varepsilon > 0$, such that the following statements hold true:

1. If $x^0 \in B_\varepsilon(\tilde{x})$ and the directions d^k generated by Algorithm (9) satisfy the condition (9.2) with $\eta_k \leq \eta$ for sufficiently small $\eta \in (0, 1)$, then either the Algorithm (9) terminates after a finite number of steps with $x^k = \tilde{x}$ or the algorithm generates a sequence (x^k) , which linearly converges to \tilde{x} .
2. If additionally $\eta_k \rightarrow 0$, then the convergence rate will be superlinear.
3. If $\eta_k = \mathcal{O}(\|\nabla f(x^k)\|)$ and $\nabla^2 f$ is locally Lipschitz continuous, then the convergence rate will be quadratic.

Proof.

1. The first part of the statement can be derived analogously to the proof of Theorem 7.6.
2. Since f is assumed to be twice continuously differentiable, the gradient ∇f is locally Lipschitz continuous, i.e. there exists $L > 0$ such that

$$\|\nabla f(x^k)\| = \|\nabla f(x^k) - \nabla f(\tilde{x})\| \leq L\|x^k - \tilde{x}\|.$$

Thus,

$$\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\| \leq \eta_k L \|x^k - \tilde{x}\| = o(\|x^k - \tilde{x}\|).$$

The linear convergence implies the existence of $l \geq 0$ and $\gamma \in (0, 1)$, such that for all $k \geq l$, it holds true that

$$\|x^k - \tilde{x}\| \leq \|x^{k+1} - x^k\| + \|x^{k+1} - \tilde{x}\| \leq \|x^{k+1} - x^k\| + \gamma \|x^k - \tilde{x}\|.$$

It follows that

$$\|x^k - \tilde{x}\| \leq \frac{1}{1 - \gamma} \|x^{k+1} - x^k\| = O(\|x^{k+1} - x^k\|)$$

and thus,

$$\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\| \leq \eta_k L \|x^k - \tilde{x}\| = o(\|x^k - \tilde{x}\|) = o(\|x^{k+1} - x^k\|),$$

i.e. superlinear convergence (by Lemma 7.4).

3. The quadratic convergence can be shown analogously. □

Inexact Newton methods and Newton-like methods are related in the following sense.

- Newton-like methods require the solution of a linear system

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

in each iteration, i.e.

$$\nabla^2 f(x^k) d^k = -\nabla f(x^k) + r^k$$

with residuum $r^k = (\nabla^2 f(x^k) - M_k) d^k$. Therefore, we can interpret the direction d^k generated by the Newton-like method as an inexact solution of the Newton system (9.1). In some cases, error bounds of the form (9.2) can be derived.

- On the other hand, the inexact Newton method generates directions d^k with

$$\nabla^2 f(x^k) d^k = -\nabla f(x^k) + r^k.$$

Setting

$$M_k = \nabla^2 f(x^k) - \frac{r^k (d^k)^\top}{\|d^k\|^2}$$

we obtain

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

Thus, the inexact solution of the Newton system can be viewed as an iteration of a Newton-like method with $M_k = \nabla^2 f(x^k) - \frac{r^k (d^k)^\top}{\|d^k\|^2}$.

10. Quasi-Newton Methods

Unlike Newton's method, quasi-Newton methods do not make use of second order derivatives of f . They approximate the second order derivatives iteratively with the help of first order derivatives, i.e. Quasi-Newton methods belong to the class of Newton-like methods (Algorithm (7) and (8)). We will denote the Hessian approximation by H_k (instead of M_k) and require H_k to be invertible and symmetric. The idea is to update H_k by gradient information in each iteration, such that the resulting method converges superlinearly.

Quasi-Newton methods use that two successive iterates x^k and x^{k+1} together with the gradients $\nabla f(x^k)$ and $\nabla f(x^{k+1})$ contain curvature (i.e. Hessian) information. Therefore, at every iteration, H_{k+1} is chosen such that the **Quasi-Newton equation** or **secant equation**

$$H_{k+1}(x^{k+1} - x^k) = \nabla f(x^{k+1}) - \nabla f(x^k). \quad (10.1)$$

is satisfied.

The local Quasi-Newton method can be described as follows.

Algorithm (10) Local Quasi-Newton Method

Input: starting point $x^0 \in \mathbb{R}^n$ and a symmetric, invertible initial matrix $H_0 \in \mathbb{R}^{n \times n}$

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: STOP, if $\nabla f(x^k) = 0$
- 3: Compute the Quasi-Newton step $d^k \in \mathbb{R}^n$ by solving the linear system

$$H_k d^k = -\nabla f(x^k) \quad (10.2)$$

- 4: Set $x^{k+1} = x^k + d^k$
- 5: Compute by an update formula a symmetric, invertible matrix

$$H_{k+1} = \mathcal{H}(H_k, x^{k+1} - x^k, \nabla f(x^{k+1}) - \nabla f(x^k))$$

such that the Quasi-Newton equation (10.1) is satisfied

- 6: **end for**
-

The Quasi-Newton equation (10.1) can be motivated as follows: The fundamental theorem of calculus yields

$$\nabla f(x^{k+1}) - \nabla f(x^k) = \underbrace{\int_0^1 \nabla^2 f(x^k + t(x^{k+1} - x^k)) dt}_{=: M(x^k, x^{k+1})} (x^{k+1} - x^k). \quad (10.3)$$

Thus, the averaged Hessian $M(x^k, x^{k+1})$ satisfies the Quasi-Newton equation.

The following theorem gives a further justification of the Quasi-Newton equation.

Theorem 10.1 Assume that the point \tilde{x} satisfies second order sufficient conditions. Further, assume that Algorithm (10) generates a convergent sequence (x^k) with limit \tilde{x} and it holds true that

$$\lim_{k \rightarrow \infty} \|H_{k+1} - H_k\| = 0.$$

Then, H_k satisfies the Dennis-Moré condition and (x^k) converges superlinearly to \tilde{x} .

Proof.  Exercise □

In the following, we set

$$s^k = x^{k+1} - x^k, \quad y^k = \nabla f(x^{k+1}) - \nabla f(x^k).$$

and discuss several update rules for the Hessian approximation:

$$H_{k+1} = \mathcal{H}(H_k, x^{k+1} - x^k, \nabla f(x^{k+1}) - \nabla f(x^k)) = \mathcal{H}(H_k, s^k, y^k).$$

The initial matrix H_0 , a symmetric, invertible matrix is chosen. A standard choice is given by $H_0 = I$, but sometimes better scaling might be necessary, e.g. by using a the exact Hessian's diagonal, usually modified to ensure positive definiteness. The matrices H_k are then chosen such that H_k is again symmetric and invertible and the Quasi-Newton equation is satisfied for all k . Some variants require only $\mathcal{O}(n^2)$ multiplications per iteration (instead of $\mathcal{O}(n^3)$ like Newton's method).

Symmetric rank-1 formula

Symmetric rank-1-formula are based on symmetric rank-1 modification:

$$H_{k+1} = H_k + \gamma_k u^k (u^k)^\top$$

with $\gamma_k \in \mathbb{R}$ and $u^k \in \mathbb{R}^n$, $\|u^k\| = 1$. Inserting this update into the Quasi-Newton equation (10.1) yields

$$H_{k+1} s^k = H_k s^k + \gamma_k \underbrace{((u^k)^\top s^k)}_{\in \mathbb{R}} u^k \stackrel{!}{=} y^k.$$

If $y^k - H_k s^k = 0$ (with $s^k = d^k$ as in Algorithm (10))

$$\nabla f(x^{k+1}) = \nabla f(x^k) + y^k = \nabla f(x^k) + H_k s^k = \nabla f(x^k) + H_k d^k \stackrel{(10.2)}{=} 0$$

and thus the stopping criterion is satisfied for x^{k+1} . If $y^k - H_k s^k \neq 0$, it follows that

$$u^k = \pm \frac{y^k - H_k s^k}{\|y^k - H_k s^k\|},$$

where we choose w.l.o.g. the solution with „+“. The scalar γ_k can be determined by

$$\gamma_k = \frac{\|y^k - H_k s^k\|^2}{(y^k - H_k s^k)^\top s^k}$$

and thus, we obtain

$$H_{k+1} = H_k + \frac{(y^k - H_k s^k)(y^k - H_k s^k)^\top}{(y^k - H_k s^k)^\top s^k}.$$

The above formula is the **symmetric rank-1 formula** (SR1-formula). Unfortunately this formula has a few drawbacks: The denominator $(y^k - H_k s^k)^\top s^k$ can become 0 or close to 0 causing numerical problems. Further, the positive definiteness of H_{k+1} gets lost in case $(y^k - H_k s^k)^\top s^k < 0$ (even if H_k is positive definite). Thus, H_{k+1} might be singular and $d^{k+1} = -(H_{k+1})^{-1} \nabla f(x^{k+1})$ cannot be guaranteed to be a descent direction.

