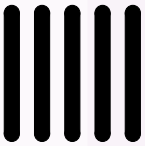
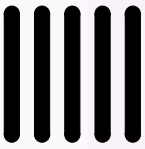


Optimization Theory and Algorithms
CIE / DDA 6010 – Fall 2020/21

Lecture Notes



Andre Milzarek
The Chinese University of Hong Kong, Shenzhen
SDS · School of Data Science

Contents

Remarks and Versions

These full lectures notes are designed for the course "CIE / DDA 6010: *Optimization Theory and Algorithms*" and are under development in the fall term 2020/21. The full notes are meant to complement and summarize the course materials including the lecture slides, weekly notes, and the notes made during the lectures. The notes are mainly based on the textbooks:

- A. Beck: *Introduction to Nonlinear Optimization*, Theory, Algorithms, and Applications with MATLAB. Vol. 19 of MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2014.
- D. P. Bertsekas: *Nonlinear Programming*. Athena Scientific, third ed., Belmont, 2016.
- M. Ulbrich and S. Ulbrich: *Nichtlineare Optimierung*. Birkhäuser, Basel, 2012.

Acknowledgements. At this point, let me thank Wenqing Ouyang for providing the example in Remark 5.28 and for valuable feedback.

Comments. Proofs followed by a purple square "■" have been discussed during the classes. Proofs followed by a red square "■" have not been presented in detail during class and are included for extended reading.

Version. This is version v1.2 (date: 2021/01/18).

1. Introduction to Optimization and Mathematical Review

1.1. Optimization: Basics, Examples, and Modeling

1.2. Mathematical Background

In this and in the following subsections, we briefly introduce and repeat the necessary mathematical background that will be required and utilized in the forthcoming lectures.

1.2.1. Vectors and Norms

A *point* or *vector* $x \in \mathbb{R}^n$ is represented by a column vector

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{with components} \quad x_i \in \mathbb{R}, \quad \forall\, i.$$

The *transposed* $x^\top$ of a vector $x \in \mathbb{R}^n$ is the row vector $x^\top = (x_1, x_2, \cdots, x_n)$.

For $x, y \in \mathbb{R}^n$, we define the *Euclidean inner product* by

$$\langle x, y \rangle := \sum_{i=1}^{n} x_i y_i = x^\top y.$$

The inner product is a function $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with the following properties:

- Positivity: $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ if and only if $x = 0$,
- Symmetry: $\langle x, y \rangle = \langle y, x \rangle$,
- Linearity: $\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle$,

for all $\alpha, \beta \in \mathbb{R}$ and $x, y, z \in \mathbb{R}^n$. It is possible to define other functions on $\mathbb{R}^n \times \mathbb{R}^n$ that satisfy the latter properties and many results involving the Euclidean inner product also hold for these other forms of inner products.

Inner products allow to measure angles and distances between vectors. Two vectors x and y are called *orthogonal* if $\langle x, y \rangle = 0$.

The *Euclidean norm* of a point $x \in \mathbb{R}^n$ is defined as:

$$\|x\|_2 = \|x\| = \sqrt{\langle x, x \rangle} = \sqrt{x^\top x} = \sqrt{x_1^2 + x_2^2 + ... + x_n^2}.$$

The Euclidean norm satisfies the following properties for all $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$:

- Positivity: $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$.
- Homogeneity: $\|\alpha x\| = |\alpha| \|x\|$.
- Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$.

- Cauchy-Schwarz inequality: $|\langle x, y\rangle| \leq \|x\|\|y\|$. Equality holds if and only if x and y are parallel (i.e., there exists $\lambda \in \mathbb{R}$ such that $x = \lambda y$).

We briefly verify the last (important) property:

Proof. If $x = 0$ or $y = 0$, the inequality is obviously satisfied. Thus, we can assume $x, y \neq 0$ without loss of generality. Let us set $\tilde{x} = x/\|x\|$ and $\tilde{y} = y/\|y\|$, then we have

$$0 \leq \|\tilde{x} - \tilde{y}\|^2 = \langle \tilde{x} - \tilde{y}, \tilde{x} - \tilde{y}\rangle = \|\tilde{x}\|^2 - 2\langle \tilde{x}, \tilde{y}\rangle + \|\tilde{y}\|^2 = 2 - 2\langle \tilde{x}, \tilde{y}\rangle,$$

which implies $\langle \tilde{x}, \tilde{y}\rangle \leq 1$ and $\langle x, y\rangle \leq \|x\|\|y\|$. Furthermore, equality only holds in the case $\tilde{x} = \tilde{y}$. We now repeat the same calculation with $\tilde{x} = -x/\|x\|$ which yields $-\langle x, y\rangle \leq \|x\|\|y\|$. This establishes the Cauchy-Schwarz inequality. Equality is only satisfied if $x/\|x\| = \pm y/\|y\|$, i.e., if $x = \lambda y$ for some $\lambda \in \mathbb{R}$. ■

A mapping $r : \mathbb{R}^n \to \mathbb{R}$ satisfying the first three mentioned properties is said to be a *norm* on $\mathbb{R}^n$. There are norms that can not be directly derived from an underlying inner product. The following functions are popular examples of norms:

- ℓ_1-norm: $\|x\|_1 = \sum_{i=1}^n |x_i|$.
- Euclidean norm: $\|x\|_2 = \left[\sum_{i=1}^n |x_i|^2\right]^{1/2}$.
- ℓ_p-norm or p-norm: $\|x\|_p = \left[\sum_{i=1}^n |x_i|^p\right]^{1/p}$ where $p \in [1, \infty)$.
- Maximum norm: $\|x\|_\infty = \max_{1\leq i\leq n} |x_i|$.
- Weighted p-norm: $\|x\|_W = \|Wx\|_p$ where $W \in \mathbb{R}^{n\times n}$ is a fixed nonsingular matrix and $p \in [1, \infty)$.

The notation $\|\cdot\|$ is often used to express a general norm function. Here – if not otherwise stated – we will always use the convention $\|\cdot\| = \|\cdot\|_2$.

In $\mathbb{R}^n$, it can be shown that all norms are *equivalent*. In particular, let $\|\cdot\|$ and $|||\cdot|||$ be two different norms on $\mathbb{R}^n$. Then there exist constants $c, C > 0$ such that

$$c \cdot |||x||| \leq \|x\| \leq C \cdot |||x|||, \quad \forall\, x \in \mathbb{R}^n.$$

The constants c, C typically depend on the dimension n.

1.2.2. Matrices and Definiteness

A *matrix* $A \in \mathbb{R}^{m\times n}$ is represented as follows

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \ddots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad \text{with components} \quad a_{ij} \in \mathbb{R}, \quad \forall\, i, j.$$

The transposed matrix $A^\top \in \mathbb{R}^{n\times m}$ is defined by swapping columns with rows of A. In particular, the i-th column of $A^\top$ is set to be the i-th row of A. We continue with several basic notations and terminologies:

- A matrix is called *square* if $m = n$.
- A square matrix A is called *symmetric* if $A^\top = A$.
- The *identity matrix* is the square matrix I with $I_{ii} = 1$ for all i and $I_{ij} = 0$ for $i \neq j$.
- A square matrix $A \in \mathbb{R}^{n\times n}$ is called *invertible* if there is $B \in \mathbb{R}^{n\times n}$ with $AB = I$. We then write $B = A^{-1}$.

The inner product allows to measure the angles between vectors. Specifically, two vectors x and y are called *orthogonal* if $\langle x, y\rangle = 0$. A matrix A is called *orthogonal* if all of its columns are pair-wise orthogonal to each other. A is called *orthonormal* if $A^\top A = I$. The *trace* of a square matrix $A \in \mathbb{R}^{n\times n}$ is defined by

$$\mathrm{tr}(A) = a_{11} + a_{22} + ... + a_{nn}.$$

In the space of $\mathbb{R}^{m\times n}$ matrices, the *inner product* is then given by:

$$\langle A, B\rangle = \mathrm{tr}(A^\top B) = \mathrm{tr}(AB^\top) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ij}.$$

We next introduce or repeat the notion of a positive (negative) semidefinite matrix and a positive (negative) definite matrix. These concepts will be employed in the next chapter to study optimality conditions.

Definition 1.1: Definiteness
Let A be a real $n \times n$ matrix.

(i) A is said to be **positive semidefinite** if $x^\top Ax \geq 0$ for all $x \in \mathbb{R}^n$.

(ii) A is said to be **positive definite** if $x^\top Ax > 0$ for all $x \in \mathbb{R}^n\backslash\{0\}$.

(iii) A is said to be **negative semidefinite** if $x^\top Ax \leq 0$ for all $x \in \mathbb{R}^n$.

(iv) A is said to be **negative definite** if $x^\top Ax < 0$ for all $x \in \mathbb{R}^n\backslash\{0\}$.

(v) A is **indefinite** if it is neither positive semidefinite nor negative semidefinite.

If the matrix A is additionally symmetric, then we can connect definiteness with positivity of the eigenvalues of A. In particular, let A be a real symmetric $n \times n$ matrix. Then, there is an *eigenvalue decomposition* of A

$$A = Q\Lambda Q^\top$$

where $Q \in \mathbb{R}^{n\times n}$ is an orthogonal matrix, $Q^\top Q = QQ^\top = I$, and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$ is a diagonal matrix. The numbers $\lambda_i \in \mathbb{R}$ are called *eigenvalues* of A. We can now formulate the following theorem.

Theorem 1.2: Eigenvalues and Definiteness

Let $A \in \mathbb{R}^{n\times n}$ be symmetric with eigenvalues λ_i, $i = 1, ..., n$. Then, it holds that:

(i) A is positive semidefinite if and only if $\lambda_i \geq 0$ for all i.

(ii) A is positive definite if and only if $\lambda_i > 0$ for all i.

(iii) A is negative semidefinite if and only if $\lambda_i \leq 0$ for all i.

(iv) A is negative definite if and only if $\lambda_i < 0$ for all i.

(v) A is indefinite if and only if there are $i, j \in \{1, ..., n\}$ with $\lambda_i > 0$ and $\lambda_j < 0$.

Example 1.3. Let us consider the matrix

$$A = \begin{pmatrix} 2 & -1 \\ 0 & 3 \end{pmatrix}$$

Then, we have

$$x^\top Ax = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1^2 - 2x_1x_2 + 3x_2^2 = (x_1 - x_2)^2 + 2x_2^2 > 0$$

for all $x \neq 0$. Hence, A is positive definite.

Example 1.4. Let us consider the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

Then, it follows $\mathrm{tr}(A) = 1 + 1 = 2 = \lambda_1 + \lambda_2$ and $\det(A) = 1 - 4 = -3 = \lambda_1 \cdot \lambda_2$. As a consequence, one of the eigenvalues needs to be positive while the other one has to be negative. This implies that A is indefinite.

We now collect further useful terminologies and connections between eigenvalues of definiteness of matrices.

- We write $A \succeq 0$ if A is positive semidefinite and symmetric. We write $A \succ 0$ if A is a positive definite, symmetric matrix.
- Let $A \in \mathbb{R}^{n\times n}$ be given with eigenvalues $\lambda_1, ..., \lambda_n$, then:

$$\mathrm{tr}(A) = \sum\nolimits_{i=1}^{n} \lambda_i, \quad \det(A) = \prod\nolimits_{i=1}^{n} \lambda_i.$$

- Let A be a symmetric 2×2 matrix. Then, A is positive (semi)definite if and only if

$$\mathrm{tr}(A)\ (\geq) > 0 \quad \text{and} \quad \det(A)\ (\geq) > 0.$$

- A symmetric matrix A is positive definite if and only if all *leading principal minors* of A have positive determinant.

The next lemma shows that the terms $x^\top Ax$ can be bounded in terms of the minimum and maximum eigenvalue of A times $\|x\|^2$. This is a rather helpful estimate which will be used throughout the lecture.

Lemma 1.5
Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Then, it holds that

$$\lambda_{\min}(A)\|x\|^2 \leq x^\top Ax \leq \lambda_{\max}(A)\|x\|^2, \quad \forall\, x \in \mathbb{R}^n,$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalue of A, respectively.

Proof. Since A is symmetric, we can write $A = Q\Lambda Q^\top$ where $Q \in \mathbb{R}^{n\times n}$ is orthogonal and Λ is a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$. Setting $y = Q^\top x$, it holds that

$$\begin{aligned} x^\top Ax = x^\top Q\Lambda Q^\top x = y^\top \Lambda y &= \sum_{i=1}^n \lambda_i y_i^2 \\ &\leq \max_{1\leq i\leq n} \lambda_i \cdot \|y\|^2 = \lambda_{\max}(A)\cdot x^\top QQ^\top x = \lambda_{\max}(A)\cdot \|x\|^2, \end{aligned}$$

where we used the orthogonality of Q in the last step. The lower bound can be shown in a similar fashion. ■

If the matrix A is non-square then we can not directly work with eigenvalue decomposition. However, by discussing eigenvalues of $AA^\top$ and $A^\top A$ a similar concept – the so-called *singular value decomposition* (SVD) – can be derived. The SVD of a non-square matrix $A = \mathbb{R}^{m\times n}$ is given by

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

where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthonormal and $\Sigma \in \mathbb{R}^{m\times n}$ is a non-square diagonal matrix. The main diagonal elements of Σ – say $\sigma_1, \sigma_2, \ldots, \sigma_r$ – with $r = \min\{m, n\}$ are the *singular values* of A. The singular values are the square-roots of the eigenvalues of $A^\top A$. Notice that the *rank* of A is equal to the number of nonzero singular values.

We finally mention some different matrix norms for non-square matrices $A \in \mathbb{R}^{m\times n}$:

- *Frobenius norm:* $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\mathrm{tr}(A^\top A)} = \sqrt{\sum_i \sigma_i^2}$.
- *Nuclear norm:* $\|A\|_* = \sum_i \sigma_i$.
- *Spectral norm:* $\|A\|_2 = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \max_i \sigma_i$.

Similar to the vector case, the matrix norms are equivalent to each other and some of the matrix are compatible with vector norms. For instance, the following useful inequalities hold:

$$\|Ax\| \leq \|A\|_2\|x\|, \quad \|A\|_* \geq \|A\|_F \geq \|A\|_2, \quad \langle A, B\rangle \leq \|A\|_F\|B\|_F.$$

Notice that the last inequality is a version of the Cauchy-Schwartz inequality for matrix spaces.

1.2.3. Sets and Set Operations

The *supremum* of a nonempty set $X \subset \mathbb{R}$ is the smallest scalar y such that

$$y \geq x \quad \text{for all} \quad x \in X.$$

The *infimum* of a set $X \subset \mathbb{R}$ is the largest scalar y such that

$$y \leq x \quad \text{for all} \quad x \in X.$$

In the case $\sup X \in X$ ($\inf X \in X$), we write $\sup X = \max X$ ($\inf X = \min X$). We now briefly consider an example to illustrate the definition of sup and inf. The supremum and infimum of the set $\{\frac{1}{n} : n \geq 1\}$ are given by

$$\sup\{1/n : n \geq 1\} = \max\{1/n : n \geq 1\} = 1, \quad \inf\{1/n : n \geq 1\} = 0.$$

For $\epsilon > 0$ and $x \in \mathbb{R}^n$ we define $B_\epsilon(x) = \{y \in \mathbb{R}^n : \|x - y\| < \epsilon\}$ to be open ball with radius ϵ and center x. Next, we collect further properties and terminologies for sets:

- A set $X \subset \mathbb{R}^n$ is called *open* if for every $x \in X$ there exists $\epsilon > 0$ such that $B_\epsilon(x) \subset X$.
- A set $X \subset \mathbb{R}^n$ is *closed* if $\mathbb{R}^n \backslash X$ is open. Alternatively, we can define closedness of set as follows: For every sequence (x^k) with $x^k \in X$ for all k and $x^k \to x$, we have $x \in X$.
- A set $X \subset \mathbb{R}^n$ is *bounded* if there exists $B \in \mathbb{R}$ with $\|x\| \leq B$ for all $x \in X$.
- A bounded and closed set is called *compact*.

1.2.4. Sequences

A *sequence* $(x^k)_{k\in\mathbb{N}}$ is a list of enumerated items

$$x^1, x^2, x^3, \cdots, x^k, x^{k+1}, \cdots,$$

where each element or item x^k can be a real number $x^k \in \mathbb{R}$ or an arbitrary vector $x^k \in \mathbb{R}^n$. A sequence can be interpreted as the *image* of a function $a : \mathbb{N} \to \mathbb{R}^n$, i.e., we have $a(k) = x^k$ for all $k \in \mathbb{N}$.

Some simple examples:

$$1, 2, 3, \cdots \rightsquigarrow x^k = k; \qquad 1, 4, 9, \cdots \rightsquigarrow x^k = k^2; \qquad 1, \frac{1}{2}, \frac{1}{3}, \cdots \rightsquigarrow x^k = \frac{1}{k}.$$

Sequences are an essential and helpful tool that form the basis of many mathematical concepts. A goal of this course is to design algorithmic approaches that solve optimization problems iteratively. In particular, those methods generate a *sequence* of iterates that ideally approaches a (local or global) solution of the problem.

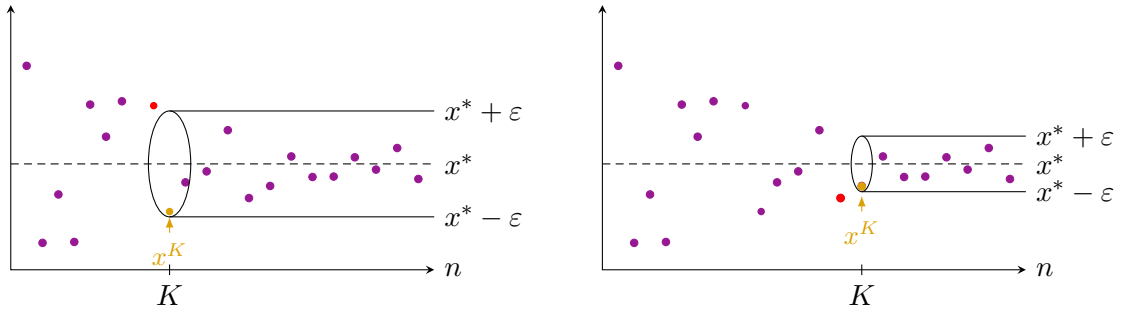


Figure 1.1: Visualization of the convergence of a sequence $(x^k)_k$ in $\mathbb{R}$.

Definition 1.6: Convergence of Sequences

A point $x^* \in \mathbb{R}^n$ is called the **limit** of the sequence $(x^k)_{k\in\mathbb{N}}$ if for all $\varepsilon > 0$ there exists $K \in \mathbb{N}$ (which may depend on ε) such that

$$\|x^k - x^*\| \leq \varepsilon, \quad \forall\, k \geq K. \tag{1.1}$$

In this case, we write $x^* = \lim_{k\to\infty} x^k$ or $x^k \to x^*$ and we say that $(x^k)_{k\in\mathbb{N}}$ is a **convergent sequence** that converges to x^*.

Let us notice that the condition (1.1) reduces to $x^* - \varepsilon \leq x^k \leq x^* + \varepsilon$ (for all $k \geq K$) in the case $n = 1$. This situation is illustrated in Figure 1.1. We summarize several more definitions and properties related to convergent sequences:

- A convergent sequence has only one (unique) limit.
- A sequence $(x^k)_k$ in $\mathbb{R}^n$ is *bounded* if there is $B \geq 0$ such that $\|x^k\| \leq B$ for all k.

  Every convergent sequence is bounded. Conversely, a bounded sequence does not need to be convergent.
- A sequence $(x^k)_k$ in $\mathbb{R}$ is *nondecreasing* (*increasing*) if we have $x^k \leq x^{k+1}$ ($x^k < x^{k+1}$) for all $k \in \mathbb{N}$. Similarly, $(x^k)_k$ is said to be *nonincreasing* (*decreasing*) if $x^{k+1} \leq x^k$ ($x^{k+1} < x^k$). Nonincreasing and nondecreasing sequences are so-called *monotone sequences*.

  Every monotone and bounded sequence in $\mathbb{R}$ converges.

Example 1.7. Let us consider the sequence $(x^k)_k$ with $x^k = 1/k$ for all k. In order to show convergence of this sequence using Definition 1.6, we first need to find a suitable candidate for the limit x^*. By calculating the first elements of $(x^k)_k$ or by sketching the sequence, we can guess the candidate $x^* = 0$. Let $\varepsilon > 0$ be arbitrary. We need to find $K \in \mathbb{N}$ such that

$$|x^k - x^*| = \frac{1}{|k|} = \frac{1}{k} \leq \varepsilon, \quad \forall\, k \geq K.$$

Rearranging the terms, we can obviously choose $K := 1/\varepsilon$. Hence, $(x^k)_k$ converges to zero.

Example 1.8. The sequence $(x^k)_k$, $x^k = (-1)^k$ is an example of a bounded sequence (with $B = 1$) that does not converge.

1.2.5. Continuous Functions

Definition 1.9: Continuous Function
A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called **continuous** at a point $x \in \mathbb{R}^n$, if for every convergent sequence $(x^k)_k$ in $\mathbb{R}^n$ with $x^k \to x$, it holds that

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

We say that f is continuous on a set $X \subset \mathbb{R}^n$, if f is continuous at all points $x \in X$.

Let us add the following remarks and additional comments:

- For a function $f : \mathbb{R} \to \mathbb{R}$, continuity at a point x is equivalent to say that the limits from the left $\lim_{y\nearrow x} f(y)$ and right $\lim_{y\searrow x} f(y)$ exist and coincide with $f(x)$.
- Intuitively, a continuous function is a mapping that does not "jump" at a certain point.
- The function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called *Lipschitz continuous* (with constant L), if there exists $L > 0$ such that $\|f(x) - f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^n$.
- Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous. Then, for all $\alpha \in \mathbb{R}$, the *level set* $L_{\leq\alpha} := \{x \in \mathbb{R}^n : f(x) \leq \alpha\}$ is a closed set.

Example 1.10. Let us consider the one-dimensional function $f : \mathbb{R} \to \mathbb{R}$, $f(x) := \sin(1/x)$. Then, f is not continuous at $x = 0$.

Proof. Let us set $x^k = \frac{2}{(2k-1)\pi}$. Then, we obviously have $x^k \to 0$. Moreover, it holds that

$$f(x^k) = \sin\left(\frac{(2k-1)\pi}{2}\right) = \sin\left(\frac{\pi}{2} + (k-1)\pi\right) = (-1)^{(k-1)}.$$

Since $(f(x^k))_k$ does not converge, f can not be continuous at $x = 0$. ■

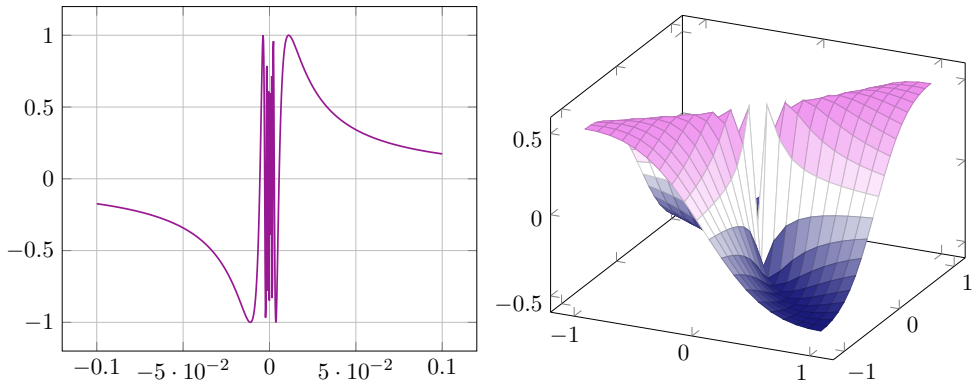
Example 1.11. We consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ with

$$f(x_1, x_2) = \frac{x_1 x_2}{x_1^2 + x_2^2} \quad \text{if} \quad (x_1, x_2) \neq (0, 0) \quad \text{and} \quad f(0, 0) = 0.$$

The function f is continuous on $\mathbb{R}^2\backslash\{0\}$, but for all $m \neq 0$ and $x_1 \neq 0$ we have

$$f(x_1, mx_1) = \frac{m}{1 + m^2} \neq 0.$$

Hence, f does not approach 0 when (x_1, mx_1) converges to $(0, 0)$. This shows that f is not continuous in $(0, 0)$. However, the component-wise functions $f_1 : \mathbb{R} \to \mathbb{R}$, $f_1(x) := f(x, x_2)$ and $f_2 : \mathbb{R} \to \mathbb{R}$, $f_2(x) := f(x_1, x)$ are continuous for all $(x_1, x_2) \in \mathbb{R}^2$!



(a) Example 1.10 (b) Example 1.11

Figure 1.2: Illustration: Discontinuous Functions.

1.2.6. Differentiability

In this section, we briefly review the gradient vector and Hessian matrix of a function f.

Definition 1.12: Differentiability

A mapping $F : \mathbb{R}^n \to \mathbb{R}^m$ is said to be **(Fréchet) differentiable** at a point x if there exists a linear function $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$\lim_{h \to 0} \frac{\|F(x+h) - F(x) - \mathcal{L}(h)\|}{\|h\|} = 0.$$

The function $\mathcal{L}$ is the **derivative** of f at x and it holds that $\mathcal{L}(h) = J_f(x) \cdot h$, where $J_f(x) \in \mathbb{R}^{m \times n}$ is the so-called **Jacobian-Matrix**.

Definition 1.13: Gradient Vector and Jacobian-Matrix

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a given and differentiable at a point x. The **gradient vector** of f at x is the column vector

$$\nabla f(x) = J_f(x)^\top = \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) \end{pmatrix},$$

where $\frac{\partial}{\partial x_i} f$ denotes the **partial derivative** of f at x with respect to x_i. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a vector-valued function with components F_1, ... F_m and suppose that F is differentiable at x. The **Fréchet derivative** or **Jabocian-Matrix** of F at x is given by

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

Example 1.14. In this example, we calculate the gradient of the Euclidean norm $f(x) = \|x\|_2$. Let us first consider $x = 0$. By the definition of differentiability, we need to find a vector $y = \nabla f(0)$ such that

$$\frac{|\|0+h\| - 0 - y^\top h|}{\|h\|} = \left|1 - \frac{y^\top h}{\|h\|}\right| \to 0, \quad \text{as } h \to 0.$$

Since the same condition must also hold for $h \to -h$, such a vector y cannot exist. Hence, f is not differentiable at $x = 0$. Now, consider $x \neq 0$. By the chain rule we have:

$$\frac{\partial f}{\partial x_i}(x) = \frac{\partial}{\partial x_i}\left[\sqrt{x_1^2 + ... + x_{i-1}^2 + x_i^2 + ...x_n^2}\right] = \frac{1}{2\sqrt{x_1^2 + ... + x_n^2}} \cdot 2x_i = \frac{x_i}{\|x\|}.$$

Thus, the gradient ∇f is given by $\nabla f(x) = \frac{x}{\|x\|}$.

Definition 1.15: Hessian matrix
Let $f : \mathbb{R}^n \to \mathbb{R}$ and $x \in \mathbb{R}^n$ be given. The **Hessian** of f at x is the $n \times n$ matrix

$$\nabla^2 f(x) = H_f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1 \partial x_1}(x) & \frac{\partial f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial f}{\partial x_1 \partial x_n}(x) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \frac{\partial f}{\partial x_n \partial x_1}(x) & \frac{\partial f}{\partial x_n \partial x_2}(x) & \cdots & \frac{\partial f}{\partial x_n \partial x_n}(x) \end{pmatrix}.$$

Remarks:

- The (i,j)-th entry of $\nabla^2 f(x)$ is the partial derivative $\frac{\partial f}{\partial x_i \partial x_j}(x)$.
- The Hessian is the derivative of the transposed gradient $\nabla f^\top$. If the gradient is Fréchet differentiable in the sense of Definition 1.12, then the Hessian matrix $\nabla f^2(x)$ is *symmetric*. This is also true if all partial derivatives are continuous.

Example 1.16. Let us consider the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) := \frac{1}{2}x^\top Ax + b^\top x + c$, where $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix and $b \in \mathbb{R}^n$, $c \in \mathbb{R}$ are given. The gradient and Hessian of f can be computed as follows:

$$\nabla f(x) = Ax + b, \quad \nabla^2 f(x) = A.$$

We can use Definition 1.12 directly to verify the last formulae. It holds that

$$\begin{aligned} f(x+h) - f(x) &= \frac{1}{2}(x+h)^\top A(x+h) + b^\top h - \frac{1}{2}x^\top Ax \\ &= \frac{1}{2}x^\top Ah + \frac{1}{2}h^\top Ax + \frac{1}{2}h^\top Ah + b^\top h \\ &= (x^\top A + b^\top)h + \frac{1}{2}h^\top Ah = (Ax+b)^\top h + \frac{1}{2}h^\top Ah, \end{aligned}$$

where we used the symmetry of A and the fact $y = y^\top$ for every real number $y \in \mathbb{R}$. As a consequence, we obtain

$$\lim_{\|h\|\to 0} \frac{|f(x+h) - f(x) - (Ax+b)^\top h|}{\|h\|} = \lim_{\|h\|\to 0} \frac{|h^\top A h|}{2\|h\|} = 0,$$

which establishes $\nabla f(x) = Ax + b$. The Hessian of f can then be computed in the same way.

Next, we present two helpful computational rules.

Theorem 1.17: Chain Rule
Let $F : \mathbb{R}^n \to \mathbb{R}^m$, $G : \mathbb{R}^p \to \mathbb{R}^n$ and $R : \mathbb{R}^p \to \mathbb{R}^m$ be given with $R(x) = F(G(x))$ be given. If F and G are differentiable, then R is also differentiable with

$$DR(x) = DF(G(x)) \cdot DG(x) \in \mathbb{R}^{m\times p}.$$

In the special case $m = 1$, $r(x) := f(G(x))$, the chain rule in Theorem 1.17 reduces to $Dr(x) = \nabla r(x)^\top = \nabla f(G(x))^\top DG(x)$. We now consider the product rule for multivariate functions.

Lemma 1.18: Product Rule
Let $F : \mathbb{R}^n \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}$ be differentiable functions and set $R(x) = F(x) \cdot g(x)$. Then, R is differentiable with

$$DR(x) = DF(x) \cdot g(x) + F(x) \cdot \nabla g(x)^\top \in \mathbb{R}^{n\times n}.$$

Finally, we present Taylor's theorem and the mean value theorem.

Theorem 1.19: Taylor's Theorem
Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Then, it holds that

$$f(x+h) = f(x) + \nabla f(x)^\top h + o(\|h\|) \quad \text{as} \quad h \to 0.$$

In addition, if f is twice continuously differentiable, then we have

$$f(x+h) = f(x) + \nabla f(x)^\top h + \frac{1}{2} h^\top \nabla^2 f(x) h + o(\|h\|^2) \quad \text{as} \quad h \to 0.$$

Theorem 1.20: Mean Value Theorem
Let $f : \mathbb{R}^n \to \mathbb{R}$ be (once or twice) continuously differentiable. Then, for every $x, h \in \mathbb{R}^n$ there exists $t \in [0, 1]$ such that:

- $f(x+h) = f(x) + \nabla f(x+th)^\top h$,
- $f(x+h) = f(x) + \nabla f(x)^\top h + \frac{1}{2} h^\top \nabla^2 f(x+th) h$.

We conclude this section with two additional remarks:

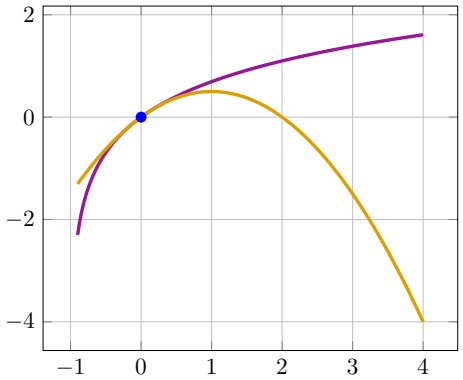


Figure 1.3: Illustration of Example 1.21. The function $f(x) = \log(1+x)$ is plotted in blue. The quadratic approximation $g(x) = x - \frac{1}{2}x^2$ is depicted in red.

- We can obtain another variant of the last theorems by setting $h = y - x$ for some $y \in \mathbb{R}^n$.
- The so-called *Landau-notation* is used to describe the asymptotic behavior of a function in a compact way. In particular, for a function $g : \mathbb{R}^n \to \mathbb{R}$, we write $g(h) = o(\|h\|)$ if and only if $\lim_{h\to 0} g(h)/\|h\| = 0$.

  We write $g(h) = O(\|h\|)$ (for $h \to 0$) if there exist $C, \epsilon > 0$ such that $|g(h)| \leq C\|h\|$ for all $h \in B_\epsilon(0)$.

Example 1.21. Let us consider $f(x) = \log(1+x)$. Then we have $f'(x) = (1+x)^{-1}$ and $f''(x) = -(1+x)^{-2}$. Taylor's theorem implies:

$$f(h) = f(0) + f'(0) \cdot h + \frac{1}{2} f''(0) \cdot h^2 + o(|h|^2) = h - \frac{1}{2}h^2 + o(|h|^2),$$

for $h \to 0$. Taylor's theorem can be used to find *local quadratic approximations* of a function.

References. See [2, Chapter 1 and Section 2.2].

2. Basic Notation: Formal Definition of Minimizer

We consider a general constrained optimization problem of the form:

$$\min_{x\in X} \; f(x), \tag{2.1}$$

where $X \subset \mathbb{R}^n$ is a given feasible set. In the following, we define the notion of local and global minimizer.

Definition 2.1: Local and Global Minimizer
Let $X \subset \mathbb{R}^n$ be a nonempty set and let $f : X \to \mathbb{R}$ be a given mapping. We define $B_\varepsilon(y) := \{x \in \mathbb{R}^n : \|x-y\| < \varepsilon\}$ to be the open ball in $\mathbb{R}^n$ with center y and radius $\varepsilon > 0$.

The point $x^* \in \mathbb{R}^n$ is said to be a:

(a) **local minimizer** of problem (2.1), if $x^* \in X$ and there is $\varepsilon > 0$ such that $f(x) \geq f(x^*)$ for all $x \in X \cap B_\varepsilon(x^*)$.

(b) **strict local minimizer** of the problem (2.1), if $x^* \in X$ and there is $\varepsilon > 0$ with $f(x) > f(x^*)$ for all $x \in (X \cap B_\varepsilon(x^*))\backslash\{x^*\}$.

(c) **global minimizer** of (2.1), if $x^* \in X$ and we have $f(x) \geq f(x^*)$ for all $x \in X$.

(d) **strict global minimizer** of (2.1), if $x^* \in X$ and it holds that $f(x) > f(x^*)$ for all $x \in X\backslash\{x^*\}$.

The definition immediately implies that every (strict) global minimizer or solution is also a (strict) local minimizer of problem (2.1). The converse direction does not hold in general and will be discussed and illustrated in the following examples.

Example 2.2. We consider the unconstrained problem

$$\min_{x\in\mathbb{R}} \; f(x) := x^4 - 9x^2 + 4x - 1.$$

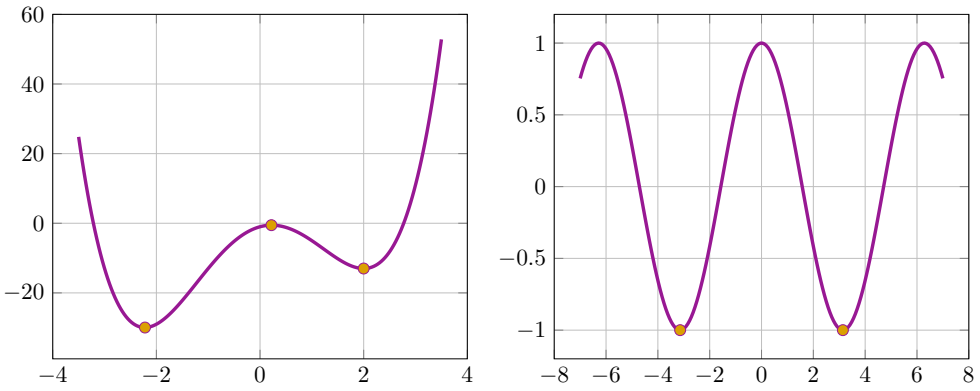
The derivative of f is given by $f'(x) = 2(2x^3 - 9x + 2)$ and it vanishes at $x_1^* = 2$, $x_2^* = -1+0.5\sqrt{6}$, and $x_3^* = -1-0.5\sqrt{6}$. The graph of f is shown in Figure 2.1 (a). We see that x_3^* is a global minimizer and x_1^* is a local minimizer that is not global. (x_2^* is a local maximizer that is not a global maximum).

Example 2.3. In this example, we consider another unconstrained problem with $f(x) = \cos(x)$ and $X = \mathbb{R}$. Apparently, the set of global minima is given by $S := \{x \in \mathbb{R}^n : x = (2k+1)\pi,\, k \in \mathbb{Z}\}$. None of the points $x^* \in S$ is a strict global minimizer.

Finally, let us discuss a constrained minimization example.

Example 2.4. We consider the following nonlinear program in $\mathbb{R}^2$:

$$\min_{x\in\mathbb{R}^n} \; f(x) := x_2 \quad \text{s.t.} \quad 0 \leq x_1 \leq 2, \quad (x_2 - 1)\cdot(x_1 + x_2 - 2) \leq 0. \tag{2.2}$$



(a) Example 2.2 (b) Example 2.3

Figure 2.1: Illustration of global and local minimizer for different objective functions.

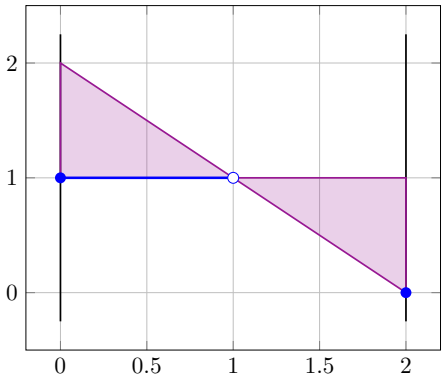


Figure 2.2: Example 2.4 – Illustration of global and local minimizer for a constrained nonlinear program.