Broyden Class

More flexible update formulas can be derived by applying symmetric rank-2-updates, i.e.

$$H_{k+1} = H_k + \gamma_k^{(1)} u^k (u^k)^\top + \gamma_k^{(2)} v^k (v^k)^\top.$$

The most popular update formulas are

- the BFGS update (Broyden, Fletcher, Goldfarb and Shanno)

$$H_{k+1}^{BFGS} = \mathcal{H}^{BFGS}(H_k, s^k, y^k) := H_k + \frac{y^k (y^k)^\top}{(y^k)^\top s^k} - \frac{H_k s^k (H_k s^k)^\top}{(s^k)^\top H_k s^k},$$

- the DFP update (Davidon, Fletcher und Powell)

$$\begin{aligned} H_{k+1}^{DFP} &= \mathcal{H}^{DFP}(H_k, s^k, y^k) \\ &:= H_k + \frac{(y^k - H_k s^k)(y^k)^\top + y^k (y^k - H_k s^k)^\top}{(y^k)^\top s^k} - \frac{(y^k - H_k s^k)^\top s^k}{((y^k)^\top s^k)^2} y^k (y^k)^\top, \end{aligned}$$

- the Broyden class

$$H_{k+1}^\lambda = (1 - \lambda) H_{k+1}^{BFGS} + \lambda H_{k+1}^{DFP}$$

with $\lambda \in \mathbb{R}$ and for $\lambda \in [0, 1]$ the convex Broyden class.

Remark As Newton's method, the Quasi-Newton methods based on updates in the Broyden class are invariant under affine transformations (in particular the BFGS method). \triangle


Theorem 10.2 Let H_k be symmetric. Further assume that

$$(y^k)^\top s^k \neq 0 \quad \text{und} \quad (s^k)^\top H_k s^k \neq 0.$$

Then, the matrices H_{k+1}^λ with $\lambda \in \mathbb{R}$ are well defined, symmetric and satisfy the Quasi-Newton equation (10.1). In addition, if H_k is positive definite and it holds true that

$$(y^k)^\top s^k > 0,$$

then H_{k+1}^λ will be positive definite for $\lambda \geq 0$.

Proof.  Exercise □

Using the **Sherman-Morrison-Woodbury formula**, we can derive explicit expressions for the update of the inverse Hessian approximation, which is a rank-2 modification as well.

Lemma 10.3 (Sherman-Morrison-Woodbury formula) Let $A \in \mathbb{R}^{n \times n}$ be an invertible matrix and $u, v \in \mathbb{R}^n$. Then, the matrix $A + uv^\top$ is invertible if and only if it holds true that $1 + v^\top A^{-1}u \neq 0$. Then, the inverse is given by

$$(A + uv^\top)^{-1} = \left(I - \frac{A^{-1}uv^\top}{1 + v^\top A^{-1}u} \right) A^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u}.$$

Remark The Shermann-Morrison-Woodbury formula can be generalized to rank-r modifications. Let $U, V \in \mathbb{R}^{n \times r}$. Then,

$$(A + UV^\top)^{-1} = A^{-1} - A^{-1}U(I + V^\top A^{-1}U)^{-1}V^\top A^{-1}.$$

△

In particular, we obtain update formulas for the inverse of the BFGS and DFP updates:

Theorem 10.4 Let H_k be positive definite and $B_k = H_k^{-1}$. Further assume that $(y^k)^\top s^k > 0$. Then, the matrices $B_{k+1}^{BFGS} = (H_{k+1}^{BFGS})^{-1}$ and $B_{k+1}^{DFP} = (H_{k+1}^{DFP})^{-1}$ can be computed via the following update formulas:

$$B_{k+1}^{BFGS} = B_k + \frac{(s^k - B_k y^k)(s^k)^\top + s^k (s^k - B_k y^k)^\top}{(s^k)^\top y^k} - \frac{(s^k - B_k y^k)^\top y^k}{((s^k)^\top y^k)^2} s^k (s^k)^\top,$$

$$B_{k+1}^{DFP} = B_k + \frac{s^k (s^k)^\top}{(s^k)^\top y^k} - \frac{B_k y^k (B_k y^k)^\top}{(y^k)^\top B_k y^k}.$$

Remark Theorem 10.4 shows that the BFGS update formula for the inverse Hessian approximation corresponds to the DFP update of the Hessian approximation with flipped roles of s^k and y^k . Vice-versa, the DFP update for the inverse Hessian approximation corresponds to the BFGS update of the Hessian approximation:

$$\begin{aligned} H_{k+1}^{BFGS} &= \mathcal{H}^{BFGS}(H_k, s^k, y^k) & H_{k+1}^{DFP} &= \mathcal{H}^{DFP}(H_k, s^k, y^k) \\ B_{k+1}^{BFGS} &= \mathcal{H}^{DFP}(B_k, y^k, s^k) & B_{k+1}^{DFP} &= \mathcal{H}^{BFGS}(B_k, y^k, s^k) \end{aligned}$$

△

BFGS Method

We can derive a version of the BFGS algorithm that works with the inverse Hessian approximation 10.4, Algorithm (11). Theorem 10.5 gives a convergence result.

Theorem 10.5 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be twice continuously differentiable with locally Lipschitz continuous Hessian $\nabla^2 f$. Furthermore, assume that sufficient second order optimality conditions are satisfied at \tilde{x} . Then, there exist $\delta > 0$ and $\varepsilon > 0$, such that, for every starting point $x^0 \in B_\delta(\tilde{x})$ and every symmetric, positive definite matrix $B_0 \in \mathbb{R}^{n \times n}$ with $\|B_0 - \nabla^2 f(\tilde{x})^{-1}\| < \varepsilon$, either

Algorithm (11) (Local) inverse BFGS method

Input: Starting point $x^0 \in \mathbb{R}^n$ and a symmetric, positive definite matrix $B_0 \in \mathbb{R}^{n \times n}$ (inverse Hessian approximation at x^0).

- 1: STOP, if $\nabla f(x^0) = 0$.
- 2: **for** $k = 0, 1, 2, \dots$ **do**
- 3: Compute the direction $d^k = -B_k \nabla f(x^k)$.
- 4: Set $x^{k+1} = x^k + d^k$.
- 5: STOP, if $\nabla f(x^{k+1}) = 0$.
- 6: Set $s^k = d^k$ and $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$.
- 7: STOP (with error message), if $(y^k)^\top s^k \leq 0$.
- 8: Compute

$$B_{k+1} = B_k + \frac{(s^k - B_k y^k)(s^k)^\top + s^k(s^k - B_k y^k)^\top}{(s^k)^\top y^k} - \frac{(s^k - B_k y^k)^\top y^k}{((s^k)^\top y^k)^2} s^k (s^k)^\top.$$

- 9: **end for**

Algorithm (11) terminates after a finite number of iterations at $x^k = \tilde{x}$ or generates a sequence $(x^k) \subset B_\delta(\tilde{x})$, which converges superlinearly to \tilde{x} .

Proof. See e.g. Geiger, Kanzow, Theorem 11.33. □

Using the Powell-Wolfe step size strategy, Algorithm (11) can be globalized as follows, cp. Algorithm (12). The condition $(y^k)^\top s^k > 0$ will be ensured by the step size strategy.

Algorithm (12) Globalized inverse BFGS method

Input: Starting point $x^0 \in \mathbb{R}^n$, parameters $\gamma \in (0, 1/2)$, $\eta \in (\gamma, 1)$ and a symmetric, positive definite matrix $B_0 \in \mathbb{R}^{n \times n}$ (inverse Hessian approximation at x^0).