Using our standard notation, the feasible set $X \subset \mathbb{R}^2$ of problem (2.2) is defined by the three inquality constraints $g_1(x) := -x_1 \leq 0$, $g_2(x) := x_1 - 2 \leq 0$ and $g_3(x) := (x_2 - 1) \cdot (x_1 + x_2 - 2) \leq 0$, i.e., we have

$$X := \{x \in \mathbb{R}^2 : g_1(x) \leq 0,\, g_2(x) \leq 0,\, g_3(x) \leq 0\}.$$

Let us note that the condition $g_3(x) \leq 0$ is equivalent to

$$x_2 \leq 1 \quad \text{and} \quad x_2 \geq 2 - x_1 \quad \text{or} \quad x_2 \geq 1 \quad \text{and} \quad x_2 \leq 2 - x_1.$$

This implies that the feasible set consists of two components that intersect at the point $(1,1)^\top$ (see also Figure 2.2). Hence, all points in the set $S := \{x \in \mathbb{R}^2 : x_1 \in [0,1),\, x_2 = 1\}$ are local minimizer and $x^* = (2,0)^\top$ is the unique, strict global minimizer. Notice that the minimizer in S are not *strict* local minimizer.

We conclude with several additional definitions and remarks:

- Let $x^* \in X$ be a global minimizer of the problem $\min_{x \in X} f(x)$. Then, x^* is also called a *global solution* of (2.1) (and the value $f(x^*)$ is called the *optimal value*).
- For a minimization problem, the objective value is said to be *unbounded* (or the optimal cost is $-\infty$), if for every $K \in \mathbb{R}$, we can find a feasible point $x \in X$ with $f(x) \leq K$.
- The following optimization problems are equivalent:

$$\max_x \ f(x) \quad \text{s.t.} \quad x \in X \quad \text{and} \quad \min_x \ -f(x) \quad \text{s.t.} \quad x \in X.$$

  A feasible point x^* is a global maximizer of the maximization problem with optimal value $f(x^*)$ if and only if it is also a global minimizer of the minimization problem with optimal objective function $-f(x^*)$.

  Hence, we will only concentrate on minimization problems.

3. Optimality Conditions for Unconstrained Problems

Throughout the next sections, we consider problem (2.1) and assume that $X \subset \mathbb{R}^n$ either coincides with the full set $\mathbb{R}^n$ or is an open set.

3.1. Necessary Conditions

Theorem 3.1: Necessary First-Order Optimality Conditions

Let $f : X \to \mathbb{R}$ be differentiable ($X = \mathbb{R}^n$ or $X \subset \mathbb{R}^n$ open) and let $x^* \in X$ be a local minimum of f. Then, it follows $\nabla f(x^*) = 0$.

Proof. Let $h \in \mathbb{R}^n$ be arbitrary. Since x^* is a local minimizer, we have $f(x^* + th) \geq f(x^*)$ for all $t > 0$ sufficiently small. Rearranging the terms, dividing by t and taking the limit, we thus obtain

$$0 \leq \lim_{t \to 0} \frac{f(x^* + th) - f(x^*)}{t} = \nabla f(x^*)^\top h.$$

Choosing $h = -\nabla f(x^*)$ finishes the proof. ■

Definition 3.2: Stationary Point

Let $f : X \to \mathbb{R}$ be differentiable and let $x^* \in X$, $X \subset \mathbb{R}^n$ open, be given with $\nabla f(x^*) = 0$. Then, x^* is called **stationary point** or **critical point** of f.

The condition $\nabla f(x^*) = 0$ is only necessary but not sufficient for a local minimum. In particular, due to $\nabla(-f) = -\nabla f$, every stationary point of f is also a stationary point of $-f$. Hence, the condition $\nabla f(x^*) = 0$ can not distinguish between minimizer and maximizer.

Definition 3.3: Saddle Point

A stationary point x^* of f that is neither a local minimizer nor a local maximizer is called a **saddle point** of f.

Example 3.4. The function $f(x) = x_1^2 - x_2^2$ has the gradient $\nabla f(x) = (2x_1, -2x_2)^\top$. Hence, the point $x^* = (0, 0)^\top$ is the single stationary point of f. Since f increases along the x_1-direction and decreases along the x_2-direction, x^* has to be a saddle point, see also Figure 3.1.

To distinguish minima, maxima and saddle points, we consider the curvature of f.

Theorem 3.5: Necessary Second-Order Optimality Conditions

Let $f : X \to \mathbb{R}$ be twice continuously differentiable on the open set $X \subset \mathbb{R}^n$ and let x^* be a local minimum of f. Then, it holds that:

(i) $\nabla f(x^*) = 0$ (i.e., x^* is a stationary point of f).

(ii) The Hessian $\nabla^2 f(x^*)$ is positive semidefinite and symmetric, i.e.,

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

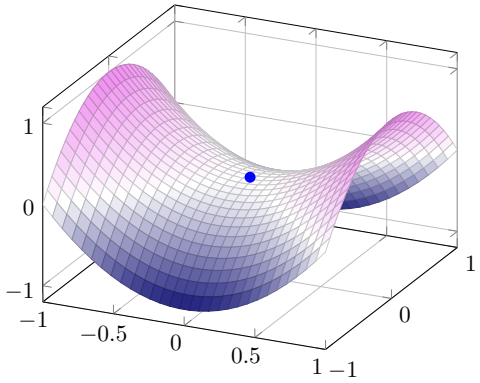


Figure 3.1: Illustration of Example 3.4

Proof. Since x^* is a local minimum, we have $\nabla f(x^*) = 0$ and there exists $\varepsilon > 0$ such that $f(y) \geq f(x^*)$ for all $y \in B_\varepsilon(x^*) \cap X$. Let $h \in \mathbb{R}^n$ be an arbitrary vector. Since X is open, we have $x^* + th \in X$ and $\|x^* + th - x^*\| = t\|h\| < \varepsilon$ for all t sufficiently small, which implies $f(x^* + th) - f(x^*) \geq 0$. Next, we use a Taylor expansion

$$\begin{aligned} 0 \leq f(x^* + th) - f(x^*) &= t\nabla f(x^*)^\top h + \frac{1}{2}(th)^\top \nabla^2 f(x^*)(th) + o(\|th\|^2) \\ &= \frac{t^2}{2} h^\top \nabla^2 f(x^*) h + o(t^2) \end{aligned}$$

for t sufficiently small. Dividing by $\frac{t^2}{2}$, we obtain $h^\top \nabla^2 f(x^*)h + \frac{2o(t^2)}{t^2} \geq 0$ and taking the limit $t \to 0$, this yields $h^\top \nabla^2 f(x^*)h \geq 0$. Since h was arbitrary, this finishes the proof. ■

If x^* is a *local maximizer*, then condition (ii) in Theorem 3.5 changes to:

(ii)′ The Hessian $\nabla^2 f(x^*)$ is *negative semidefinite* and symmetric, i.e., $h^\top \nabla^2 f(x^*)h \leq 0$ for all $h \in \mathbb{R}^n$.

Corollary 3.6

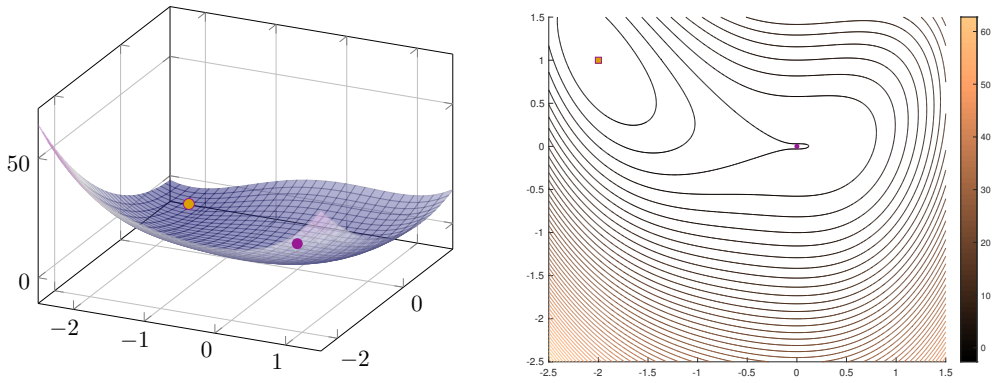
Let $x^* \in X$ be stationary point of f. If $\nabla^2 f(x^*)$ is indefinite, then x^* is a saddle point.

The conditions in Theorem 3.5 are only necessary but not sufficient! In particular, there are stationary points with positive semidefinite Hessian that are not local minimizer!

Example 3.7. The following two examples illustrate that a stationary point with positive semidefinite Hessian can either be a local minimum or a saddle point.

(a) Consider $f : \mathbb{R} \to \mathbb{R}$, $f(x) := x^3$. Then, we have $f'(x) = 3x^2$ and $f''(x) = 6x$. The point $x^* = 0$ is the only stationary point with $f''(0) = 0$ ($f''(0)$ is pos. semidefinite). But x^* is a saddle point.

(b) Let us consider $f : \mathbb{R} \to \mathbb{R}$, $f(x) := x^4$. Then, we have $f'(x) = 4x^3$ and $f''(x) = 12x^2$. The point $x^* = 0$ is the only stationary point with $f''(0) = 0$. In this case, x^* is a local minimizer.



(a) Plot of f (b) Contour Plot

Figure 3.2: Illustration of Example 3.9

Theorem 3.1 and Theorem 3.5 imply that we only need to check all stationary points with $\nabla f(x) = 0$ in order to find local minima and maxima of f. Specifically, if the problem has a global solution, we can find such a solution by comparing the objective function values of all stationary points (with positive semidefinite Hessian).

The necessary second-order optimality conditions can be used to show that a stationary point is **not** a local minimizer or maximizer.

3.2. Sufficient Conditions

Theorem 3.8: Sufficient Second-Order Optimality Conditions

Let $f : X \to \mathbb{R}$ be twice continuously differentiable on the open set $X \subset \mathbb{R}^n$ and let $x^* \in X$ be a point such that

(i) $\nabla f(x^*) = 0$ (i.e., x^* is a stationary point of f).

(ii) The Hessian $\nabla^2 f(x^*)$ is positive definite, i.e.,

$$h^\top \nabla^2 f(x^*) h > 0, \quad \forall\, h \in \mathbb{R}^n \backslash \{0\}.$$

Then, x^* is a strict local minimum of f.

Proof. Since the Hessian $\nabla^2 f(x^*)$ is positive definite, the eigenvalue $\mu := \lambda_{\min}(\nabla^2 f(x^*)) > 0$ is positive. Using the estimate

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

and a Taylor expansion, it follows

$$f(x^* + h) - f(x^*) = \nabla f(x^*)^\top h + \frac{1}{2} h^\top \nabla^2 f(x^*) h + o(\|h\|^2) \geq \|h\|^2 \cdot \left[\frac{\mu}{2} + \frac{o(\|h\|^2)}{\|h\|^2} \right]$$

for $h \to 0$ (since X is open, we can ensure $x^* + h \in X$ for all h sufficiently close to 0). Since $\mu > 0$ and $o(\|h\|^2)/\|h\|^2 \to 0$ as $h \to 0$, we have $o(\|h\|^2)/\|h\|^2 \geq -\frac{\mu}{4}$ for h all sufficiently small. Hence, this shows

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

for all h sufficiently small, which implies that x^* is a strict local minimum. ■

If x^* is a stationary point and the Hessian $\nabla^2 f(x^*)$ is negative definite, then x^* is a strict local maximizer.

Example 3.9. We consider the two-dimensional optimization problem

$$\min_{x \in \mathbb{R}^2} \; f(x) = x_1^4 + 2(x_1 - x_2)x_1^2 + 4x_2^2.$$

Task: Find all local minimizer, local maximizer and saddle points of f. It holds that

$$\nabla f(x) = \begin{pmatrix} 4x_1^3 + 6x_1^2 - 4x_1x_2 \\ -2x_1^2 + 8x_2 \end{pmatrix}, \quad \nabla^2 f(x) = \begin{pmatrix} 12x_1^2 + 12x_1 - 4x_2 & -4x_1 \\ -4x_1 & 8 \end{pmatrix}.$$

Step 1: Calculate all stationary points of f by solving $\nabla f(x) = 0$:

$$\nabla f(x) = 0 \quad \Longleftrightarrow \quad \begin{bmatrix} 4x_2 = x_1^2 \;\text{ and }\; 3x_1^2(x_1 + 2) = 0, \\ x_1 = 0 \;\text{ or }\; x_1 = -2. \end{bmatrix}$$

Thus, the stationary points are: $x_1^* = (0, 0)^\top$ and $x_2^* = (-2, 1)^\top$.

Step 2: Determine the definiteness of the Hessian $\nabla^2 f(x^*)$ to decide whether the stationary points x^* are local minima, maxima or saddle points:

$$\nabla^2 f(x_1^*) = \begin{pmatrix} 0 & 0 \\ 0 & 8 \end{pmatrix} \quad \text{and} \quad \nabla^2 f(x_2^*) = \begin{pmatrix} 20 & 8 \\ 8 & 8 \end{pmatrix}.$$

Due to $\mathrm{tr}(\nabla^2 f(x_2^*)) = 28 > 0$ and $\det(\nabla^2 f(x_2^*)) = 160 - 64 > 0$, the Hessian $\nabla^2 f(x_2^*)$ is positive definite and x_2^* is a strict local minimizer. The Hessian $\nabla^2 f(x_1^*)$ has the eigenvalues 0 and 8 and we can not use the second-order optimality conditions to decide whether it is a saddle-point or local minimum. We consider the function f directly around x_1^*:

$$f(\pm|t|, 0) = t^4 + 2(\pm|t|)t^2 = |t|^3(|t| \pm 2).$$

Consequently, f is increasing and decreasing around $(0, 0)$ and x_1^* has to be a saddle point.

The result in Theorem 3.8 can also be strengthened as follows.

Theorem 3.10: Quadratic Growth Condition

Let $f : X \to \mathbb{R}$ be twice continuously differentiable on the open set $X \subset \mathbb{R}^n$. Then, the conditions (i)–(ii) in Theorem 3.8 for $x^* \in X$ are equivalent to:

$$\exists\, \alpha, \epsilon > 0 \quad \text{such that} \quad f(x) - f(x^*) \geq \frac{\alpha}{2}\|x - x^*\|^2 \quad \forall\, x \in B_\epsilon(x^*) \cap X.$$

Proof. Following the proof of Theorem 3.8, we only need to verify the second direction (quadratic growth implies the second-order sufficient conditions). The quadratic growth condition implies that x^* is a (strict) local minimum and hence, we have $\nabla f(x^*) = 0$. The result can then be shown by using a second-order Taylor expansion of f at x^*. ■

3.3. Existence of Global Minimizer: Weierstraß & Coercivity

We first state the Weierstraß Theorem for continuous functions.

Theorem 3.11: Weierstraß Theorem
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function and let $X \subset \mathbb{R}^n$ be a bounded, closed, and nonempty set. Then, f has a global maximum and global minimum on the set X.

The Weierstraß Theorem guarantees existence of global minima if we minimize a continuous function on a compact set. For unconstrained problems ($X = \mathbb{R}^n$), this result is not immediately or directly applicable. We will now introduce a property of the objective function f that will allow us to ensure existence of global solutions.

Definition 3.12: Coercivity
A continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be **coercive** if

$$\lim_{\|x\| \to \infty} f(x) = +\infty.$$

Geometrically, coercivity means that the function values $f(x)$ increases as x moves away from the origin in any possible direction. Mathematically, coercivity means:

$$\forall\, B > 0 \quad \exists\, r > 0 \quad \text{such that} \quad \|x\| > r \quad \Longrightarrow \quad f(x) > B.$$

Example 3.13. The mappings $f(x) = x^2$, $f(x) = x^4$, and $f(x) = |x|^3$ are simple examples of coercive functions. The functions $f(x) = x$, $f(x) = x^3$, $f(x) = e^x$, and $f(x) = 1$ are not coercive.

Coercivity is often established by estimating the function and by finding a suitable *lower bound* for sufficiently large x. (What are the dominating terms in f?).

Example 3.14. The function $f(x) = e^x - x$ is coercive.

Theorem 3.15: Coercivity and Compact Level Sets
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous and coercive function. Then, for all $\alpha \in \mathbb{R}$, the level set

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

is compact and the problem $\min_{x \in \mathbb{R}^n} f(x)$ has at least one global minimizer.

Proof. Since f is continuous, the set $L_{\leq\alpha}$ is closed. Let us now assume that $L_{\leq\alpha}$ is not bounded, i.e., there exists a sequence $(x^k)_k$ with $x^k \in L_{\leq\alpha}$ for all k and $\|x^k\| \to \infty$. But then the coercivity of f implies $f(x^k) \to \infty$ as $k \to \infty$. This is a contradiction to $x^k \in L_{\leq\alpha}$! Hence, our assumption must have been wrong and the set $L_{\leq\alpha}$ is indeed bounded. Let $y \in \mathbb{R}^n$ be arbitrary. Then, the two problems:

$$\min_{x\in\mathbb{R}^n} \; f(x) \quad \text{and} \quad \min_{x} \; f(x) \quad \text{s.t.} \quad x \in L_{\leq f(y)}$$

obviously have the same global solutions. But now the Weierstraß Theorem can be applied to the second problem. This shows that f has at least one global minimizer. ■

References. See [2, Chapter 2] or [5, Section 1.1].

4. Convexity

We now consider an important class of functions for which local minima are always global solutions: the class of convex functions. We will start with the definition of convex sets that form the foundation of convex functions.

4.1. Convex Sets

Definition 4.1: Convex Sets
A set $X \subset \mathbb{R}^n$ is called **convex** if for all $x, y \in X$ and $\lambda \in [0,1]$, we have

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

If x and y are two points in X, then the whole line connecting x and y has to lie in X.

Example 4.2 (Half Space). Let $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ be given, then the half space $H := \{x \in \mathbb{R}^n : a^\top x \leq b\}$ is a convex set.

Proof. Let $x, y \in H$ and $\lambda \in [0,1]$ be given. We need to show $(1-\lambda)x + \lambda y \in H$:

$$a^\top((1-\lambda)x + \lambda y) = (1-\lambda)a^\top x + \lambda a^\top y \leq (1-\lambda)b + \lambda b = b.$$

Thus, it follows $(1-\lambda)x + \lambda y \in H$ for all $x, y \in H$ and $\lambda \in [0,1]$. ■

Example 4.3. For given $a \in \mathbb{R}^n$, $r > 0$, the closed ball $\bar{B}_r(a) = \{x \in \mathbb{R}^n : \|x - a\| \leq r\}$ is a convex set.

Proof. We again start with $x, y \in \bar{B}_r(a)$ and $\lambda \in [0,1]$. Then we obtain

$$\begin{aligned}\|[(1-\lambda)x + \lambda y] - a\| &= \|(1-\lambda)[x-a] + \lambda[y-a]\| \\ &\leq |1-\lambda|\|x-a\| + |\lambda|\|y-a\| \leq (1-\lambda)r + \lambda r = r.\end{aligned}$$

This shows $(1-\lambda)x + \lambda y \in \bar{B}_r(a)$ and establishes convexity of $\bar{B}_r(a)$. ■
Examples of sets that are not convex:

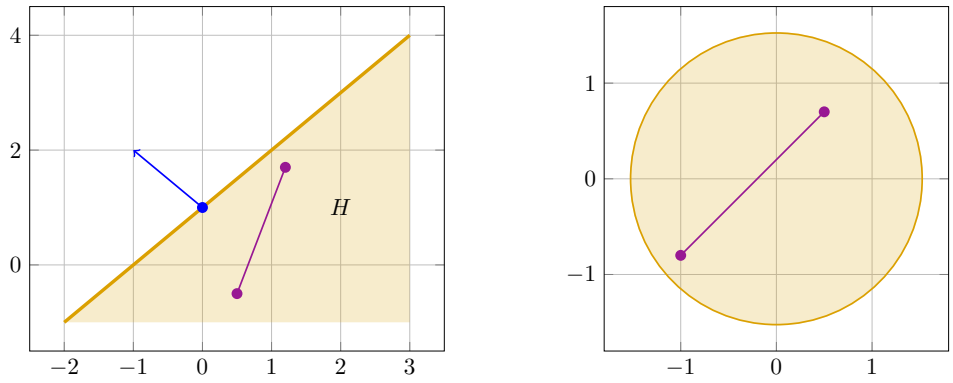
$$S_1 = \{x \in \mathbb{R}^n : \|x - a\| \geq r\}, \quad (a \in \mathbb{R}^n,\, r > 0), \quad S_2 = \{(x,y) \in \mathbb{R}^2 : y = x^2\}, \quad \ldots$$

Lemma 4.4
If S_1,S_2, ..., S_m are convex sets in $\mathbb{R}^n$, then $S := \bigcap_{i=1}^m S_i$ is also convex.

Proof. Let $x, y \in S$ and $\lambda \in [0,1]$ be given. This implies that $x, y \in S_i$ for all i. Using the convexity of the sets S_i, we thus have $(1-\lambda)x + \lambda y \in S_i$ for all i. Hence, it follows $(1-\lambda)x + \lambda y \in S$. Therefore, S is a convex set. ■

Example 4.5 (Polyhedral Sets). Let the matrix $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ be given. Then, the *polyhedral set* $\{x \in \mathbb{R}^n : Ax \leq b\}$ is convex.

This can be shown by combining Example 4.2 and Lemma 4.4.



(a) Plot of $H = \{x : (-1,1)^\top x \leq 1\}$. (b) Example 4.3

Figure 4.1: Illustration of Example 4.2 and Example 4.3

4.2. Convex and Concave Functions

Definition 4.6: Convex Functions

Let $f : X \to \mathbb{R}$ be given and let $X \subset \mathbb{R}^n$ be a convex set. The function f is called

(i) **convex** if for all $x, y \in X$ and $\lambda \in [0,1]$, it holds that

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

(ii) **strictly convex** if for all $x, y \in X$ with $x \neq y$ and $\lambda \in (0,1)$, we have

$$f((1-\lambda)x + \lambda y) < (1-\lambda)f(x) + \lambda f(y).$$

(iii) **strongly convex** (with parameter μ) if there exists $\mu > 0$ such that for all $x, y \in X$ and all $\lambda \in [0,1]$, we have

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

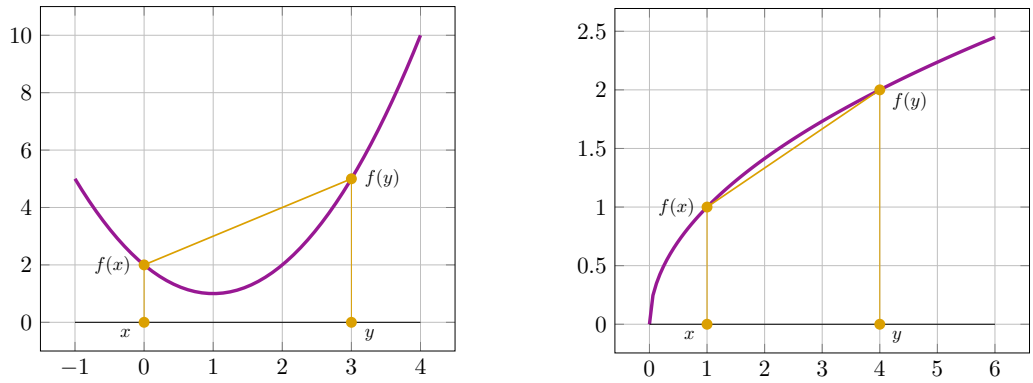
A function $f : X \to \mathbb{R}$ on a convex set $X \subset \mathbb{R}^n$ is called **(strictly, strongly) concave** if $-f$ is (strictly, strongly) convex, i.e., for all $x, y \in X$ and $\lambda \in [0,1]$, we have

$$f((1-\lambda)x + \lambda y) \geq (1-\lambda)f(x) + \lambda f(y).$$

Question: Why do we need to assume convexity of X?

Geometric Interpretation:

1. For a convex function f, the line segment connecting $f(x)$ and $f(y)$ lies **above** the graph of f in $[x; y]$, see also Figure 4.2.
2. For a concave function f, the line segment connecting $f(x)$ and $f(y)$ lies **below** the graph of f in $[x; y]$.



(a) Plot of $f(x) = (x-1)^2 + 1$.

(b) Plot of $f(x) = \sqrt{x}$

Figure 4.2: Examples illustrating convexity and concavity

Example 4.7. We consider the Euclidean norm $f(x) = \|x\| = \sqrt{x^\top x}$. Then, f is convex.

Proof. Let $x, y \in \mathbb{R}^n$ and $\lambda \in [0,1]$ be arbitrary. Using $\lambda, 1-\lambda \geq 0$ and the triangle inequality, we have

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

This shows that the Euclidean norm is a convex function. ■

Example 4.8. Affine-linear functions of the form $f(x) = a^\top x + b$, $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ are convex **and** concave.

Example 4.9. We show that the function $f(x) = \sqrt{x}$ is strictly concave on $[0, \infty)$. We need to show

$$\sqrt{(1-\lambda)x + \lambda y} > (1-\lambda)\sqrt{x} + \lambda\sqrt{y}, \quad \forall\, x, y \in [0,\infty), \quad x \neq y, \quad \lambda \in (0,1).$$

We start from the right side and take squares:

$$\begin{aligned}
[(1-\lambda)\sqrt{x} + \lambda\sqrt{y}]^2 &= (1-\lambda)^2 x + 2\lambda(1-\lambda)\sqrt{xy} + \lambda^2 y \\
&= (1-\lambda)x + \lambda y + (1-\lambda)(1-\lambda-1)x + 2\lambda(1-\lambda)\sqrt{xy} + \lambda(\lambda-1)y \\
&= (1-\lambda)x + \lambda y - \lambda(1-\lambda)(\sqrt{x} - \sqrt{y})^2 < (1-\lambda)x + \lambda y.
\end{aligned}$$

Lemma 4.10: Sum of Convex Functions
Let $f_1, ..., f_m : X \to \mathbb{R}$ be convex functions on the convex set $X \subset \mathbb{R}^n$. Then, the mapping

$$f : X \to \mathbb{R}, \quad f(x) := \sum_{i=1}^{m} \alpha_i f_i(x), \quad \alpha_i \geq 0, \quad \forall\, i,$$

is also a convex function on X. Moreover, if at least one function f_j is strictly convex on X and we have $\alpha_j > 0$, then f is strictly convex.

Proof. We only need to show the case $m = 2$. Let $x, y \in X$ and $\lambda \in [0,1]$ be arbitrary. Then due to $\alpha_1, \alpha_2 \geq 0$ and using the convexity of f_1 and f_2, it follows

$$\begin{aligned} f((1-\lambda)x + \lambda y) &= \alpha_1 f_1((1-\lambda)x + \lambda y) + \alpha_2 f_2((1-\lambda)x + \lambda y) \\ &\leq (1-\lambda)[\alpha_1 f_2(x) + \alpha_2 f_2(x)] + \lambda[\alpha_1 f_1(y) + \alpha_2 f_2(y)] = (1-\lambda)f(x) + \lambda f(y). \end{aligned}$$

This shows that f is convex. ■

We now analyze convexity of the composition of two mappings.

Lemma 4.11: Composition of Convex Functions

Let $h : X \to \mathbb{R}$ be convex and let $g : Y \to \mathbb{R}$ be convex and non-decreasing. Suppose that $X \subset \mathbb{R}^n$ and $Y \subset \mathbb{R}$ are convex sets with $h(X) \subset Y$. Then, the composite function $f = g \circ h : X \to \mathbb{R}$, $f(x) = g(h(x))$ is convex.

Proof. Let $x, y \in X$ and $\lambda \in [0,1]$ be arbitrary. Using the convexity of h and the monotonicity of g, it holds that:

$$\begin{aligned} f((1-\lambda)x + \lambda y) &= g(h((1-\lambda)x + \lambda y)) \leq g((1-\lambda)h(x) + \lambda h(y)) \\ &\leq (1-\lambda)g(h(x)) + \lambda g(h(y)) = (1-\lambda)f(x) + \lambda f(y). \end{aligned}$$

(In the last step we also used the convexity of g). ■

Lemma 4.12: Convexity of Max-Functions

Let $f_1, ..., f_m : X \to \mathbb{R}$ be convex functions on the convex set $X \subset \mathbb{R}^n$. Then, the mapping $f : X \to \mathbb{R}$, $f(x) = \max\{f_1(x), f_2(x), ..., f_m(x)\}$ is convex.

Proof. Again, we only need to verify the case $m = 2$. (For $m = 3$ or $m > 3$, we can use $\max\{f_1(x), f_2(x), f_3(x)\} = \max\{f_1(x), \max\{f_2(x), f_3(x)\}\}$ and apply the result for $m = 2$ twice to establish convexity). Let $x, y \in X$ and $\lambda \in [0,1]$ be arbitrary. We have

$$\begin{aligned} f_i((1-\lambda)x + \lambda y) &\leq (1-\lambda)f_i(x) + \lambda f_i(y) \\ &\leq (1-\lambda)\max\{f_1(x), f_2(x)\} + \lambda \max\{f_1(x), f_2(x)\} = (1-\lambda)f(x) + \lambda f(y) \end{aligned}$$

for all $i \in \{1, 2\}$. Since the inequality holds for all i, we can take the maximum with respect to i which proves convexity of f. ■

Notice that the last result can also be extended to $\sup_{i \in \mathcal{I}} f_i$ where $(f_i)_{i \in \mathcal{I}}$ is a family of convex functions. The next lemma establishes a helpful relationship between convexity of a function f and convexity of associated level sets $L_{\leq \alpha}$.

Lemma 4.13: Convexity and Level Sets

Suppose that $X \subset \mathbb{R}^n$ is a convex set and $f : X \to \mathbb{R}$ is convex. Then, for all $\alpha \in \mathbb{R}$, the level set $L_{\leq \alpha} = \{x \in X : f(x) \leq \alpha\}$ is convex.

Proof. Let $x, y \in L_{\leq\alpha}$ and $\lambda \in [0,1]$ be arbitrary. Due to the convexity of f, we have

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

This shows $(1-\lambda)x + \lambda y \in L_{\leq\alpha}$ and establishes convexity of $L_{\leq\alpha}$. ■

4.3. Convexity and Differentiability

If the mapping f is differentiable, then convexity can also be characterized differently:

Theorem 4.14: Convexity and Differentiability

Let $f : X \to \mathbb{R}$ be C^1 on an open set that contains the convex set $X \subset \mathbb{R}^n$. It follows:

(i) The function f is convex if and only if we have $f(y) - f(x) \geq \nabla f(x)^\top (y - x)$ for all $x, y \in X$.

(ii) The mapping f is strictly convex if and only if

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

(iii) The function f is μ-strongly convex if and only if there exists $\mu > 0$ such that

$$f(y) - f(x) \geq \nabla f(x)^\top (y - x) + 0.5\mu\|y - x\|^2, \quad \forall\, x, y \in X.$$

Proof. Proof of part (i). We start with showing "⇒". Using the convexity of f, we have

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

for all $x, y \in X$ and $\lambda \in [0,1]$. Taking the limit $\lambda \to 0$, we immediately obtain $f(y) - f(x) \geq \nabla f(x)^\top (y - x)$. To prove the second direction, we set $x_\lambda = (1-\lambda)x + \lambda y$ and use

$$\begin{aligned} f(x) - f(x_\lambda) &\geq \nabla f(x_\lambda)^\top (x - x_\lambda) = \lambda \nabla f(x_\lambda)^\top (x - y) \\ f(y) - f(x_\lambda) &\geq \nabla f(x_\lambda)^\top (y - x_\lambda) = (1-\lambda) \nabla f(x_\lambda)^\top (y - x) \end{aligned} \tag{4.1}$$

for all $x, y \in X$ and $\lambda \in [0,1]$. Now, it follows

$$0 \leq (1-\lambda)(f(x) - f(x_\lambda)) + \lambda(f(y) - f(x_\lambda)) = (1-\lambda)f(x) + \lambda f(y) - f(x_\lambda),$$

which shows that f is convex. *Proof of part (ii).* Suppose that f is strictly convex. Then for $x, y \in X$, $x \neq y$ and $z = \frac{1}{2}(x+y)$ we have

$$f(z) < \frac{1}{2}f(x) + \frac{1}{2}f(y) \quad \Longrightarrow \quad f(z) - f(x) < \frac{1}{2}(f(y) - f(x)).$$

Utilizing (i), it holds that $f(z)-f(x) \geq \nabla f(x)^\top(z-x) = \frac{1}{2}\nabla f(x)^\top(y-x)$ and together this yields

$$\nabla f(x)^\top(y-x) \leq 2(f(z)-f(x)) < f(y)-f(x).$$

The direction "$\Leftarrow$" follows as before using ">" in (4.1). *Proof of part (iii).* We mimic the proof of part (i). Defining $x_\lambda = (1-\lambda)x + \lambda y$, it holds that

$$\begin{aligned}\nabla f(x)^\top(y-x) &= \lim_{\lambda \downarrow 0} \frac{f(x_\lambda)-f(x)}{\lambda} \leq \lim_{\lambda \downarrow 0} \frac{(1-\lambda)f(x)+\lambda f(y) - \frac{\mu}{2}\lambda(1-\lambda)\|y-x\|^2 - f(x)}{\lambda} \\ &= f(y)-f(x)-0.5\mu\|y-x\|^2.\end{aligned}$$

In order to show the second direction, we utilize

$$\|x-x_\lambda\| = \lambda\|y-x\| \quad \text{and} \quad \|y-x_\lambda\| = (1-\lambda)\|y-x\|.$$

We then obtain

$$\begin{aligned}(1-\lambda)f(x)+\lambda f(y)-f(x_\lambda) &= (1-\lambda)(f(x)-f(x_\lambda)) + \lambda(f(y)-f(x_\lambda)) \\ &\geq (1-\lambda)[\nabla f(x_\lambda)^\top(x-x_\lambda) - 0.5\mu\|x-x_\lambda\|^2] + \lambda[\nabla f(x_\lambda)^\top(y-x_\lambda) - 0.5\mu\|y-x_\lambda\|^2] \\ &= \nabla f(x_\lambda)^\top[(1-\lambda x)+\lambda y - x_\lambda] + 0.5\mu[(1-\lambda)\lambda^2 + \lambda(1-\lambda)^2]\|y-x\|^2 \\ &= 0.5\mu\lambda(1-\lambda)\|y-x\|^2\end{aligned}$$

which verifies strong convexity of f. ■

Remark: The last theorem illustrates that a function is convex if and only if all of its tangent planes "support" the graph of the function from below.

If f is twice continuously differentiable, then there is a connection between convexity and the definiteness of the Hessian of f.

Theorem 4.15: Convexity and Definiteness

Let $f : X \to \mathbb{R}$ be twice continuously differentiable on the open and convex set $X \subset \mathbb{R}^n$.

(i) The function f is convex if and only if the Hessian $\nabla^2 f(x)$ is positive semidefinite for all $x \in X$, i.e., it holds that

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

(ii) If the Hessian $\nabla^2 f(x)$ is positive definite for all $x \in X$, i.e., if we have

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

then f is strictly convex.

(iii) The mapping f is strongly convex (with parameter μ) if and only if $\nabla^2 f(x)$ is uniformly positive definite (with parameter μ), i.e., there is $\mu > 0$ such that

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

Proof. Proof of part (i). Let us first assume that $\nabla^2 f(x)$ is positive semidefinite for all $x \in X$. Then, the mean value theorem implies that there exist $t \in [0,1]$ such that

$$f(y) - f(x) = \nabla f(x)^\top (y-x) + \frac{1}{2}(y-x)^\top \nabla^2 f(x+t(y-x))(y-x) \geq \nabla f(x)^\top (y-x). \tag{4.2}$$

By Theorem 4.14 (ii), this establishes convexity of f. To prove the second direction, we use Taylor's theorem. Let $x \in X$ and $h \in \mathbb{R}^n$ be arbitrary. Since X is open there is $\epsilon = \epsilon(x,h) > 0$ such that $x + th \in X$ for all $t \in [0,\epsilon]$. It then holds that

$$f(x+th) = f(x) + t\nabla f(x)^\top h + \frac{t^2}{2} h^\top \nabla^2 f(x) h + o(t^2) \tag{4.3}$$

for t sufficiently small and $t \to 0$. Since f is convex, it follows $f(x+th) - f(x) - t\nabla f(x)^\top h \geq 0$ which implies

$$\frac{t^2}{2} h^\top \nabla^2 f(x) h + o(t^2) \geq 0, \quad (t \to 0).$$

Dividing both sides by $t^2/2$ and taking the limit $t \to 0$ yields $h^\top \nabla^2 f(x) h \geq 0$. Since both $h \in \mathbb{R}^n$ and $x \in X$ are arbitrary, this shows that $\nabla^2 f(x)$ is positive semidefinite for all x. *Proof of (ii).* This follows from (4.2) (with ">") and Theorem 4.14 (ii). *Proof of (iii).* As in the proof of part (i), the uniform positive definiteness of $\nabla^2 f$ implies

$$\begin{aligned} f(y) - f(x) &= \nabla f(x)^\top (y-x) + \frac{1}{2}(y-x)^\top \nabla^2 f(x+t(y-x))(y-x) \\ &\geq \nabla f(x)^\top (y-x) + \frac{\mu}{2}\|y-x\|^2. \end{aligned}$$

Applying Theorem 4.14 (iii), this establishes strong convexity (with parameter μ) of f. Conversely, using the strong convexity of f, i.e.,

$$f(x+th) - f(x) \geq t\nabla f(x)^\top h + \frac{\mu t^2}{2}\|h\|^2$$

and (4.3), it follows $\frac{t^2}{2} h^\top \nabla^2 f(x) h + o(t^2) \geq \frac{\mu t^2}{2}\|h\|^2$, $(t \to 0)$. This finishes the proof. ■

Remark: Similar results can also be shown for concave mappings. In particular, the function f is concave if and only if the Hessian $\nabla^2 f(x)$ is negative semidefinite for all x. If $\nabla^2 f(x)$ is negative definite for all x, then f is strictly concave. If the Hessian $\nabla^2 f(x)$ is indefinite at some point $x \in X$, then f is neither convex nor concave on X.