- 1: STOP, if $\nabla f(x^0) = 0$.
- 2: **for** $k = 0, 1, 2, \dots$ **do**
- 3: Compute the direction $d^k = -B_k \nabla f(x^k)$.
- 4: Compute the step size $\sigma_k > 0$ with the Powell-Wolfe rule (Alg. (4)).
- 5: Set $x^{k+1} = x^k + \sigma_k d^k$.
- 6: STOP, if $\nabla f(x^{k+1}) = 0$.
- 7: Set $s^k = x^{k+1} - x^k$ and $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$.
- 8: Compute

$$B_{k+1} = B_k + \frac{(s^k - B_k y^k)(s^k)^\top + s^k(s^k - B_k y^k)^\top}{(s^k)^\top y^k} - \frac{(s^k - B_k y^k)^\top y^k}{((s^k)^\top y^k)^2} s^k (s^k)^\top.$$

- 9: **end for**

We first show that the condition $(y^k)^\top s^k > 0$ is indeed ensured by the Powell-Wolfe step size strategy:

Lemma 10.6 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable. If B_k is positive definite and the Powell-Wolfe step size strategy generates a feasible step size $\sigma_k > 0$ in Algorithm (12), then $(y^k)^\top s^k > 0$ and B_{k+1} will be positive definite.

Proof. The slope condition (6.3) yields

$$\begin{aligned} (y^k)^\top s^k &= \sigma_k (\nabla f(x^{k+1})^\top d^k - \nabla f(x^k)^\top d^k) \stackrel{(6.3)}{\geq} \sigma_k (\eta \nabla f(x^k)^\top d^k - \nabla f(x^k)^\top d^k) \\ &= -\sigma_k (1 - \eta) \nabla f(x^k)^\top d^k > 0. \end{aligned}$$

Since B_k is positive definite by assumption, $H_k = B_k^{-1}$ is positive definite and due to $(y^k)^\top s^k > 0$, H_{k+1}^{BFGS} is also positive definite (by Theorem 10.2). Thus, $B_{k+1} = (H_{k+1}^{BFGS})^{-1}$ is positive definite, too. \square

The following convergence result can be shown for Algorithm (12):

Theorem 10.7 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable and $x^0 \in \mathbb{R}^n$ such that the level set $N_f(x^0)$ is compact. Then, Algorithm (12) is well defined. Further, if the condition number of the matrices B_k is uniformly bounded, then every accumulation point of (x^k) will be a stationary point.

Proof.  Exercise \square

11. Optimality Conditions for Constrained Optimization Problems

In the second part of the course, we will focus on constrained optimization problems, i.e.

$$\min_{x \in X} f(x) \quad (11.1)$$

with objective function $f : X \rightarrow \mathbb{R}$ and **feasible set** $X \subseteq \mathbb{R}^n$, which is assumed to be defined via a system of equality and inequality constraints

$$X = \{x \in \mathbb{R}^n : g(x) \leq 0, h(x) = 0\} \quad (11.2)$$

with continuous functions $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $h : \mathbb{R}^n \rightarrow \mathbb{R}^p$.

In this chapter, we derive mathematical characterizations of the solutions of (11.1). As in the unconstrained case, we discuss optimality conditions of two types, necessary and sufficient conditions. We start by noting one important item of terminology that recurs throughout the rest of the notes.

Definition 11.1 The index sets of the inequality and equality constraints will be denoted by

$$\mathcal{G} = \{1, \dots, m\} \quad \text{and} \quad \mathcal{H} = \{1, \dots, p\}.$$

The **index set of active inequality constraints** $\mathcal{A}(x)$ at any feasible point $x \in X$ is defined by

$$\mathcal{A}(x) = \{i \in \mathcal{G} : g_i(x) = 0\}$$

and the **index set of inactive inequality constraints**

$$\mathcal{I}(x) = \{i \in \mathcal{G} : g_i(x) < 0\} = \mathcal{G} \setminus \mathcal{A}(x).$$

The following example illustrates the basic principles behind the characterization of solutions for constrained optimization problems, in particular the need for more general optimality conditions.

Example 11.2 We consider the minimization of the function $f(x) = x^3 + x$ such that $x \geq -1$. By inspection of the first order necessary optimality condition $f'(x) = 0$, we see that $f'(x) = 3x^2 + 1 > 0$. The minimizer is obviously $\tilde{x} = -1$, i.e. lies on the **boundary** of the feasible set. Starting at $\tilde{x} = -1$, we observe that the point $x = \tilde{x} + d$ remains feasible for all positive directions $d \geq 0$. For $d > 0$, it holds true that

$$f'(\tilde{x}) \cdot d = 4d > 0,$$

i.e. the function values increase along feasible directions. Thus, $\tilde{x} = -1$ is a local minimizer. \triangle

Definition 11.3 Let $M \subseteq \mathbb{R}^n$ and $x \in M$. Then,

$$\mathcal{T}(M, x) := \{d \in \mathbb{R}^n : \exists(x^k) \subset M, \exists(\eta_k) \subset \mathbb{R}_{>0} : \lim_{k \rightarrow \infty} x^k = x, \lim_{k \rightarrow \infty} \eta_k(x^k - x) = d\}$$

is called **tangent cone** to M at x .

Remark It is easy to see that $0 \in \mathcal{T}(M, x)$ and the tangent cone is indeed a cone, i.e. $d \in \mathcal{T}(M, x)$ implies $\alpha d \in \mathcal{T}(M, x)$ for $\alpha \in \mathbb{R}_{>0}$. Furthermore, the tangent cone $\mathcal{T}(M, x)$ is a closed set for $x \in M$. \triangle

Example 11.4  Exercise: Determine the tangent cone $\mathcal{T}(X, x)$.

1. $X := \{x \in \mathbb{R}^2 : x_1 \leq 1, x_2 \geq 0, x_2 \leq x_1^2 + x_1\}$, and $x = (0, 0)^\top$
2. $X := \{x \in \mathbb{R}^2 : x_1 \geq x_2^2, x_2 \geq x_1^2\}$, and $x = (0, 0)^\top$

\triangle

The necessary conditions defined in the following theorem are called first order conditions because they are concerned with properties of the gradients of the objective and constraint functions.

Theorem 11.5 Let f be continuously differentiable and $\tilde{x} \in X$ be a local minimizer of (11.1). Then, it holds true that

$$\nabla f(\tilde{x})^T d \geq 0 \quad \forall d \in \mathcal{T}(X, \tilde{x}).$$

Proof. Let $d \in \mathcal{T}(X, \tilde{x})$ arbitrary and

$$X \ni (x^k) \rightarrow \tilde{x}, \quad (\eta_k) \in \mathbb{R}_{>0}, \quad d^k := \eta_k(x^k - \tilde{x}) \rightarrow d,$$

where $x^k \neq \tilde{x}$ w.l.o.g..

Due to the local optimality of \tilde{x} , it holds true that

$$f(x^k) - f(\tilde{x}) \geq 0$$

for sufficiently large k . Thus, Taylor expansion yields (for sufficiently large k)

$$\begin{aligned} 0 &\leq \eta_k(f(x^k) - f(\tilde{x})) \\ &= \eta_k \nabla f(\tilde{x})^T (x^k - \tilde{x}) + \eta_k o(\|x^k - \tilde{x}\|) \\ &= \nabla f(\tilde{x})^T d^k + \|d^k\| \frac{o(\|x^k - \tilde{x}\|)}{\|x^k - \tilde{x}\|} \xrightarrow{k \rightarrow \infty} \nabla f(\tilde{x})^T d. \end{aligned}$$

\square

Theorem 11.5 includes as a special case the first order optimality condition for unconstrained optimization problems, Theorem 2.1, since for points \tilde{x} in the interior of $X \subseteq \mathbb{R}^n$, it holds true that $\mathcal{T}(X, \tilde{x}) = \mathbb{R}^n$. The first order optimality condition for constrained optimization problems cannot be easily verified unless we have a simple representation of the tangent cone. We will linearize the (active) constraints to form a local approximation of the feasible set.

Definition 11.6 The set

$$\mathcal{T}_l(g, h, x) = \{d \in \mathbb{R}^n : \nabla g_i(x)^T d \leq 0 \quad \forall i \in \mathcal{A}(x), \nabla h(x)^T d = 0\}$$

is called **linearized tangent cone** at $x \in X$.