Example 4.16. Let us consider $f(x) = e^x$. Then, we have $f''(x) = e^x > 0$ for all $x \in \mathbb{R}$. Hence, the exponential map is strictly convex but not strongly convex!

Example 4.17. Let us consider $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x) = 4x_1^2 + 4x_1x_2 + x_2^2$. Then, it holds that

$$\nabla f(x) = \begin{pmatrix} 8x_1 + 4x_2 \\ 4x_1 + 2x_2 \end{pmatrix}, \quad \nabla^2 f(x) = \begin{pmatrix} 8 & 4 \\ 4 & 2 \end{pmatrix}.$$

Consequently, we have $\lambda_1 + \lambda_2 = \mathrm{tr}(\nabla^2 f(x)) = 10$, $\lambda_1\lambda_2 = \det(\nabla^2 f(x)) = 0$ which implies that the eigenvalues of $\nabla^2 f(x)$ are 0 and 10 for all x. Hence, f is a convex function.

Example 4.18. We consider the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) := \frac{1}{2}x^\top Ax + b^\top x + c$, where $A \in \mathbb{R}^{n\times n}$ is symmetric and $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$ are given. Then, it holds that

$$\nabla f(x) = Ax + b, \quad \nabla^2 f(x) = A.$$

Hence, due to Theorem 4.15, f is convex if and only if A is positive semidefinite and f is strongly convex with parameter $\mu = \lambda_{\min}(A)$ if and only if A is positive definite.

4.4. Convex Optimization Problems

The importance of convex functions and convexity for optimization problems stems from the following result:

Theorem 4.19: Optimality and Convex Problems

Let $f : X \to \mathbb{R}$ be convex and let $X \subset \mathbb{R}^n$ be a convex set. We consider the problem

$$\min \ f(x) \quad \text{s.t.} \quad x \in X. \tag{4.4}$$

(i) Every local minimizer of problem (4.4) is also a global minimizer of (4.4).

(ii) If f is strictly convex, then problem (4.4) possesses at most one local minimizer which (if it exists) is also the strict global minimum of f on X.

(iii) Let X be open (or $X = \mathbb{R}^n$) and suppose that f is continuously differentiable and x^* is a stationary point of f. Then, x^* is a global minimizer of f on X.

Proof. Let x^* be a local minimizer of (4.4), i.e., there exists $\varepsilon > 0$ such that

$$f(x^*) \leq f(x) \quad \forall\, x \in B_\varepsilon(x^*) \cap X.$$

Now, let $y \in X$ be arbitrary and consider the point $y_\lambda = \lambda y + (1-\lambda)x^*$ for some $\lambda \in [0,1]$. Since X is convex, we have $y_\lambda \in X$ and it holds that

$$\|y_\lambda - x^*\| = \lambda \|y - x^*\|.$$

Hence, for all λ with $\lambda \leq \varepsilon \|y - x^*\|^{-1}$, it follows $y_\lambda \in B_\varepsilon(x^*) \cap X$. Combining this observation with the convexity of f, we obtain

$$f(x^*) \leq f(y_\lambda) \leq \lambda f(y) + (1-\lambda) f(x^*) \quad \Longrightarrow \quad f(x^*) \leq f(y).$$

Since $y \in X$ is arbitrary, this shows that x^* is a global minimizer of (4.4).

We show the second part by contradiction: Suppose that f possesses two local minimizer $x^* \neq y^*$. As shown in part (i), these minimizer are then already global minimizer, i.e., we have $f(x^*) = f(y^*) \leq f(x)$ for all $x \in X$. Let us now define $z^* = \frac{1}{2}x^* + \frac{1}{2}y^*$, then by the strict convexity of f, it follows

$$f(z^*) < \frac{1}{2} f(x^*) + \frac{1}{2} f(y^*) = f(x^*).$$

But this is a contradiction to the fact that x^* is a global minimizer. Hence, f can only have at most one local minimizer.

In order to prove the last part, let x^* be a stationary point of f with $\nabla f(x^*) = 0$. Then Theorem 4.14 (i) implies

$$f(y) \geq f(x^*) + \nabla f(x^*)^\top (y - x^*) = f(x^*), \quad \forall\, y \in X.$$

Hence, x^* is a global solution of (4.4). ■

References. See [2, Chapter 6 and 7].

5. The Gradient Method

5.1. Descent Direction Methods

In this section we consider the unconstrained minimization problem

$$\min_x \ f(x) \quad \text{s.t.} \quad x \in \mathbb{R}^n.$$

We assume that the objective function is continuously differentiable on $\mathbb{R}^n$. We have already seen that a first order necessary optimality condition is that the gradient vanishes at optimal points:

$$x^* \text{ is a local minimizer or maximizer} \quad \Longrightarrow \quad \nabla f(x^*) = 0.$$

In principle, the optimal solution of the problem can be obtained by calculating *all* stationary points of f and by comparing the objective function values. Unfortunately, such a procedure has several drawbacks:

- It might not be possible or too difficult to solve the equation $\nabla f(x) = 0$ analytically.
- There might be infinitely many stationary points and finding the lowest function value is another (maybe challenging) optimization problem.

We will consider iterative algorithms for finding stationary points. In this section, we will consider algorithms of the form:

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, 2, ...$$

where d^k is the so-called *direction* and α_k is the *step size*. We will mainly work with descent directions:

Definition 5.1: Descent Directions
Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. A vector $d \in \mathbb{R}^n \backslash \{0\}$ is called **descent direction** of f at x if

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

The most important property of descent directions is that taking a small enough step along these directions leads to a decrease of the objective function.

Lemma 5.2: Descent Property
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function and suppose that d is a descent direction of f at x. Then, there is $\varepsilon > 0$ such that

$$f(x + \alpha d) < f(x) \quad \forall\ \alpha \in (0, \varepsilon].$$

Proof. Using the directional derivative condition, we have

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

Algorithm 5.1: Schematic Descent Direction Method

Initialization: Select an initial point $x^0 \in \mathbb{R}^n$.
for $k = 0, 1, ...$ **do**
  Pick a descent direction d^k.
  Find a stepsize α_k satisfying $f(x^k + \alpha_k d^k) < f(x^k)$.
  Set $x^{k+1} = x^k + \alpha_k d^k$.
  If a stopping criterion is satisfied, then STOP and x^{k+1} is the output.

Since the mapping $\alpha \mapsto \alpha^{-1}[f(x + \alpha d) - f(x)]$ is cont. on $(0, \infty)$, there exists $\varepsilon > 0$ with $\alpha^{-1}[f(x + \alpha d) - f(x)] < 0$ for all $\alpha \in (0, \varepsilon]$. (Otherwise the limit cannot be negative)! ■

We are now ready to define an abstract framework for a general descent direction method. A pseudocode of the full algorithm is shown in Algorithm 5.1.

Some open questions and missing details related to Algorithm 5.1 are summarized below:

- What is the initial point x^0?
- How to choose the descent direction? What step size should be taken?
- What is the stopping criterion?

Unless we have a good guess of the location of local or global minima, the initial point x^0 is usually chosen arbitrarily. The norm of the gradient $\nabla f(x^{k+1})$ is a popular stopping criterion:

$$\|\nabla f(x^{k+1})\| \leq \texttt{tol} \quad \text{with tolerance } \texttt{tol} > 0.$$

We will now discuss the second question on possible choices of d^k and α_k in more detail.

5.2. Descent Directions and the Direction of Steepest Descent

The steepest descent direction is probably the most obvious choice for a descent direction:

Definition 5.3: Steepest Descent Directions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let $x \in \mathbb{R}^n$ be given with $\nabla f(x) \neq 0$. Let $d \in \mathbb{R}^n$ denote the solution of the optimization problem

$$\min_{\|d\|=1} \nabla f(x)^\top d. \tag{5.1}$$

Every vector of the form $s = \lambda d$, $\lambda > 0$, is called **steepest descent direction** of f in x.

The optimization problem (5.1) has the unique optimal solution

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

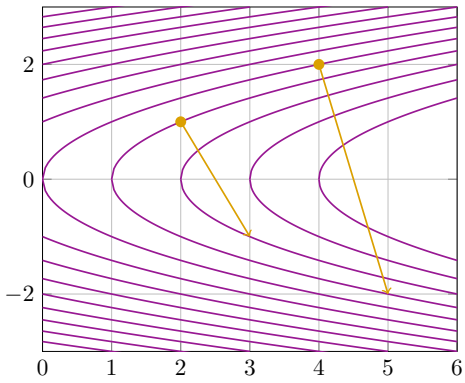


Figure 5.1: Contour lines of the function $f(x) = x_1 - x_2^2$. Plot of x and the vector $x + \nabla f(x)$ for $x = (2,1)^\top$ and $x = (4,2)^\top$.

By the Cauchy-Schwarz inequality, we have $|\nabla f(x)^\top d| \leq \|\nabla f(x)\|\|d\|$ and equality is satisfied if and only if $\nabla f(x)$ and d are linearly dependent. Hence, for every $d \in \mathbb{R}^n$ with $\|d\| = 1$, it follows

$$\nabla f(x)^\top d \geq -\|\nabla f(x)\|\|d\| = -\|\nabla f(x)\|$$

and the latter estimate holds with equality if and only if $d = -\nabla f(x)/\|\nabla f(x)\|$. Thus, every direction $s = -\lambda \nabla f(x)$, $\lambda > 0$ is a steepest descent direction.

Remark 5.4. Notice that the derivation of the steepest descent direction after Definition 5.3 is only correct if the Euclidean norm $\|x\| = \sqrt{x^\top x}$ is used in (5.1). Any other norm yields a different type of steepest descent direction. For instance, if $A \in \mathbb{R}^{n\times n}$ is a symmetric, positive definite matrix, then we can utilize the scaled A-norm $\|x\|_A := \sqrt{x^\top A x}$. The steepest descent directions of f at x w.r.t. the norm $\|\cdot\|_A$ are then given by $s = -\lambda A^{-1}\nabla f(x)$, $\lambda > 0$.

Remark 5.4 motivates different choices of descent directions. In particular, we can also consider directions of the type

$$d = -D\nabla f(x),$$

where $D \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix. It is not hard to see that d is a descent direction of f at x if $\nabla f(x) \neq 0$. We continue with a geometric illustration of the steepest descent direction.

Remark 5.5. Let $y \in \mathbb{R}^n$ be given with $\nabla f(y) \neq 0$. Then, the gradient $\nabla f(y)$ is *perpendicular to the level set* $L_{=f(y)} := \{x \in \mathbb{R}^n : f(x) = f(y)\}$. Specifically, let $d \in \mathbb{R}^n$ be an arbitrary tangent vector to $L_{=f(y)}$ at y and let us consider a C^1-curve $\gamma : (-\delta, \delta) \to L_{=f(y)}$ with $\gamma(0) = y$ and $\gamma'(0) = d$. Then, differentiating the identity $f(\gamma(t)) = f(y)$, we obtain

$$\nabla f(\gamma(t))^\top \gamma'(t) = 0 \quad \forall\, t \in (-\delta, \delta) \qquad \Longrightarrow \qquad \nabla f(y)^\top d = \nabla f(\gamma(0))^\top \gamma'(0) = 0.$$

Hence, $\nabla f(y)$ is perpendicular to every tangent vector. This also holds for the direction of steepest descent $-\nabla f(y)/\|\nabla f(y)\|$. A visualization of this fact is shown in Figure 5.1.

5.3. Step Size Strategies

In order to complete the schematic descent approach, we need to determine a suitable step size strategy. There are many possible choices and we present some common step size rules in the following:

- **Constant step size**: $\alpha_k = \bar{\alpha}$ for all k.
- **Exact line search**: α_k is a minimizer of f along $\alpha \mapsto x^k + \alpha d^k$:

$$\alpha_k \in \operatorname*{arg\,min}_{\alpha \geq 0} \; f(x^k + \alpha d^k).$$

- **Backtracking/Armijo Line Search**: Select some parameter $\sigma, \gamma \in (0,1)$. The step size α_k is determined by the following procedure. First choose $\alpha_k = 1$ (or $\alpha_k = s$ where $s > 0$ is an initial guess). Then, as long as

$$f(x^k + \alpha_k d^k) - f(x^k) > \gamma \alpha_k \nabla f(x^k)^\top d^k,$$

  we set $\alpha_k \leftarrow \sigma \alpha_k$. The step size is chosen as $\alpha_k = \sigma^{i_k}$ (or $\alpha_k = s\sigma^{i_k}$) where i_k is the smallest nonnegative integer for which the condition

$$f(x^k + \sigma^{i_k} d^k) - f(x^k) \leq \gamma \sigma^{i_k} \nabla f(x^k)^\top d^k \tag{5.2}$$

  (or $f(x^k + s\sigma^{i_k} d^k) - f(x^k) \leq \gamma s \sigma^{i_k} \nabla f(x^k)^\top d^k$) is satisfied. The condition (5.2) is called *Armijo condition*.

- **Diminishing step sizes:** We choose the step sizes $(\alpha_k)_k$ such that $\alpha_k \to 0$ and

$$\sum_{k=0}^{\infty} \alpha_k = \infty.$$

  Such choices are especially popular and important for stochastic optimization methods.

The constant step size strategy is very simple but it is not immediately clear how to choose the constant. (A large constant might cause divergence of the method). The exact line search seems to be attractive as it yields the most descent along d^k. However, an optimization problem needs to be solved to determine α_k in this way which can be expensive. Backtracking is a compromise that calculates α_k iteratively.

A visualization of the Armijo condition is given in Figure 5.2. If we define $\phi_k(\alpha) := f(x^k + \alpha d^k) - f(x^k)$, then we have $\phi_k'(0) = \nabla f(x^k)^\top d^k < 0$. Hence, the Armijo condition means

$$\text{find } \; \alpha > 0 \; \text{ such that : } \quad \phi_k(\alpha) \leq \gamma \alpha \phi_k'(0).$$

The next result shows that the Armijo line search algorithm always stops after a finite number of steps with an appropriate step size.

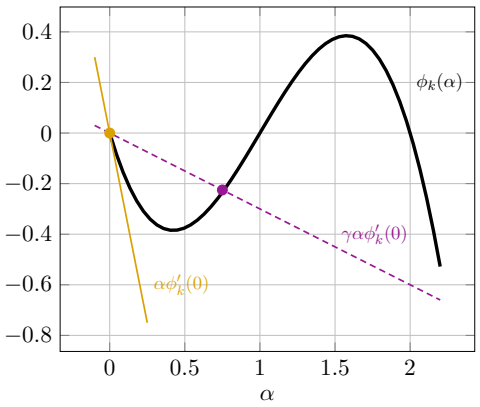


Figure 5.2: Illustration of the Armijo condition. All step sizes $\alpha > 0$ on the left of the purple point satisfy the Armijo condition.

Lemma 5.6: Armijo Condition

Let f be continuously differentiable and let $d \in \mathbb{R}^n \backslash \{0\}$ be a descent direction of f at x. Let $\gamma \in (0,1)$ be given. Then there exists $\varepsilon > 0$ such that the inequality

$$f(x + \alpha d) - f(x) \leq \gamma \alpha \nabla f(x)^\top d$$

holds for all $\alpha \in [0, \varepsilon]$.

Proof. Since f is continuously differentiable, we can use Taylor's expansion to get

$$f(x + \alpha d) - f(x) = \alpha \nabla f(x)^\top d + o(\|\alpha d\|)$$

for $\alpha \to 0$ and hence,

$$f(x + \alpha d) - f(x) - \gamma \alpha \nabla f(x)^\top d = (1 - \gamma)\alpha \nabla f(x)^\top d + o(\|\alpha d\|).$$

Since d is a descent direction, we obtain

$$\lim_{\alpha \downarrow 0} \frac{(1 - \gamma)\alpha \nabla f(x)^\top d + o(\|\alpha d\|)}{\alpha} = (1 - \gamma)\nabla f(x)^\top d < 0.$$

Thus, there exists $\varepsilon > 0$ such that for all $\alpha \in (0, \varepsilon]$, the inequality

$$(1 - \gamma)\alpha \nabla f(x)^\top d + o(\|\alpha d\|) < 0$$

holds, which implies the desired result. ■

We now discuss the exact line search rule for an example.

Example 5.7 (Exact Line Search for Quadratic Functions). Let $f(x) = \frac{1}{2} x^\top A x + b^\top x + c$, where A is a $n \times n$ symmetric and positive definite matrix, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$. Let d be a descent direction of f at x. We now compute an explicit formula for step sizes

generated by exact line search, i.e., we need to solve the problem $\min_{\alpha\geq 0}\ f(x+\alpha d)$.We have

$$\begin{aligned}\phi(\alpha) = f(x+\alpha d) &= \frac{1}{2}(x+\alpha d)^\top A(x+\alpha d) + b^\top(x+\alpha d) + c\\ &= \frac{1}{2}\alpha^2\cdot d^\top Ad + \alpha\cdot x^\top Ad + \alpha\cdot b^\top d + \frac{1}{2}x^\top Ax + b^\top x + c\\ &= \frac{1}{2}\alpha^2\cdot d^\top Ad + \alpha\cdot \nabla f(x)^\top d + f(x).\end{aligned}$$

Since $\phi'(\alpha) = \alpha\cdot d^\top Ad + \nabla f(x)^\top d$, it follows that $\phi'(\alpha) = 0$ if and only if

$$\alpha = \bar{\alpha} = -\frac{\nabla f(x)^\top d}{d^\top Ad}.$$

Notice that ϕ is a parabola that opens up, so the global minimum of ϕ over $[0,\infty)$ is either attained at $\alpha = 0$ or $\alpha = \bar{\alpha}$. Since d is a descent direction of f at x, it follows $\phi'(0) = \nabla f(x)^\top d < 0$ and hence ϕ decreases around $\alpha = 0$ and the exact step size is given by

$$\bar{\alpha} = -\frac{\nabla f(x)^\top d}{d^\top Ad} = \operatorname*{arg\,min}_{\alpha\geq 0}\ f(x+\alpha d).$$

5.4. The Full Gradient Method

In the gradient method the descent direction is chosen to be the negative gradient at the current point: $d^k = -\nabla f(x^k)$. As we have seen, the negative gradient is the steepest descent direction and points in the direction where the objective function decreases most rapidly. The resulting algorithmic scheme – the basic gradient method – is summarized in Algorithm 5.2.

Algorithm 5.2: The Gradient Method

1 Initialization: Select an initial point $x^0 \in \mathbb{R}^n$.
for $k = 0, 1, \ldots$ **do**
2 Pick a step size α_k by a line search procedure on $\phi(\alpha) = f(x^k - \alpha\nabla f(x^k))$ or by a different step size strategy.
3 Set $x^{k+1} = x^k - \alpha_k\nabla f(x^k)$.
4 If $\|\nabla f(x^{k+1})\| \leq \varepsilon$, then STOP and x^{k+1} is the output.

Before investigating the theoretical properties of the gradient method, we will now consider two simple numerical examples that illustrate the performance and implementation of the gradient descent method.

Example 5.8 (Exact Line Search). We implement a `MATLAB` function `gm_quadratic` that finds the optimal solution of the quadratic problem

$$\min_{x\in\mathbb{R}^n}\ \frac{1}{2}x^\top Ax + b^\top x$$

using the gradient method with exact line search. Here, $A \in \mathbb{R}^{n\times n}$ is a given positive definite

and symmetric matrix and $b \in \mathbb{R}^n$. As shown in Example 5.7, the exact step size α_k is given by

$$\alpha_k = \frac{\|\nabla f(x^k)\|^2}{\nabla f(x^k)^\top A \nabla f(x^k)}.$$

The code is shown in Appendix A in the Listing 1. We now want to test the method and solve the problem

$$\min_x \; f(x) = x_1^2 + 2x_2^2 = \frac{1}{2} x^\top \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix} x.$$

Using the initial point $x^0 = (2,1)^\top$ and the tolerance $\varepsilon = 10^{-5}$, we can execute the following `MATLAB` command:

- `[x,obj] = gm_quadratic([2,0;0,4],[0;0],[2;1],1e-5).`

The output is:

```
- - - gradient method for quadratic programs; n = 2
ITER ;   OBJ.VAL ;    G.NORM ;     STEP | ITER ;   OBJ.VAL ;    G.NORM ;     STEP
[ 1] ;   0.666667 ;   1.885618 ;   0.33 | [ 8] ;   0.000000 ;   0.000862 ;   0.33
[ 2] ;   0.074074 ;   0.628539 ;   0.33 | [ 9] ;   0.000000 ;   0.000287 ;   0.33
[ 3] ;   0.008230 ;   0.209513 ;   0.33 | [ 10] ;   0.000000 ;   0.000096 ;   0.33
[ 4] ;   0.000914 ;   0.069838 ;   0.33 | [ 11] ;   0.000000 ;   0.000032 ;   0.33
[ 5] ;   0.000102 ;   0.023279 ;   0.33 | [ 12] ;   0.000000 ;   0.000011 ;   0.33
[ 6] ;   0.000011 ;   0.007760 ;   0.33 | [ 13] ;   0.000000 ;   0.000004 ;   0.33
[ 7] ;   0.000001 ;   0.002587 ;   0.33 |
```

The method stops after 13 iterations with a solution that is already very close to the optimal value `x = 1.0e-05 * (0.1254 ; -0.0627)`. The iterates and the contour plots of the objective function are given in Figure 5.3.

The gradient method follows a "zig-zag" path towards the solution x^*, i.e., the direction found at the kth iteration $x^{k+1} - x^k$ is orthogonal to the direction found at the $(k+1)$th iteration $x^{k+2} - x^{k+1}$. We have the following observation:

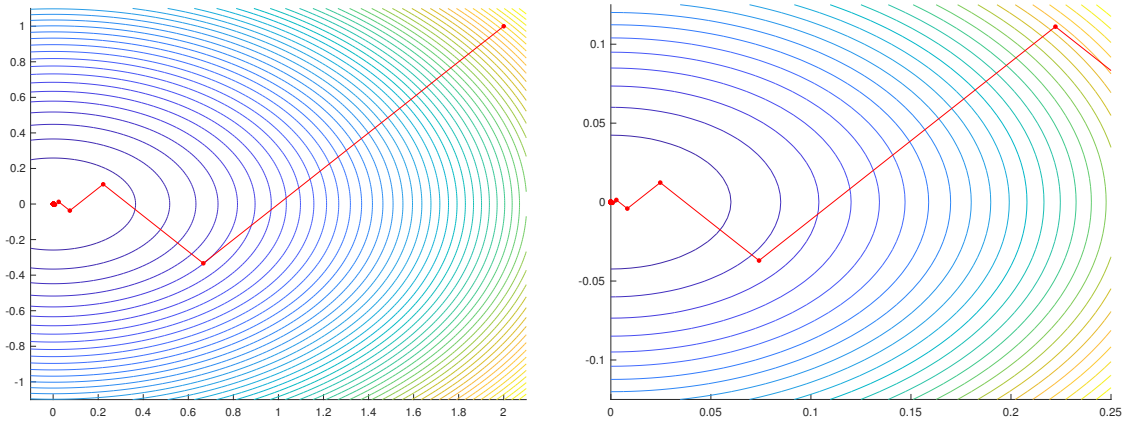
Lemma 5.9: Perpendicular Steps

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let $(x^k)_k$ be the sequence generated by the gradient method with exact line search. Then, for any $k = 0, 1, 2, ...$

$$(x^{k+2} - x^{k+1})^\top (x^{k+1} - x^k) = 0.$$

Proof. By the definition of the gradient method, we have $x^{k+2} - x^{k+1} = -\alpha_{k+1} \nabla f(x^{k+1})$ and $x^{k+1} - x^k = -\alpha_k \nabla f(x^k)$. Hence, we want to show $\nabla f(x^{k+1})^\top \nabla f(x^k) = 0$. Since α_k is determined by exact line search, it follows:

$$\alpha_k \in \arg\min_{\alpha \geq 0} \; \phi(\alpha) = f(x^k - \alpha \nabla f(x^k)).$$



(a) (b)

Figure 5.3: Visualization of the path of iterates of the gradient method (with exact line search) along the contour lines of the objective function.

Moreover, we have $\phi'(\alpha) = -\nabla f(x^k - \alpha \nabla f(x^k))^\top \nabla f(x^k)$ and $\phi'(0) < 0$. Thus, the optimal solution of the latter problem is not 0 and it holds that $\phi'(\alpha_k) = 0$. This implies

$$0 = \phi'(\alpha_k) = -\nabla f(x^k - \alpha_k \nabla f(x^k))^\top \nabla f(x^k) = -\nabla f(x^{k+1})^\top \nabla f(x^k),$$

as desired. ■

Example 5.10 (Backtracking). We implement a `MATLAB` function `gm_armijo` that utilizes the backtracking procedure to calculate the step sizes α_k. The code can be found in Listing 2 in the appendix. As before, we define `A = [2,0;0,4]`, `opts.maxit = 10000`, `opts.tol = 1e-5`, `opts.s = 1`, `opts.gamma = 0.1`, and

```
f.obj = @(x) 0.5*x'*A*x;   f.grad = @(x) A*x.
```

Executing the `MATLAB` command `gm_armijo(f,[2;1],opts)` yields the output:

```
- - - gradient method with backtracking; n = 2
ITER ;   OBJ.VAL ;    G.NORM ;     STEP
[ 0] ;   6.000000 ;   5.656854 ;   0.00
[ 1] ;   2.000000 ;   4.000000 ;   0.50
[ 2] ;   0.000000 ;   0.000000 ;   0.25
```

Thus, the gradient method with backtracking terminated only after 2 iterations. In fact, in this case, it converged to the exact optimal solution (we are just very lucky). The gradient method can also behave very differently. Let us consider the problem

$$\min_x \; x_1^2 + \frac{1}{100} x_2^2$$

and let us use the backtracking gradient method with initial point $x^0 = (\frac{1}{100}, 1)^\top$:

- `A = [1,0;0,0.01]`, `f.obj = @(x) x'*A*x;  f.grad = @(x) 2*A*x`, `opts.maxit = 10000`, `opts.tol = 1e-5`, `opts.s = 1`, `opts.gamma = 0.1`.

```
- - - gradient method with backtracking; n = 2
ITER ;   OBJ.VAL ;   G.NORM ;    STEP
[ 0] ;   0.010100 ;  0.028284 ;  0.00
[ 1] ;   0.009704 ;  0.028003 ;  1.00
[ 2] ;   0.009324 ;  0.027730 ;  1.00
[ 3] ;   0.008958 ;  0.027465 ;  1.00
[ 4] ;   0.008608 ;  0.027209 ;  1.00
          ⋮
[377] ;  0.000000 ;  0.000010 ;  1.00
```

(Exact line search requires 398 iterations). We will later revisit and discuss this phenomenon in more detail.

5.5. Convergence Analysis of the Gradient Method

In this section, we will present a comprehensive convergence analysis of the gradient method summarized in Algorithm 5.2. In particular, we will investigate the following aspects:

- Which type of convergence can we expect for the different introduced step size strategies? Which assumptions are required to establish such convergence results?
- Can we generally ensure convergence of the sequence $(x^k)_k$ (to stationary points)?
- Can we derive complexity bounds to characterize the global behavior of the algorithm?
- Can we allow different descent directions besides the gradient direction $d^k = -\nabla f(x^k)$?
- Under which conditions can we establish local rates of convergence and what type of performance can we expect locally?

5.5.1. Basic Global Convergence Results

Theorem 5.11: Global Convergence

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let $(x^k)_k$ be the sequence generated by the gradient method for solving $\min_{x\in\mathbb{R}^n} f(x)$ with one of the following step size strategies:

- exact line search,
- Armijo line search (backtracking) with $\sigma, \gamma \in (0, 1)$ and $s > 0$.

Then, $(f(x^k))_k$ is nonincreasing and every cluster point of $(x^k)_k$ is a stationary point.

Proof. Let x^* be an arbitrary accumulation point of $(x^k)_k$. Both the exact line search condition and the Armijo condition imply

$$f(x^{k+1}) \leq f(x^k) \quad \forall\, k.$$

Hence, the sequence $(f(x^k))_k$ is nonincreasing and converges to some limit $\xi \in \mathbb{R} \cup \{-\infty\}$. Let $(x^{k_\ell})_{\ell \in \mathbb{N}}$ be a subsequence of $(x^k)_k$ that converges to x^*, i.e., we have $x^{k_\ell} \to x^*$ as $\ell \to \infty$. Then, by the continuity of f, it holds that

$$f(x^{k_\ell}) \to f(x^*) \quad \text{as} \quad \ell \to \infty.$$

Since we also have $f(x^k) \to \xi$, this shows $\xi = f(x^*)$. We now first assume that the step sizes $(\alpha_k)_k$ are generated by backtracking. It follows

$$f(x^0) - f(x^*) = \lim_{k\to\infty} f(x^0) - f(x^k) = \lim_{k\to\infty} \sum_{i=0}^{k-1} f(x^i) - f(x^{i+1}) \geq \sum_{i=0}^{\infty} \gamma \alpha_i \|\nabla f(x^i)\|^2,$$

where we used the Armijo condition with $d^i = -\nabla f(x^i)$. This implies that the sequence $(\alpha_k \|\nabla f(x^k)\|^2)_k$ converges to zero. Let us now assume $\|\nabla f(x^*)\| \neq 0$. Then there exists $\varepsilon > 0$ and $L \in \mathbb{N}$ with $\|\nabla f(x^{k_\ell})\| > \varepsilon$ for all $\ell \geq L$. Furthermore, the sequence $(\alpha_{k_\ell})_\ell$ then has to converge to zero. By the construction of the backtracking strategy, the larger step size $\sigma^{-1}\alpha_{k_\ell}$ does not satisfy the Armijo condition, i.e., we have

$$f(x^{k_\ell} + \sigma^{-1}\alpha_{k_\ell} d^{k_\ell}) - f(x^{k_\ell}) \geq -\gamma\sigma^{-1}\alpha_{k_\ell} \|\nabla f(x^{k_\ell})\|^2.$$

Using the mean value theorem, there exists $t_{k_\ell} \in [0,1]$ such that

$$f(x^{k_\ell} + h^{k_\ell}) = f(x^{k_\ell}) + \nabla f(x^{k_\ell} + t_{k_\ell} h^{k_\ell})^\top [-\sigma^{-1}\alpha_{k_\ell} \nabla f(x^{k_\ell})],$$

where we set $h^{k_\ell} := -\sigma^{-1}\alpha_{k_\ell} \nabla f(x^{k_\ell})$. Combining the last two estimates, we obtain

$$\nabla f(x^{k_\ell} + t_{k_\ell} h^{k_\ell})^\top \nabla f(x^{k_\ell}) \leq \gamma \|\nabla f(x^{k_\ell})\|^2$$

and taking the limit $\ell \to \infty$ this yields $(1-\gamma)\|\nabla f(x^*)\| \leq 0$. Here, we utilized $h^{k_\ell} \to 0$ as $\ell \to \infty$ (notice that we assumed $\alpha_{k_\ell} \to 0$, $\ell \geq L$). Due to $\gamma \in (0,1)$, this is a contradiction to $\nabla f(x^*) \neq 0$. Consequently, this assumption must be wrong and we can infer $\nabla f(x^*) = 0$. Let α_k^e and α_k^a denote the step sizes determined by exact line search and by backtracking in iteration k, respectively. By definition of the exact step size, it follows

$$\begin{aligned} f(x^k + \alpha_k^e d^k) - f(x^k) &= \min_{\alpha \geq 0} f(x^k + \alpha d^k) - f(x^k) \\ &\leq f(x^k + \alpha_k^a d^k) - f(x^k) \leq -\gamma \alpha_k^a \cdot \|\nabla f(x^k)\|^2. \end{aligned} \tag{5.3}$$

Consequently, the step produced by exact line search also satisfies the Armijo condition (with a different step size) and hence, we can reuse the same proof for exact line search. ■

Let us now introduce the following sets that are connected to our convergence analysis:

$$\mathcal{X}^* := \{x \in \mathbb{R}^n : f(x) = \inf_{x\in\mathbb{R}^n} f(x)\}, \quad \mathcal{S} := \{x \in \mathbb{R}^n : \nabla f(x) = 0\},$$
$$\mathfrak{A} := \{x \in \mathbb{R}^n : \exists\ (k_\ell)_\ell \text{ such that } x^{k_\ell} \to x, \ell \to \infty\}.$$

Here, the set $\mathcal{X}^*$ denotes the solution set of the problem $\min_{x\in\mathbb{R}^n} f(x)$, $\mathcal{S}$ denotes the set of

associated stationary points, and $\mathfrak{A}$ denotes the set of accumulation points of the sequence $(x^k)_k$ generated by the gradient descent method. Consequently, the result in Theorem 5.11 can be compactly written as $\mathfrak{A} \subseteq \mathcal{S}$. Notice that Theorem 5.11 does not guarantee existence of cluster points (we might have $\mathfrak{A} = \emptyset$). Theorem 5.11 also does not ensure convergence of the *whole* sequence $(x^k)_k$. Next, we study convergence under a convexity assumption.

Theorem 5.12: Global Convergence Under Convexity

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and continuously differentiable and let $(x^k)_k$ be generated by the gradient method utilizing the following step size strategy:

- Armijo line search (backtracking) with $\sigma, \gamma \in (0, 1)$ and $s > 0$.

Suppose that the solution set $\mathcal{X}^*$ is nonempty. Then, $(x^k)_k$ converges to a global solution of the problem $\min_{x \in \mathbb{R}^n} f(x)$.

Proof. By Theorem 5.11, the sequence is $(f(x^k))_k$ is nonincreasing and due to $\mathcal{X}^* \neq \emptyset$, we can infer that $(f(x^k))_k$ converges to some limit $\xi \in \mathbb{R}$. For every $x \in \mathcal{X}^*$, we have

$$\begin{aligned} \|x^{k+1} - x\|^2 &= \|x^k - x\|^2 - 2\alpha_k \langle \nabla f(x^k), x^k - x \rangle + \alpha_k^2 \|\nabla f(x^k)\|^2 \\ &\leq \|x^k - x\|^2 + 2\alpha_k [f(x) - f(x^k)] + s\alpha_k \|\nabla f(x^k)\|^2 \\ &\leq \|x^k - x\|^2 + \frac{s}{\gamma}[f(x^k) - f(x^{k+1})]. \end{aligned} \tag{5.4}$$

Here, we used the convexity of f, $\alpha_k \leq s$, the global optimality of $x \in \mathcal{X}^*$, and the Armijo condition $f(x^{k+1}) - f(x^k) \leq -\gamma\alpha_k \|\nabla f(x^k)\|^2$. Summing the latter expression for $k = 0, ..., \ell$ yields

$$\|x^{\ell+1} - x\|^2 \leq \|x^0 - x\|^2 + \frac{s}{\gamma}[f(x^0) - f(x^{\ell+1})] \leq \|x^0 - x\|^2 + \frac{s}{\gamma}[f(x^0) - \xi] \quad \forall\, \ell \geq 0.$$

Consequently, $(x^k)_k$ is bounded and there exists an accumulation point $\bar{x} \in \mathfrak{A}$ of $(x^k)_k$. By Theorem 5.11, we have $\bar{x} \in \mathcal{S} = \mathcal{X}^*$. Let $(x^{k_\ell})_\ell$ be a subsequence converging to $\bar{x}$. Then, for every $\epsilon > 0$ there exists $L \in \mathbb{N}$ such that

$$\|x^{k_\ell} - \bar{x}\|^2 < \frac{\epsilon}{2} \quad \text{and} \quad f(x^{k_\ell}) - \xi < \frac{\gamma\epsilon}{2s} \quad \forall\, \ell \geq L.$$

The second estimate follows from the convergence $f(x^k) \to \xi$ as $k \to \infty$. Now, setting $x = \bar{x}$ in (5.4), we obtain:

$$\|x^{k+1} - \bar{x}\|^2 \leq \|x^{k_\ell} - \bar{x}\|^2 + \frac{s}{\gamma} \sum_{i=k_\ell}^{k} [f(x^i) - f(x^{i+1})] \leq \|x^{k_\ell} - \bar{x}\|^2 + \frac{s}{\gamma}[f(x^{k_\ell}) - \xi] < \varepsilon$$

for all $k \geq k_L$. But this establishes convergence $\lim_{k\to\infty} x^k = \bar{x}$. ∎

The proof presented in Theorem 5.12 is prototypical for the analysis of convex optimization

methods and to obtain stronger convergence results for the whole sequence of iterates. The property in (5.4) is connected to quasi Fejér monotonicity – a concept that we will now briefly introduce in more detail.

Definition 5.13: Quasi Fejér Monotonicity

A sequence $(y^k)_k$ is called **quasi Fejér convergent** to a set $\mathcal{Y} \subset \mathbb{R}^n$ if it holds that

$$\|y^{k+1} - y\|^2 \leq (1 + \beta_k)\|y^k - y\|^2 + \gamma_k \quad \forall\, k \in \mathbb{N}, \quad \forall\, y \in \mathcal{Y},$$

where (β_k), (γ_k) are nonnegative sequences satisfying $\sum_{k=0}^{\infty} \beta_k < \infty$ and $\sum_{k=0}^{\infty} \gamma_k < \infty$.

The proof of Theorem 5.12 is then a special case of the following theorem.

Theorem 5.14: Convergence of Quasi Fejér Sequences

Let the sequence $(y^k)_k$ be quasi Fejér convergent to a nonempty set $\mathcal{Y} \subset \mathbb{R}^n$ in the sense of Definition 5.13. Then, $(y^k)_k$ is bounded. Moreover, if an accumulation point y of $(y^k)_k$ belongs to $\mathcal{Y}$, then we have $\lim_{k\to\infty} y^k = y$.