Since the verification of $d \in \mathcal{T}_l(g, h, x)$ (in contrast to $d \in \mathcal{T}(X, x)$) is straightforward, the idea is to replace the tangent cone by the linearized tangent cone in the optimality conditions. Obviously, it holds true that

$$\mathcal{T}(X, x) \subseteq \mathcal{T}_l(g, h, x).$$

In order to formulate optimality conditions using the linearized tangent cone, we will require that

$$\{v \in \mathbb{R}^n : v^T d \geq 0 \quad \forall d \in \mathcal{T}(X, x)\} = \{v \in \mathbb{R}^n : v^T d \geq 0 \quad \forall d \in \mathcal{T}_l(g, h, x)\}, \quad (11.3)$$

since

$$\nabla f(x)^T d \geq 0 \quad \forall d \in \mathcal{T}(X, x)$$

is equivalent to

$$\nabla f(x) \in \{v \in \mathbb{R}^n : v^T d \geq 0 \quad \forall d \in \mathcal{T}(X, x)\}.$$

Note that (11.3) is in particular satisfied, if $\mathcal{T}(X, x) = \mathcal{T}_l(g, h, x)$.

Definition 11.7 Let $x \in X$. A condition that implies (11.3) is called **constraint qualification**.

Corollary 11.8 Let f be continuously differentiable and $\tilde{x} \in X$ be a local minimizer such that a constraint qualification is satisfied. Then, it holds true that

$$\nabla f(\tilde{x})^T d \geq 0 \quad \forall d \in \mathcal{T}_l(g, h, \tilde{x}).$$

In the following, we give an overview of the most important constraint qualifications:

Theorem 11.9 Each of the following conditions is a constraint qualification at $x \in X$:

1. The functions g_i for $i \in \mathcal{A}(x)$ are concave and h is affine linear, i.e.

$$g_i(y) \leq g_i(x) + \nabla g_i(x)^T (y - x) \quad \forall i \in \mathcal{A}(x) \quad \text{and} \quad h(x) = Bx - b.$$

2. **Mangasarian-Fromovitz:** $\nabla h(x)$ has full column rank (or h is affine linear) and there exists $d \in \mathbb{R}^n$ such that

$$\nabla g_i(x)^T d < 0 \quad \forall i \in \mathcal{A}(x) \quad \text{and} \quad \nabla h(x)^T d = 0.$$

3. **Regularity:** The columns of the matrix

$$(\nabla g_{\mathcal{A}(x)}(x), \nabla h(x))$$

are linearly independent.

This CQ is known as *Linear Independence Constraint Qualification (LICQ)*.

The simple structure of the set $\mathcal{T}_l(g, h, \tilde{x})$ in Corollary 11.8 allows the application of the following lemma to reformulate the optimality conditions.

Lemma 11.10 (Lemma of Farkas) Let $A \in \mathbb{R}^{n \times m}$, $B \in \mathbb{R}^{n \times p}$, $c \in \mathbb{R}^n$. Then, the following two statements are equivalent:

1. For all $d \in \mathbb{R}^n$ with $A^T d \leq 0$ and $B^T d = 0$, it holds true that $c^T d \leq 0$.
2. There exists $u \in \mathbb{R}_{\geq 0}^m$ and $v \in \mathbb{R}^p$ with $c = Au + Bv$.

By Farkas' Lemma and Corollary 11.8, we obtain the following first order optimality condition:

Theorem 11.11 (First order optimality conditions) Let \tilde{x} be a local minimizer of (11.1) and let a constraint qualification be satisfied in \tilde{x} . Then, the following conditions are satisfied (**Karush-Kuhn-Tucker conditions, KKT conditions**):

There exist **Lagrange multipliers** $\tilde{\lambda} \in \mathbb{R}^m$ and $\tilde{\mu} \in \mathbb{R}^p$, such that the following hold:

1. multiplier rule:

$$\nabla f(\tilde{x}) + \sum_{i=1}^m \tilde{\lambda}_i \nabla g_i(\tilde{x}) + \sum_{j=1}^p \tilde{\mu}_j \nabla h_j(\tilde{x}) = \nabla f(\tilde{x}) + \nabla g(\tilde{x})\tilde{\lambda} + \nabla h(\tilde{x})\tilde{\mu} = 0$$

2. feasibility: $h(\tilde{x}) = 0$, $g(\tilde{x}) \leq 0$,
3. complementary conditions: $\tilde{\lambda} \geq 0$, $\tilde{\lambda}^T g(\tilde{x}) = 0$.

Proof. Let \tilde{x} be a local minimizer of (11.1) and let a constraint qualification be satisfied in \tilde{x} . Then, feasibility (2) follows directly by the assumptions. Furthermore, by Corollary 11.8, it holds true that

$$-\nabla f(\tilde{x})^T d \leq 0 \quad \forall d \in \mathcal{T}_l(g, h, \tilde{x}),$$

and for all $d \in \mathbb{R}^n$ satisfying

$$\nabla g_i(\tilde{x})^T d \leq 0 \quad \text{for } i \in \mathcal{A}(\tilde{x}) \quad \text{and} \quad \nabla h(\tilde{x})^T d = 0.$$

Farkas' Lemma with

$$c = -\nabla f(\tilde{x}), \quad A = \nabla g_{\mathcal{A}(\tilde{x})}(\tilde{x}), \quad B = \nabla h(\tilde{x})$$

implies the existence of vectors $u \in \mathbb{R}_{\geq 0}^{|\mathcal{A}(\tilde{x})|}$ and $v \in \mathbb{R}^p$ with

$$c = Au + Bv.$$

Choosing $\tilde{\lambda} \in \mathbb{R}^m$ with $\tilde{\lambda}_{\mathcal{A}(\tilde{x})} = u$, $\tilde{\lambda}_{\mathcal{I}(\tilde{x})} = 0$ and $\tilde{\mu} = v$. yields 1) with multipliers satisfying the complementary conditions 3). \square

Remark The KKT conditions say that the negative gradient of the objective function lies in the cone of the gradients of the active constraints. \triangle

Definition 11.12 Suppose that $(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ satisfies the KKT conditions. Then, \tilde{x} is called a **KKT point** of (11.1) and $(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$ a **KKT triple**. For a KKT triple, the strict complementary condition is satisfied, if it holds true that

$$\tilde{\lambda}_i > 0 \quad \forall i \in \mathcal{A}(\tilde{x}).$$

Definition 11.13 The function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ given by

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

is called **Lagrangian function**.

Hence, the first KKT condition is equivalent to

$$\nabla_x \mathcal{L}(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) = 0.$$

Analogously to Theorem 3.6, first order necessary conditions are sufficient for convex optimization problems.

Definition 11.14 The nonlinear optimization problem (11.1) is called **convex**, if the functions f and g_i (for all $i \in \mathcal{G}$) are convex and the function h is affine linear. Note that these conditions in particular imply that the feasible set X is convex.

Theorem 11.15 Suppose that (11.1) is convex. Then, it holds:

1. Every local minimizer of (11.1) is a global minimizer.
2. If a constraint qualification is satisfied at a local/global minimizer $\tilde{x} \in X$ of (11.1), then the KKT conditions are satisfied at \tilde{x} .
3. If the KKT conditions are satisfied at a point \tilde{x} , then \tilde{x} will be a global minimizer of (11.1).