The proof of this theorem is similar to the proof strategy presented in Theorem 5.12 and will be part of the exercises. At this point, we may wonder why convergence of the whole sequence of iterates $(x^k)_k$ as shown in Theorem 5.12 is a desirable property. We will discuss this question now briefly:

- Convergence of the sequence $(x^k)_k$ is typically a very basic requirement for studying rates of convergence. If x^* is the limit of $(x^k)_k$, then these rates are stated as follows:

  $$\|x^k - x^*\| \leq p_k \quad \text{or} \quad \|x^{k+1} - x^*\| \leq q_k\|x^k - x^*\|,$$

  where (often) $p_k \to 0$ and q_k is suitably chosen. Such conditions can not hold if $(x^k)_k$ does not converge to some x^*.
- Convergence of the full sequence is very desirable in stochastic settings or when the gradient is corrupted by some noise. In this case, the termination criterion $\|\nabla f(x^{k+1})\| \leq \varepsilon$ might not be suitable or is too expensive to evaluate. However, if we can still guarantee or generalize convergence of $(x^k)_k$ in such cases, then stopping the algorithm at some iteration k and using the last iterate x^k as output is still an appropriate strategy that is supported by the theoretical results. One goal of this section is to discuss and introduce different theoretical tools and concepts that can be utilized in the convergence analysis of optimization algorithm.

5.5.2. Convergence Analysis under Lipschitz Continuity

In the following, we assume that the objective function f is continuously differentiable and its gradient ∇f is *Lipschitz continuous* over $\mathbb{R}^n$, i.e., we have

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall\, x, y \in \mathbb{R}^n,$$

where $L > 0$ is the associated *Lipschitz constant*. The class of functions with Lipschitz gradient with constant L is denoted by $C_L^{1,1}(\mathbb{R}^n)$ or $C_L^{1,1}$. The concept of Lipschitz continuity is an extremely important tool in the convergence analysis of optimization methods. If ∇f is Lipschitz continuous, then we can derive a quadratic upper bound on the function f:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \quad \forall\, x, y \in \mathbb{R}^n.$$

This property is known as the celebrated *descent lemma*. An illustration is given in Figure 5.4.

Lemma 5.15: Descent Lemma

Suppose that $f \in C_L^{1,1}(\mathbb{R}^n)$, then it follows:

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2}\|y - x\|^2 \quad \forall\, x, y \in \mathbb{R}^n.$$

Proof. Let us define $\phi : \mathbb{R} \to \mathbb{R}$, $\phi(t) = f(x + t(y - x))$. Then, by the fundamental theorem of calculus, we have for all $x, y \in \mathbb{R}^n$

$$f(y) - f(x) = \phi(1) - \phi(0) = \int_0^1 \phi'(t)\mathrm{d}t = \int_0^1 \nabla f(x + t(y - x))^\top (y - x)\mathrm{d}t,$$

where we used the chain rule to calculate ϕ'. We obtain:

$$\begin{aligned} |f(y) - f(x) - \langle \nabla f(x), y - x \rangle| &= \left| \int_0^1 [\nabla f(x + t(y - x)) - \nabla f(x)]^\top (y - x)\mathrm{d}t \right| \\ &\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \|y - x\| \mathrm{d}t \\ &\leq L\|y - x\|^2 \cdot \int_0^1 t\mathrm{d}t = \frac{L}{2}\|y - x\|^2 \end{aligned}$$

for all $x, y \in \mathbb{R}^n$. ■

If f is twice continuously differentiable, then Lipschitz continuity of the gradient mapping is equivalent to boundedness of the Hessian. This is verified in detail in the next lemma.

Lemma 5.16: Lipschitz Continuity and Boundedness

Let f be a twice continuously differentiable function. The next two claims are equivalent:

(i) $f \in C_L^{1,1}(\mathbb{R}^n)$.

(ii) $\|\nabla^2 f(x)\| \leq L$ for any $x \in \mathbb{R}^n$.

Proof. Let us define $\Phi : \mathbb{R} \to \mathbb{R}^n$, $\Phi(t) = \nabla f(x + t(y - x))$. Then, by the fundamental theorem of calculus, we have for all $x, y \in \mathbb{R}^n$

$$\nabla f(y) - \nabla f(x) = \Phi(1) - \Phi(0) = \int_0^1 \Phi'(t)\mathrm{d}t = \int_0^1 \nabla^2 f(x + t(y - x))(y - x)\mathrm{d}t, \tag{5.5}$$

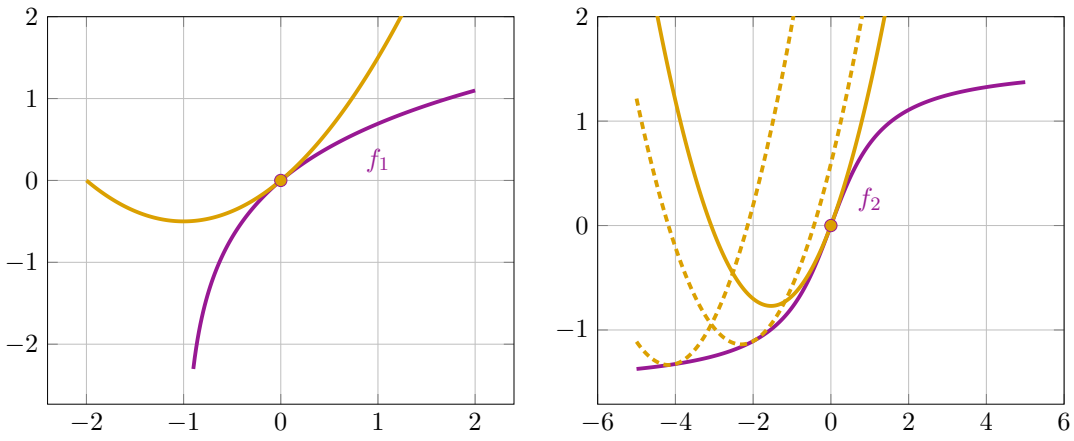


Figure 5.4: Plot of $f_1(x) = \log(1+x)$ and $f_2(x) = \arctan(x)$ and of the quadratic upper surrogate functions $q_1(y) = y + \frac{1}{2}y^2$ and $q_2(y) = y + \frac{3\sqrt{3}}{16}y^2$ at $x = 0$. (The Lipschitz constant of the derivatives $f_1'(x) = (1+x)^{-1}$ and $f_2'(x) = (1+x^2)^{-1}$ is given by $L = 1$ and $L = 3\sqrt{3}/8$, respectively).

where we used the chain rule to calculate Φ'. Now, if $\|\nabla^2 f(x)\| \leq L$ for all x, we obtain

$$\begin{aligned}\|\nabla f(y) - \nabla f(x)\| &= \left\| \int_0^1 \nabla^2 f(x + t(y-x))(y-x) \mathrm{d}t \right\| \\ &\leq \int_0^1 \|\nabla^2 f(x + t(y-x))(y-x)\| \mathrm{d}t \leq \int_0^1 L \cdot \|y - x\| \mathrm{d}t = L\|y-x\|,\end{aligned}$$

which establishes $f \in C_L^{1,1}(\mathbb{R}^n)$. Notice that the inequality $\| \int_0^1 V(t)\mathrm{d}t\| \leq \int_0^1 \|V(t)\|\mathrm{d}t$ for some continuous mapping $V : [0,1] \to \mathbb{R}$ can be explained and shown by approximating the integral with Riemann sums and by applying the triangle inequality. Next, on the other hand, let us suppose $f \in C_L^{1,1}(\mathbb{R}^n)$. Let $h \in \mathbb{R}^n$ be arbitrary, then a Taylor expansion of ∇f and the Lipschitz continuity of ∇f yield

$$Lt\|h\| \geq \|\nabla f(x+th) - \nabla f(x)\| = \|t\nabla^2 f(x)h + o(t\|h\|)\|, \quad \text{for} \quad t \to 0.$$

Dividing both sides by t and taking the limit $t \to 0$, this implies $\|\nabla^2 f(x)h\| \leq L\|h\|$ for all h and x and thus, since the spectral norm is an induced norm, we have $\|\nabla^2 f(x)\| \leq L$. ■

Let us briefly discuss an exemplary application of this theorem.

Let $q(x) = \frac{1}{2}x^\top Ax + b^\top x + c$ be a quadratic function with given symmetric matrix $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, $c \in \mathbb{R}$. Then, the last theorem shows that ∇q is Lipschitz continuous with $L = \|A\|_2$.

Exploiting the Lipschitz continuity of ∇f will allow us to strengthen Theorem 5.11 and to establish global convergence for variants of the gradient method using constant or diminishing step sizes.

Theorem 5.17: Global Convergence Under Lipschitz Continuity

Let $f \in C_L^{1,1}(\mathbb{R}^n)$ and let $(x^k)_k$ be the sequence generated by the gradient method with one of the following step size strategies:

- constant step size $\bar{\alpha} \in (0, \frac{2}{L})$,
- exact line search,
- Armijo line search (backtracking) with $\sigma, \gamma \in (0,1)$ and $s > 0$,
- diminishing step sizes, i.e., $\alpha_k \to 0$ and $\sum_{k=0}^{\infty} \alpha_k = \infty$.

Then, we either have $f(x^k) \to -\infty$ or $(f(x^k))_k$ converges to a finite value and it holds that $\nabla f(x^k) \to 0$ as $k \to \infty$. (In particular, every accumulation point of $(x^k)_k$ is a stationary point in this case).

Proof. We first discuss the convergence behavior of the gradient method using a constant step size $\bar{\alpha}$. Applying the descent lemma, we obtain

$$f(x^{k+1}) - f(x^k) \leq -\bar{\alpha}\|\nabla f(x^k)\|^2 + \frac{L\bar{\alpha}^2}{2}\|\nabla f(x^k)\|^2 = \left[\frac{L\bar{\alpha}}{2} - 1\right]\bar{\alpha}\|\nabla f(x^k)\|^2. \tag{5.6}$$

Consequently, if $\bar{\alpha} \in (0, \frac{2}{L})$, then the sequence $(f(x^k))_k$ is decreasing and it converges to some $\xi \in \mathbb{R} \cup \{-\infty\}$. We only need to consider the case $\xi > -\infty$. As before by summing the last estimate for $k = 0, ..., M$, it follows

$$\left[1 - \frac{L\bar{\alpha}}{2}\right]\bar{\alpha}\sum_{k=0}^{M}\|\nabla f(x^k)\|^2 \leq \sum_{k=0}^{M} f(x^k) - f(x^{k+1}) = f(x^0) - f(x^{M+1}) \leq f(x^0) - \xi.$$

Thus, taking the limit $M \to \infty$, the series $\sum_{k=0}^{\infty}\|\nabla f(x^k)\|^2$ is convergent and we can infer $\nabla f(x^k) \to 0$ as $k \to \infty$. Our proof also shows that the Armijo condition is satisfied for α with

$$\frac{L\alpha}{2} - 1 \leq -\gamma \quad \iff \quad \alpha \leq \frac{2(1-\gamma)}{L}$$

Thus, every step size α_k generated by backtracking satisfies $\alpha_k \geq \min\{s, \frac{2\sigma(1-\gamma)}{L}\} =: \tau$. Under the assumption $\xi > -\infty$ and mimicking the steps in the proof of Theorem 5.11, we obtain

$$f(x^0) - \xi = \lim_{k\to\infty}\sum_{i=0}^{k-1} f(x^i) - f(x^{i+1}) \geq \sum_{i=0}^{\infty}\gamma\alpha_i\|\nabla f(x^i)\|^2 \geq \gamma\tau\sum_{i=0}^{\infty}\|\nabla f(x^i)\|^2.$$

Consequently, under Lipschitz continuity of ∇f we can strengthen the convergence statement in Theorem 5.11 to $\nabla f(x^k) \to 0$ as $k \to \infty$. (The proof for exact line search is identical using (5.3)).

When utilizing diminishing step sizes, the estimate (5.6) still holds with $\bar{\alpha}$ being substituted by α_k. Since $(\alpha_k)_k$ converges to zero, there exists $c > 0$ such that we have $f(x^{k+1}) - f(x^k) \leq$

$-c\alpha_k\|\nabla f(x^k)\|^2$ for all k sufficiently large. This shows that $(f(x^k))_k$ is again monotonically decreasing and thus, the function values need to converge to some $\xi \in \mathbb{R} \cup \{-\infty\}$. In the case $\xi > -\infty$, summation of the descent estimate yields

$$\sum_{k=0}^{\infty} \alpha_k \|\nabla f(x^k)\|^2 < \infty \tag{5.7}$$

and $\lim_{k\to\infty} \alpha_k\|\nabla f(x^k)\|^2 = 0$. Using $\sum_{k=0}^{\infty} \alpha_k = \infty$, this implies $\liminf_{k\to\infty} \|\nabla f(x^k)\| = 0$. Otherwise, there exists $\varepsilon > 0$ and $K \in \mathbb{N}$ such that $\|\nabla f(x^k)\| \geq \varepsilon$ for all $k \geq K$. But then we obtain the contradiction

$$\infty > \sum_{k=K}^{\infty} \alpha_k \|\nabla f(x^k)\|^2 \geq \varepsilon^2 \sum_{k=K}^{\infty} \alpha_k = \infty.$$

We still need to show $\lim_{k\to\infty} \|\nabla f(x^k)\| = 0$. Assume on the contrary that the sequence $(\|\nabla f(x^k)\|)_k$ does not converge to zero. Then there exists $\varepsilon > 0$ and infinite sequences $(t_i)_i$ and $(\ell_i)_i$ such that $\ell_i > t_i$ for all i and

$$\|\nabla f(x^{t_i})\| \geq 2\varepsilon, \quad \|\nabla f(x^{\ell_i})\| < \varepsilon, \quad \text{and} \quad \|\nabla f(x^k)\| \geq \varepsilon$$

for all $k = t_i + 1, ..., \ell_i - 1$. Using (5.7), it follows

$$\infty > \sum_{k=0}^{\infty} \alpha_k \|\nabla f(x^k)\|^2 \geq \sum_{i=0}^{\infty} \sum_{k=t_i}^{\ell_i - 1} \alpha_k \|\nabla f(x^k)\|^2 \geq \varepsilon^2 \sum_{i=0}^{\infty} \sum_{k=t_i}^{\ell_i - 1} \alpha_k$$

and consequently, setting $\beta_i := \sum_{k=t_i}^{\ell_i - 1} \alpha_k$, we have $\beta_i \to 0$ as $i \to \infty$. Applying the Cauchy-Schwarz inequality, i.e.,

$$\sum_{k=1}^{K} |a_k b_k| \leq \left[\sum_{k=1}^{K} a_k^2\right]^{\frac{1}{2}} \left[\sum_{k=1}^{K} b_k^2\right]^{\frac{1}{2}},$$

we obtain

$$\begin{aligned} \|x^{\ell_i} - x^{t_i}\| \leq \sum_{k=t_i}^{\ell_i - 1} \|x^{k+1} - x^k\| &= \sum_{k=t_i}^{\ell_i - 1} \sqrt{\alpha_k}\sqrt{\alpha_k}\|\nabla f(x^k)\| \\ &\leq \sqrt{\beta_i}\left[\sum_{k=t_i}^{\ell_i - 1} \alpha_k \|\nabla f(x^k)\|^2\right]^{\frac{1}{2}} \leq \sqrt{\beta_i}\left[\sum_{k=0}^{\infty} \alpha_k \|\nabla f(x^k)\|^2\right]^{\frac{1}{2}}. \end{aligned}$$

Consequently, combining (5.7) and $\beta_i \to 0$, we can infer $\|x^{\ell_i} - x^{t_i}\| \to 0$. Finally, using the special definition of the sequences $(t_i)_i$ and $(\ell_i)_i$, the inverse triangle inequality, and the Lipschitz continuity of ∇f, we have

$$\varepsilon \leq |\|\nabla f(x^{\ell_i})\| - \|\nabla f(x^{t_i})\|| \leq \|\nabla f(x^{\ell_i}) - \nabla f(x^{t_i})\| \leq L\|x^{\ell_i} - x^{t_i}\| \to 0$$

as $i \to \infty$, which is a contradiction. Consequently, our assumption was wrong and it follows $\nabla f(x^k) \to 0$ as $k \to \infty$. ■

5.5.3. Isolated Accumulation Points

Before investigating complexity results and convergence rates, we want to briefly discuss a situation that allows to guarantee convergence of the full sequence $(x^k)_k$ of iterates. Specifically, we will now assume that there exists an isolated accumulation point $\bar{x}$ of the sequence $(x^k)_k$, i.e., there is a neighborhood of $\bar{x}$ that does not contain any other accumulation points of $(x^k)_k$. The full result is summarized in the following theorem.

Theorem 5.18: Convergence to Isolated Accumulation Points

Let f be continuously differentiable and let $(x^k)_k$ be generated by the gradient method. Suppose that one of the following scenarios is satisfied:

- Backtracking with $\sigma, \gamma \in (0,1)$ and $s > 0$ is used;
- It additionally holds $f \in C_L^{1,1}(\mathbb{R}^n)$ and a constant step size $\bar{\alpha} \in (0, \frac{2}{L})$ or diminishing step sizes, i.e., $\alpha_k \to 0$ and $\sum_{k=0}^{\infty} \alpha_k = \infty$, are used.

Assume that $(x^k)_k$ has an isolated accumulation point $\bar{x}$. Then, $(x^k)_k$ converges to $\bar{x}$ and we have $\nabla f(x^k) \to \nabla f(\bar{x}) = 0$ as $k \to \infty$.

Proof. Let $\varepsilon > 0$ be chosen such that the set $B_\varepsilon(\bar{x})$ does not contain any other accumulation points of $(x^k)_k$. We now assume that the sequence $(x^k)_k$ does not converge to $\bar{x}$. We define the subset $L \subset \mathbb{N}$ as follows

$$\ell \in L \quad :\Longleftrightarrow \quad \|x^{\ell-1} - \bar{x}\| \leq \frac{\varepsilon}{2} \quad \text{and} \quad \|x^{\ell} - \bar{x}\| > \varepsilon.$$

Since $\bar{x}$ is an accumulation point, $(x^k)_k$ does not converge to $\bar{x}$, and no other accumulation points are in $B_\varepsilon(\bar{x})$, the set L must contain infinitely many elements. We obtain

$$\|x^{\ell} - \bar{x}\| \leq \|x^{\ell} - x^{\ell-1}\| + \|x^{\ell-1} - \bar{x}\| \leq \alpha_{\ell-1}\|\nabla f(x^{\ell-1})\| + \frac{\varepsilon}{2} \quad \forall\, \ell \in L. \tag{5.8}$$

In all mentioned step size strategies, the step sizes $(\alpha_k)_k$ can be upper bounded by some $c_\alpha > 0$. Moreover, in the second scenario, Theorem 5.17 is applicable and due to $\bar{x} \in \mathfrak{A}$, we can infer $\nabla f(x^k) \to 0$ as $k \to \infty$. But this yields a contradiction in (5.8). Similarly, applying Theorem 5.11, it follows $\nabla f(\bar{x}) = 0$ and by continuity of ∇f, it is possible to choose ε sufficiently small to reach a contradiction in (5.8). ■

Notice that the last step of the proof can also be slightly modified. Due to $x^{\ell-1} \in \bar{B}_\varepsilon(\bar{x})$ for all $\ell \in L$, the sequence $(x^{\ell-1})_{\ell \in L}$ is bounded and we can select another subsequence $\tilde{L} \subset L$ such that $(x^{\ell-1})_{\ell \in \tilde{L}}$ converges. Since $\bar{x}$ is the only accumulation point of $(x^k)_k$ in $B_\varepsilon(\bar{x})$, the limit of $(x^{\ell-1})_{\ell \in \tilde{L}}$ can only be $\bar{x}$. We can then apply Theorem 5.11 and Theorem 5.17 directly. Theorem 5.18 is typically applied based on the following result.

Lemma 5.19: Isolated Stationary Points

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is C^2 in a neighborhood of a stationary point $\bar{x}$ of f and assume that the Hessian $\nabla^2 f(\bar{x})$ is invertible. Then there exists $\varepsilon, \eta > 0$ such that

$$\|\nabla f(x)\| \geq \eta \|x - \bar{x}\| \quad \forall\, x \in B_\varepsilon(\bar{x}).$$

In particular, $\bar{x}$ is an isolated zero of the gradient mapping ∇f.

Proof. First of all, we have

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

Setting $\eta := 1/(2\|\nabla^2 f(\bar{x})^{-1}\|)$, this implies:

$$\|\nabla^2 f(\bar{x})(x - \bar{x})\| \geq 2\eta \|x - \bar{x}\|.$$

By the definition of differentiability there exists $\varepsilon > 0$ such that

$$\|\nabla f(x) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x - \bar{x})\| \leq \eta \|x - \bar{x}\| \quad \forall\, x \in B_\varepsilon(\bar{x}).$$

Utilizing $\nabla f(\bar{x}) = 0$ and the triangle inequality, we obtain

$$\begin{aligned} 2\eta \|x - \bar{x}\| &\leq \|\nabla^2 f(\bar{x})(x - \bar{x})\| = \|\nabla f(x) - [\nabla f(x) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x - \bar{x})]\| \\ &\leq \|\nabla f(x)\| + \|\nabla f(x) - \nabla f(\bar{x}) - \nabla^2 f(\bar{x})(x - \bar{x})\| \leq \|\nabla f(x)\| + \eta \|x - \bar{x}\| \end{aligned}$$

for all $x \in B_\varepsilon(\bar{x})$. ■

Next, let us assume that the Hessian $\nabla^2 f(\bar{x})$ is positive definite at one of the accumulation points $\bar{x}$ of $(x^k)_k$. By the sufficient second-order optimality conditions we can then infer that $\bar{x}$ is a strict local minimum and by Lemma 5.19, $\bar{x}$ is also an isolated stationary point. Since all accumulation points of $(x^k)_k$ are stationary points, this implies that $\bar{x}$ is an isolated accumulation point of $(x^k)_k$. In this setting, Theorem 5.18 is applicable and the whole sequence $(x^k)_k$ has to converge to the local minimum $\bar{x}$. Hence, Theorem 5.18 and Lemma 5.19 partially explain why we typically experience full convergence when applying the gradient method. Theorem 5.18 is also known as the "capture theorem" and the presented proof technique also works more generally.

Lemma 5.20: Isolated Accumulation Points

Let $(x^k)_k$ be a sequence in $\mathbb{R}^n$ with an isolated accumulation point $\bar{x}$. Then, the following conditions are equivalent:

(i) For every subsequence $(x^{k_\ell})_\ell$ converging to $\bar{x}$, it holds that $\lim_{\ell \to \infty} \|x^{k_\ell + 1} - x^{k_\ell}\| = 0$.

(ii) The sequence $(x^k)_k$ converges to $\bar{x}$.

Remark 5.21. Let us finally comment on the choice of descent directions and potential generalizations. Most of the proofs presented in the last sections also work for more general

descent directions that are *gradient related*. Specifically, if there exist $c_1, c_2 > 0$ such that

$$c_1\|\nabla f(x^k)\|^2 \leq -\nabla f(x^k)^\top d^k \quad \text{and} \quad \|d^k\|^2 \leq c_2\|\nabla f(x^k)\|^2$$

for all $k \in \mathbb{N}$, then the different convergence statements in Theorem 5.11, Theorem 5.17, and Theorem 5.18 are still valid in general. (And only some of the step size bounds need to be adjusted).

5.5.4. Complexity Results and Convergence Rates

We now analyze convergence rates and complexity results for the gradient method. We start with several complexity results. If the gradient of f is Lipschitz continuous, we can derive a general but rather slow rate of convergence in terms of the gradient norm.

Theorem 5.22: Complexity Result

Let $f \in C_L^{1,1}(\mathbb{R}^n)$ be given and let $(x^k)_k$ be generated by the gradient method. Suppose that $(f(x^k))_k$ converges to some $f^* \in \mathbb{R}$. Then, for any $k = 0, 1, 2, ...$

$$\min_{i=0,1,\ldots,k} \|\nabla f(x^i)\| \leq \sqrt{\frac{f(x^0) - f^*}{\mathcal{C}(k+1)}}, \quad \mathcal{C} = \begin{cases} \bar{\alpha}(1 - \frac{L\bar{\alpha}}{2}) & \text{constant step size,} \\ \frac{1}{2L} & \text{exact line search,} \\ \gamma \min\{s, \frac{2\sigma(1-\gamma)}{L}\} & \text{Armijo line search.} \end{cases}$$

Proof. The proof of Theorem 5.17 implies

$$\mathcal{C}\sum_{i=0}^{k} \|\nabla f(x^i)\|^2 \leq f(x^0) - f(x^{k+1}) \leq f(x^0) - f^* \tag{5.9}$$

for all k where $\mathcal{C} = (1 - \frac{L\bar{\alpha}}{2})\bar{\alpha}$ and $\mathcal{C} = \gamma\tau$ if a constant step size $\bar{\alpha}$ or backtracking is used, respectively. Similarly, in the case we utilize an exact step size, it follows

$$\begin{aligned} f(x^{k+1}) - f(x^k) &= \min_{\alpha \geq 0} f(x^k - \alpha\nabla f(x^k)) - f(x^k) \leq f(x^k - \nabla f(x^k)/L) - f(x^k) \\ &\leq -\frac{1}{L}\|\nabla f(x^k)\|^2 + \frac{L}{2}\|\nabla f(x^k)/L\|^2 = -\frac{1}{2L}\|\nabla f(x^k)\|^2. \end{aligned}$$

Notice that $\alpha = \frac{1}{L}$ minimizes the quadratic function $\alpha \mapsto (\frac{L\alpha}{2} - 1)\alpha$ which explains this specific choice. Hence, when using exact line search, (5.9) holds with $\mathcal{C} = 1/2L$. Now, (5.9) immediately implies

$$\min_{i=0,1,\ldots,k} \|\nabla f(x^i)\|^2 \cdot \mathcal{C}(k+1) \leq f(x^0) - f^*,$$

which establishes the desired complexity bound. ■

Theorem 5.22 tells us that we need to run the gradient method for approximately

$$k \geq \frac{f(x^0) - f^*}{\mathcal{C}\varepsilon^2} = \mathcal{O}(\varepsilon^{-2})$$

iterations in order to guarantee $\min_{i=0,\dots,k-1} \|\nabla f(x^i)\| \leq \varepsilon$. Notice that this is a quite weak result. (Specifically, it does not guarantee $\|\nabla f(x^{k-1})\| \leq \varepsilon$ for the last iterate x^{k-1}). In the convex case, Theorem 5.22 can be significantly improved.

Theorem 5.23: Complexity Under Convexity

Let $f \in C_L^{1,1}(\mathbb{R}^n)$ be given and suppose that f is convex with $\mathcal{X}^* \neq \emptyset$. Let $(x^k)_k$ be generated by the gradient method utilizing

(i) a constant step size $\bar{\alpha} \in (0, \frac{2}{L})$ or Armijo line search with $\sigma, \gamma \in (0,1)$ and $s > 0$. Then, (x^k) converges to some $x^* \in \mathcal{X}^*$ and we have

$$f(x^k) - f^* = f(x^k) - f(x^*) = o(k^{-1}).$$

(ii) a constant step size $\bar{\alpha} \in (0, \frac{1}{L}]$ or backtracking with $\sigma \in (0,1)$, $\gamma \in [\frac{1}{2}, 1)$, and $s > 0$. Then, in addition to the results in part (i), it holds that:

$$f(x^k) - f^* \leq \frac{\|x^0 - x^*\|^2}{2\mathcal{B}k}, \quad \mathcal{B} = \begin{cases} \bar{\alpha} & \text{constant step size,} \\ \tau = \min\{s, \frac{2\sigma(1-\gamma)}{L}\} & \text{Armijo line search.} \end{cases}$$

Proof. As before, utilizing the descent lemma or the Armijo condition, we have

$$f(x^{k+1}) \leq f(x^k) - \mathcal{C}\|\nabla f(x^k)\|^2. \tag{5.10}$$

where $\mathcal{C} = (2 - L\bar{\alpha})\frac{\bar{\alpha}}{2}$ and $\mathcal{C} = \gamma\tau$, $\tau = \min\{s, \frac{2\sigma(1-\gamma)}{L}\}$, if a constant step size $\bar{\alpha}$ or backtracking is used. Furthermore, we have $\alpha_k \geq \tau$ for all $k \in \mathbb{N}$ and for step sizes generated by backtracking. In addition, convergence of the sequence $(x^k)_k$ is ensured by Theorem 5.12 if Armijo line search is used. In order to show convergence of the full sequence $(x^k)_k$ for constant step sizes, we can mimic the proof of Theorem 5.12. In particular, due to the convexity of f, we obtain

$$\begin{aligned} \|x^{k+1} - x\|^2 &= \|x^k - x\|^2 - 2\alpha_k \nabla f(x^k)^\top (x^k - x) + \alpha_k^2 \|\nabla f(x^k)\|^2 \\ &\leq \|x^k - x\|^2 + 2\alpha_k [f(x) - f(x^k)] + \alpha_k^2 \|\nabla f(x^k)\|^2 \end{aligned} \tag{5.11}$$

for every $x \in \mathcal{X}^*$ and for $\alpha_k = \bar{\alpha}$. The last estimate can be combined with inequality (5.10) as in the proof of Theorem 5.12 which allows to apply quasi Fejér monotonicity. Next, we assume that $(x^k)_k$ converges to some global solution $x^* \in \mathcal{X}^*$. Setting $\Delta_k = f(x^k) - f^* = f(x^k) - f(x^*)$ and $d_k := \|x^k - x^*\|^2$, the convexity of f yields

$$\Delta_k = f(x^k) - f^* \leq \nabla f(x^k)^\top (x^k - x^*) \leq \|\nabla f(x^k)\| \|x^k - x^*\| = \|\nabla f(x^k)\| d_k.$$

Hence, we can infer

$$\Delta_{k+1} \leq \Delta_k - \frac{\mathcal{C}}{d_k^2} \Delta_k^2 \leq \Delta_k.$$

We next divide this estimate by $\Delta_k\Delta_{k+1}$:

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\mathcal{C}}{d_k^2}\frac{\Delta_k}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{\mathcal{C}}{d_k^2}.$$

Summing the resulting inequality over all k, we obtain

$$\Delta_{k+1} \leq \left[\frac{1}{\Delta_0} + \sum_{i=0}^{k}\frac{\mathcal{C}}{d_i^2}\right]^{-1} \quad \Longrightarrow \quad (k+1)\Delta_{k+1} \leq \left[\frac{1}{(k+1)\Delta_0} + \frac{\mathcal{C}}{k+1}\sum_{i=0}^{k}\frac{1}{d_i^2}\right]^{-1}.$$

Due to $x^k \to x^*$, it now follows $d_k \to 0$ and $d_k^{-2} \to \infty$. Consequently, we have

$$\frac{\mathcal{C}}{k+1}\sum_{i=0}^{k} d_i^{-2} \to \infty$$

which establishes $\Delta_{k+1} = o(1/(k+1))$ (This can be shown by conducting a proof of contradiction). This finishes the proof of the first part. We continue to verify the second part. We first consider backtracking. Setting $x = x^*$ in (5.11), we have

$$\Delta_k \leq \frac{1}{2\alpha_k}\left[\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right] + \frac{\alpha_k}{2}\|\nabla f(x^k)\|^2.$$

Together with the Armijo condition, this yields

$$\begin{aligned}\Delta_{k+1} = f(x^{k+1}) - f(x^k) + \Delta_k &\leq \frac{\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2}{2\alpha_k} + \left[\frac{1}{2} - \gamma\right]\alpha_k\|\nabla f(x^k)\|^2 \\ &\leq \frac{1}{2\tau}\left[\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right].\end{aligned}$$

Summing this estimate for all k and using the monotonicity of $(f(x^k))_k$, we can finally infer

$$k\Delta_k \leq \sum_{i=0}^{k-1}\Delta_{i+1} \leq \frac{1}{2\tau}\|x^0 - x^*\|^2.$$

The proof for constant step sizes is essentially identical. Notice that the factor $\frac{1}{2} - \gamma$ before $\|\nabla f(x^k)\|^2$ takes the form $\frac{1}{2}(L\bar{\alpha} - 1) \leq 0$ in this case. ■

Here and in contrast to Theorem 5.22, we require approximately

$$k \geq \frac{\|x^0 - x^*\|^2}{2\mathcal{B}\varepsilon} = \mathcal{O}(\varepsilon^{-1})$$

iterations in order to ensure $f(x^k) - f^* \leq \varepsilon$. This is a significant improvement compared to the previous nonconvex case. We further notice that the last two complexity results do not give any insight on how close the iterates $(x^k)_k$ are to a potential stationary point or solution of the problem. In the remainder of this section, we will investigate the local convergence behavior of $(x^k)_k$ in more detail.

The following definition characterizes the "speed of convergence" of a sequence $(x^k)_k$.

Definition 5.24: Convergence Rates

We say that the sequence $(x^k)_k$

(i) **converges q-linear** with rate $\eta \in (0,1)$ to $x^* \in \mathbb{R}^n$ if there is $\ell \geq 0$ such that

$$\|x^{k+1} - x^*\| \leq \eta \cdot \|x^k - x^*\|, \quad \forall\, k \geq \ell.$$

(ii) **converges q-superlinear** to $x^* \in \mathbb{R}^n$ if $x^k \to x^*$ and

$$\|x^{k+1} - x^*\| = o(\|x^k - x^*\|), \quad k \to \infty \quad \iff \quad \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} \to 0, \quad k \to \infty.$$

(iii) **converges q-quadratic** to x^* if $x^k \to x^*$ and we have

$$\|x^{k+1} - x^*\| = O(\|x^k - x^*\|^2), \quad k \to \infty \quad \iff \quad \|x^{k+1} - x^*\| \leq C\|x^k - x^*\|^2$$

for some $C > 0$ and all $k \geq 0$.

(iv) **converges r-linear** with rate $\eta \in (0,1)$ to x^* if there is a sequence $(\beta_k)_k \subset (0,\infty)$ that converges q-linearly to 0 with rate η and we have $\|x^k - x^*\| \leq \beta_k$ for $k \to \infty$.

(v) **converges r-superlinear** to $x^* \in \mathbb{R}^n$ if there is a sequence $(\beta_k)_k \subset (0,\infty)$ that converges q-superlinearly to 0 and we have $\|x^k - x^*\| \leq \beta_k$ for $k \to \infty$.

Example 5.25. In this example, we discuss and present sequences with different convergence properties to illustrate Definition 5.24.

- Let $\rho \in (0,1)$ be given, then the sequence $(x^k)_k$ with $x^k := \rho^k$ converges q-linear to $x^* = 0$ with rate $\eta = \rho$. In fact, we have:

$$\frac{|x^{k+1} - x^*|}{|x^k - x^*|} = \frac{\rho^{k+1}}{\rho^k} = \rho, \quad \forall\, k \geq 0.$$

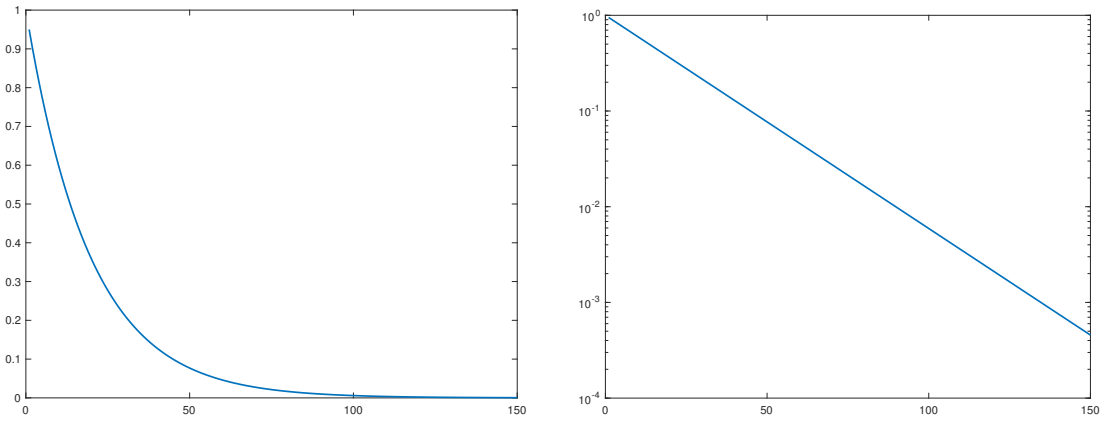
- Obviously, the sequence $(x^k)_k$ does not converge q-superlinearly! Let us define $y^k := \rho^{k^2}$ for all k. Then, $(y^k)_k$ converges to 0 and it holds that

$$\frac{|y^{k+1}|}{|y^k|} = \frac{\rho^{k^2+2k+1}}{\rho^{k^2}} = \rho^{2k+1} \to 0 \quad \text{as} \quad k \to \infty.$$

  Hence $(y^k)_k$ converges q-superlinearly to 0.

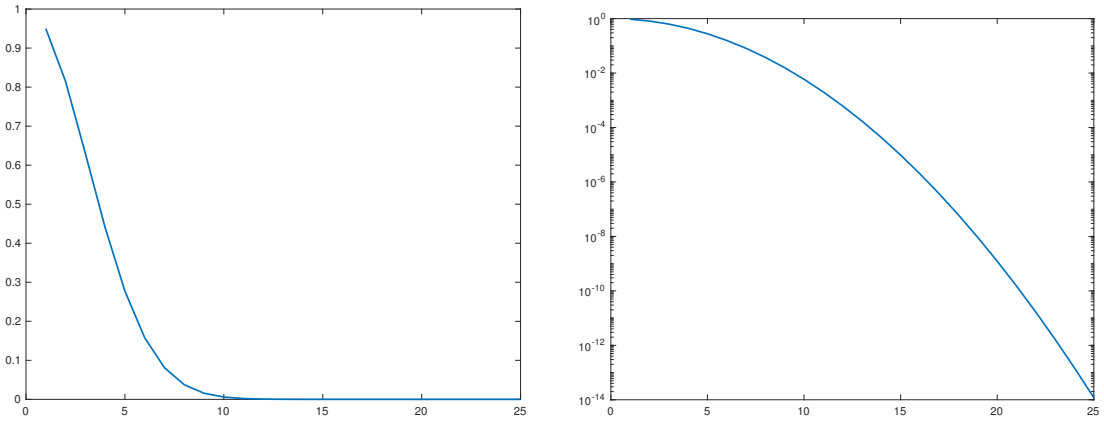
- The sequence $(y^k)_k$ does not converge q-quadratically:

$$\frac{|y^{k+1}|}{|y^k|^2} = \frac{\rho^{k^2+2k+1}}{\rho^{2k^2}} = \frac{\rho^{2k+1}}{\rho^{k^2}} \to \infty \quad \text{as} \quad k \to \infty.$$



(a) Plot of $(0.95^k)_k$ (b) Logarithmic plot of $(0.95^k)_k$

Figure 5.5: Plot of the sequence $(x^k)_k$ for $x^k = \rho^k$, $\rho = 0.95$, and $k \in \{1, ..., 150\}$. Since the value $x^k = \rho^k$ can be very small for large k, often a logarithmic plot is considered. The plot in (b) shows convergence of the sequence $\tilde{x}^k = \log_{10}(x^k) = \log_{10}(\rho) \cdot k \approx -0.022 \cdot k$ and the labels of the y-axis are given by $10^{\tilde{x}^k}$. We see that q-linear convergence corresponds to linear behavior with slope $\log_{10}(\rho)$ in this case.



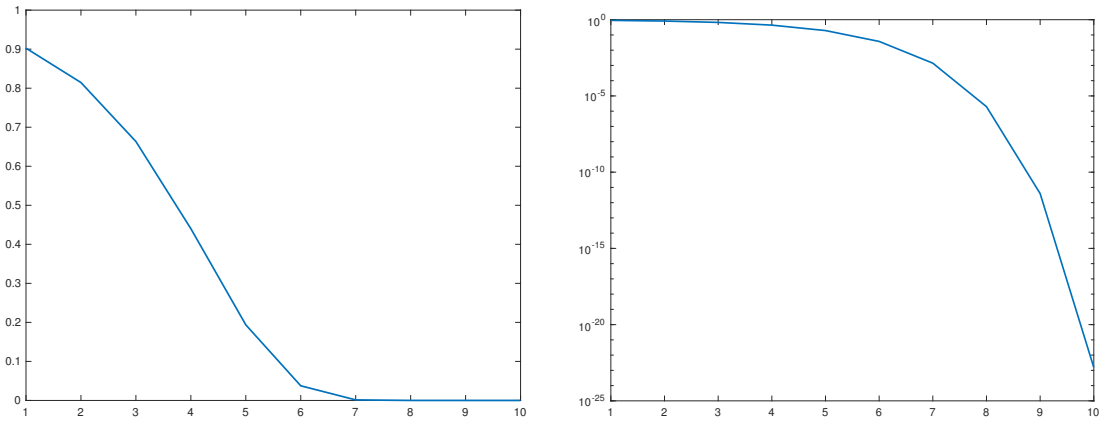
(a) Plot of $(0.95^{k^2})_k$ (b) Logarithmic plot of $(0.95^{k^2})_k$

Figure 5.6: Standard and logarithmic plot of the sequence $(y^k)_k$ for $y^k = \rho^{k^2}$, $\rho = 0.95$, and $k \in \{1, ..., 25\}$. Superlinear convergence is much faster than linear convergence. The graph in the logarithmic plot always has a (small nonzero) "negative curvature", i.e., it does not approach a linear line. This leads to fast convergence.

An example of a q-quadratically converging sequence is $(z^k)_k$ with $z^k := \rho^{2^k}$. We have $z^k \to 0$ and:

$$\frac{|z^{k+1}|}{|z^k|^2} = \frac{\rho^{2^{k+1}}}{\rho^{2 \cdot 2^k}} = 1, \quad \forall\, k \geq 0.$$

Exemplary plots of the sequence discussed in Example 5.25 are given in Figure 5.5–Figure 5.7.



(a) Plot of $(0.95^{2^k})_k$ (b) Logarithmic plot of $(0.95^{2^k})_k$

Figure 5.7: Standard and logarithmic plot of the sequence $(z^k)_k$ for $z^k = \rho^{2^k}$, $\rho = 0.95$, and $k \in \{1, ..., 10\}$. Quadratic convergence implies that the number of correct digits (i.e., the digits that coincide with the limit $z^* = 0$) doubles after each iteration. Hence, the logarithmic plot in (b) is reminiscent of a quadratic function that opens downward.

We first present two refined properties of convex functions that have Lipschitz continuous gradients which will turn out to be helpful in our subsequent analysis.

Lemma 5.26: Convexity and Lipschitz Continuity

Let $f \in C_L^{1,1}(\mathbb{R}^n)$ be a given function and assume that f is convex. Then, we have

(i) The gradient mapping ∇f is $1/L$-**cocoercive**, i.e., it holds that $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2$ for all $x, y \in \mathbb{R}^n$.

(ii) If f is additionally μ-strongly convex, then it follows:

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{L\mu}{L+\mu}\|x - y\|^2 + \frac{1}{L+\mu}\|\nabla f(x) - \nabla f(y)\|^2$$

for all $x, y \in \mathbb{R}^n$.

Proof. Let $x \in \mathbb{R}^n$ be given and consider the function $g(y) := f(y) - \langle \nabla f(x), y\rangle$. This function is still convex and its gradient $\nabla g(y) = \nabla f(y) - \nabla f(x)$ is Lipschitz continuous with constant L (notice that x is fixed here). We also see that x is a global minimum of g satisfying $\nabla g(x) = 0$. Hence, using the descent lemma, it follows

$$\begin{aligned} g(x) = \min_{y\in\mathbb{R}^n} g(y) \leq \min_{\alpha\in\mathbb{R}}\ g(y - \alpha\nabla g(y)) &\leq \min_{\alpha\in\mathbb{R}}\ g(y) - \alpha\|\nabla g(y)\|^2 + \frac{L\alpha^2}{2}\|\nabla g(y)\|^2 \\ &= g(y) - \frac{1}{2L}\|\nabla g(y)\|^2. \end{aligned}$$

By exchanging the role of x and y, the same proof also shows $f(y) - \langle \nabla f(y), y\rangle \leq f(x) - \langle \nabla f(y), x\rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2$. Summing the latter two inequalities establishes $(1/L)$-cocoercivity and finishes the proof of part (i). In order to verify the second part, let us define $h(x) = f(x) - \frac{\mu}{2}\|x\|^2$. Then, we have $\nabla h(x) = \nabla f(x) - \mu x$ and

$$\begin{aligned} &h(y) - h(x) - \langle \nabla h(x), y - x\rangle \\ &\quad = f(y) - f(x) - \langle \nabla f(x), y - x\rangle - \frac{\mu}{2}[\|y\|^2 - 2\langle x, y - x\rangle - \|x\|^2] \leq \frac{L-\mu}{2}\|y - x\|^2. \end{aligned}$$

Moreover, it holds that $h(y) - h(x) - \langle \nabla h(x), y - x\rangle \geq 0$ by using the μ-strong convexity of f, i.e., h is a convex function. Exchanging the role of x and y and adding the resulting inequalities yields

$$0 \leq \langle \nabla h(x) - \nabla h(y), x - y\rangle \leq (L - \mu)\|y - x\|^2.$$

Since this estimate holds for all x and y, this implies that h satisfies the descent lemma property in Lemma 5.15 (which follows by mimicking the proof of Lemma 5.15). Since the proof of part (i) only relied on the descent lemma, this shows that ∇h is $1/(L-\mu)$-cocoercive which implies that $h \in C^{1,1}_{L-\mu}(\mathbb{R}^n)$. The cocoercivity of ∇h now implies

$$\begin{aligned} \langle \nabla h(x) - \nabla h(y), x - y\rangle &\geq \frac{1}{L-\mu}\|\nabla h(x) - \nabla h(y)\|^2 \\ &= \frac{1}{L-\mu}\left[\|\nabla f(x) - \nabla f(y)\|^2 - 2\mu\langle \nabla f(x) - \nabla f(y), x - y\rangle + \mu^2\|x - y\|^2\right]. \end{aligned}$$

Furthermore, we have

$$\langle \nabla h(x) - \nabla h(y), x - y\rangle = \langle \nabla f(x) - \nabla f(y), x - y\rangle - \mu\|x - y\|^2.$$

Combining these expressions and rearranging the terms finally establishes the bound stated in part (ii). ■

If the function f is additionally strongly convex, we can further strengthen our convergence results and establish q-linear convergence of the sequence of iterates $(x^k)_k$.

Before we state the theorem, let us notice that μ-strong convexity of f implies $f(y) - f(x) \geq \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$ for all $x, y \in \mathbb{R}^n$ and that f has a unique global minimizer $x^* \in \mathbb{R}^n$. (This was shown in Theorem 4.14 (iii) and in Assignment A1.6). Specifically, it follows

$$f(x) - f(x^*) \geq \frac{\mu}{2}\|x - x^*\|^2 \quad \text{and} \quad \|\nabla f(x)\|^2 \geq 2\mu[f(x) - f(x^*)]. \tag{5.12}$$

Here, the last estimate follows from minimizing $f(y) - f(x)$ and $\langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$ separately with respect to y (the minimum of the latter expression is attained at $y = x - \frac{1}{\mu}\nabla f(x)$). We now present a linear convergence theorem for strongly convex problems.

Theorem 5.27: Strong Convexity and Linear Convergence

Suppose that $f \in C_L^{1,1}(\mathbb{R}^n)$ and let f be strongly convex with parameter $\mu > 0$. Let x^* be the unique global minimizer of $\min_x f(x)$ and let the sequence $(x^k)_k$ be generated by the gradient method using

(i) exact line search, backtracking with $\sigma, \gamma \in (0,1)$ and $s > 0$, or a constant step size $\bar{\alpha} \in (0, \frac{2}{L})$. Then, $(x^k)_k$ converges r-linearly to x^* and $(f(x^k))_k$ converges q-linearly to $f(x^*)$ with

$$f(x^{k+1}) - f(x^*) \leq \eta(f(x^k) - f(x^*)) \quad \text{where} \quad \eta = 1 - 2\mathcal{C}\mu$$

and $\mathcal{C}$ is given as in Theorem 5.22.

(ii) Armijo line search with $\sigma \in (0,1)$, $\gamma \in [\frac{1}{2}, 1)$, and $s > 0$ or a constant step size $\bar{\alpha} \in (0, \frac{2}{L+\mu}]$. Then, the iterates $(x^k)_k$ converge q-linearly to x^* and it follows

$$\|x^{k+1} - x^*\| \leq \rho \cdot \|x^k - x^*\|, \quad \rho = \begin{cases} \sqrt{1 - \frac{2L\mu\bar{\alpha}}{L+\mu}} & \text{constant step size,} \\ \sqrt{\frac{1-\tau\mu-(2\gamma-1)\tau^2\mu^2}{1+\tau\mu}} & \text{backtracking.} \end{cases}$$

Proof. We start with the proof of part (i) and we mimic our earlier proofs. In particular, we have

$$f(x^{k+1}) - f(x^*) \leq f(x^k) - f(x^*) - \mathcal{C}\|\nabla f(x^k)\|^2$$

where $\mathcal{C}$ depends on the chosen step size strategy and is defined as in Theorem 5.22. Applying (5.12), this yields

$$f(x^{k+1}) - f(x^*) \leq (1 - 2\mathcal{C}\mu)(f(x^k) - f(x^*)).$$

Repeating this estimate and using (5.12), we obtain

$$\frac{\mu}{2}\|x^k - x^*\|^2 \leq f(x^k) - f(x^*) \leq [1 - 2\mathcal{C}\mu]^k (f(x^0) - f(x^*)) \leq [1 - 2\mathcal{C}\mu]^k \frac{L}{2}\|x^0 - x^*\|^2$$

which establishes r-linear convergence of $(x^k)_k$ and finishes the proof of part (i). To show the second part, we first consider the case when a fixed step size is used. Applying Lemma 5.26 then yields:

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - 2\bar{\alpha}\nabla f(x^k)^\top (x^k - x^*) + \bar{\alpha}^2\|\nabla f(x^k)\|^2 \\ &\leq \left[1 - \frac{2L\mu\bar{\alpha}}{L+\mu}\right]\|x^k - x^*\|^2 + \left[\bar{\alpha} - \frac{2}{L+\mu}\right]\bar{\alpha}\|\nabla f(x^k)\|^2. \end{aligned}$$

Hence, the result follows immediately from the choice of $\bar{\alpha}$. Next, assume that backtracking is used to generate the step sizes $(\alpha_k)_k$. By the strong convexity of f, it follows

$$\begin{aligned} f(x^*) - f(x^k) &\geq \langle \nabla f(x^k), x^* - x^k \rangle + \frac{\mu}{2}\|x^k - x^*\|^2 \\ &= -\frac{1}{\alpha_k}\langle x^{k+1} - x^k, x^* - x^k \rangle + \frac{\mu}{2}\|x^k - x^*\|^2 \end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{\alpha_k}\langle x^{k+1} - x^*, x^* - x^k\rangle + \left[\frac{\mu}{2} - \frac{1}{\alpha_k}\right]\|x^k - x^*\|^2 \\
&= -\frac{1}{2\alpha_k}\|x^{k+1} - x^k\|^2 + \frac{1}{2\alpha_k}\|x^{k+1} - x^*\|^2 + \left[\frac{\mu}{2} - \frac{1}{2\alpha_k}\right]\|x^k - x^*\|^2 \\
&= \frac{1}{2\alpha_k}\|x^{k+1} - x^*\|^2 + \left[\frac{\mu}{2} - \frac{1}{2\alpha_k}\right]\|x^k - x^*\|^2 - \frac{\alpha_k}{2}\|\nabla f(x^k)\|^2.
\end{aligned}
$$

Combining this with the Armijo condition, rearranging the terms, and applying (5.12) and $\gamma \geq 1/2$, we obtain

$$
\begin{aligned}
\left[\frac{1}{2\alpha_k} - \frac{\mu}{2}\right]\|x^k - x^*\|^2 &\geq f(x^k) - f(x^*) + \frac{1}{2\alpha_k}\|x^{k+1} - x^*\|^2 - \frac{\alpha_k}{2}\|\nabla f(x^k)\|^2 \\
&\geq f(x^{k+1}) - f(x^*) + \frac{1}{2\alpha_k}\|x^{k+1} - x^*\|^2 + \left[\gamma - \frac{1}{2}\right]\alpha_k\|\nabla f(x^k)\|^2 \\
&\geq \left[\frac{\mu}{2} + \frac{1}{2\alpha_k}\right]\|x^{k+1} - x^*\|^2 + \left[\gamma - \frac{1}{2}\right]\alpha_k\mu^2\|x^k - x^*\|^2.
\end{aligned}
$$

Overall, we then have

$$
\|x^{k+1} - x^*\|^2 \leq \frac{1 - \alpha_k\mu - (2\gamma - 1)\alpha_k^2\mu^2}{1 + \alpha_k\mu}\|x^k - x^*\|^2.
$$

The function $\rho(\alpha) = \frac{1-\alpha\mu-(2\gamma-1)\alpha^2\mu^2}{1+\alpha\mu} = 1 - \left[2\gamma - 1 + \frac{3-2\gamma}{1+\alpha\mu}\right]\alpha\mu$ is decreasing in α and hence, as a consequence, it follows $\rho(\alpha_k) \leq \rho(\tau)$ for all k. This finishes the proof of Theorem 5.27. ■

Consequently, in the strongly convex case, we require approximately

$$
k \geq \log\left[\frac{f(x^0) - f(x^*)}{\varepsilon}\right] / \log(\eta^{-1}) = \mathcal{O}(\log(\varepsilon^{-1}))
$$

iterations to ensure $f(x^k) - f^* \leq \varepsilon$. Furthermore, $(f(x^k))_k$ and $(x^k)_k$ converge to $f(x^*)$ and x^* at a *geometric rate*. The speed of convergence is faster if the rates η and ρ are as small as possible. For the constant step size, the optimal rate in Theorem 5.27 (ii) is achieved if we choose $\bar{\alpha} = \frac{2}{L+\mu}$. In this case, we obtain:

$$
\|x^{k+1} - x^*\| \leq \frac{L - \mu}{L + \mu} \cdot \|x^k - x^*\| = \frac{\kappa - 1}{\kappa + 1} \cdot \|x^k - x^*\|
$$

where $\kappa = L/\mu$. In addition, in the case $\gamma = \frac{1}{2}$, $s = 1$, $L > 1$, the constant τ reduces to $\tau = \frac{\sigma}{L}$. The rate ρ when utilizing Armijo line search is then given by:

$$
\rho = \sqrt{\frac{1 - \tau\mu}{1 + \tau\mu}} = \sqrt{\frac{\kappa - \sigma}{\kappa + \sigma}}
$$

which is worse than the rate for the gradient method with constant step size. We also see that the rate of convergence is faster when κ is as small as possible and close to one.

Remark 5.28. It is possible to improve the result in Theorem 5.27 (ii) and to establish q-linear convergence for $\gamma \in (\bar{\gamma}, \frac{1}{2})$, i.e., for $\gamma < \frac{1}{2}$. (By applying the estimate $\|\nabla f(x^k)\| \leq L\|x^k - x^*\|$). However, in general, we can not guarantee q-linear convergence of $(x^k)_k$ when using the gradient method with backtracking for arbitrary choices of γ as the following example demonstrates. Consider the mapping

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) = \begin{cases} x^2 & \text{if } x < 0, \\ \frac{2}{3}x^2 & \text{if } x \geq 0. \end{cases}$$

Due to $f(x) \geq \frac{2}{3}x^2$, this function is strongly convex with parameter $\mu = \frac{2}{3}$. Moreover, the derivative satisfies

$$|f'(x) - f'(y)| = \begin{cases} 2|x-y| & \text{if } x, y < 0, \\ \frac{2}{3}|x-y| & \text{if } x, y \geq 0 \end{cases}$$

and $|f'(x) - f'(y)| \leq \frac{2}{3}|x-y| + \frac{4}{3}|x| \leq 2|x-y|$ or $|f'(x) - f'(y)| \leq \frac{2}{3}|x-y| + \frac{4}{3}|y| \leq 2|x-y|$ if $xy \leq 0$. Hence, f' is Lipschitz continuous with constant $L = 2$. Let us now consider an arbitrary initial point $x^0 < 0$. Then, we have $f(x^0) = x_0^2$ and $x^1 = x^0 - f'(x^0) = -x^0 > 0$, $f(x^1) = \frac{2}{3}x_0^2$. Consequently, the iterate x^1 satisfies

$$f(x^1) - f(x^0) = -\frac{1}{3}x_0^2 = -\frac{1}{12}\|f'(x^0)\|^2.$$

Similarly, if we set $x^2 = x^1 - f'(x^1) = -\frac{1}{3}x^1 < 0$, we obtain $f(x^2) = \frac{1}{9}x_1^2 = \frac{2}{3}x_1^2 - \frac{5}{9}x_1^2 = f(x^1) - \frac{5}{16}\|f'(x^1)\|^2$. Consequently, choosing $\gamma = \frac{1}{12}$, we see that the Armijo condition is always satisfied for the full step size $\alpha_k = 1$ and iteratively we obtain

$$x^{k+1} = -x^k \quad \text{if } x^k < 0 \quad \text{and} \quad x^{k+1} = -\frac{1}{3}x^k \quad \text{if } x^k > 0.$$

But then $(x^k)_k$ can not converge q-linearly to 0.

5.5.5. Numerical Examples

We now continue with several numerical examples and experiments.

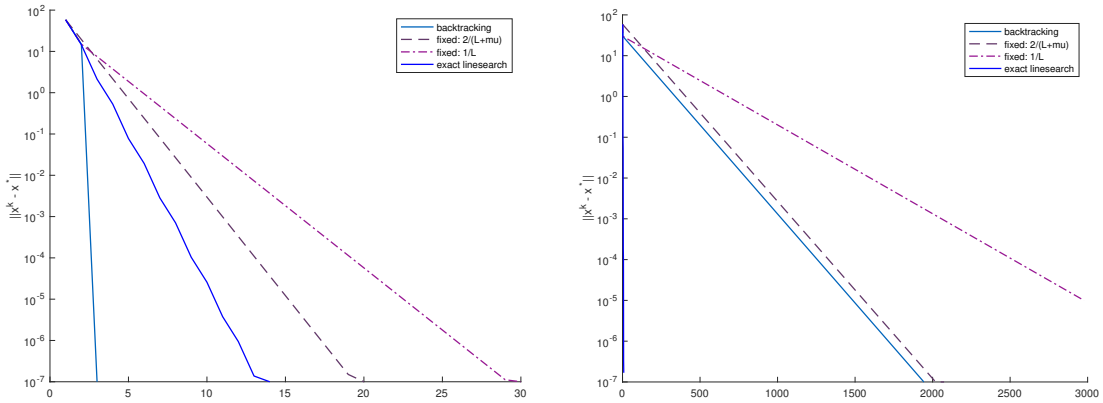
Example 5.29. We first consider the simple quadratic problem

$$\min_{x \in \mathbb{R}^n} \; f(x) = \frac{1}{2}x^\top Ax + b^\top x + c, \tag{5.13}$$

where $A \in \mathbb{R}^{n \times n}$ is a symmetric, positive definite matrix and $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$ are given. In this case, we have $A = \nabla^2 f(x)$ and

$$\lambda_{\min}(A)\|h\|^2 \leq h^\top Ah = h^\top \nabla^2 f(x) h \leq \lambda_{\max}(A)\|h\|^2 \quad \forall\, h.$$

Hence, f is strongly convex with $\mu = \lambda_{\min}(A) > 0$ and its gradient is Lipschitz continuous with $L = \lambda_{\max}(A)$. According to Theorem 5.27, we expect the gradient method to converge



(a) Problem 1: $\kappa = 2$ (b) Problem 2: $\kappa = 200$

Figure 5.8: Logarithmic plot of sequences $(\|x^k - x^*\|)_k = (\|x^k\|)_k$. The iterates are generated using the gradient method with different step size rule to solve the quadratic problems described in Example 5.29.

to the unique solution $x^* = -A^{-1}b$ at a linear convergence rate. As discussed before, the rates η and ρ significantly depend on the factor

$$\kappa = \frac{L}{\mu} = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

Thus, in the quadratic case, κ coincides with the condition number $\kappa(A) = \|A^{-1}\|\|A\|$ of the matrix A. (If A is symmetric and positive definite, then we have $\kappa(A) = \lambda_{\max}(A)/\lambda_{\min}(A) = \kappa$). Furthermore, using this connection we can restate our theoretical observations in terms of the condition number: if the condition number $\kappa(A) = \kappa$ is large, then η and ρ are close to 1 and we expect slow convergence.

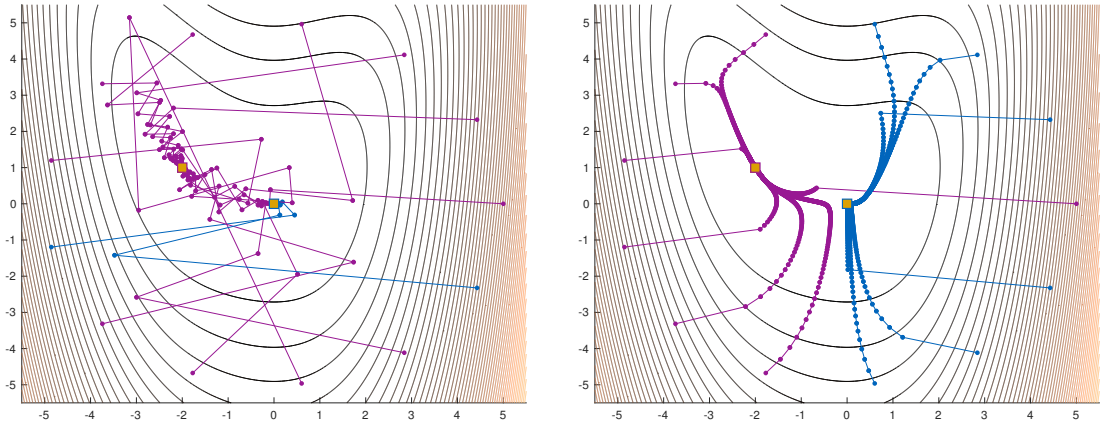
We now run the gradient method with backtracking ($\gamma = \sigma = \frac{1}{2}$, $s = 1$), constant step sizes ($\bar{\alpha} = \frac{2}{L+\mu}$, $\bar{\alpha} = \frac{1}{L}$), and exact line search on (5.13) with

$$A_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{100} \end{pmatrix}, \quad b_1 = b_2 = 0, \quad c_1 = c_2 = 0.$$

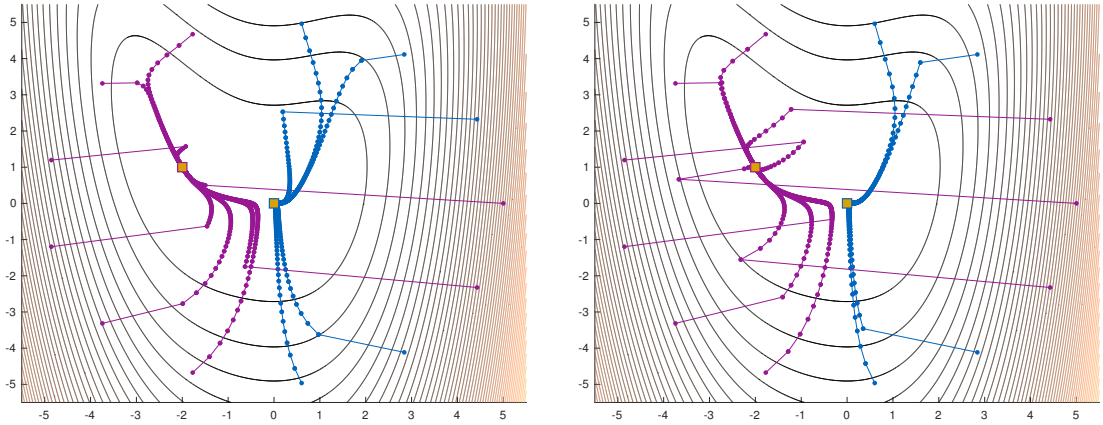
In the first case, the condition number is given by $\kappa(A_1) = 2$ and in the second example, we have $\kappa(A_2) = 200$. The methods are run starting at the initial point $x^0 = (5, 3)^\top$. In Figure 5.8, we plot the sequence $(\|x^k - x^*\|)_k = (\|x^k\|)_k$ using a logarithmic plot for the different step size strategies. The figures clearly demonstrate how the rate of convergence is affected by the condition number κ. In the more general non-quadratic case, a large condition number of the Hessian $\nabla^2 f$ typically indicates and leads to slow convergence of the gradient method.

Example 5.30. We again consider the optimization problem

$$\min_{x \in \mathbb{R}^2} \; f(x) = x_1^4 + 2(x_1 - x_2)x_1^2 + 4x_2^2.$$



(a) Backtracking (b) Diminishing Step Sizes



(c) Fixed Step Size $\bar{\alpha} = \frac{1}{100}$ (d) Fixed Step Size $\bar{\alpha} = \frac{1}{75}$

Figure 5.9: Solution path of the gradient method using different step sizes and initial points when applied to the minimization problem in Example 5.30.

As verified in Example 3.9, this problem has two stationary points $x_1^* = (0,0)^\top$ and $x_2^* = (-2,1)^\top$. Furthermore, we have shown that x_1^* is a saddle point (with positive semi-definite Hessian) and x_2^* is a strict local minimizer. We now want to apply the gradient method with different step size strategies and investigate its performance. On a theoretical level, we expect the following behavior:

- The mapping f is coercive (why?). Hence, the level sets of f are compact. In particular, if we use a step size rule guaranteeing descent, then the sequence of generated iterates $(x^k)_k$ has to be bounded.
- The two stationary points of f are isolated. Using the boundedness of $(x^k)_k$ and our convergence results, $(x^k)_k$ has accumulation points that need to be stationary points. Hence, every accumulation point of $(x^k)_k$ needs to be isolated. This implies that the whole sequence has to either converge to x_1^* or x_2^* (by the capture theorem).
- Since f is nonconvex, the weak complexity result for the gradient norms $(\|\nabla f(x^k)\|)_k$

is only applicable.

We are now testing the gradient descent method with:

- backtracking with $\gamma = 0.1$, $\sigma = 0.5$, and $s = 1$.
- diminishing step sizes $\alpha_k = 0.01(k+2)^{-\frac{1}{8}}$.
- constant step size $\bar{\alpha} = \frac{1}{100}$ and $\bar{\alpha} = \frac{1}{75}$.

As initial point, we choose 13 different points on the sphere $\{x \in \mathbb{R}^2 : \|x\| = 5\}$. The results and solutions paths are summarized in Figure 5.9.

References. This section mainly follows [5, Section 1.2 and 1.3] and [2, Chapter 4]. The quasi Féjer analysis presented in Theorem 5.12 is based on [6].

6. Newton's Method

6.1. Pure Newton's Method

In this section, we consider the unconstrained optimization problem

$$\min_{x\in\mathbb{R}^n} \; f(x)$$

under the assumption that f is twice continuously differentiable. We will now discuss another traditional approach for unconstrained problems – Newton's method. Newton's method is a *second order* approach, i.e., a method that – besides function and gradient values – also uses Hessian information. Newton's method for optimization problems is an iterative scheme that tries to find a zero of the nonlinear system of equations:

$$\nabla f(x) = 0. \tag{6.1}$$

Given an iterate x^k, this equation is equivalent to

$$\nabla f(x^k + d) = 0, \tag{6.2}$$

i.e., $d = d^k$ is a solution of (6.2) if and only if $x = x^k + d^k$ is a solution of the system (6.1). The main idea is to approximate $\nabla f(x^k + d)$ using a Taylor approximation and to linearize $\nabla f(x^k + d)$:

$$\nabla f(x^k + d) = \nabla f(x^k) + \nabla^2 f(x^k)d + o(\|d\|) \quad d \to 0.$$

A step of Newton's method is then given by solving the linearized system of equations:

$$\nabla f(x^k) + \nabla^2 f(x^k)d^k = 0 \quad \Longrightarrow \quad x^{k+1} = x^k + d^k = x^k - (\nabla^2 f(x^k))^{-1}\nabla f(x^k). \tag{6.3}$$

It is also possible to derive Newton's step using a different perspective. Let x^k be a given iterate and consider the following quadratic Taylor approximation of f:

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

The idea is now to define the next iterate x^{k+1} as minimizer of the latter quadratic approximation of f around x^k (neglecting the remainder terms $o(\|x - x^k\|^2)$):

$$x^{k+1} = \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \; f(x^k) + \nabla f(x^k)^\top (x - x^k) + \frac{1}{2}(x - x^k)^\top \nabla^2 f(x^k)(x - x^k).$$

In fact, if $\nabla^2 f(x^k)$ is positive definite, then there exists a unique global minimizer x^{k+1} and the latter update formula is well-defined. In this case, the solution of the quadratic problem is exactly Newton's update:

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

The vector $d^k = -(\nabla^2 f(x^k))^{-1}\nabla f(x^k)$ is called the *Newton direction* and the procedure in (6.3) defines the basic *Newton's method*. If $\nabla^2 f(x^k)$ is positive definite then d^k is a descent

direction. The full method is shown in Algorithm 6.1.

Algorithm 6.1: Pure Newton's Method for Optimization Problems

1 Initialization: Select an initial point $x^0 \in \mathbb{R}^n$.
for $k = 0, 1, ...$ **do**
2 Compute the Newton direction d^k which is the solution of the linear system

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

3 Set $x^{k+1} = x^k + d^k$.
4 If $\|\nabla f(x^{k+1})\| \leq \varepsilon$, then STOP and x^{k+1} is the output.

We now verify that Newton's method has favorable local convergence properties and converges locally q-superlinearly or q-quadratically.

Theorem 6.1: Local Convergence of Newton's Method

Let f be a twice continuously differentiable function and let x^* be a local minimizer of f that satisfies the second-order sufficient optimality conditions. Then there exists $\varepsilon > 0$ and $\mu > 0$ such that

- The minimizer x^* is the only stationary point in $B_\varepsilon(x^*)$.
- We have $\lambda_{\min}(\nabla^2 f(x)) \geq \mu$ for all $x \in B_\varepsilon(x^*)$.
- For every $x^0 \in B_\varepsilon(x^*)$, Algorithm 6.1 either terminates with $x^k = x^*$ or it generates a sequence $(x^k)_k \subset B_\varepsilon(x^*)$ that converges q-superlinearly to x^*.
- In addition, if $\nabla^2 f$ is Lipschitz continuous on $B_\delta(x^*)$ with constant L, then the rate of convergence is q-quadratic (if the algorithm does not terminate after finitely many steps) and it holds that:

$$\|x^{k+1} - x^*\| \leq \frac{L}{2\mu}\|x^k - x^*\|^2 \quad \forall\ k.$$

The proof of Theorem 6.1 is partly based on continuity properties of invertible matrices and definite matrices. We now first present a preparatory result.

Lemma 6.2: Maintaining Positive Definiteness

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric and positive definite matrix. Then, for all $\mu \in (0, \lambda_{\min}(A))$ and all symmetric matrices $B \in \mathbb{R}^{n\times n}$ with $\|B\| \leq \lambda_{\min}(A) - \mu$, we have:

$$\lambda_{\min}(A + B) \geq \mu.$$

Proof. It holds that

$$\lambda_{\min}(A + B) = \min_{\|h\|=1} h^\top (A + B)h \geq \min_{\|h\|=1} h^\top Ah - \|B\| = \lambda_{\min}(A) - \|B\|,$$

as desired. ■

We can now present a proof of Theorem 6.1.

Proof. The second-order sufficient conditions imply that $\nabla^2 f(x^*)$ is positive definite. Hence, applying Lemma 5.19, x^* is an isolated stationary point, i.e., there exist $\varepsilon_1, \eta_1 > 0$ such that $\|\nabla f(x)\| \geq \eta_1 \|x - x^*\|$ for all $x \in B_{\varepsilon_1}(x^*)$. Furthermore, due to the continuity of $\nabla^2 f$ and positive definiteness of $\nabla^2 f(x^*)$ there exist $0 < \varepsilon_2 < \varepsilon_1$ and $\mu > 0$ such that

$$\lambda_{\min}(\nabla^2 f(x)) = \lambda_{\min}(\nabla^2 f(x^*) + \nabla^2 f(x) - \nabla^2 f(x^*)) \geq \mu > 0 \quad \forall\, x \in B_{\varepsilon_2}(x^*).$$

This follows from Lemma 6.2 and it implies $\|\nabla^2 f(x)^{-1}\| = (\lambda_{\min}(\nabla^2 f(x)))^{-1} \leq \mu^{-1}$ for all $x \in B_{\varepsilon_2}(x^*)$. We now have

$$\begin{aligned}
\nabla^2 f(x^k)(x^{k+1} - x^*) &= \nabla^2 f(x^k) d^k + \nabla^2 f(x^k)(x^k - x^*) \\
&= \nabla f(x^*) - \nabla f(x^k) + \nabla^2 f(x^k)(x^k - x^*) \\
&= \int_0^1 [\nabla^2 f(x^k + t(x^* - x^k)) - \nabla^2 f(x^k)](x^* - x^k) \mathrm{d}t =: R(x^k, x^*).
\end{aligned}$$

Notice that by the continuity of $\nabla^2 f$ it follows

$$\frac{\|R(x, x^*)\|}{\|x - x^*\|} \leq \int_0^1 \|\nabla^2 f(x + t(x^* - x)) - \nabla^2 f(x)\| \mathrm{d}t \to 0 \quad \text{as} \quad x \to x^*.$$

Consequently, for some arbitrary but fixed $\delta \in (0, 1)$ there exists $0 < \varepsilon \leq \varepsilon_2$ such that $\|R(x, x^*)\| \leq \delta\mu \|x - x^*\|$ for all $x \in B_\varepsilon(x^*)$. And thus, for every $x^k \in B_\varepsilon(x^*)$ we obtain

$$\|x^{k+1} - x^*\| = \|\nabla^2 f(x^k)^{-1} R(x^k, x^*)\| \leq \delta\mu \|\nabla^2 f(x^k)^{-1}\| \|x^k - x^*\| \leq \delta \|x^k - x^*\|.$$

Hence, if $x^0 \in B_\varepsilon(x^*)$, this shows $x^k \in B_{\delta^k \varepsilon}(x^*) \subset B_\varepsilon(x^*)$ inductively. In the case $\nabla f(x^k) = 0$, the algorithm terminates. Due to $B_\varepsilon(x^*) \subset B_{\varepsilon_1}(x^*)$ this can only happen when $x^k = x^*$. Otherwise, the algorithm generates a sequence that converges to x^* (as just shown). The q-superlinear convergence now follows from

$$\begin{aligned}
\frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} &\leq \|\nabla^2 f(x^k)^{-1}\| \int_0^1 \|\nabla^2 f(x^k + t(x^* - x^k)) - \nabla^2 f(x^k)]\| \mathrm{d}t \\
&\leq \frac{1}{\mu} \int_0^1 \|\nabla^2 f(x^k + t(x^* - x^k)) - \nabla^2 f(x^k)\| \mathrm{d}t \to 0,
\end{aligned}$$

as $k \to \infty$. If $\nabla^2 f$ is additionally Lipschitz continuous on $B_\varepsilon(x^*)$, then we obtain:

$$\int_0^1 \|\nabla^2 f(x^k + t(x^* - x^k)) - \nabla^2 f(x^k)\| \mathrm{d}t \leq \int_0^1 Lt \|x^k - x^*\| \, \mathrm{d}t = \frac{L}{2} \|x^k - x^*\|.$$

This finishes the proof of Theorem 6.1. ■

We continue with an important remark:

- Newton's method can also be applied to solve nonlinear equations of the form:

$$F : \mathbb{R}^n \to \mathbb{R}^n, \quad \text{find } x \text{ such that:} \quad F(x) = 0.$$

  (This is actually its original purpose). In this case, we require the point x^* to be a solution of the equation $F(x) = 0$ and the derivative $DF(x^*)$ to be a nonsingular matrix. Furthermore, one can show that x^* is an isolated zero of F and that $DF(x)$ is boundedly invertible in a neighborhood of x^*. If the initial point x^0 is sufficiently close to x^*, then Newton's method for $F(x) = 0$ generates a sequence $(x^k)_k$ that converges q-superlinearly or q-quadratically (under a corresponding Lipschitz condition) to x^*.

  This generalization of Newton's method is discussed in more detail in Appendix B.

Next, we discuss an example that demonstrates the local nature of Theorem 6.1.