Proof. The first statement directly follows from Theorem 3.6. The second statement has been proven in Theorem 11.11. It remains to show that the KKT conditions are sufficient for optimality. Let $(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$ be a KKT triple and $x \in X$ arbitrary. For $d = x - \tilde{x}$, we obtain (for all $i \in \mathcal{G}$)

$$\tilde{\lambda}_i \nabla g_i(\tilde{x})^T d \leq \tilde{\lambda}_i (g_i(x) - g_i(\tilde{x})) = \tilde{\lambda}_i g_i(x) \leq 0.$$

Since h is assumed to be affine linear, it holds true that $\nabla h(\tilde{x})^T d = h(x) - h(\tilde{x}) = 0$ and thus,

$$f(x) - f(\tilde{x}) \geq \nabla f(\tilde{x})^T d = -\tilde{\lambda}^T \nabla g(\tilde{x})^T d - \tilde{\mu}^T \nabla h(\tilde{x})^T d = -\tilde{\lambda}^T \nabla g(\tilde{x})^T d \geq 0.$$

□

To formulate second order optimality conditions, we consider the following cone:

Definition 11.16 Given $x \in X$ and $\lambda \in \mathbb{R}_{\geq 0}^m$, we define

$$\mathcal{T}_+(g, h, x, \lambda) = \left\{ d \in \mathbb{R}^n : \nabla h(x)^T d = 0, \nabla g_i(x)^T d \begin{cases} = 0, & \text{if } i \in \mathcal{A}(x) \text{ and } \lambda_i > 0 \\ \leq 0, & \text{if } i \in \mathcal{A}(x) \text{ and } \lambda_i = 0 \end{cases} \right\}$$

Theorem 11.17 (Second order sufficient optimality conditions)

Let f, g , and h be twice continuously differentiable. Suppose that the point \tilde{x} satisfies the KKT conditions with multipliers $\tilde{\lambda} \in \mathbb{R}^m$ and $\tilde{\mu} \in \mathbb{R}^p$. Furthermore, assume that

$$d^T \nabla_{xx}^2 \mathcal{L}(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) d > 0 \quad \forall d \in \mathcal{T}_+(g, h, \tilde{x}, \tilde{\lambda}) \setminus \{0\}. \quad (11.4)$$

Then, \tilde{x} is a strict local minimizer of (11.1).

Theorem 11.18 (Necessary second order optimality conditions)

Let f, g, h be twice continuously differentiable, \tilde{x} be a local solution of (11.1) and the columns of the matrix

$$(\nabla g_{\mathcal{A}(\tilde{x})}(\tilde{x}), \nabla h(\tilde{x}))$$

be linearly independent.

Then, there exist **Lagrange multipliers** $\tilde{\lambda} \in \mathbb{R}^m$ and $\tilde{\mu} \in \mathbb{R}^p$, such that

- $\nabla f(\tilde{x}) + \nabla g(\tilde{x})\tilde{\lambda} + \nabla h(\tilde{x})\tilde{\mu} = 0$ (stationarity of the Lagrangian)
- $h(\tilde{x}) = 0, g(\tilde{x}) \leq 0$ (feasibility)
- $\tilde{\lambda} \geq 0, \tilde{\lambda}^T g(\tilde{x}) = 0$ (complementary conditions)
- $d^T \nabla_{xx}^2 L(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) d \geq 0 \quad \forall d \in \mathcal{T}_+(g, h, \tilde{x}, \tilde{\lambda})$

12. Penalty Method

An important class of methods for constrained optimization seeks the solution by replacing the original constrained problem by a sequence of unconstrained subproblems.

Most approaches define a sequence of such penalty functions, in which the penalty terms for the constraint violations are multiplied by some positive coefficient. By making this coefficient larger and larger, we penalize constraint violations more and more severely, thereby forcing the minimizer of the penalty function closer and closer to the feasible region for the constrained problem.

Such approaches are sometimes known as exterior penalty methods, because the penalty term for each constraint is nonzero only when x is infeasible with respect to that constraint. Often, the minimizers of the penalty functions are infeasible with respect to the original problem, and approach feasibility only in the limit as the penalty parameter grows increasingly large.

We consider subproblems of the form

$$\min_{x \in \mathbb{R}^n} P_\alpha(x) = f(x) + \alpha\pi(x)$$

with $\alpha > 0$ and a penalization function $\pi : \mathbb{R}^n \rightarrow \mathbb{R}$, where $\pi(x) = 0$ for all $x \in X$ and $\pi(x) > 0$ for all $x \in \mathbb{R}^n \setminus X$. The function P_α is called a *penalty function*.

A general framework for algorithms based on the penalty functions can be specified as in the following Algorithm (13):

Algorithm (13) Penalty Method

Input: Starting point $x^0 \in \mathbb{R}^n$, penalty parameter $\alpha_1 > 0$.

- 1: **for** $k = 1, 2, \dots$ **do**
 - 2: Compute the solution x^k of the subproblem $\min_{x \in \mathbb{R}^n} P_{\alpha_k}(x)$ (use x^{k-1} as starting point)
 - 3: **STOP**, if $x^k \in X$
 - 4: Choose $\alpha_{k+1} > \alpha_k$
 - 5: **end for**
-

12.1 The Quadratic Penalty Method

The simplest penalty function of this type is the quadratic penalty function, in which the penalty terms are the squares of the constraint violations

$$P_\alpha(x) = f(x) + \frac{\alpha}{2} \sum_{i=1}^m (g_i(x)^+)^2 + \frac{\alpha}{2} \sum_{j=1}^p h_j(x)^2 = f(x) + \frac{\alpha}{2} \|g(x)^+\|^2 + \frac{\alpha}{2} \|h(x)\|^2 \quad (12.1)$$

with $g_i(x)^+ := \max\{0, g_i(x)\}$ and $g(x)^+ = (g_i(x)^+)_{i=1..m}$.

Let f, g, h be continuously differentiable, then the quadratic penalty function (12.1) is continuously differentiable too,

$$\nabla P_\alpha(x) = \nabla f(x) + \alpha \sum_{i=1}^m g_i(x)^+ \cdot \nabla g_i(x) + \alpha \sum_{j=1}^p h_j(x) \nabla h_j(x).$$

Under the assumption that global minimizers are computed for each subproblem of Algorithm (13), then the following convergence result can be proven.

Theorem 12.1 Let f, g, h be continuously differentiable and X be nonempty. Further, assume that the sequence $(\alpha_k) \subset \mathbb{R}_{>0}$ is strictly increasing with $\alpha_k \rightarrow \infty$ for $k \rightarrow \infty$ and that Algorithm (13) generates a sequence (x^k) of global minimizers of the corresponding subproblems. With

$$\lambda_i^k = \alpha_k \max\{0, g_i(x^k)\} \quad \text{and} \quad \mu_j^k = \alpha_k h_j(x^k),$$

the following statements hold true:

1. If $(x^k, \lambda^k, \mu^k)_K$ is a convergent subsequence of (x^k, λ^k, μ^k) with limit $(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$, then \tilde{x} will be a global solution of (11.1) and $(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$ will be a KKT triple of (11.1).
2. Let \tilde{x} be an accumulation point of (x^k) and $(x^k)_K$ be a convergent subsequence with limit \tilde{x} . Further assume that the columns of the matrix

$$(\nabla g_{\mathcal{A}(\tilde{x})}(\tilde{x}), \nabla h(\tilde{x}))$$

are linearly independent. Then the sequence $(x^k, \lambda^k, \mu^k)_K$ converges to a KKT triple of (11.1) and \tilde{x} is a global solution of (11.1).

The differentiability of the penalty function P_α is a desirable property. However, it holds true that for all $x \in X$ (and all $i \in \mathcal{G}$ and $j \in \mathcal{H}$)

$$(g_i(x))_+ = 0 \quad \text{and} \quad h_j(x) = 0.$$

This yields

$$\nabla P_\alpha(x) = \nabla f(x) \quad \forall x \in X$$

and thus

$$\nabla P_\alpha(x) = 0 \quad \xleftrightarrow{x \in X} \quad \nabla f(x) = 0$$

for all $x \in X$. Thus, the minimizers of P_α are infeasible in general, since, in general, $\nabla f(\tilde{x}) \neq 0$ for solutions \tilde{x} of the (constrained) problem (11.1).

Remark For $\alpha \rightarrow \infty$, the subproblems become increasingly ill conditioned causing a slow rate of convergence for Newton(-like) methods. \triangle

Exact Penalty Method

Definition 12.2 Let \tilde{x} be the local minimizer of (11.1). The penalty function $P : \mathbb{R}^n \rightarrow \mathbb{R}$ is called *exact* at \tilde{x} , if \tilde{x} is a local minimizer of P .

The l_1 -penalty function

$$P_\alpha^1(x) = f(x) + \alpha \sum_{i=1}^m g_i(x)^+ + \alpha \sum_{j=1}^p |h_j(x)| = f(x) + \alpha \|g(x)^+\|_1 + \alpha \|h(x)\|_1 \quad (12.2)$$

is exact under suitable assumptions and for sufficient large $\alpha > 0$. Note that the l_1 -penalty function is not differentiable.