Example 6.3. Let us consider the function $f(x) = \sqrt{1+x^2}$ and the problem $\min_x \ f(x)$. Then, the derivatives of f are given by

$$f'(x) = \frac{x}{\sqrt{1+x^2}}, \quad f''(x) = \frac{1}{(1+x^2)\sqrt{1+x^2}}$$

and hence, a Newton step can be calculated as follows

$$x^{k+1} = x^k - \frac{f'(x^k)}{f''(x^k)} = x^k - x^k(1 + (x^k)^2) = -(x^k)^3.$$

We see that the method diverges if $|x^0| \geq 1$ and it converges rapidly (at a cubic rate) to the global solution (and unique stationary point) $x^* = 0$ if $|x^0| < 1$.

6.2. Globalized Newton's Method

The theoretical results in the last section and Example 6.3 illustrate that the pure Newton method is a conceptual algorithm and it is only guaranteed to converge locally and under strong assumptions if the initial point x^0 is chosen sufficiently close to a solution x^* of the equation

$$\nabla f(x) = 0,$$

i.e., close to a stationary point of f. In order to make the method more robust and practical, we need to address the following issues:

- Introduce a globalization strategy that allows arbitrary choices of x^0.
- If $\nabla^2 f(x^k)$ is not positive definite, the Newton direction $d^k = -\nabla^2 f(x^k)^{-1}\nabla f(x^k)$ might not exist or might not be a descent direction. How can we ensure that the sequence of function values $(f(x^k))_k$ decreases?
- Can we find a design that does not destroy the favorable local properties of the Newton method?

A possible globalized version of Newton's method is presented in Algorithm 6.2. It introduces the following changes:

Algorithm 6.2: Globalized Newton's Method

1 Initialization: Select an initial point $x^0 \in \mathbb{R}^n$. Choose parameters $\sigma, \gamma \in (0,1)$, $s > 0$, $\beta_1, \beta_2 > 0$, and $p > 0$.
for $k = 0, 1, \ldots$ **do**
2 Compute the Newton direction s^k as solution of the system $\nabla^2 f(x^k)s^k = -\nabla f(x^k)$. If this is possible and if s^k satisfies

$$-\nabla f(x^k)^\top s^k \geq \min\{\beta_1, \beta_2 \|s^k\|^p\}\|s^k\|^2,$$

then set $d^k = s^k$. Otherwise set $d^k = -\nabla f(x^k)$.
3 Calculate a step size α_k using backtracking and set $x^{k+1} = x^k + \alpha_k d^k$.
4 If $\|\nabla f(x^{k+1})\| \leq \varepsilon$, then STOP and x^{k+1} is the output.

- After the computation of the Newton step, we check the quality of the Newton direction. If it is a sufficiently good descent direction, we accept it. Otherwise, we choose $d^k = -\nabla f(x^k)$ as our direction.
- We perform an additional Armijo line search to guarantee descent along the objective function and to damp the Newton steps.
- Typically, in order to accept as many Newton steps as possible, the parameters β_1, β_2 and p in Algorithm 6.2 are chosen rather small. A common choice is $\beta_1 = \beta_2 = 10^{-6}$ and $p = \frac{1}{10}$. Notice that also other types of acceptance mechanisms and tests are possible. For instance, we can also use

$$-\nabla f(x^k)^\top s^k \geq \min\{\beta_1, \beta_2 \|\nabla f(x^k)\|^p\}\|\nabla f(x^k)\|\|s^k\|$$

in step 2 of Algorithm 6.2 to monitor the quality of the Newton steps s^k.

It is possible to derive the following convergence result:

Theorem 6.4: Convergence of the Globalized Newton Method

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and let $(x^k)_k$ be generated by Algorithm 6.2. Then, every accumulation point of $(x^k)_k$ is a stationary point.
Additionally, if we have $\gamma \in (0, \frac{1}{2})$, $s = 1$, and if there exists an accumulation point x^* of $(x^k)_k$ such that $\nabla^2 f(x^*)$ is positive definite, then it follows:

(i) The point x^* is a strict local minimum of f.

(ii) The whole sequence $(x^k)_k$ converges to x^*.

(iii) There is $K \in \mathbb{N}$ such that for all $k \geq K$, the algorithm turns into the pure Newton's method with step size $\alpha_k = 1$. In particular, $(x^k)_k$ converges q-superlinearly to x^*. If $\nabla^2 f$ is Lipschitz continuous in a neighborhood of x^*, the rate of convergence is q-quadratic.

Proof. The proof of the global convergence is similar to our previous results. We will state it here for the sake of completeness. We first notice that

$$-\nabla f(x^k)^\top d^k \begin{cases} \geq \min\{\beta_1, \beta_2 \|s^k\|^p\}\|s^k\|^2 & \text{if a Newton step is performed,} \\ = \|\nabla f(x^k)\|^2 & \text{if a gradient step is performed.} \end{cases}$$

Hence, the directions $(d^k)_k$ are descent directions. Since we use backtracking (which is well-defined for descent directions), the sequence of function values $(f(x^k))_k$ is monotonically decreasing and has to converge to some $\xi \in \mathbb{R} \cup \{-\infty\}$. Let x^* now be an arbitrary accumulation point of $(x^k)_k$ and let $(x^{k_\ell})_\ell$ be a corresponding subsequence converging to x^*. By the continuity of f, we can infer $f(x^{k_\ell}) \to f(x^*)$ as $\ell \to \infty$ and by the uniqueness of limits, this implies $\xi = f(x^*) \in \mathbb{R}$. Thus, using the Armijo condition and a telescoping sum, we obtain

$$f(x^0) - f(x^*) = \lim_{k\to\infty} f(x^0) - f(x^k) = \sum_{k=0}^{\infty} f(x^k) - f(x^{k+1}) \geq -\gamma \sum_{k=0}^{\infty} \alpha_k \nabla f(x^k)^\top d^k.$$

As a consequence, since the expressions $-\nabla f(x^k)^\top d^k$ are nonnegative for all k, we can infer $-\sum_{\ell=0}^{\infty} \alpha_{k_\ell} \nabla f(x^{k_\ell})^\top d^{k_\ell} < \infty$. Moreover, since $\nabla^2 f$ is continuous and $(x^{k_\ell})_\ell$ is bounded, there exists $C > 0$ such that $\|\nabla^2 f(x^{k_\ell})\| \leq C$ for all ℓ. Hence, for every successful Newton step, we have

$$\|\nabla f(x^{k_\ell})\| = \|\nabla^2 f(x^{k_\ell}) s^{k_\ell}\| \leq C\|s^{k_\ell}\|$$

and

$$-\nabla f(x^{k_\ell})^\top d^{k_\ell} \geq \min\{\beta_1, \beta_2\|s^k\|^p\}\|s^k\|^2 \geq \frac{1}{C^2} \min\left\{\beta_1, \frac{\beta_2}{C^p}\|\nabla f(x^{k_\ell})\|^p\right\} \|\nabla f(x^{k_\ell})\|^2$$

for all ℓ and $d^{k_\ell} = s^{k_\ell}$. Defining the factor

$$\gamma_{k_\ell} = \begin{cases} \frac{1}{C^2} \min\{\beta_1, \frac{\beta_2}{C^p}\|\nabla f(x^{k_\ell})\|^p\} & \text{if } d^{k_\ell} = s^{k_\ell}, \\ 1 & \text{otherwise,} \end{cases}$$

this shows $\alpha_{k_\ell} \nabla f(x^{k_\ell})^\top d^{k_\ell} \to 0$ and $\alpha_{k_\ell}\gamma_{k_\ell}\|\nabla f(x^{k_\ell})\|^2 \to 0$ as $\ell \to \infty$. Similar to the proof of Theorem 5.11, we assume $\nabla f(x^*) \neq 0$ which implies $\alpha_{k_\ell} \to 0$. Furthermore, there then exist $\varepsilon > 0$ and $L \in \mathbb{N}$ such that

$$\frac{-\nabla f(x^{k_\ell})^\top d^{k_\ell}}{\|d^{k_\ell}\|} \geq \min\{1, C\}\gamma_{k_\ell}\|\nabla f(x^{k_\ell})\| > \varepsilon \tag{6.4}$$

for all $\ell \geq L$. Due to $\alpha_{k_\ell} \to 0$, the step size $\sigma^{-1}\alpha_{k_\ell}$ does not satisfy the Armijo condition, i.e., we have

$$f(x^{k_\ell} + \sigma^{-1}\alpha_{k_\ell} d^{k_\ell}) - f(x^{k_\ell}) \geq \gamma\sigma^{-1}\alpha_{k_\ell} \nabla f(x^{k_\ell})^\top d^{k_\ell} \tag{6.5}$$

for all ℓ sufficiently large. In addition, applying the mean value theorem, there exists $t_{k_\ell} \in$

$[0,1]$ such that

$$\frac{f(x^{k_\ell}+\sigma^{-1}\alpha_{k_\ell}d^{k_\ell})-f(x^{k_\ell})}{\sigma^{-1}\alpha_{k_\ell}\|d^{k_\ell}\|} = \frac{\nabla f(x^{k_\ell}+\sigma^{-1}\alpha_{k_\ell}t_{k_\ell}d^{k_\ell})^\top d^{k_\ell}}{\|d^{k_\ell}\|}$$
$$\leq \|\nabla f(x^{k_\ell}+\sigma^{-1}\alpha_{k_\ell}t_{k_\ell}d^{k_\ell})-\nabla f(x^{k_\ell})\| + \frac{\nabla f(x^{k_\ell})^\top d^{k_\ell}}{\|d^{k_\ell}\|}.$$

Since $(x^{k_\ell})_\ell$ converges to x^* and the continuous function ∇f is uniformly continuous on every compact set, there exist $\rho>0$ and $L'\geq L$ such that

$$\|\nabla f(x^{k_\ell}+d)-\nabla f(x^{k_\ell})\| < (1-\gamma)\varepsilon \quad \forall\, d\in B_\rho(0) \quad \forall\, \ell \geq L'.$$

However, utilizing (6.4), we readily obtain $\alpha_{k_\ell}\|d^{k_\ell}\|\to 0$. Together with (6.5) this yields the contradiction

$$(1-\gamma)\varepsilon > -(1-\gamma)\frac{\nabla f(x^{k_\ell})^\top d^{k_\ell}}{\|d^{k_\ell}\|} > (1-\gamma)\varepsilon$$

for all ℓ sufficiently large. Thus, our assumption was wrong and it follows $\nabla f(x^*)=0$. This finishes the proof of the first statement in the theorem.

Part (i) is an immediate consequence of the second-order sufficient conditions.

In order to show part (ii), we first notice that due to the positive definiteness of $\nabla^2 f(x^*)$ there exist $\varepsilon>0$ and $\mu>0$ such that $\lambda_{\min}(\nabla^2 f(x))\geq\mu$ and $\|\nabla^2 f(x)^{-1}\|\leq\mu^{-1}$ for all $x\in B_\varepsilon(x^*)$. (See the proof of Theorem 6.1 and Lemma 6.2 for comparison). We further note that the positive definiteness of $\nabla^2 f(x^*)$ implies that x^* is an isolated stationary point (this is Lemma 5.19). Since every accumulation point of the sequence $(x^k)_k$ generated by Algorithm 6.2 corresponds to a stationary point, this shows that x^* is an isolated accumulation point of $(x^k)_k$. Our goal is now to apply the capture theorem formulated in Lemma 5.20. Thus, let $(x^{k_\ell})_\ell$ be an arbitrary subsequence converging to x^*. Then for all ℓ sufficiently large, the matrix $\nabla^2 f(x^{k_\ell})$ is invertible and it holds that

$$\|x^{k_\ell+1}-x^{k_\ell}\| = \alpha_{k_\ell}\|d^{k_\ell}\| \leq \begin{cases} \|\nabla^2 f(x^{k_\ell})^{-1}\nabla f(x^{k_\ell})\| \leq \mu^{-1}\|\nabla f(x^{k_\ell})\| & \text{if } d^{k_\ell}=s^{k_\ell}, \\ \|\nabla f(x^{k_\ell})\| & \text{otherwise,} \end{cases}$$
$$\to 0 \quad \ell\to\infty.$$

Hence, by Lemma 5.20, we can infer $x^k\to x^*$ as $k\to\infty$.

We now finally prove part (iii). We need to verify that the Newton direction s^k satisfies the growth condition stated in step 2 of the algorithm and that the full step size $\alpha_k=1$ is always accepted eventually. Since $(x^k)_k$ converges to x^*, there exists K_1 with $x^k\in B_\varepsilon(x^*)$ for all $k\geq K_1$. Moreover, due to $\|\nabla f(x^k)\|\to 0$, we can find $K_2\geq K_1$ such that

$$\|\nabla f(x^k)\| \leq \mu[\min\{\beta_1,\mu\}/\beta_2]^{\frac{1}{p}} \quad \forall\, k\geq K_2.$$

This implies

$$\|s^k\| \leq \|\nabla f(x^k)^{-1}\|\|\nabla f(x^k)\| \leq [\min\{\beta_1,\mu\}/\beta_2]^{\frac{1}{p}} \quad \Longrightarrow\ \beta_2\|s^k\|^p \leq \min\{\beta_1,\mu\}$$

and $\min\{\beta_1, \beta_2\|s^k\|^p\} \leq \mu$ for all $k \geq K_2$. Thus, we obtain

$$-\nabla f(x^k)^\top s^k = (s^k)^\top \nabla^2 f(x^k) s^k \geq \mu \|s^k\|^2 \geq \min\{\beta_1, \beta_2\|s^k\|^p\}\|s^k\|^2$$

for all $k \geq K_2$. Applying the mean value theorem, there exists $\tau_k = \tau_k(\alpha) \in [0,1]$ such that

$$\begin{aligned}
\frac{f(x^k + \alpha s^k) - f(x^k)}{\alpha} - \gamma \nabla f(x^k)^\top s^k &= (1-\gamma)\nabla f(x^k)^\top s^k + \frac{\alpha}{2}(s^k)^\top \nabla^2 f(x^k + \tau_k \alpha s^k) s^k \\
&= -(1-\gamma)(s^k)^\top \nabla^2 f(x^k) s^k + \frac{\alpha}{2}(s^k)^\top \nabla^2 f(x^k + \tau_k \alpha s^k) s^k \\
&\leq -\left[1 - \gamma - \frac{\alpha}{2}\right](s^k)^\top \nabla^2 f(x^k) s^k + \frac{\alpha}{2}\|\nabla^2 f(x^k + \tau_k \alpha s^k) - \nabla^2 f(x^k)\|\|s^k\|^2 \\
&\leq -\left[\frac{1}{2} - \gamma\right]\mu \|s^k\|^2 + \frac{1}{2}\|\nabla^2 f(x^k + \tau_k \alpha s^k) - \nabla^2 f(x^k)\|\|s^k\|^2
\end{aligned}$$

Due to $x^k \to x^*$ and $\|s^k\| \leq \mu^{-1}\|\nabla f(x^k)\| \to 0$ as $k \to \infty$, we can again use the uniform continuity of the Hessian and find $K_3 \geq K_2$ such that

$$\|\nabla^2 f(x^k + \tau_k \alpha s^k) - \nabla^2 f(x^k)\| \leq \left[\frac{1}{2} - \gamma\right]\mu$$

for all $\alpha \in [0,1]$, $\tau_k = \tau_k(\alpha)$ and $k \geq K_3$. Notice that the restriction $\gamma < \frac{1}{2}$ is crucial in our last steps. This shows that the Armijo condition is satisfied for the step size $\alpha_k = 1$ for all $k \geq K_3$. Altogether, we have $d^k = s^k$ for all $k \geq K_3$ and the algorithm turns into a pure Newton method after K_3 many iterations. Q-superlinear and q-quadratic convergence follows from Theorem 6.1. ■

The result in Theorem 6.4 is an important global-local result. It ensures that the globalized Newton method converges globally in the sense that every accumulation point corresponds to a stationary point of the problem – independent of the chosen initial point. Furthermore, under suitable additional local assumptions, we can establish a transition result, i.e., the approach locally turns into the pure Newton method and we can maintain fast local convergence.

6.3. Numerical Experiments

In this final subsection, we want to compare and investigate the performance of Newton's method. We consider the following nonconvex optimization problem:

$$\min_{x \in \mathbb{R}^2} \; f_1(x)^2 + f_2(x)^2$$

where $f_1 : \mathbb{R}^2 \to \mathbb{R}$ and $f_2 : \mathbb{R}^2 \to \mathbb{R}$ are given by:

$$f_1(x) := -1 + x_1 + ((5 - x_2)x_2 - 2)x_2, \quad f_2(x) := -1 + x_1 + ((x_2 + 1)x_2 - 10)x_2.$$

Notice that this problem was analyzed in Assignment A2.3. We compare the gradient method with backtracking ($\gamma = 0.1$, $s = 1$, $\sigma = 0.5$) with the globalized version of Newton's method

	Gradient Method with Backtracking: $s = 1$, $\gamma = 0.1$, $\sigma = 0.5$			
`tol`	`iter`–avg.	`iter`–min	`iter`–max	`time`–avg.
10^{-5}	3953.4	840	8284	0.06 s
10^{-7}	5448.4	1202	11355	0.08 s
10^{-9}	6945.6	1564	14385	0.10 s
	Newton's Method			
`tol`	`iter`–avg.	`iter`–min	`iter`–max	`time`–avg.
10^{-5}	26.6	2	319	0.00 s
10^{-7}	27	2	320	0.00 s
10^{-9}	27.2	2	320	0.00 s

Table 6.1: Numerical comparison of Newton's method and of the gradient method.

presented in Algorithm 6.2. We use the following parameter:

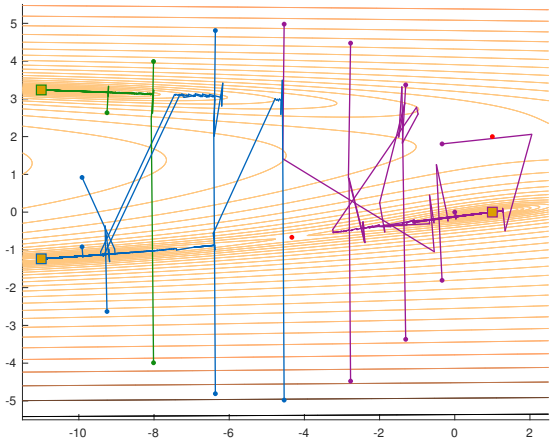
- $\beta_1 = \beta_2 = 10^{-6}$, $p = 0.1$, and $s = 1$, $\gamma = 0.1$, $\sigma = 0.5$.

The methods stop whenever the stopping criterion $\|\nabla f(x^k)\| \leq \texttt{tol}$, $\texttt{tol} \in \{10^{-5}, 10^{-7}, 10^{-9}\}$ is satisfied. To analyze the robustness with respect to the initial points, we run the methods using 17 different initial points generated via

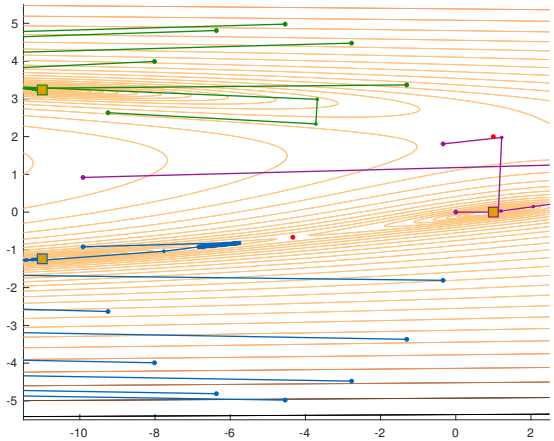
$$x^0 = 5\begin{pmatrix}\cos(2\pi(i-1)/17) & \sin(2\pi(i-1)/17)\end{pmatrix}^\top + \begin{pmatrix}-5 & 0\end{pmatrix}^\top \quad i = 1, ..., 17.$$

The different solution paths are shown in Figure 6.1. Further numerical results are reported in Table 6.1 and Figure 6.2. The experiments demonstrate the favorable performance of Newton's method. In almost every case, the method converges rapidely within a few number of iterations to the (global) solution of the problem. This behavior is different from the gradient method and the exploitation of the additional curvature information allows Newton's method to clearly outperform the gradient method in this example.

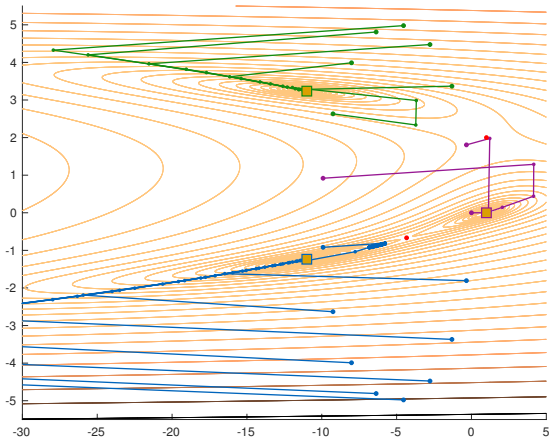
References. This section of the manuscript closely follows [9, Chapter 10]. See also [2, Chapter 5] or [5, Section 1.4].



(a) Gradient Method with Backtracking

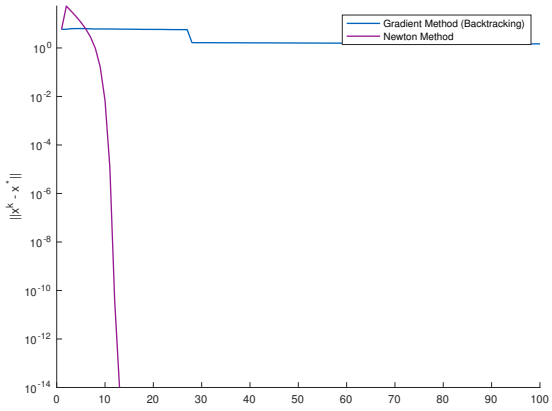


(b) Newton's Method

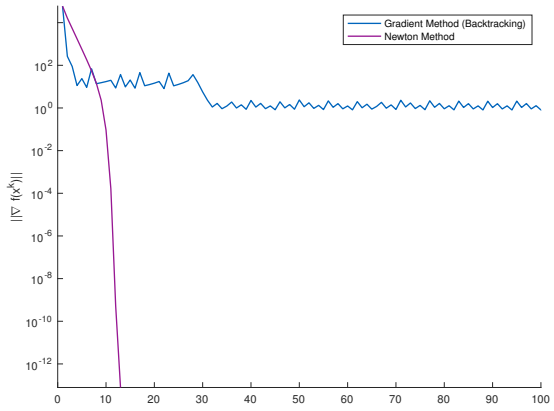


(c) Newton's Method (Larger Scale)

Figure 6.1: Solution paths for the gradient method and Newton's method for different x^0.



(a) Logarithmic Plot of $(\|x^k - x^*\|)_k$



(b) Logarithmic Plot of $(\|\nabla f(x^k)\|)_k$

Figure 6.2: Plot and comparison of the convergence rates.

7. Newton-Type and Quasi-Newton Methods

The Newton method requires to calculate Newton steps as solutions of the linear system of equations:

$$\nabla^2 f(x^k) \cdot d = -\nabla f(x^k), \quad k = 0, 1, 2, ...$$

So far, we have always assumed that this system can be and is solved accurately by Gaussian elimination and numerical algorithms which efficiently implement this procedure, such as, e.g., the LU-decomposition. However, if the dimension n of the problem is large, we can face the following computational issues:

- We need to store the full Hessian matrix $\nabla^2 f(x^k)$ which requires to store $n \times n = n^2$ numbers.
- Solving the linear system requires up to $\mathcal{O}(n^3)$ operations, this can be time-consuming!

  For instance, calculating the inner product $x^\top y = x_1 \cdot y_1 + x_2 \cdot y_2 + ... + x_n \cdot y_n$ requires $2n - 1$ operations.
- For complex applications, the computation of $\nabla^2 f$ might not be possible. In addition, the matrix $\nabla^2 f(x^k)$ might not be invertible causing warnings or numerical instabilities.

7.1. Newton-Type Methods

The principle idea of *Newton-type methods* is to substitute the matrix $\nabla^2 f(x^k)$ by a suitable approximation $M_k \in \mathbb{R}^{n \times n}$ and to solve the equation:

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

This general procedure already defines the following local Newton-type approach:

Algorithm 7.1: Local Newton-Type Method

1 Initialization: Select an initial point $x^0 \in \mathbb{R}^n$.
for $k = 0, 1, ...$ **do**
2 Choose an invertible matrix $M_k \in \mathbb{R}^{n \times n}$ and compute the direction d^k via solving the linear system
$$M_k d^k = -\nabla f(x^k).$$
3 Set $x^{k+1} = x^k + d^k$.
4 If $\|\nabla f(x^{k+1})\| \leq \varepsilon$, then STOP and x^{k+1} is the output.

7.1.1. Characterizing Fast Convergence: Dennis-Moré Conditions

We now study a necessary and sufficient condition that ensures q-superlinear convergence of Algorithm 7.1. This characterization will then allow us to design specific approximations $(M_k)_k$ that can still guarantee fast local convergence.

Theorem 7.1: Characterization of Q-Superlinear Convergence

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable and let $\bar{x}$ be a point such that $\nabla^2 f(\bar{x})$ is invertible. Furthermore, let $(x^k)_k$ be a given sequence that converges to $\bar{x}$ with $x^k \neq \bar{x}$ for all k. Then the following statements are equivalent:

(i) $(x^k)_k$ converges q-superlinearly to $\bar{x}$ and we have $\nabla f(\bar{x}) = 0$.

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

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

Proof. Let us set $s^k = x^{k+1} - x^k$. Applying the fundamental theorem of differentiation and integration, we have

$$\nabla f(x^{k+1}) = \int_0^1 [\nabla^2 f(x^k + ts^k) - \nabla^2 f(\bar{x})]s^k \, \mathrm{d}t + \nabla f(x^k) + \nabla^2 f(\bar{x})s^k. \tag{7.1}$$

(ii) $\implies$ (i). Due to $x^k \to \bar{x}$, the condition $\nabla f(\bar{x}) = 0$ follows from (ii) by taking the limit $k \to \infty$. Using (7.1), (ii), and the continuity of $\nabla^2 f$, we obtain

$$\|\nabla f(x^{k+1})\| \leq \int_0^1 \|\nabla^2 f(x^k + ts^k) - \nabla^2 f(\bar{x})\| \, \mathrm{d}t \cdot \|s^k\| + o(\|s^k\|) = o(\|s^k\|). \tag{7.2}$$

Since $\nabla^2 f(\bar{x})$ is invertible, we can apply Lemma 5.19, i.e., there exists $\eta > 0$ such that

$$\|\nabla f(x^k)\| \geq \eta \|x^k - \bar{x}\|$$

for all k sufficiently large. Hence, using $\|s^k\| \leq \|x^{k+1} - \bar{x}\| + \|x^k - \bar{x}\|$, we have

$$\frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|} \leq \frac{1}{\eta} \frac{\|\nabla f(x^{k+1})\|}{\|x^k - \bar{x}\|} \leq \frac{1}{\eta} \frac{\|\nabla f(x^{k+1})\|}{\|s^k\|} \left[\frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|} + 1 \right].$$

Q-superlinear convergence then follows from (7.2).

(i) $\implies$ (ii). Due to the superlinear convergence, there is $K \in \mathbb{N}$ such that

$$\|x^k - \bar{x}\| \leq \|s^k\| + \|x^{k+1} - \bar{x}\| \leq \|s^k\| + \frac{1}{2}\|x^{k+1} - \bar{x}\| \quad \forall\, k \geq K.$$

Hence, we obtain $\|x^k - \bar{x}\| \leq 2\|s^k\|$ for all $k \geq K$. Since the sequence $(x^k)_k$ converges to $\bar{x}$, the sequence $(x^k)_k$ needs to be contained in a sufficiently large compact set X. Using the continuous differentiability of ∇f and Lemma 5.16, we can infer that ∇f is Lipschitz continuous on X with some constant $L_X > 0$. Consequently, it holds that

$$\|\nabla f(x^{k+1})\| = \|\nabla f(x^{k+1}) - \nabla f(\bar{x})\| \leq L_X \|x^{k+1} - \bar{x}\| \quad \forall\, k.$$

Combining the last results and using (7.1), we then obtain

$$\begin{aligned}\|\nabla f(x^k) + \nabla^2 f(\bar{x})s^k)\| &\leq \|\nabla f(x^{k+1})\| + \int_0^1 \|\nabla^2 f(x^k + ts^k) - \nabla^2 f(\bar{x})\|\,\mathrm{d}t \cdot \|s^k\| \\ &= O(\|x^{k+1} - \bar{x}\|) + o(\|s^k\|) = o(\|x^k - \bar{x}\|) + o(\|s^k\|) = o(\|s^k\|).\end{aligned}$$

(ii) $\Longrightarrow$ (iii). Using (ii) and the convergence $x^k \to \bar{x}$, we have

$$\begin{aligned}\|\nabla f(x^k) + \nabla^2 f(x^k)s^k)\| &\leq \|\nabla f(x^k) + \nabla^2 f(\bar{x})s^k)\| + \|\nabla^2 f(x^k) - \nabla^2 f(\bar{x})\|\|s^k\| \\ &= o(\|s^k\|) + o(\|s^k\|) = o(\|s^k\|).\end{aligned}$$

(iii) $\Longrightarrow$ (ii). Using (iii) and $x^k \to \bar{x}$, we obtain

$$\begin{aligned}\|\nabla f(x^k) + \nabla^2 f(\bar{x})s^k)\| &\leq \|\nabla f(x^k) + \nabla^2 f(x^k)s^k)\| + \|\nabla^2 f(x^k) - \nabla^2 f(\bar{x})\|\|s^k\| \\ &= o(\|s^k\|) + o(\|s^k\|) = o(\|s^k\|).\end{aligned}$$

This finishes the proof of Theorem 7.1. ■

We now specifically consider sequences $(x^k)_k$ generated by Algorithm 7.1.

Corollary 7.2: Dennis-Moré-Conditions

Suppose that the sequence $(x^k)_k$ is generated by Algorithm 7.1. We further assume that $(x^k)_k$ converges to some point $\bar{x}$ at which the Hessian $\nabla^2 f(\bar{x})$ is invertible. Then the following statements are equivalent:

(i) $(x^k)_k$ converges q-superlinearly to $\bar{x}$ and we have $\nabla f(\bar{x}) = 0$.

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

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

Proof. This immediately follows from Theorem 7.1, by noticing

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

Hence, the conditions in Theorem 7.1 and Corollary 7.2 coincide in this case. ■

The characterization in (ii) of q-superlinear convergence was first established by Dennis and Moré and hence, it is called Dennis-Moré-condition. This condition shows that q-superlinear convergence can be ensured if $M_k d^k$ is sufficiently close to $\nabla^2 f(\bar{x})d^k$ (or $\nabla^2 f(x^k)d^k$). More specifically, for $v \notin \{td^k : t \in \mathbb{R}\}$, the directions $M_k v$ and $\nabla^2 f(\bar{x})v$ (or $\nabla^2 f(x^k)v$) can arbitrarily differ from each other without affecting the rate of convergence as long as the condition (ii) or (iii) is satisfied.

Example 7.3. Let $(x^k)_k$ and $(M_k)_k$ be generated by Algorithm 7.1 and assume that $x^k \to \bar{x}$

Algorithm 7.2: Globalized Newton-Type Method

1 Initialization: Select an initial point $x^0 \in \mathbb{R}^n$. Choose parameters $\sigma, \gamma \in (0,1)$, $\beta_1, \beta_2 > 0$, and $p > 0$.
for $k = 0, 1, \ldots$ **do**
2 Select an invertible symmetric matrix $M_k \in \mathbb{R}^{n\times n}$ and compute the direction s^k as solution of the system $M_k s^k = -\nabla f(x^k)$. If this is possible and if s^k satisfies

$$-\nabla f(x^k)^\top s^k \geq \min\{\beta_1, \beta_2 \|s^k\|^p\}\|s^k\|^2,$$

then set $d^k = s^k$. Otherwise set $d^k = -\nabla f(x^k)$.
3 Calculate a step size α_k using backtracking and set $x^{k+1} = x^k + \alpha_k d^k$.
4 If $\|\nabla f(x^{k+1})\| \leq \varepsilon$, then STOP and x^{k+1} is the output.

and $M_k \to \nabla^2 f(\bar{x})$. Then, it follows

$$\|(M_k - \nabla^2 f(\bar{x}))(x^{k+1} - x^k)\| \leq \underbrace{\|M_k - \nabla^2 f(\bar{x})\|}_{\to 0} \|x^{k+1} - x^k\| = o(\|x^{k+1} - x^k\|),$$

i.e., condition (ii) in Corollary 7.2 is satisfied. Thus, if $\nabla^2 f(\bar{x})$ is invertible, then we can expect q-superlinear convergence to $\bar{x}$ in such case.

7.1.2. Globalized Newton-Type Methods

We can mimic the globalization strategy and mechanism introduced in the last section to build a globally convergent version of Algorithm 7.1.

A careful inspection of the proof of Theorem 6.4 reveals that a similar result can also be established for Algorithm 7.2 if the matrices $(M_k)_k$ are bounded. Transition to fast local convergence can then be shown based on the Dennis-Moré-conditions.

7.1.3. Inexact Newton Methods

The idea of *inexact Newton methods* is to solve the linear system

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

only approximately up to a certain accuracy and to generate inexact and cheaper Newton steps. This can also be combined with *regularization techniques* and ideas. In particular, we can consider direction d^k satisfying

$$\|[\nabla^2 f(x^k) + \delta_k I]d^k + \nabla f(x^k)\| \leq \rho_k \|\nabla f(x^k)\|, \tag{7.3}$$

where $\delta_k > 0$, $\rho_k \in (0,1)$ are the regularization parameter and tolerance, respectively. Regularizations of the form $\nabla^2 f(x^k) + \delta_k I$ can help to robustify and stabilize the Newton step when $\nabla^2 f(x^k)$ is positive semidefinite but (approximately) nonsingular around a stationary point. We now discuss the application of the Dennis-Moré conditions in this case. Setting

$M_k = \nabla^2 f(x^k) + \delta_k I$, we have

$$\|\nabla f(x^k)\| \leq \|M_k d^k + \nabla f(x^k)\| + \|M_k d^k\| \leq \rho_k \|\nabla f(x^k)\| + \|M_k d^k\|,$$

i.e., $\|\nabla f(x^k)\| \leq (1-\rho_k)^{-1}\|M_k d^k\|$. This implies

$$\begin{aligned}\|\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k)\| &\leq \|\nabla f(x^k) + M_k d^k\| + \delta_k \|x^{k+1} - x^k\| \\ &\leq \rho_k \|\nabla f(x^k)\| + \delta_k \|x^{k+1} - x^k\| \\ &\leq \frac{\rho_k}{1-\rho_k}\|M_k\|\|x^{k+1} - x^k\| + \delta_k \|x^{k+1} - x^k\|\end{aligned}$$

Hence, if $(x^k)_k$ converges to some $\bar{x}$ with $\nabla^2 f(\bar{x})$ is invertible and if $\delta_k \to 0$ and $\rho_k \to 0$, then Theorem 7.1 ensures q-superlinear convergence of the sequence $(x^k)_k$.

Notice that the condition (7.3) just means that we want to find an approximate solution d^k of the equation

$$(\nabla^2 f(x^k) + \delta_k I) d^k \approx -\nabla f(x^k),$$

satisfying the accuracy bound given in (7.3). Typically, an *iterative approach*, such as the *CG-method*, is used to solve the latter linear system. The CG-method is an iterative method that solves a linear system of equations

$$Ax = b,$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and positive definite and $b \in \mathbb{R}^n$ is a given vector. The method generates a sequence of iterates $(y^\ell)_\ell$ and stops if an iterate satisfies the criterion

$$\|Ay^\ell - b\| \leq \texttt{tol}.$$

We will discuss the details of this approach in one of the next subsections. In our context, we can use the CG-method and set $A = \nabla^2 f(x^k) + \delta_k I$, $b = -\nabla f(x^k)$, and

$$\texttt{tol} = \rho_k \|\nabla f(x^k)\|.$$

If $(\rho_k)_k$ and $(\delta_k)_k$ are chosen to converge to zero as $k \to \infty$, this strategy can ensure local q-superlinear convergence of the inexact globalized Newton method – as we have just shown. An additional advantage of the CG-method is that storage of the full Hessian matrix $\nabla^2 f(x^k)$ is not necessary. Instead, the method only requires "Hessian $\times$ vector" applications, i.e., it only uses terms of the form $\nabla^2 f(x^k) \cdot h$. The combination of the Newton and CG-method is known as *line search Newton-CG method* or *truncated Newton method*.

7.1.4. A Large-Scale Optimization Problem: Inpainting

In order to motivate inexact Newton approaches and to demonstrate the necessity of inexact and iterative linear solvers, we now investigate a large-scale image reconstruction problem. Image *inpainting* describes the task of recovering an image $U \in \mathbb{R}^{m\times n}$ from partial data and observations. In particular, parts of the image U are missing or damaged, (e.g., due to scratches, stains, or compression), and the aim is to reconstruct this missing or damaged

Figure 7.1: Examples of different damaged images. We want to recover the missing image information in the purple areas.

information via solving a suitable inpainting or optimization problem.

The images in Figure 7.1 show several typical situations where inpainting techniques can be applied. Our overall task is to reconstruct the purple target areas. Here, we assume that these target areas are known, i.e., we have access to a binary mask $\texttt{Ind} \in \mathbb{R}^{m\times n}$ with

$$\texttt{Ind}_{ij} = \begin{cases} 1 & \text{the pixel } (i,j) \text{ is not damaged,} \\ 0 & \text{the pixel } (i,j) \text{ is damaged.} \end{cases}$$

This mask contains the relevant information about missing or damaged parts in the image.