Theorem 12.3 Let f, g_i be convex and continuously differentiable, h be affine linear and $(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$ be a KKT triple of (11.1). Then, \tilde{x} is a global minimizer of (11.1) and \tilde{x} is a global minimizer of P_α^1 for all

$$\alpha \geq \max\{\tilde{\lambda}_1, \dots, \tilde{\lambda}_m, |\tilde{\mu}_1|, \dots, |\tilde{\mu}_p|\}.$$

Remark

- For exact penalty functions, it suffices to solve only a single penalized problem, if the penalty parameter α is chosen properly (difficult a priori, but after a solution is computed, one can verify if the condition of theorem 12.3 is fulfilled).

- Similar/related methods:

– *Barrier methods*: (here: only inequality constraints)

$$\min f(x) - \mu \sum_{i=1}^m \log(-g_i(x)) \quad \text{started with feasible } x^0 \in X, \mu > 0$$

– *Augmented Lagrangian methods*: (here: only equality constraints)

$$\mathcal{L}_A(x, \alpha, \beta) = f(x) - \sum_{j=1}^p \alpha_j h_j(x) + \frac{1}{2\beta} \sum_{j=1}^p h_j^2(x)$$

- SQP methods may use an exact penalty function as *merit function* for globalization.

△

13. Sequential Quadratic Programming

One of the most effective methods for nonlinearly constrained optimization generates steps by solving quadratic subproblems.

We first focus on equality constrained optimization problems, i.e.

$$\min f(x) \quad \text{s.t.} \quad h(x) = 0. \quad (13.1)$$

The essential idea of SQP is to model (13.1) at the current iterate x^k by a quadratic programming subproblem and to use the minimizer of this subproblem to define a new iterate x^{k+1} . The challenge is to design the quadratic subproblem so that it yields a good step for the underlying constrained optimization problem and so that the overall SQP algorithm has good convergence properties and good practical performance. Perhaps the simplest derivation of SQP methods, which we now present, views them as an application of Newton's method to the KKT optimality conditions

$$F(x, \mu) := \begin{pmatrix} \nabla_x \mathcal{L}(x, \mu) \\ h(x) \end{pmatrix} = 0. \quad (13.2)$$

To determine \tilde{x} with multipliers $\tilde{\mu}$, we use Newton's method, i.e. in every iteration, the following linear systems has to be solved

$$F'(x^k, \mu^k) d^k = -F(x^k, \mu^k) \quad (13.3)$$

with

$$F'(x, \mu) = \begin{pmatrix} \nabla_{xx}^2 \mathcal{L}(x, \mu) & \nabla_{x\mu}^2 \mathcal{L}(x, \mu) \\ \nabla h(x)^T & 0 \end{pmatrix} = \begin{pmatrix} \nabla_{xx}^2 \mathcal{L}(x, \mu) & \nabla h(x) \\ \nabla h(x)^T & 0 \end{pmatrix}$$

Split $d^k = \begin{pmatrix} d_x^k \\ d_\mu^k \end{pmatrix} \in \mathbb{R}^{(n+p)}$ to derive

$$\begin{pmatrix} \nabla_{xx}^2 \mathcal{L}(x, \mu) & \nabla h(x) \\ \nabla h(x)^T & 0 \end{pmatrix} \begin{pmatrix} d_x^k \\ d_\mu^k \end{pmatrix} = \begin{pmatrix} -\nabla_x \mathcal{L}(x^k, \mu^k) \\ -h(x^k) \end{pmatrix}$$

This idea is shown in Algorithm (14). It is straightforward to establish a local convergence result.

Algorithm (14) Local SQP Method

Input: Starting point $x^0 \in \mathbb{R}^n$ and $\mu^0 \in \mathbb{R}^p$

- 1: **for** $k = 0, 1, 2, \dots$ **do**
 - 2: STOP, if $h(x^k) = 0$ and $\nabla_x \mathcal{L}(x^k, \mu^k) = 0$ (i.e. (x^k, μ^k) is a KKT pair)
 - 3: Compute d^k by solving (13.3)
 - 4: Set $x^{k+1} = x^k + d_x^k$ and $\mu^{k+1} = \mu^k + d_\mu^k$
 - 5: **end for**
-

Theorem 13.1 Let f and h be twice continuously differentiable and $(\tilde{x}, \tilde{\mu})$ be a KKT pair. Further assume that the following conditions hold:

- $\text{rank } \nabla h(\tilde{x}) = p$ (regularity)
- $s^T \nabla_{xx}^2 L(\tilde{x}, \tilde{\mu}) s > 0 \forall s \in \mathbb{R}^n \setminus \{0\}$ with $\nabla h(\tilde{x})^T s = 0$ (second order SOC)

Then, there exists $\delta > 0$, such that, for all $(x^0, \mu^0) \in B_\delta(\tilde{x}, \tilde{\mu})$, Algorithm (14) either terminates after a finite number of iterations or generates a sequence (x^k, μ^k) which converges superlinearly to $(\tilde{x}, \tilde{\mu})$. If $\nabla^2 f$ and $\nabla^2 h_j$ are Lipschitz continuous in $B_\delta(\tilde{x})$ (for all $j \in \mathcal{H}$), the rate of convergence will be quadratic.

The algorithm can be also derived by making a quadratic approximation of the Lagrangian and a linear approximation of the constraint, which leads to an alternative motivation of the SQP method:

$$\begin{aligned} \min_{d \in \mathbb{R}^n} \quad & q_k(d) = f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d \\ \text{s.t.} \quad & h(x^k) + \nabla h(x^k)^T d = 0 \end{aligned} \quad (13.4)$$

with $H_k = \nabla_{xx}^2 \mathcal{L}(x^k, \mu^k)$. The pair $(d^k, \mu_{qp}^k) = (d_x^k, \mu^k + d_\mu^k)$ is a KKT pair of (13.4), if and only if $d^k = (d_x^k, d_\mu^k)$ is a solution of (13.3). ($\text{\textcircled{E}}$ Exercise).

Remark The Hessians H_k can be modified to make them positive definite (possibly replacing it by a quasi-Newton approximation). \triangle

Algorithm (15) Local SQP Method II

Input: Starting point $x^0 \in \mathbb{R}^n$ and $\mu^0 \in \mathbb{R}^p$

- 1: **for** $k = 0, 1, 2, \dots$ **do**
 - 2: STOP, if (x^k, μ^k) is a KKT pair of (13.1)
 - 3: Compute the solution d^k of (13.4) with corresponding multipliers μ_{qp}^k
 - 4: Set $x^{k+1} = x^k + d^k$ and $\mu^{k+1} = \mu_{qp}^k$
 - 5: **end for**
-

The SQP framework can be extended to the general nonlinear programming problem by linearizing both the inequality and equality constraints

$$\begin{aligned} \min_{d \in \mathbb{R}^n} \quad & q_k(d) = f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d \\ \text{s.t.} \quad & g(x^k) + \nabla g(x^k)^T d \leq 0 \\ & h(x^k) + \nabla h(x^k)^T d = 0 \end{aligned} \quad (13.5)$$

with $H_k = \nabla_{xx}^2 \mathcal{L}(x^k, \lambda^k, \mu^k)$.

Theorem 13.2 Suppose that the following conditions are satisfied:

1. The functions f, g, h are twice continuously differentiable
2. Let $H_k = \nabla_{xx}^2 L(x^k, \lambda^k, \mu^k)$
3. $(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$ is a KKT triple of (11.1)
4. Strict complementary conditions: $g_i(\tilde{x}) = 0 \Rightarrow \tilde{\lambda}_i > 0 \forall i \in \mathcal{G}$

Algorithm (16) Local SQP Method III**Input:** Starting point $x^0 \in \mathbb{R}^n$, $\lambda^0 \in \mathbb{R}^m$ and $\mu^0 \in \mathbb{R}^p$.

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: STOP, if (x^k, λ^k, μ^k) is a KKT triple of (11.1).
- 3: Compute the solution d^k of (13.5) with corresponding multipliers $\lambda_{qp}^k, \mu_{qp}^k$.
- 4: Set $x^{k+1} = x^k + d^k$, $\lambda^{k+1} = \lambda_{qp}^k$ and $\mu^{k+1} = \mu_{qp}^k$.
- 5: **end for**

5. Regularity: $(\nabla g_{\mathcal{A}(\tilde{x})}(\tilde{x}), \nabla h(\tilde{x}))$ has full column rank

6. Sufficient second order optimality conditions:

$$s^T \nabla_{xx}^2 L(\tilde{x}, \tilde{\lambda}, \tilde{\mu}) s > 0 \quad \forall s \in \mathbb{R}^n \setminus \{0\} \text{ with } \nabla g_{\mathcal{A}(\tilde{x})}(\tilde{x})^T s = 0 \text{ and } \nabla h(\tilde{x})^T s = 0$$

7. Among all KKT triples $(d^k, \lambda_{qp}^k, \mu_{qp}^k)$ of (13.5), one chooses the triple with the smallest distance

$$\|(x^k + d^k, \lambda_{qp}^k, \mu_{qp}^k) - (x^k, \lambda^k, \mu^k)\|$$

in each iteration 3 of Algorithm (16).

Then, there exists $\delta > 0$, such that, for all $(x^0, \lambda^0, \mu^0) \in B_\delta(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$, either Algorithm (16) terminates after a finite number of iterations or generates a sequence (x^k, λ^k, μ^k) , which converges superlinearly to $(\tilde{x}, \tilde{\lambda}, \tilde{\mu})$. If $\nabla^2 f$, $\nabla^2 g_i$ (for all $i \in \mathcal{G}$), $\nabla^2 h_j$ (for all $j \in \mathcal{H}$) are Lipschitz continuous in $B_\delta(\tilde{x})$, then the rate of convergence will be quadratic.

Proof: (idea) Apply Newton's method to the system

$$F(x, \lambda, \mu) := \begin{pmatrix} \nabla_x \mathcal{L}(x, \lambda, \mu) \\ \lambda_{\mathcal{I}(\tilde{x})} \\ g_{\mathcal{A}(\tilde{x})}(x) \\ h(x) \end{pmatrix} = 0.$$

and show that the iterates generated by Newton's method applied to $F(x, \lambda, \mu)$ and the iterates generated by the SQP method are identical.

To be practical, an SQP method must be able to converge from remote starting points and on nonconvex problems. We now outline how the local SQP strategy can be adapted to meet these goals. We consider the l_1 -penalty function

$$P_\alpha^1(x) = f(x) + \alpha(\|g(x)^+\|_1 + \|h(x)\|_1)$$


and the Armijo rule. The l_1 -penalty function is not differentiable. However, directional derivatives exist provided that f, g, h are continuously differentiable.

Definition 13.3 The continuous function $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$ is called *directionally differentiable* at $x \in \mathbb{R}^n$, if the directional derivative exists for all $d \in \mathbb{R}^n$:

$$D_+ \phi(x; d) := \lim_{t \rightarrow 0^+} \frac{\phi(x + td) - \phi(x)}{t} \in \mathbb{R}.$$

Example 13.4 Consider the function

$$f(x_1, x_2) = \begin{cases} \frac{x_1^3}{x_1^2 + x_2^2} & \text{for } (x_1, x_2) \neq (0, 0) \\ 0 & \text{for } (x_1, x_2) = (0, 0) \end{cases}$$

which is not differentiable but directionally differentiable at the origin.  Exercise. △

Theorem 13.5 Let f, g, h be continuously differentiable and $\alpha > 0$. Then, P_α^1 is directionally differentiable at each point $x \in \mathbb{R}^n$ with

$$\begin{aligned} D_+ P_\alpha^1(x; d) = & \nabla f(x)^T d + \alpha \sum_{g_i(x) > 0} \nabla g_i(x)^T d + \alpha \sum_{g_i(x) = 0} (\nabla g_i(x)^T d)^+ \\ & + \alpha \sum_{h_j(x) > 0} \nabla h_j(x)^T d - \alpha \sum_{h_j(x) < 0} \nabla h_j(x)^T d + \alpha \sum_{h_j(x) = 0} |\nabla h_j(x)^T d|. \end{aligned}$$

Proof. (idea) Apply the definition of directional derivative to every summand of P_α^1 , then use continuity of g and h (e.g. if $g_i(\bar{x}) < 0$, then $g_i(x)^+ = 0$ in an environment around \bar{x}). □

Theorem 13.6 Let f, g, h be continuously differentiable and $(d^k, \lambda_{qp}^k, \mu_{qp}^k)$ be a KKT triple of (13.5). Further assume that

$$\alpha \geq \max\{(\lambda_{qp}^k)_1, \dots, (\lambda_{qp}^k)_m, |(\mu_{qp}^k)_1|, \dots, |(\mu_{qp}^k)_p|\}.$$

Then, it holds true that

$$D_+ P_\alpha^1(x^k; d^k) \leq -d^{kT} H_k d^k.$$

In particular, d^k will be a descent direction, if H_k is positive definite.

We can use these results to formulate a globalized version of the SQP method:

Algorithm (17) Globalized SQP Method

Input: Starting point $x^0 \in \mathbb{R}^n$, $\lambda^0 \in \mathbb{R}^m$, $\mu^0 \in \mathbb{R}^p$, a symmetric matrix $H_0 \in \mathbb{R}^{n \times n}$, $\alpha > 0$ sufficiently large, $0 < \gamma < 1/2$.

- 1: **for** $k = 0, 1, 2, \dots$ **do**
- 2: STOP, if (x^k, λ^k, μ^k) is a KKT triple of (11.1)
- 3: Compute the solution d^k of (13.5) with corresponding multipliers $\lambda_{qp}^k, \mu_{qp}^k$
- 4: Compute the largest $\sigma_k \in \{1, 2^{-1}, 2^{-2}, \dots\}$ such that

$$P_\alpha^1(x^k + \sigma_k d^k) - P_\alpha^1(x^k) \leq \gamma \sigma_k D_+ P_\alpha^1(x^k, d^k)$$

- 5: Set $x^{k+1} = x^k + \sigma_k d^k$.
 - 6: Compute the multipliers λ^{k+1} and μ^{k+1} (e.g., $\lambda^{k+1} = \lambda_{qp}^k$ and $\mu^{k+1} = \mu_{qp}^k$)
 - 7: Compute the symmetric matrix $H_{k+1} \in \mathbb{R}^{n \times n}$
 - 8: **end for**
-

Remark

- In Algorithm (17), the exact penalty function P_α^1 is used as so-called *merit function*.

- In practical implementations, also $\alpha = \alpha_k$ will be adjusted during the computations.

△

Damped BFGS updates

The Hessian approximation H_k can be updated using (modified) BFGS updates that ensure the QN condition

$$H_{k+1}d^k = y^k$$

with $d^k := x^{k+1} - x^k$ and $y^k := \nabla_x \mathcal{L}(x^{k+1}, \lambda^k, \mu^k) - \nabla_x \mathcal{L}(x^k, \lambda^k, \mu^k)$.

To ensure $(d^k)^\top y^k > 0$ and thus positive definiteness of H_{k+1} , one uses the *damped BFGS update*:

$$H_{k+1} := \mathcal{H}^{BFGS}(H_k, d^k, y_{mod}^k)$$

with modified y^k :

$$y_{mod}^k := \theta_k y^k + (1 - \theta_k) H_k d^k$$

and

$$\theta_k := \begin{cases} 1 & \text{if } (d^k)^\top y^k \geq 0.2 \cdot (d^k)^\top H_k d^k \\ \frac{0.8 \cdot (d^k)^\top H_k d^k}{(d^k)^\top H_k d^k - (d^k)^\top y^k} & \text{otherwise} \end{cases}$$

The above damped BFGS update is frequently used in SQP methods and can be shown to result in superlinear convergence under appropriate conditions.

The Maratos Effect

The globalization of the SQP method in Algorithm (17) uses an approximate line search on the exact penalty function P_α^1 (in line 4).

Every globalization method that relies on reducing such a *merit function* suffers from the Maratos effect: Many good search directions are unacceptable for the constrained problem as they may increase both objective value and constraint violation.

Consider the minimization problem on the unit circle:

$$\begin{aligned} \min \quad & f(x) = 2 \cdot (x_1^2 + x_2^2 - 1) - x_1 \\ \text{s.t.} \quad & h(x) = x_1^2 + x_2^2 - 1 = 0 \end{aligned} \tag{13.6}$$

with solution $(\tilde{x}, \tilde{\mu}) = ((1, 0)^\top, -\frac{3}{2})$ and $\nabla_{xx}^2(\tilde{x}, \tilde{\mu}) = I$.