Let us set $\texttt{ind} := \text{vec}(\texttt{Ind}) \in \mathbb{R}^{mn}$, $s := \sum_{i=1}^{mn} \texttt{ind}_i$, and $\mathcal{I} := \{i : \texttt{ind}_i = 1\}$. Here, the vectorization $\text{vec}(\cdot)$ transforms a matrix into a column vector by stacking all of the column vectors on top of each other, i.e., we have

$$U = (u_{[1]}, u_{[2]}, \cdots, u_{[n]}) \in \mathbb{R}^{m\times n} \quad \Longrightarrow \quad u = \text{vec}(U) = ((u_{[1]}^\top, u_{[2]}^\top, \cdots, u_{[n]}^\top)^\top \in \mathbb{R}^{mn}.$$

Let $\mathcal{I} = \{q_1, q_2, ..., q_s\}$ denote the different elements of the index set $\mathcal{I}$. Then, we can define the following *selection matrix*

$$A = \begin{pmatrix} e_{q_1}^\top \\ \vdots \\ e_{q_s}^\top \end{pmatrix} \in \mathbb{R}^{s\times mn}, \quad e_j \in \mathbb{R}^{mn}, \quad [e_j]_i = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases} \quad \forall\, j \in \mathcal{I}.$$

The vectors e_j are unit vectors in $\mathbb{R}^{mn}$ and the matrix A selects all undamaged pixels of a stacked image according to the mask $\texttt{ind}$. The vector

$$b = Au \in \mathbb{R}^s$$

contains the (color) information of all undamaged pixels of the original image U. Hence, we now want to find a reconstruction $y \in \mathbb{R}^{mn}$ of u such that:

(i) Original information of the undamaged parts in U is maintained, i.e., y satisfies

$$Ay = b \quad \text{or} \quad Ay \approx b. \tag{7.4}$$

(ii) The image y can recover the missing parts in U in a "suitable" way.

The linear system of equations (7.4) is underdetermined and has infinitely many possible solutions. In order to recover solutions with a "natural image structure", we consider a so-called ℓ_1-regularized image reconstruction problem. This optimization problem utilizes an ℓ_1-regularization to solve the inpainting task (7.4); it is given by:

$$\min_{x\in\mathbb{R}^{mn}} \|\Psi x\|_1 \quad \text{s. t.} \quad Ax = b. \tag{7.5}$$

Here, the matrix $\Psi \in \mathbb{R}^{mn\times mn}$ changes the basis and transfers the image x to the *frequency domain*. In this new basis or representation, many images tend to be very sparse and many of the components $[\Psi x]_i$ are zero or close to zero. This motivates the choice of the ℓ_1-norm, $\|x\|_1 = \sum_{i=1}^{mn} |x_i|$, in the model (7.5). In order to apply the inexact Newton method, we consider the following smooth variant of problem (7.5)

$$\min_{x\in\mathbb{R}^{mn}} \frac{1}{2}\|A\Psi^{-1}x - b\|^2 + \mu \sum_{i=1}^{mn} \log(1 + \nu^{-1}x_i^2) \tag{7.6}$$

with parameters $\mu, \nu > 0$. The mapping $y \mapsto \log(1 + \nu^{-1}y^2)$ is a nonconvex smooth approximation of the absolute value $|y|$ that also promotes sparsity. In the following, we want to use a discrete cosine transformation (DCT) as sparse basis for the images, i.e., we have $\Psi x = \Psi(x) = \texttt{dct}(x)$ and $\Psi^{-1}x = \Psi^\top x = \texttt{idct}(x)$. Although the DCT is a linear transformation, we typically can not access the full matrices $\Psi, \Psi^{-1} \in \mathbb{R}^{mn\times mn}$. This is primarily caused by the huge dimensions of the problem: for an input image with resolution 512×512, the matrix Ψ would have dimension $262\,144 \times 262\,144$ and requires 500 GB of memory. However, Ψ and Ψ^{-1} can be accessed as functions and there are fast implementations to calculate and evaluate $\texttt{dct}(x)$ and $\texttt{idct}(x)$. Similarly, the matrix A can be either build as a sparse matrix or we implement it as a function extracting components from a vector x according to the mask `ind` and $\mathcal{I}$.

The Hessian of the objective function f in (7.6) is given by

$$\nabla^2 f(x) = \Psi A^\top A \Psi^\top + \mu D(x), \quad D(x) = \operatorname{diag}(\delta_1, ..., \delta_n), \quad \delta_i = \frac{2(\nu - x_i^2)}{(\nu + x_i^2)^2}.$$

We now want to discuss the performance of an inexact Newton method for (7.6) that utilizes a CG-method to approximately solve the Newton equation $\nabla f(x^k)d^k = -\nabla f(x^k)$. Our implementation is based on the following parameters and strategies:

- We use the CG-method with tolerance $\texttt{tol} = \min\{0.01, \|\nabla f(x^k)\|^{1.1}\}$. The maximum number of CG steps is limited to 10.
- All other parameters are standard: $\gamma = 0.1$, $s = 1$, $\sigma = 0.5$, $\beta_1 = \beta_2 = 10^{-6}$, $p = 0.1$.

We stop whenever the termination criterion $\|\nabla f(x^k)\| \le 10^{-6}$ is satisfied. The parameters in the model are chosen as $\mu = 5 \cdot 10^{-4}$ and $\nu = 0.015$. The computational results are shown

Figure 7.2: Examples of different reconstructed images obtained from solving the model (7.6) via an inexact Newton method.

Image	Size	Mask	[Time / Iter.]	CG-tot	$\|\nabla f(x^*)\|$	PSNR
`circles`	512×512	`random70`	2.30 sec / 26	168	$5.3823 \cdot 10^{-7}$	25.01
`eagle`	640×640	`mesh`	3.93 sec / 19	129	$5.8945 \cdot 10^{-7}$	32.57
`carousel`	640×640	`handwriting`	4.00 sec / 18	140	$4.6395 \cdot 10^{-7}$	28.57

Table 7.1: Numerical results. The PSNR value is a popular and simple image quality measure. Let $u^* = \mathrm{vec}(U^*)$ be the original true (undamaged) image, then it is defined via:

$$\mathrm{PSNR} := 10 \cdot \log_{10}\left[\frac{mn}{\|y - u^*\|^2}\right] \quad \text{where} \quad y = \texttt{idct}(x).$$

In the column CG-tot, we report the total number of CG-iterations required by the inexact Newton method.

in Table 7.1 and Figure 7.2. Let us note that the gradient with backtracking requires 1955 iterations, 40.59 sec for `circles`, 1846 iterations, 95.83 sec for `eagle`, and 2840 iterations, 138.18 sec for `carousel` to reach an iterate satisfying $\|\nabla f(x^k)\| \leq 10^{-4}$. Hence, the inexact Newton method is around 20-30 times faster than the gradient method!

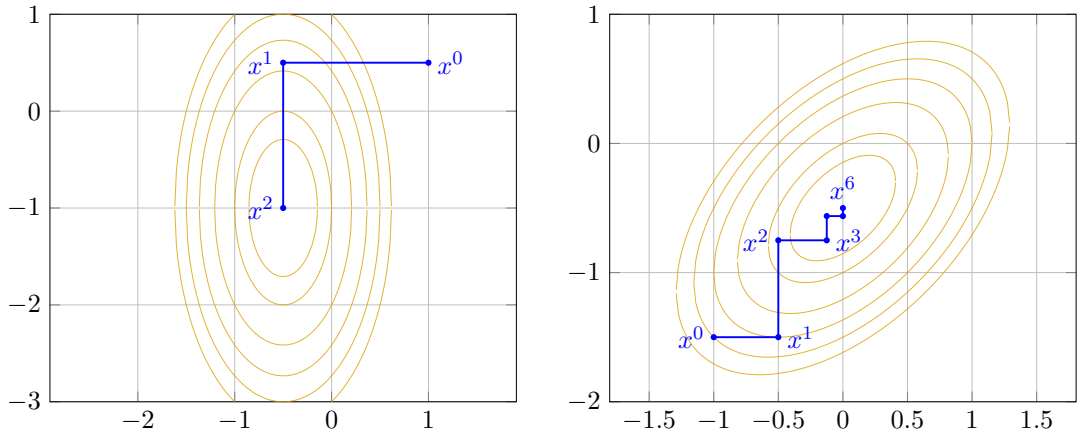
7.1.5. The CG-Method

As mentioned the conjugate gradient method (CG-method) is an iterative approach for solving linear systems of equations

$$Ax = b, \tag{7.7}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric and positive definite. This problem can be equivalently restated as a minimization problem

$$\min_{x\in\mathbb{R}^n} \ \phi(x) = \frac{1}{2}x^\top Ax - b^\top x. \tag{7.8}$$



(a) Diagonal Case (b) Non-Diagonal Case

Figure 7.3: Illustration of the conjugate direction method using the unit vectors e_1 and e_2. If A is diagonal, $\{e_1, e_2\}$ is a set of conjugate vectors and the procedure terminates after only two steps. In the non-diagonal case, this two-step convergence behavior can not be guaranteed since $\{e_1, e_2\}$ is no longer an A-conjugate set of vectors. The two subfigures depict the contour lines in the diagonal and non-diagonal case and the corresponding solution paths generated by the conjugated direction updates defined in (7.9).

This equivalence allows to treat the CG-method as a technique to minimize (strongly) convex quadratic functions. Based on the optimality condition $\nabla\phi(x) = Ax - b$, we are interested in the convergence of the residuals:

$$r^k := Ax^k - b$$

at the k-th iteration of the CG-method. The principle idea of the conjugated gradient method is to generate a sequence of *conjugate* vectors. A set of nonzero vectors $\{p^0, p^1, ..., p^\ell\}$ is said to be *conjugate* with respect to a symmetric positive matrix A if

$$(p^i)^\top A p^j = 0 \quad \forall\, i \neq j.$$

This property also implies that the vectors $p^0, p^1, ..., p^\ell$ are linearly independent. Given a set of n conjugate directions, we can minimize ϕ in n steps by successively minimizing it along the individual directions p^i, $i = 0, ..., n-1$. In particular, given an initial point $x^0 \in \mathbb{R}^n$ and a set of conjugate directions $\{p^0, p^1, ..., p^{n-1}\}$, we can consider the approach

$$x^{k+1} = x^k + \alpha_k p^k, \quad \alpha_k = -\frac{(r^k)^\top p^k}{(p^k)^\top A p^k}. \tag{7.9}$$

Here α_k is chosen to minimize ϕ along $x^k + \alpha p^k$, i.e., α_k coincides with performing the exact line search $\alpha_k = \arg\min_\alpha \phi(x^k + \alpha p^k)$. We have the following result.

Algorithm 7.3: The (Linear) Conjugate Gradient Method

Initialization: Select an initial point $x^0 \in \mathbb{R}^n$ and set $r^0 = Ax^0 - b$, $p^0 = -r^0$, $\texttt{tol} \geq 0$.
for $k = 0, 1, ...$ **do**
  Compute the updates $\alpha_k = -\frac{(r^k)^\top p^k}{(p^k)^\top A p^k}$, $x^{k+1} = x^k + \alpha_k p^k$, $r^{k+1} = r^k + \alpha_k A p^k$.
  If $\|r^{k+1}\| \leq \texttt{tol}$, then STOP and x^{k+1} is the output.
  Calculate $\beta_{k+1} = \frac{(r^{k+1})^\top A p^k}{(p^k)^\top A p^k}$ and $p^{k+1} = -r^{k+1} + \beta_{k+1} p^k$.

Theorem 7.4: Conjugate Directions and Convergence

For any initial point $x^0 \in \mathbb{R}^n$, the sequence generated by the conjugate direction algorithm (7.9) converges to a solution x^* of (7.8) satisfying $Ax^* = b$ in at most n steps.

Proof. Since the directions $(p^i)_{i=0}^{n-1}$ are linearly independent there exist $\sigma_0, ..., \sigma_{n-1} \in \mathbb{R}$ such that

$$x^* - x^0 = \sum_{i=0}^{n-1} \sigma_i p^i$$

Multiplying this equation with $(p^k)^\top A$ yields $(p^k)^\top A(x^* - x^0) = \sigma_k \cdot (p^k)^\top A p^k$ by the A-conjugacy of the directions $(p^i)_{i=0}^{n-1}$. Hence, we have

$$\sigma_k = \frac{(p^k)^\top A(x^* - x^0)}{(p^k)^\top A p^k}.$$

Let x^k now be generated by the conjugate direction method (7.9), then it follows $x^k = x^0 + \sum_{i=0}^{k-1} \alpha_i p^i$ and we obtain $(p^k)^\top A(x^k - x^0) = 0$. This implies

$$(p^k)^\top A(x^* - x^0) = (p^k)^\top A(x^* - x^k) = (p^k)^\top (b - Ax^k) = -(p^k)^\top r^k.$$

Consequently, we have $\sigma_k = \alpha_k$ for all k which establishes $x^* = x^{n-1}$. ■

The conjugate gradient method is a special conjugate direction method that iteratively generates the set of conjugate vectors $\{p^0, p^1, ..., p^{n-1}\}$ while solving the optimization problem (7.8). As an important feature of the method, the computation of p^k is solely based on p^{k-1}, i.e., we do not need to store all of the past conjugate vectors $p^0, p^1, ..., p^{k-2}$. Each direction p^k is chosen as a linear combination of the negative residual $-r^k$ (this is the steepest descent direction for ϕ at x^k) and the previous direction p^{k-1}. We set

$$p^k = -r^k + \beta_k p^{k-1}.$$

By multiplying this equation with $(p^{k-1})^\top A$ and by imposing the condition $(p^{k-1})^\top A p^k = 0$, we obtain

$$\beta_k = \frac{(r^k)^\top A p^{k-1}}{(p^{k-1})^\top A p^{k-1}}.$$

As in the conceptual conjugate direction method described in (7.9), we perform successive one-dimensional minimizations along each of the conjugate directions (via exact line search). The algorithm is summarized in Algorithm 7.3

We now present a convergence result for Algorithm 7.3 in which we verify that the directions $(p^k)_k$ generated by the CG-method are actually conjugate directions. We can then apply Theorem 7.4 to guarantee convergence within at most n steps.

Theorem 7.5: Properties of the CG-Method

Let the sequences $(p^k)_{k=0}^{\ell}$, $(r^k)_{k=0}^{\ell}$, and $(x^k)_{k=0}^{\ell}$ be generated by Algorithm 7.3 and assume that the method does not stop in step 3. Then, we have:

(i) For all $k = 1, ..., \ell$, it holds that

$$(p^k)^\top A p^i = 0, \quad (r^k)^\top r^i = 0, \quad (r^k)^\top p^i = 0, \quad \forall\, i = 0, ..., k-1. \tag{7.10}$$

(ii) The method stops after at most n steps returning a solution of the linear system of equations (7.7).

Let us now consider the specific choice $x^0 = 0$. Then, it additionally holds that:

(iii) The sequence $(\phi(x^k))_{k=0}^{\ell}$ is strictly decreasing.

(iv) The sequence $(\|x^k\|)_{k=0}^{\ell}$ is strictly increasing.

Proof. Let us define the Krylov space $\mathcal{K}^m(A, r^0) := \operatorname{span}\{r^0, Ar^0, ..., A^m r^0\}$ for some $m \in \{1, ..., n-1\}$. We now prove the statements in part (i) together with the two additional claims

$$\operatorname{span}\{r^0, r^1, ..., r^k\} = \mathcal{K}^k(A, r^0), \quad \operatorname{span}\{p^0, p^1, ..., p^k\} = \mathcal{K}^k(A, r^0) \tag{7.11}$$

by induction over k. For $k = 0$, we have

$$\operatorname{span}\{r^0\} = \operatorname{span}\{p^0\} = \mathcal{K}^0(A, r^0).$$

Hence, all statements are satisfied in the base case $k = 0$. Let us now suppose, that the equations in part (i) and in (7.11) hold for some $k \in \{1, ..., \ell - 1\}$. Due to $r^{k+1} = r^k + \alpha_k A p^k$ and $r^k, p^k \in \mathcal{K}^k(A, r^0)$, it follows $r^{k+1} \in \mathcal{K}^{k+1}(A, r^0)$ and thus, using (7.11) for k, we obtain

$$\operatorname{span}\{r^0, r^1, ..., r^{k+1}\} \subseteq \mathcal{K}^{k+1}(A, r^0).$$

To establish the reserve direction, we note that by (7.11), we have $A^{k+1} r^0 = A(A^k r^0) \in \operatorname{span}\{Ap^0, Ap^1, ..., Ap^k\}$. Moreover, for all $i = 0, ..., k$, we can write $Ap^i = \frac{1}{\alpha_i}(r^{i+1} - r^i)$. This shows $A^{k+1} r^0 \in \operatorname{span}\{r^0, r^1, ..., r^{k+1}\}$ and $\mathcal{K}^{k+1}(A, r^0) \subseteq \operatorname{span}\{r^0, r^1, ..., r^{k+1}\}$. Now, due to $p^{k+1} \in \operatorname{span}\{p^k, r^{k+1}\}$, $r^{k+1} \in \operatorname{span}\{p^k, p^{k+1}\}$, and (7.11), we have

$$\begin{aligned} \operatorname{span}\{p^0, p^1, ..., p^{k+1}\} &= \operatorname{span}\{p^0, ..., p^k, r^{k+1}\} \\ &= \operatorname{span}\{r^0, Ar^0, ..., A^k r^0, r^{k+1}\} \end{aligned}$$

$$= \text{span}\{r^0, ..., r^k, r^{k+1}\} = \mathcal{K}^{k+1}(A, r^0).$$

Next, it holds that

$$(r^{k+1})^\top p^k = (r^k)^\top p^k + \alpha_k (p^k)^\top A p^k = 0,$$

where we used the definition of α_k. Similarly, we obtain $(r^{k+1})^\top p^i = (r^k)^\top p^i + \alpha_k (p^k)^\top A p^i = 0$ for all $i = 0, ..., k-1$ by applying (7.10) for $i = 0, ..., k-1$. Utilizing these results, we also get

$$(r^{k+1})^\top r^i = (r^{k+1})^\top (\beta_i p^{i-1} - p^i) = 0, \quad (r^{k+1})^\top r^0 = -(r^{k+1})^\top p^0 = 0,$$

for all $i = 1, ..., k$. By the definition of β_{k+1} and p^{k+1}, we have

$$(p^{k+1})^\top A p^k = -(r^{k+1})^\top A p^k + \beta_{k+1} (p^k)^\top A p^k = 0.$$

By induction and as shown, it holds that $(p^j)^\top A p^i = 0$ and $(r^{k+1})^\top p^j = 0$ for all $j = 1, ..., k$ and $i < j$. Consequently, it follows

$$Ap^i \in A\mathcal{K}^i(A, r^0) = \text{span}\{Ar^0, ...A^{i+1} r^0\} \subseteq \text{span}\{p^0, ..., p^{i+1}\}, \quad \forall\, i = 0, ..., k-1.$$

This implies $(r^{k+1})^\top A p^i = 0$ for all $i = 0, ..., k-1$ and hence, we have $(p^{k+1})^\top A p^i = 0$ for all $i = 0, ..., k$. This finishes the proof of part (i).

The second part now follows easily from Theorem 7.4. We continue with the proof of part (iii). Due to $r^{k+1} = b + Ax^{k+1}$ and $x^{k+1} = \sum_{i=0}^k \alpha_i p^i \in \text{span}\{p^0, ..., p^k\}$ (notice that this uses $x^0 = 0$), it follows

$$\begin{aligned}\phi(x^{k+1}) &= (b + Ax^{k+1})^\top x^{k+1} - \frac{1}{2}(x^{k+1})^\top A x^{k+1} \\ &= -\frac{1}{2}(x^{k+1})^\top A x^{k+1} = -\frac{1}{2}(x^k)^\top A x^k - \alpha_k (x^k)^\top A p^k - \frac{\alpha_k^2}{2}(p^k)^\top A p^k,\end{aligned}$$

where we used (7.10). By induction this establishes $\phi(x^k) = -\frac{1}{2}(x^k)^\top A x^k$ for all $k = 0, ..., \ell$. Consequently, we obtain

$$\phi(x^{k+1}) = \phi(x^k) - \frac{1}{2}\frac{((r^k)^\top p^k)^2}{(p^k)^\top A p^k}, \quad \forall\, k = 0, ..., \ell - 1$$

which implies that the sequence $(\phi(x^k))_{k=0}^\ell$ is strictly decreasing.

In order to show the last part, we first note

$$\|x^{k+1}\|^2 = \|x^k\|^2 + 2\alpha_k (x^k)^\top p^k + \alpha_k^2 \|p^k\|^2.$$

Moreover, due to $(x^k)^\top r^k = \sum_{i=0}^{k-1} \alpha_i (p^i)^\top r^k = 0$ (for all k), it follows

$$(x^k)^\top p^k = -(x^k)^\top r^k + \beta_k (x^k)^\top p^{k-1} = \beta_k (x^{k-1})^\top p^{k-1} + \alpha_{k-1}\beta_k \|p^{k-1}\|^2.$$

Inductively, this implies $(x^k)^\top p^k > 0$ for all k, if the base case $k = 1$ satisfies $(x^1)^\top p^1 > 0$. However, we have $(x^1)^\top p^1 = (\alpha_0 p^0)^\top (-r^1 + \beta_1 p^0) = \alpha_0 \beta_1 \|p^0\|^2 > 0$. Thus, combining the last results, we can infer $\|x^{k+1}\|^2 > \|x^k\|^2$ which establishes part (iv). ■

Algorithm 7.4: Truncated Newton Method

Initialization: Select an initial point $x^0 \in \mathbb{R}^n$. Choose parameters $\sigma, \gamma \in (0,1)$, $s > 0$, $(\rho_k)_k \in (0,1)$.
for $k = 0, 1, 2, \ldots$ **do**
  Set $A = \nabla^2 f(x^k)$, $v^0 = 0$, $r^0 = \nabla f(x^k)$, $p^0 = -r^0$, and $\texttt{tol} = \rho_k \|\nabla f(x^k)\|$.
  for $j = 0, 1, 2, \ldots$ **do**
    If $(p^j)^\top A p^j \leq 0$ return $d^k = v^j$ (or $d^k = -\nabla f(x^k)$ if $j = 0$).
    Compute the updates $\sigma_j = \frac{\|r^j\|^2}{(p^j)^\top A p^j}$, $v^{j+1} = v^j + \sigma_j p^j$, $r^{j+1} = r^j + \sigma_j A p^j$.
    If $\|r^{j+1}\| \leq \texttt{tol}$, then STOP and return $d^k = v^{j+1}$.
    Calculate $\beta_{j+1} = \frac{\|r^{j+1}\|^2}{\|r^j\|^2}$ and $p^{j+1} = -r^{j+1} + \beta_{j+1} p^j$.
  Calculate a step size α_k using backtracking and set $x^{k+1} = x^k + \alpha_k d^k$.
  If $\|\nabla f(x^{k+1})\| \leq \varepsilon$, then STOP and x^{k+1} is the output.

Using the properties in Theorem 7.5, it is possible to slightly simplify the update rules in the CG-method. First of all, due to $-(r^0)^\top p^0 = \|r^0\|^2$ and $p^i = -r^i + \beta_i p^{i-1}$, we have $-(r^k)^\top p^k = \|r^k\|^2$. This implies

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

Moreover, since $\alpha_k A p^k = r^{k+1} - r^k$, we can simplify β_{k+1} to

$$\beta_{k+1} = \frac{\|r^{k+1}\|^2}{\|r^k\|^2}.$$

The standard form of the CG-method uses this way to calculate α_k and β_{k+1}.

In the line search Newton-CG method, we typically add the safeguard

$$(p^k)^\top A p^k \leq 0 \;\text{ or }\; (p^k)^\top A p^k \leq \delta_k \quad \Longrightarrow \quad \texttt{break}$$

in case the Hessian $\nabla^2 f(x^k)$ or the matrix A is not positive (semi)definite. The full combination of the CG-method and the Newton-method is described and shown in Algorithm 7.4. We set $x^k \equiv v^k$ and $\alpha_k \equiv \sigma_k$ in the CG-method to avoid confusion. Notice that we again use the specific initial point $x^0 \equiv v^0 = 0$. This ensures that the CG-method returns a descent direction satisfying $d^k = -\nabla f(x^k)$ or $d^k = v^\ell$ for some $\ell \in \mathbb{N}$ and we have

$$\nabla f(x^k)^\top v^\ell = -\sum_{i=0}^{\ell-1} \sigma_i (p^0)^\top p^i.$$

In particular and inductively, we obtain $(p^0)^\top p^i = -(p^0)^\top r^i + \beta_i (p^0)^\top p^{i-1} = \ldots = \beta_i \beta_{i-1} \cdot \ldots \cdot \beta_1 \|p^0\|^2 = \|r^i\|^2$. Hence, it follows $\nabla f(x^k)^\top v^\ell = -\sum_{i=0}^{\ell-1} \sigma_i \|r^i\|^2 < 0$.

Using this special descent property, we can establish global-local convergence without requiring an additional acceptance mechanism as in the globalized Newton method.

Theorem 7.6: Global-Local Convergence of the Truncated Newton Method

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and let $(x^k)_k$ be generated by Algorithm 7.4. Then, every accumulation point of $(x^k)_k$ is a stationary point.
Additionally, if we have $\gamma \in (0, \frac{1}{2})$, $s = 1$, and if there exists an accumulation point x^* of $(x^k)_k$ satisfying the second-order sufficient conditions, then it follows:

(i) The whole sequence $(x^k)_k$ converges to the strict local minimum x^*.

(ii) Suppose that $\rho_k \to 0$. Then, there is $K \in \mathbb{N}$ such that $\alpha_k = 1$ for all $k \geq K$, i.e., the full truncated Newton step is always performed eventually. Furthermore, $(x^k)_k$ converges q-superlinearly to x^*.

Proof. Let x^* be an accumulation point and let $(x^{k_\ell})_\ell$ be a subsequence converging to x^*. Mimicking our previous results this implies

$$\alpha_k \nabla f(x^k)^\top d^k \to 0 \quad \text{as} \quad k \to \infty.$$

As just shown, the directions d^k either satisfy $\nabla f(x^k)^\top d^k = -\|\nabla f(x^k)\|^2$ or

$$\nabla f(x^k)^\top d^k \leq -\frac{\|\nabla f(x^k)\|^4}{\nabla f(x^k)^\top \nabla^2 f(x^k) \nabla f(x^k)}.$$

Since $(x^{k_\ell})_\ell$ is bounded, there exists $C > 0$ such that $\|\nabla^2 f(x^{k_\ell})\| \leq C$ for all ℓ and hence, it follows $-\nabla f(x^{k_\ell})^\top d^{k_\ell} \geq \min\{1, C^{-1}\}\|\nabla f(x^{k_\ell})\|^2$. At this point, we can utilize the same proof technique as in the proof of Theorem 5.11 to infer $\nabla f(x^*) = 0$.

The second-order sufficient conditions imply that $\nabla^2 f(x^*)$ is positive definite and that x^* is a strict local minimum of f. Furthermore, there exist $\epsilon, \mu > 0$ such that $\lambda_{\min}(\nabla^2 f(x)) \geq \mu$ for all $x \in B_\varepsilon(x^*)$. By Lemma 5.19, x^* is an isolated stationary point of f. Since every accumulation point of $(x^k)_k$ is a stationary point, this again shows that x^* is an isolated accumulation point of $(x^k)_k$. Let $(x^{k_\ell})_\ell$ be an arbitrary subsequence converging to x^*. At iteration k_ℓ, the CG-method either returns $d^{k_\ell} = -\nabla f(x^\ell)$ or $d^{k_\ell} = v^\kappa = \sum_{i=0}^{\kappa-1} \sigma_i p^i$ for some $\kappa \in \{1, ..., n-1\}$. Iteratively and applying Theorem 7.5, we now obtain

$$(r^i)^\top p^i = (r^{i-1} + \sigma_{i-1} A p^{i-1})^\top p^i = (r^{i-1})^\top p^i = ... = (r^0)^\top p^i$$

and for all ℓ sufficiently large, it follows

$$\|d^{k_\ell}\| = \|v^\kappa\| \leq \sum_{i=0}^{\kappa-1} |\sigma_i| \|p^i\| \leq \sum_{i=0}^{\kappa-1} \frac{\|r^0\| \|p^i\|}{(p^i)^\top \nabla^2 f(x^{k_\ell}) p^i} \|p^i\| \leq \frac{n}{\mu} \cdot \|\nabla f(x^{k_\ell})\|.$$

This shows $\|x^{k_\ell+1} - x^{k_\ell}\| \to 0$ and implies convergence of the whole sequence $(x^k)_k$ to x^*.

Since $(x^k)_k$ converges to x^* and $\nabla^2 f(x)$ is positive definite for all $B_\epsilon(x^*)$, we can infer that the CG-method stops with $\|\nabla^2 f(x^k) d^k + \nabla f(x^k)\| \leq \rho_k \|\nabla f(x^k)\|$ for all k sufficiently large. As in the proof of the globalized Newton method, there exists $t_k(\alpha) \in [0, 1]$ for all $\alpha \in [0, 1]$

and k sufficiently large such that

$$\begin{aligned}\frac{f(x^k+\alpha d^k)-f(x^k)}{\alpha} - \gamma\nabla f(x^k)^\top d^k &= (1-\gamma)\nabla f(x^k)^\top d^k + \frac{\alpha}{2}(d^k)^\top\nabla^2 f(x^k+t_k(\alpha)\alpha d^k)d^k\\ &\leq -(1-\gamma)(d^k)^\top\nabla^2 f(x^k)d^k + \frac{\alpha}{2}(d^k)^\top\nabla^2 f(x^k+t_k(\alpha)\alpha d^k)d^k\\ &\quad + (1-\gamma)\rho_k\|\nabla f(x^k)\|\|d^k\|\end{aligned}$$

This follows from applying the mean-value theorem. By the $\nabla^2 f(x^k)$-conjugacy of the directions $(p^i)_{i=0}^{\kappa}$, we further have

$$(d^k)^\top\nabla^2 f(x^k)d^k = \sum_{i=0}^{\kappa-1}\sigma_i^2(p^i)^\top\nabla^2 f(x^k)p^i \geq \mu\sigma_0^2\|p^0\|^2$$

for all k sufficiently large. Again there exists $C>0$ such that $\|\nabla^2 f(x^k)\|\leq C$ for all k and using $\sigma_0 = \|\nabla f(x^k)\|^2/(\nabla f(x^k)^\top\nabla^2 f(x^k)\nabla f(x^k))$ and $p^0 = -\nabla f(x^k)$, it follows

$$\|\nabla f(x^k)\|^2 \leq \frac{C^3}{\mu}\|d^k\|^2$$

for all k sufficiently large. Due to $\rho_k\to 0$, we can now proceed as in the proof of Theorem 6.4 to show that the full step size $\alpha_k = 1$ will always be accepted eventually. The q-superlinear convergence then follows from the Dennis-Moré condition. ■

7.2. Quasi-Newton Methods

The principle idea of quasi-Newton methods is similar to the inexact Newton-type methods discussed in the last section. Again we want to approximate the Hessian by a suitable and "easier", invertible matrix such that:

- Less memory storage is required. In particular, we do not need to store a full $n\times n$ matrix in each step. (This is slightly different from the inexact Newton version)!
- The resulting quasi-Newton step is much cheaper.
- We still can guarantee reasonable convergence properties.

Since the Newton step was based on minimizing the Taylor expansion $f(x^k+d) = f(x^k) + \nabla f(x^k)^\top d + \frac{1}{2}d^\top\nabla^2 f(x^k)d$, the idea is now to use the approximate model

$$f(x^k+d)\approx q_k(d) = f(x^k) + \nabla f(x^k)^\top d + \frac{1}{2}d^\top B_k d.$$

If B_k is positive definite, this yields a step of the form $x^{k+1} = x^k + d = x^k - B_k^{-1}\nabla f(x^k)$. Let us consider the new model

$$f(x^{k+1}+d)\approx q_{k+1}(d) = f(x^{k+1}) + \nabla f(x^{k+1})^\top d + \frac{1}{2}d^\top B_{k+1}d.$$

We want to determine the new approximation B_{k+1} based on the knowledge gained during the last step. A reasonable requirement is to assume that the gradient of q_{k+1} matches the gradient of f at the latest two iterates x^k and x^{k+1}. Since $\nabla q_{k+1}(0) = \nabla f(x^{k+1})$, this already holds for the iterate x^{k+1}. The condition for x^k can be expressed as follows:

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

Hence, we obtain the so-called *secant equation* or *quasi-Newton equation*:

$$\nabla f(x^{k+1}) - \nabla f(x^k) = B_{k+1}(x^{k+1} - x^k). \tag{7.12}$$

Quasi-Newton methods now generate sequences of approximations $(B_k)_k$ using specific update formulae that usually follow the framework:

$$\begin{aligned} &\text{available information at step } k: \quad s^k = x^{k+1} - x^k,\ y^k = \nabla f(x^{k+1}) - \nabla f(x^k),\ B_k \\ \Longrightarrow\ &\text{generate}: \quad B_{k+1}. \end{aligned}$$

The updated matrix B_{k+1} should be symmetric and it should satisfy the quasi-Newton equation $y^k = B_{k+1}s^k$. We can make the following motivating observation:

Lemma 7.7: Quasi-Newton Steps and Superlinear Convergence

Let $x^* \in \mathbb{R}^n$ be a point satisfying the second-order sufficient conditions and let the iterates $(x^k)_k$ be generated by the local quasi-Newton method

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

where $(B_k)_k \subset \mathbb{R}^{n\times n}$ is a sequence of symmetric, nonsingular matrices fulfilling the quasi-Newton condition (7.12). Suppose that $(x^k)_k$ converges to x^* and it holds that

$$\lim_{k\to\infty} \|B_{k+1} - B_k\| = 0.$$

Then, the matrices B_k, $k \in \mathbb{N}$, satisfy the Dennis-Moré condition and the sequence $(x^k)_k$ converges q-superlinearly to x^*.

Proof. Taylor expansion and the quasi-Newton equation (7.12) imply

$$\begin{aligned} \|(B_k - \nabla^2 f(x^k))s^k\| &\leq \|(B_k - B_{k+1})s^k\| + \|(B_{k+1} - \nabla^2 f(x^k))s^k\| \\ &= o(\|s^k\|) + \|\nabla f(x^{k+1}) - \nabla f(x^k) - \nabla^2 f(x^k)s^k\| = o(\|s^k\|), \end{aligned}$$

as desired. ■

Consequently, we should try to seek quasi-Newton updates such that the matrices B_{k+1} and B_k are close to each other. Next, we discuss several update strategies based on these insights and observations.

7.2.1. The Symmetric Rank-1 (SR-1) Update

One of the simplest ideas is to modify B_k by a rank-1 update to get the new approximation B_{k+1}, i.e., we set

$$B_{k+1} = B_k + \rho u u^\top, \quad \rho \in \{+1, -1\}, \quad u \in \mathbb{R}^n.$$

Clearly, B_{k+1} is again symmetric and the secant equation requires the following condition:

$$y^k = B_{k+1}s^k = B_k s^k + \rho u u^\top s^k = B_k s^k + [\rho u^\top s^k] \cdot u.$$

The latter equation implies that there exists δ such that $u = \delta[y^k - B_k s^k]$ which yields

$$y^k - B_k s^k = \rho\delta^2[(s^k)^\top(y^k - B_k s^k)](y^k - B_k s^k).$$

Using the restriction $\rho \in \{+1, -1\}$, we obtain

$$\rho = \operatorname{sign}((s^k)^\top(y^k - B_k s^k)) \quad \text{and} \quad \delta = [|(s^k)^\top(y^k - B_k s^k)|]^{-\frac{1}{2}}$$

and hence, the *SR-1 update* is given by

$$B_{k+1}^{\texttt{SR-1}} = B_k + \frac{(y^k - B_k s^k)(y^k - B_k s^k)^\top}{(s^k)^\top(y^k - B_k s^k)}. \tag{7.13}$$

If B_k is invertible, we can calculate the inverse of B_{k+1} explicitly by the Sherman-Morrison-Woodbury formula.

Lemma 7.8: Sherman-Morrison-Woodbury

Suppose that $A \in \mathbb{R}^{n\times n}$ is an invertible matrix and let $u, v \in \mathbb{R}^n$ be given. Then $A + uv^\top$ is invertible if and only if $1 + v^\top A^{-1}u \neq 0$ and it holds that

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

Setting $H_k := B_k^{-1}$, we then can compute the inverse update rule for $H_{k+1}^{\texttt{SR-1}} = (B_{k+1}^{\texttt{SR-1}})^{-1}$ as follows:

$$\begin{aligned} H_{k+1}^{\texttt{SR-1}} &= B_k^{-1} - \frac{B_k^{-1}(y^k - B_k s^k)(y^k - B_k s^k)^\top B_k^{-1}}{(s^k)^\top(y^k - B_k s^k) \cdot (1 + \frac{(y^k - B_k s^k)^\top B_k^{-1}(y^k - B_k s^k)}{(s^k)^\top(y^k - B_k s^k)})} \\ &= H_k - \frac{(H_k y^k - s^k)(H_k y^k - s^k)^\top}{(s^k)^\top(y^k - B_k s^k) \cdot (1 + \frac{(y^k - B_k s^k)^\top(H_k y^k - s^k)}{(s^k)^\top(y^k - B_k s^k)})} \\ &= H_k - \frac{(H_k y^k - s^k)(H_k y^k - s^k)^\top}{(H_k y^k - s^k)^\top y^k} = H_k + \frac{(s^k - H_k y^k)(s^k - H_k y^k)^\top}{(s^k - H_k y^k)^\top y^k}. \end{aligned}$$

We continue with several comments:

- The SR-1 update is not well-defined if $(s^k)^\top(y^k - B_k s^k) \approx 0$ or $(y^k)^\top(s^k - H_k y^k) \approx 0$. In practice, in order to prevent the SR-1 method from breaking down, one usually skips

the update if the denominator is too small. Specifically, the update is applied only if

$$|(y^k)^\top(s^k - H_k y^k)| \geq 10^{-8}\|y^k\|\|s^k - H_k y^k\|.$$

If this condition is not satisfied, we can simply set $H_{k+1} = H_k$.

- The symmetric rank-1 update rule does not maintain positive definiteness, i.e., if H_k is positive definite then H_{k+1} does not necessarily need to be positive definite. As a consequence, the direction $d^k = -H_k\nabla f(x^k)$ does not need to be a descent direction.
- Basically, the SR-1 update can be used to substitute the Hessian $\nabla^2 f(x^k)$. However, the convergence properties of the resulting approach are not well understood and the algorithm might be unstable (slow progress and oscillating gradient and SR-1 steps). There are other more suitable optimization methods that are not based on line search that allow to utilize SR-1 updates.