Every point $x^k = (\cos \theta, \sin \theta)^\top$ is feasible for (13.6) for arbitrary θ . Assume that a minimization algorithm determines a search direction

$$d^k = \begin{pmatrix} \sin^2 \theta \\ -\sin \theta \cdot \cos \theta \end{pmatrix}$$

and the trial point

$$x^{k+1} = x^k + d^k = \begin{pmatrix} \cos \theta + \sin^2 \theta \\ \sin \theta \cdot (1 - \cos \theta) \end{pmatrix}$$

for the next step. Provided $\sin \theta \neq 0$, one can show by trigonometric transformations:

$$\begin{aligned} \|x^{k+1} - \tilde{x}\|_2 &= \|x^k + d^k - \tilde{x}\|_2 = 4 \cdot \sin^2(\theta/2) \\ \|x^k - \tilde{x}\|_2 &= 2 \cdot |\sin(\theta/2)| \end{aligned}$$

and thus $\frac{\|x^{k+1} - \bar{x}\|_2}{\|x^k - \bar{x}\|_2} = \frac{1}{2}$, i.e. the above direction is consistent with quadratic convergence.

However, since $f(x^{k+1}) = \sin^2 \theta - \cos \theta > -\cos \theta = f(x^k)$ and $h(x^{k+1}) = \sin^2 \theta > 0 = h(x^k)$ both the objective value and the constraint violation would increase in this direction, i.e. the merit function P_α^1 would increase and thus this search direction cannot be accepted.

Remark Remedies to avoid the Maratos effect:

- Fletcher's Augmented Lagrangian as merit function (computationally costly)
- Watchdog techniques: Allow increase in merit function if that leads to significant progress in further iterations.
- Additional Second Order Correction (SOC) step

△

Second order correction (SOC)

$$d_{SOC}^k := -\nabla h(x^k) \cdot \left(\nabla h(x^k)^\top \nabla h(x^k) \right)^{-1} \cdot h(x^k + d^k)$$

One can show that the SOC step is small compared to d^k , namely $\|d_{SOC}^k\| = O(\|d^k\|^2)$, but drastically improves feasibility:

$$\|h(x^k + d^k + d_{SOC}^k)\| = O(\|d^k\|^3) \quad \text{instead of} \quad \|h(x^k + d^k)\| = O(\|d^k\|^2)$$

and, under mild conditions, the step $d^k + d_{SOC}^k$ can be shown to fulfill the Armijo condition with step size $\sigma_k = 1$ in the area of fast contraction. Since the correction is small, also the fast convergence is maintained.

14. Quadratic Optimization Problems

An optimization problem with a quadratic objective function and linear constraints is called a quadratic program. Problems of this type are important in their own right, and they also arise as subproblems in methods for general constrained optimization, such as sequential quadratic programming. The general quadratic program (QP) can be stated as

$$\begin{aligned} \min \quad & q(x) = d + c^\top x + \frac{1}{2}x^\top Hx \\ \text{s.t.} \quad & g(x) := A^\top x + \alpha \leq 0 \\ & h(x) := B^\top x + \beta = 0 \end{aligned} \tag{QP}$$

with $d \in \mathbb{R}$, $c \in \mathbb{R}^n$, $H = H^\top \in \mathbb{R}^{n \times n}$, $A \in \mathbb{R}^{n \times m}$, $\alpha \in \mathbb{R}^m$, $B \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^p$. We focus on the case that H is positive definite. Then, (QP) is a convex optimization problem and the KKT conditions are necessary and sufficient.

We begin our discussion of algorithms for quadratic programming by considering the case where only equality constraints are present. Then, the solution is given by the solution of the following linear system

$$\begin{pmatrix} H & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} \tilde{x} \\ \tilde{\mu} \end{pmatrix} = \begin{pmatrix} -c \\ -\beta \end{pmatrix}.$$

To solve general quadratic optimization problems, we consider an active set strategy, which solves an equality constrained QP at each iteration

$$\begin{aligned} \min \quad & q(x) \\ \text{s.t.} \quad & A_{\mathcal{A}_k}^\top x + \alpha_{\mathcal{A}_k} = 0 \\ & B^\top x + \beta = 0 \end{aligned} \tag{QP}_k$$

where $A_{\mathcal{A}_k}$ contains the columns a_i of $A = (a_1, \dots, a_m)$ with the indices $i \in \mathcal{A}_k$, where $\mathcal{A}_k \subseteq \mathcal{A}(x^k)$, i.e. we consider a subset of the active constraints for the subproblems $(QP)_k$. This idea gives rise to the algorithm (18) (active set strategy), whose properties are collected in theorem 14.1.

Theorem 14.1 (Properties of the active set strategy)

1. The point x^k is feasible for $(QP)_k$ and for (QP), for all k .
2. If $d^k \neq 0$ and \hat{x}^{k+1} is not feasible for (QP), then there exists the step size σ_k (line 20) and the index j (line 22), and it holds:

$$\sigma_k = \min \left\{ -\frac{a_i^\top x^k + \alpha_i}{a_i^\top d^k} : i \in \mathcal{I}_k, a_i^\top d^k > 0 \right\} \in [0, 1).$$

3. If $d^k \neq 0$ and $\lambda^{k+1} \geq 0$, then $(x^{k+1}, \lambda^{k+1}, \mu^{k+1})$ is a KKT triple of (QP).

Algorithm (18) Active Set Strategy**Input:** Feasible starting point $x^0 \in \mathbb{R}^n$ for (QP).

```

1:  $\mathcal{A}_0 = \mathcal{A}(x^0)$ 
2: for  $k = 0, 1, 2, \dots$  do
3:    $\mathcal{I}_k := \mathcal{G} \setminus \mathcal{A}_k$  ;  $\lambda_{\mathcal{I}_k}^{k+1} := 0$ 
4:   Compute KKT triple  $(\hat{x}^{k+1}, \lambda_{\mathcal{A}_k}^{k+1}, \mu^{k+1})$  of  $(QP_k)$ 
5:    $d^k := \hat{x}^{k+1} - x^k$ 
6:   if  $d^k = 0$  then
7:     if  $\lambda^{k+1} \geq 0$  then
8:        $x^{k+1} := x^k$ 
9:       stop with KKT triple  $(x^{k+1}, \lambda^{k+1}, \mu^{k+1})$  of (QP)
10:    else
11:       $j := \arg \min_{i \in \mathcal{A}_k} \lambda_i^{k+1}$ 
12:       $x^{k+1} := x^k$  ;  $\mathcal{A}_{k+1} := \mathcal{A}_k \setminus \{j\}$ 
13:      continue
14:    end if
15:    else
16:      if  $\hat{x}^{k+1}$  is feasible for (QP) then
17:         $x^{k+1} := \hat{x}^{k+1}$  ;  $\mathcal{A}_{k+1} := \mathcal{A}_k$ 
18:        continue
19:      else
20:         $\sigma_k := \max\{\sigma \geq 0 : x^k + \sigma d^k \text{ feasible for (QP)}\}$ 
21:         $x^{k+1} := x^k + \sigma_k d^k$ 
22:        determine index  $j \in \mathcal{I}_k$  with  $a_j^\top x^{k+1} + \alpha_j = 0$ 
23:         $\mathcal{A}_{k+1} := \mathcal{A}_k \cup \{j\}$ 
24:      end if
25:    end if
26:  end for

```

4. If $d^k \neq 0$, then $\nabla q(x^k)^\top d^k < 0$, i.e. d^k is a descent direction for q at x^k . In particular, it holds: $q(x^{k+1}) < q(x^k)$ if $x^{k+1} \neq x^k$.
5. For every x^k generated by Algorithm (18), there exists $l \geq k$, such that x^l is the unique global solution of (QP_l) .
6. If the algorithm does not terminate after a finite number of iterations, then there exists $l \geq 0$ with $x^k = x^l$ for all $k \geq l$.
7. If the columns of the matrix $(A_{\mathcal{A}_k}, B)$ are linearly independent, then the columns of the matrix $(A_{\mathcal{A}_{k+1}}, B)$ are linearly independent, too.

Remark The case (6) of Theorem 14.1 is called *cycling*: The algorithm stalls at iterate x^l and then the active set changes in a periodic way. For anti-cycling strategies, see, e.g., the excellent textbook of Nocedal & Wright, Chapter 16.5. \triangle