7.2.2. Symmetric Rank-2 Updates

The most important and successful quasi-Newton methods use symmetric rank-2 updates

$$B_{k+1} = B_k + \gamma_k u^k (u^k)^\top + \delta_k v^k (v^k)^\top.$$

There are many possibilities to define such an update. Here, we list several possible update formulae. (As before we set $s^k = x^{k+1} - x^k$ and $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$).

- The *Broyden-Fletcher-Goldfarb-Shanno-update* (BFGS):

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

- The *Davidon-Fletcher-Powell-update* (DFP):

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

- The *Broyden-class*:

$$B_{k+1}^\lambda = (1-\lambda)B_{k+1}^{\texttt{BFGS}} + \lambda B_{k+1}^{\texttt{DFP}} = B_{k+1}^{\texttt{BFGS}} + \lambda (s^k)^\top B_k s^k \cdot v^k (v^k)^\top, \quad \lambda \in \mathbb{R}.$$

Here, we have $v^k = \frac{y^k}{(y^k)^\top s^k} - \frac{B_k s^k}{(s^k)^\top B_k s^k}$ and it holds that $B_{k+1}^0 = B_{k+1}^{\texttt{BFGS}}$, $B_{k+1}^1 = B_{k+1}^{\texttt{DFP}}$, and $B_{k+1}^\lambda = B_{k+1}^{\texttt{SR-1}}$ for $\lambda = (s^k)^\top y^k / ((s^k)^\top y^k - (s^k)^\top B_k s^k)$.

- The *convex Broyden-class*: B_{k+1}^λ for $\lambda \in [0,1]$.

The BFGS-method which is based on the BFGS-update is one of the most efficient quasi-Newton methods in practice. We now discuss several important properties of the introduced quasi-Newton updates. First, we show that the DFP and BFGS-update are optimal in the sense that they define symmetric updates with smallest change of B_k and B_k^{-1}.

Theorem 7.9

Let B_k be positive definite with $(y^k)^\top s^k > 0$. Then there exists a symmetric, positive definite matrix W satisfying $W^2 s^k = y^k$. For each of such matrix W, it holds that

- The matrix $B = B_{k+1}^{\texttt{DFP}}$ is a solution of the problem

$$\min_B \|W^{-1}(B - B_k)W^{-1}\|_F \quad \text{s.t.} \quad B = B^\top, \quad Bs^k = y^k.$$

- The matrix $B = B_{k+1}^{\texttt{BFGS}}$ is a solution of the problem

$$\min_B \|W(B^{-1} - B_k^{-1})W\|_F \quad \text{s.t.} \quad B = B^\top, \quad Bs^k = y^k.$$

Remark 7.10. The optimality property with respect to the weighted norm stated in Theorem 7.9 guarantees that the DFP- and BFGS-method are invariant under affine-linear transformations. This important property is a favorable feature of the Broyden-class.

The convex Broyden-class always generates symmetric, positive definite matrices B_k under suitable assumptions.

Theorem 7.11: Properties of Broyden-Updates

The Broyden-updates satisfy the following properties:

(i) In the case $(s^k)^\top y^k \neq 0$ and $(s^k)^\top B_k s^k \neq 0$, the matrix B_{k+1}^λ, $\lambda \in \mathbb{R}$ is well-defined, symmetric, and the quasi-Newton equation (7.12) holds.

(ii) If B_k is positive definite and we have $(s^k)^\top y^k > 0$, then B_{k+1}^λ is positive definite for all $\lambda \geq 0$.

Proof. Well-definedness and symmetry of B_{k+1}^λ are obvious. Moreover, due to $(uv^\top)w = (v^\top w)u$, we have

$$B_{k+1}^{\texttt{BFGS}} s^k = B_k s^k + \frac{(y^k)^\top s^k}{(y^k)^\top s^k} y^k - \frac{(s^k)^\top B_k s^k}{(s^k)^\top B_k s^k} B_k s^k = B_k s^k + y^k - B_k s^k = y^k,$$
$$(v^k)^\top s^k = \frac{(y^k)^\top s^k}{(y^k)^\top s^k} - \frac{(s^k)^\top B_k s^k}{(s^k)^\top B_k s^k} = 0.$$

Consequently, it follows $B_{k+1}^\lambda s^k = B_{k+1}^{\texttt{BFGS}} s^k + \lambda((s^k)^\top B_k s^k)((v^k)^\top s^k)v^k = y^k$. We now continue with part (ii). Using $\lambda \geq 0$ and $(s^k)^\top B_k s^k \geq 0$, it holds that

$$h^\top B_{k+1}^\lambda h = h^\top B_{k+1}^{\texttt{BFGS}} h + \lambda((s^k)^\top B_k s^k)((v^k)^\top h)^2 \geq h^\top B_{k+1}^{\texttt{BFGS}} h.$$

Hence, it suffices to verify positive definiteness of $B_{k+1}^{\texttt{BFGS}}$. Since the matrix B_k is positive definite, the square root $R_k = B_k^{1/2}$ exists and we can write $B_k = R_k R_k$. For all $h \in \mathbb{R}^n \backslash \{0\}$,

we then have

$$\begin{aligned} h^\top B_{k+1}^{\texttt{BFGS}} h &= h^\top B_k h + \frac{((y^k)^\top h)^2}{(y^k)^\top s^k} - \frac{((B_k s^k)^\top h)^2}{(s^k)^\top B_k s^k} \\ &= \|R_k h\|^2 + \frac{((y^k)^\top h)^2}{(y^k)^\top s^k} - \frac{((R_k s^k)^\top (R_k h))^2}{\|R_k s^k\|^2} \\ &\geq \|R_k h\|^2 + \frac{((y^k)^\top h)^2}{(y^k)^\top s^k} - \frac{\|R_k s^k\|^2 \|R_k h\|^2}{\|R_k s^k\|^2} = \frac{((y^k)^\top h)^2}{(y^k)^\top s^k} \geq 0. \end{aligned}$$

In the case $h \notin \mathbb{R} s^k$, it holds that $R_k h \notin \mathbb{R} R_k s^k$ and hence, the Cauchy-Schwarz inequality in the second last estimate is satisfied with "$>$". If h and s^k are linearly dependent, there is $t \in \mathbb{R}\backslash\{0\}$ with $h = ts^k$. In this case, it follows $((y^k)^\top h)^2 = t^2((y^k)^\top s^k) > 0$. This yields "$>$" in the last estimate. ■

Consequently, in contrast to the SR-1 update, the Broyden-updates allow to maintain positive definiteness of the approximation which can be very useful. We note that the curvature condition

$$(s^k)^\top y^k > 0 \tag{7.14}$$

is automatically satisfied if the function f is strictly convex.

By applying the Sherman-Morrison-Woodbury formula Lemma 7.8 twice, it is possible to calculate the inverse quasi-Newton update rules. In particular, suppose that B_k is positive definite and set $H_k = B_k^{-1}$. If $(s^k)^\top y^k > 0$, then the inverse quasi-Newton updates for $H_{k+1}^{\texttt{BFGS}} = (B_{k+1}^{\texttt{BFGS}})^{-1}$ and $H_{k+1}^{\texttt{DFP}} = (B_{k+1}^{\texttt{DFP}})^{-1}$ are given by:

$$\begin{aligned} H_{k+1}^{\texttt{BFGS}} &= H_k + \frac{(s^k - H_k y^k)(s^k)^\top + s^k(s^k - H_k y^k)^\top}{(s^k)^\top y^k} - \frac{(s^k - H_k y^k)^\top y^k}{((s^k)^\top y^k)^2} \cdot s^k (s^k)^\top \\ H_{k+1}^{\texttt{DFP}} &= H_k + \frac{s^k (s^k)^\top}{(s^k)^\top y^k} - \frac{H_k y^k (H_k y^k)^\top}{(y^k)^\top H_k y^k}. \end{aligned}$$

Hence, the inverse BFGS-update rule uses the (regular) DFP-update formula when substituting (B_k, s^k, y^k) by (H_k, y^k, s^k). Similarly, we obtain the inverse DFP-update rule by applying the (regular) BFGS update formula using (H_k, y^k, s^k) instead of (B_k, s^k, y^k).

7.2.3. The BFGS Method

A full algorithm based on the BFGS-updates is presented in Algorithm 7.5. We have the following remarks:

- Compared to Newton's method we do not need to access the Hessian of f to calculate the direction d^k. Moreover, the computation of d^k only requires a matrix–vector multiplication and not the solution of a linear system of equations!
- The skipping mechanism in step 5 is usually not recommended, since too many updates might be skipped if the problem is nonconvex. In such a case, H_k can not capture

Algorithm 7.5: Globalized BFGS-Method

Initialization: Select an initial point $x^0 \in \mathbb{R}^n$ and a symmetric, positive definite matrix $H_0 \in \mathbb{R}^{n\times n}$. Choose parameters $\sigma, \gamma \in (0,1)$, $s = 1$, and tolerances $\delta, \varepsilon > 0$.

for $k = 0, 1, \ldots$ **do**

Compute the quasi-Newton direction $d^k = -H_k \nabla f(x^k)$.

Calculate a step size α_k using backtracking.

Set $x^{k+1} = x^k + \alpha_k d^k$. If $\|\nabla f(x^{k+1})\| \leq \varepsilon$, then STOP and x^{k+1} is the output.

Set $s^k = x^{k+1} - x^k$ and $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$. If $(s^k)^\top y^k \leq \delta$ set $H_{k+1} = H_k$, otherwise compute

$$H_{k+1} = H_k + \frac{(s^k - H_k y^k)(s^k)^\top + s^k(s^k - H_k y^k)^\top}{(s^k)^\top y^k} - \frac{(s^k - H_k y^k)^\top y^k}{((s^k)^\top y^k)^2} \cdot s^k (s^k)^\top.$$

important curvature information of the objective function f and the performance of the BFGS method can degrade.

Typically, a different type of line search algorithm is used which can ensure that the curvature condition (7.14) is satisfied throughout the iterative process. The line search procedure is based on the *Powell-Wolfe conditions* or *Wolfe conditions*: find a step size $\alpha_k > 0$ such that

$$\begin{aligned} f(x^k + \alpha_k d^k) - f(x^k) &\leq \gamma \alpha_k \cdot \nabla f(x^k)^\top d^k, \\ \nabla f(x^k + \alpha d^k)^\top d^k &\geq \eta \cdot \nabla f(x^k)^\top d^k, \end{aligned} \tag{7.15}$$

where $\gamma \in (0, \frac{1}{2})$ and $\eta \in (\gamma, 1)$.

If d^k is a descent direction of f at x^k and f is lower bounded along d^k, i.e., $\inf_{\alpha \geq 0} f(x^k + \alpha d^k) > \infty$, then it can be shown that the Powell-Wolfe conditions are well-defined and there exists such a step size. Algorithmically, the Powell-Wolfe step size can be found as follows: we first compute an interval $[\alpha_-, \alpha_+]$ such that the Armijo condition is satisfied for α_- and violated for α_+. Setting

$$\psi(\alpha) = f(x^k + \alpha d^k) - f(x^k) - \gamma \alpha \nabla f(x^k)^\top d^k,$$

this means $\psi(\alpha_-) \leq 0$ and $\psi(\alpha_+) > 0$. Hence, there exists $\alpha^* \in [\alpha_-, \alpha_+]$ with $\psi(\alpha^*) = 0$. We can now try to find this α^* and the step size $\alpha \in (0, \alpha^*]$ that is sufficiently close to α^* by bisection. Indeed, we have

$$0 = \psi'(\alpha^*) = \nabla f(x^k + \alpha^* d^k)^\top d^k - \gamma \nabla f(x^k)^\top d^k < \nabla f(x^k + \alpha^* d^k)^\top d^k - \eta \nabla f(x^k)^\top d^k$$

and hence, by a continuity argument, the Powell-Wolfe conditions are satisfied for all such α sufficiently close to α^*.

- A good choice of the initial matrix H_0 can be crucial for fast convergence. Common choices are $H_0 = \rho I$, where ρ is chosen appropriately (dependent on the problem).

Algorithm 7.6: L-BFGS two-loop recursion

Set $q = \nabla f(x^k)$.
for $i = k-1, k-2, \ldots, k-m$ **do**
  $a_i = \rho_i \cdot (s^i)^\top q$.
  $q = q - a_i y^i$.
Set $r = H_k^0 q$.
for $i = k-m, k-m+1, \ldots, k-1$ **do**
  $\beta = \rho_i \cdot (y^i)^\top r$.
  $r = r + (a_i - \beta) s^i$.
STOP with result $r = H_k \nabla f(x^k)$.

The convergence analysis of the BFGS method is not easy. If the function f is strongly convex, then for every symmetric, positive definite initial matrix $H_0 \in \mathbb{R}^{n\times n}$ and every initial point $x^0 \in \mathbb{R}^n$, the sequence (x^k) generated by Algorithm 7.5 (using the Powell-Wolfe conditions) converges to the unique minimizer of f. In addition, the rate of convergence can be shown to be q-superlinear.

Theorem 7.12: Convergence of the BFGS method

Let $x^0 \in \mathbb{R}^n$ be an arbitrary initial point and let H_0 be a given symmetric, positive definite initial matrix. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and that the level set $L_{\leq f(x^0)} := \{x \in \mathbb{R}^n : f(x) \leq f(x^0)\}$ is convex. Further assume that there are constants $m, M > 0$ such that

$$m\|h\|^2 \leq h^\top \nabla^2 f(x) h \leq M\|h\|^2 \quad \forall\, h \in \mathbb{R}^n, \quad \forall\, x \in L_{\leq f(x^0)}.$$

Let the sequence $(x^k)_k$ be generated by Algorithm 7.5 using $\delta = \epsilon = 0$ and the Powell-Wolfe conditions (7.15) in step 4. Then $(x^k)_k$ converges to the unique global minimizer of f. In addition, if $\nabla^2 f$ is Lipschitz continuous in a neighborhood of x^*, then $(x^k)_k$ converges to x^* at a q-superlinear rate.

We refer to section 6 in "Nocedal & Wright: Numerical Optimization, 2006" for a detailed proof of this result. We can also establish a purely local result:

Theorem 7.13: Convergence of the Local BFGS method

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable with locally Lipschitz continuous Hessian. Furthermore, assume that the second order sufficient conditions are satisfied at $x^* \in \mathbb{R}^n$. Then there exists $\epsilon, \delta > 0$ such that for every initial point $x^0 \in B_\epsilon(x^*)$ and symmetric positive definite initial matrix $H_0 \in \mathbb{R}^{n\times n}$ with $\|H_0 - \nabla^2 f(x^*)^{-1}\| < \delta$, Algorithm 7.5 (using full step sizes $\alpha_k = 1$) either terminates with $x^k = x^*$ or it generates a sequence $(x^k)_k \subset B_\epsilon(x^*)$ that converges q-superlinearly to x^*.

We refer to the paper "Broyden, Dennis & Moré: On the local and superlinear convergence of quasi-Newton methods, 1973" for further details.

7.2.4. Limited Memory BFGS Updates

So far, we have built the approximations B_k and H_k of the Hessian and inverse of the Hessian as full $n \times n$ matrices. If the dimension n is large, this might not be possible or too expensive. Let us notice that the matrix H_k is not required explicitly, in fact we are only interested in $d^k = -H_k \nabla f(x^k)$. More specifically, for an arbitrary vector $v \in \mathbb{R}^n$, it holds that

$$H_{k+1}v = H_k\left[v - \frac{(s^k)^\top v}{(s^k)^\top y^k}\cdot y^k\right] + \frac{(s^k)^\top v}{(s^k)^\top y^k}s^k - \frac{(y^k)^\top H_k[v - \frac{(s^k)^\top v}{(s^k)^\top y^k}\cdot y^k]}{(s^k)^\top y^k}s^k.$$

Hence, $H_{k+1}v$ can be computed recursively:

$$\beta_k = \frac{(s^k)^\top v}{(s^k)^\top y^k}, \quad q^k = v - \beta_k y^k, \quad p^k = H_k q^k, \quad H_{k+1}v = p^k + \left[\beta_k - \frac{(y^k)^\top p^k}{(s^k)^\top y^k}\right] s^k$$

In this procedure, we only need to store the pairs

$$\{y^0, y^1, y^2, ..., y^k\} \quad \text{and} \quad \{s^0, s^1, s^2, ..., s^k\}$$

and not built H_{k+1} explicitly. However, as the number of total iterations k grows, this strategy becomes inefficient and still requires significant memory storage. The idea of the limited memory BFGS method is now to store only the last m pairs

$$\{y^{k-m}, y^{k-m+1}, ..., y^k\} \quad \text{and} \quad \{s^{k-m}, s^{k-m+1}, ..., s^k\} \tag{7.16}$$

and use this information to build a much cheaper BFGS-type approximation of the Hessian.

Since the pairs (7.16) can change in each iteration, the main difference between the standard BFGS and the L-BFGS iteration is that the inverse Hessian approximation H_k is built from scratch in every iteration in the L-BFGS method! In particular, in each iteration, an initial approximation H_k^0 is chosen (this initial approximation is allowed to change from iteration to iteration) and the product $H_k \nabla f(x^k)$ is then computed recursively following our outlined procedure. More specifically, defining

$$\rho_k := \frac{1}{(s^k)^\top y^k},$$

the product $H_k \nabla f(x^k)$ can be computed efficiently by the two-loop recursion shown in Algorithm 7.6.

Without considering the multiplication $H_k^0 q$, the two-loop recursion scheme requires $4mn$ multiplications and the storage requirements are

$$\{y^{k-m}, y^{k-m+1}, ..., y^k\} \rightsquigarrow m \times n, \quad \{s^{k-m}, s^{k-m+1}, ..., s^k\} \rightsquigarrow m \times n,$$
$$\{\rho_{k-m}, \rho_{k-m+1}, ..., \rho_k\} \rightsquigarrow m,$$

i.e., we need to store $2mn + m$ numbers. If H_k^0 is diagonal, then n additional multiplications are needed. Notice that the initial matrix H_k^0 can be chosen freely and can vary between

	BFGS Method			
`tol`	`iter`–avg.	`iter`–min	`iter`–max	`time`–avg.
10^{-5}	12.5	8	19	0.00 s
10^{-7}	13.5	8	20	0.00 s
10^{-9}	14.3	9	20	0.00 s

Table 7.2: Numerical results of the BFGS method.

iterations. Typically the matrix H_k^0 is set to $H_k^0 = \gamma_k I$, where

$$\gamma_k = \frac{(s^{k-1})^\top y^{k-1}}{(y^{k-1})^\top y^{k-1}} \quad \text{or} \quad \gamma_k = \frac{(s^{k-1})^\top s^{k-1}}{(s^{k-1})^\top y^{k-1}}$$

and the memory m is chosen as $m \in [5, 25] \cap \mathbb{N}$.

7.2.5. Numerical Experiments

We now briefly discuss the performance of the BFGS method.

A First Nonconvex Example – Continued. We consider the nonconvex optimization problem studied in subsection 6.3:

$$\min_{x \in \mathbb{R}^2} \; f_1(x)^2 + f_2(x)^2$$

where $f_1 : \mathbb{R}^2 \to \mathbb{R}$ and $f_2 : \mathbb{R}^2 \to \mathbb{R}$ are given by:

$$f_1(x) := -1 + x_1 + ((5 - x_2)x_2 - 2)x_2, \quad f_2(x) := -1 + x_1 + ((x_2 + 1)x_2 - 10)x_2.$$

We use the following parameter in the implementation of the BFGS method:

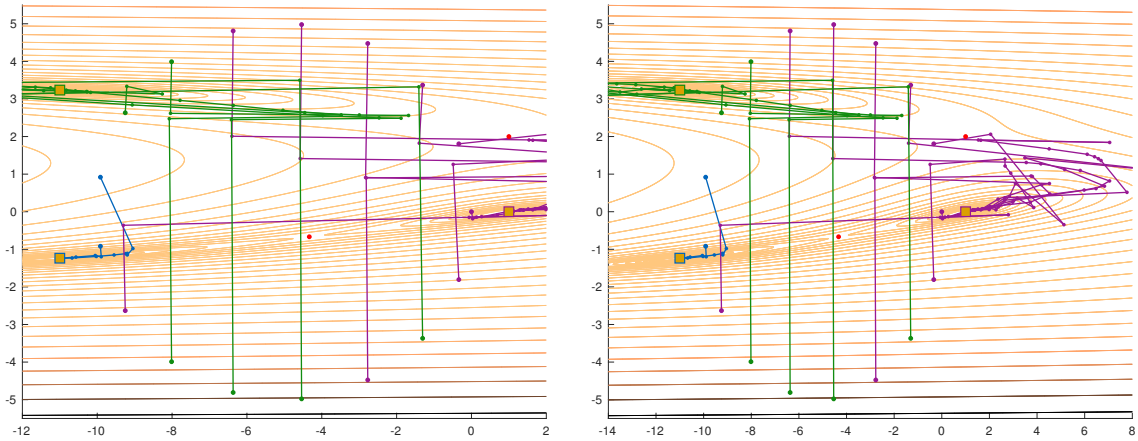
- $s = 1$, $\gamma = 0.1$, $\sigma = 0.5$ and $H_0 = I$.
- We skip the curvature update whenever $(s^k)^\top y^k < 10^{-14}$.

The method stops whenever the stopping criterion $\|\nabla f(x^k)\| \leq \texttt{tol}$, $\texttt{tol} \in \{10^{-5}, 10^{-7}, 10^{-9}\}$ is satisfied. As described in subsection 6.3, we again run the BFGS method for 17 different initial points. The solution paths are shown in Figure 7.4. Further numerical results are reported in Table 7.2 and Figure 7.5.

Logistic Regression. We now consider an optimization model for binary classification – the so-called *logistic regression model*. We focus on classification problems with two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. We further assume that we have some training data which consists of feature vectors $a_i \in \mathbb{R}^n$, $i \in \{1, 2, ..., m\}$ and corresponding class labels $b_i \in \{-1, 1\}$. Here, each label b_i indicates whether the data point a_i belongs to the class $\mathcal{C}_1$ or $\mathcal{C}_2$. The corresponding minimization problem is given by:

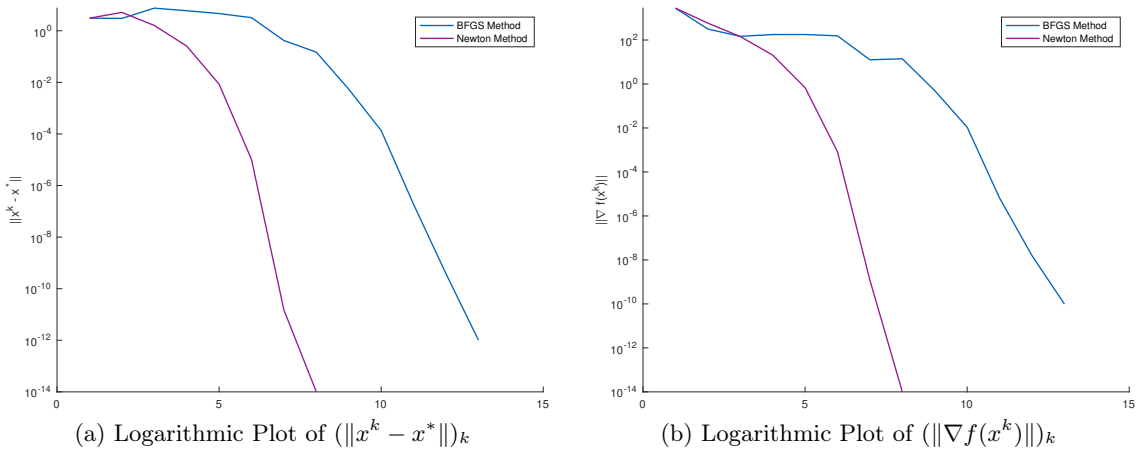
$$\min_{x,y} \; f_{\texttt{log}}(x, y) = \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-b_i \cdot (a_i^\top x + y))) + \frac{\lambda}{2}\|x\|^2, \tag{7.17}$$

where $\lambda > 0$ is a model parameter. The overall methodology is based on a probabilistic idea.



(a) BFGS Method (b) BFGS Method (Larger Scale)

Figure 7.4: Solution paths for the BFGS method for different x^0.



(a) Logarithmic Plot of $(\|x^k - x^*\|)_k$ (b) Logarithmic Plot of $(\|\nabla f(x^k)\|)_k$

Figure 7.5: Plot and comparison of the convergence rates of the BFGS and Newton method.

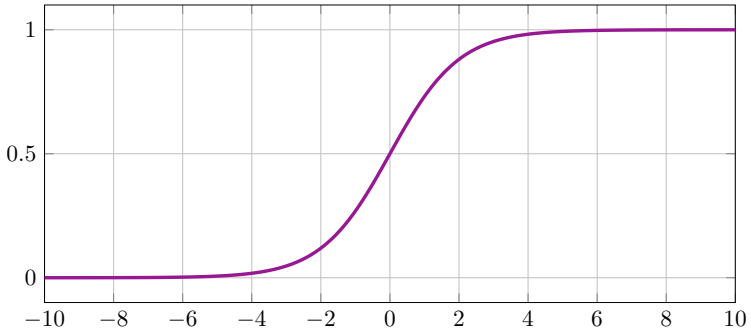


Figure 7.6: Plot of the sigmoid function σ.

Specifically, we try to find a good parametric model of the probability that a given feature vector belongs to the first class $\mathcal{C}_1$. The parameters x, y of the model function then have to be chosen such that the predicted probability is close to 1 for every feature vector which was assigned to the class $\mathcal{C}_1$. After the training has been finished, the model can be used to classify any feature vector according to the probabilistic estimate.

In the case of logistic regression, the *sigmoid function* $\sigma : \mathbb{R} \to \mathbb{R}$, $\sigma(a) = \frac{1}{1+\exp(-a)}$ is chosen as underlying probability model. A plot of the sigmoid function σ is shown in Figure 7.6. In logistic regression, we want to train the linear model $\ell_{(x,y)}(a) = a^\top x + y$ such that

$$\sigma(\ell_{(x,y)}(a_i)) = \sigma(a_i^\top x + y) \approx \begin{cases} 1 & \text{if } a_i \text{ belongs to class } \mathcal{C}_1, \text{ i.e., } b_i = +1, \\ 0 & \text{if } a_i \text{ belongs to class } \mathcal{C}_2, \text{ i.e., } b_i = -1. \end{cases}$$

A new data point $a \in \mathbb{R}^n$ can then be classified via

$$\begin{cases} +1 & \text{if } \sigma(\ell_{(x,y)}(a)) > \frac{1}{2}, \\ -1 & \text{if } \sigma(\ell_{(x,y)}(a)) \leq \frac{1}{2} \end{cases} \quad \text{or} \quad \begin{cases} \mathcal{C}_1 & \text{if } \sigma(\ell_{(x,y)}(a)) > \frac{1}{2}, \\ \mathcal{C}_2 & \text{if } \sigma(\ell_{(x,y)}(a)) \leq \frac{1}{2}. \end{cases}$$

The optimization problem (7.17) is based on a corresponding maximum likelihood approach to estimate the optimal model parameters x and y for this probabilistic strategy.

8. Acceleration and Momentum Techniques

In this section, we investigate and present accelerated and more robust gradient methods to solve the unconstrained problem

$$\min_x \ f(x) \tag{8.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

8.1. Accelerated Gradient Methods

Acceleration techniques are important tools to improve the performance and complexity results of the basic gradient method. These techniques are especially powerful if the optimization problem (8.1) is convex, i.e., we will add the assumption:

- The smooth mapping $f : \mathbb{R}^n \to \mathbb{R}$ is convex.

The principle idea of many acceleration techniques is to perform a so-called *extrapolation step*

$$y^{k+1} = x^k + \beta_k(x^k - x^{k-1}), \qquad \beta_k > 0$$

to approximate and extrapolate the next iterate $y^{k+1} \approx x^{k+1}$. The extrapolation step y^{k+1} uses the information of the last two iterates x^k and x^{k-1} to first predict "x^{k+1}". We then perform a gradient step based on the predicted information.

This idea was first proposed by Nesterov (1983, 1988, 2005) and is an essential component of the accelerated gradient method that we will discuss here. The challenging and difficult question is now: how should we choose β_k to guarantee (faster) convergence?

We present a first abstract accelerated gradient method in Algorithm Algorithm 8.1.

Algorithm 8.1: An Abstract Accelerated Gradient Method

1 Initialization: Choose an initial point $x^0 \in \mathbb{R}^n$ and set $x^{-1} = x^0$.
for $k = 0, 1, 2, \ldots$ **do**
2 Select an extrapolation parameter β_k and compute $y^{k+1} = x^k + \beta_k(x^k - x^{k-1})$.
3 Select a step size $\alpha_k > 0$ and set $x^{k+1} = y^{k+1} - \alpha_k \nabla f(y^{k+1})$.

In the case $\beta_k = 0$, the accelerated gradient method obviously reduces to the standard gradient method (without line search). We are specifically interested in extrapolation parameters of the form:

$$\beta_k = \frac{\theta_k(1 - \theta_{k-1})}{\theta_{k-1}} = \frac{\theta_k}{\theta_{k-1}} - \theta_k \quad \text{and} \quad \theta_{-1} = \theta_0 = 1. \tag{8.2}$$

The following convergence result can be established:

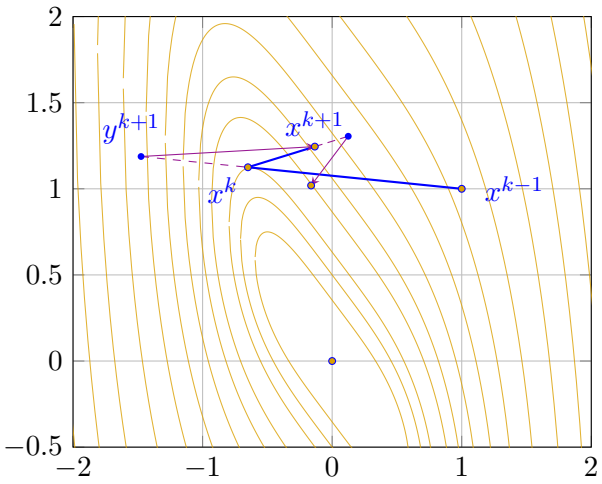


Figure 8.1: Illustration of the extrapolation procedure. The plot depicts the level sets of the convex function $f(x) = x_1^4 + x_1^2 + x_2^2 + 2x_1x_2 + 0.25$ and two exemplary steps of Algorithm 8.1 using $\beta_k = 0.5$ and $\alpha_k = 0.1$.

Theorem 8.1: Convergence of the Accelerated Gradient Method

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex mapping satisfying $f \in C_L^{1,1}(\mathbb{R}^n)$ and assume that the solution set $\mathcal{X}^*$ of the problem (8.1) is nonempty. Let $(x^k)_k$, $(\alpha_k)_k$, and $(\beta_k)_k$ be generated by Algorithm 8.1 with $\alpha_k = \bar{\alpha} \in (0, \frac{1}{L}]$,

$$\beta_k \text{ satisfies (8.2)} \quad \text{and we have} \quad 0 \leq \theta_k \leq \frac{2}{k+2}, \quad \frac{1-\theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}$$

for all $k \in \mathbb{N}$. Then, for every $x^* \in \mathcal{X}^*$, it follows

$$f(x^k) - f(x^*) \leq \frac{2\|x^0 - x^*\|^2}{\bar{\alpha}(k+1)^2} \quad \forall\, k \in \mathbb{N}.$$

Proof. We first introduce the auxiliary sequence $z^k := x^{k-1} + \theta_{k-1}^{-1}(x^k - x^{k-1})$. Using $y^{k+1} = x^k + \beta_k(x^k - x^{k-1})$ and (8.2), we can express z^k also differently, i.e., it holds that:

$$z^k = x^k + \theta_k^{-1}(y^{k+1} - x^k).$$

We further note that we have $z^{k+1} - z^k = \theta_k^{-1}(x^{k+1} - y^{k+1}) = -\bar{\alpha}\theta_k^{-1}\nabla f(y^{k+1})$. By the descent lemma Lemma 5.15, it follows:

$$\begin{aligned} f(x^{k+1}) &\leq f(y^{k+1}) + \nabla f(y^{k+1})^\top (x^{k+1} - y^{k+1}) + \frac{L}{2}\|y^{k+1} - x^{k+1}\|^2 \\ &= f(y^{k+1}) - \bar{\alpha}\left[1 - \frac{L\bar{\alpha}}{2}\right]\|\nabla f(y^{k+1})\|^2. \end{aligned} \tag{8.3}$$

Let $x^* \in \mathcal{X}^*$ be arbitrary and let us set $y^* = (1-\theta_k)x^k + \theta_k x^*$. Using the convexity and

differentiability of f and $\theta_k \in [0,1]$, we obtain

$$\begin{aligned} f(x^*) - f(y^{k+1}) &= f(x^*) - f(y^*) + f(y^*) - f(y^{k+1}) \\ &\geq (1-\theta_k)[f(x^*) - f(x^k)] + \langle \nabla f(y^{k+1}), y^* - y^{k+1}\rangle \\ &= (1-\theta_k)[f(x^*) - f(x^k)] - \theta_k \bar{\alpha}^{-1}\langle z^{k+1} - z^k, \theta_k(x^* - x^k) + x^k - y^{k+1}\rangle \\ &= (1-\theta_k)[f(x^*) - f(x^k)] - \theta_k^2 \bar{\alpha}^{-1}\langle z^{k+1} - z^k, x^* - z^k\rangle. \end{aligned}$$

The last term in this estimate can now be reformulated similar as in the proof of Theorem 5.27:

$$\begin{aligned} \langle z^{k+1} - z^k, x^* - z^k\rangle &= \langle z^{k+1} - x^*, x^* - z^k\rangle + \|z^k - x^*\|^2 \\ &= \frac{1}{2}\|z^{k+1} - z^k\|^2 - \frac{1}{2}\|z^{k+1} - x^*\|^2 + \frac{1}{2}\|z^k - x^*\|^2. \end{aligned}$$

Combining the last results, it thus follows:

$$\begin{aligned} f(y^{k+1}) - f(x^*) \leq (1-\theta_k)[f(x^k) - f(x^*)] + \frac{\bar{\alpha}}{2}\|\nabla f(y^{k+1})\|^2 \\ - \frac{\theta_k^2}{2\bar{\alpha}}\|z^{k+1} - x^*\|^2 + \frac{\theta_k^2}{2\bar{\alpha}}\|z^k - x^*\|^2. \end{aligned}$$

Adding this estimate and (8.3), we obtain

$$\frac{1}{\theta_k^2}[f(x^{k+1}) - f(x^*)] \leq \frac{1-\theta_k}{\theta_k^2}[f(x^k) - f(x^*)] - \frac{1}{2\bar{\alpha}}\|z^{k+1} - x^*\|^2 + \frac{1}{2\bar{\alpha}}\|z^k - x^*\|^2.$$

Summing this inequality and applying the bound $(1-\theta_{k+1})\theta_{k+1}^{-2} \leq \theta_k^{-2}$, we can readily infer

$$f(x^k) - f(x^*) \leq \frac{\theta_{k-1}^2}{2\bar{\alpha}}\|z^0 - x^*\|^2 = \frac{\theta_{k-1}^2}{2\bar{\alpha}}\|x^0 - x^*\|^2$$

which finishes the proof of Theorem 8.1. ■

Compared to the complexity bound for the gradient method derived in Theorem 5.23, the accelerated gradient method has an improved dependency on the k. Specifically, we only require

$$k \geq \frac{\sqrt{2}\|x^0 - x^*\|}{\sqrt{\bar{\alpha}\varepsilon}} - 1 = \mathcal{O}(\sqrt{\varepsilon^{-1}})$$

iterations to reach an iterate satisfying $f(x^k) - f(x^*) \leq \varepsilon$. In comparison, the standard gradient method requires $\mathcal{O}(\varepsilon^{-1})$ steps. We continue with several remarks and present parameter strategies.

- The choice $\theta_k = \frac{2}{k+2}$ satisfies

$$\frac{1-\theta_{k+1}}{\theta_{k+1}^2} = \frac{(k+3)^2 - 2(k+3)}{4} = \frac{k^2+4k+3}{4} = \frac{(k+2)^2 - 1}{4} < \frac{1}{\theta_k^2}.$$

In this case, the extrapolation parameter β_k reduces to $\beta_k = \frac{k-1}{k+2}$.

- Another popular choice is $\theta_k = t_k^{-1}$ and

$$t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}, \quad \beta_k = \frac{t_{k-1}-1}{t_k}, \quad t_{-1} = t_0 = 1.$$

This choice satisfies $(1-\theta_{k+1})\theta_{k+1}^{-2} = \theta_k^{-2}$.

- If the Lipschitz constant L is unknown, then the step size α_k can be determined by the following line search procedure: Choose $\sigma \in (0,1)$:
  - Set $\alpha_k = \alpha_{k-1}$ and $x^{k+1} = y^{k+1} - \alpha_k \nabla f(y^{k+1})$.
  - `while` $f(x^{k+1}) > f(y^{k+1}) - \frac{\alpha_k}{2}\|\nabla f(y^{k+1})\|^2$ `do`:
    Set $\alpha_k = \sigma\alpha_k$ and compute $x^{k+1} = y^{k+1} - \alpha_k \nabla f(y^{k+1})$.

8.2. Momentum and the Inertial Gradient Method

In this section, we discuss a second acceleration strategy for the optimization problem (8.1) that is similar to Nesterov's acceleration technique. We again consider an extrapolation step

$$y^{k+1} = x^k + \beta_k(x^k - x^{k-1}), \quad \beta_k > 0.$$

However, the new iterate x^{k+1} is now updated as follows:

$$x^{k+1} = y^{k+1} - \alpha_k \nabla f(x^k) = x^k - \alpha_k \nabla f(x^k) + \beta_k(x^k - x^{k-1}). \tag{8.4}$$

Here, α_k is a suitable step size. In contrast to the accelerated gradient method described in the last section, the gradient is now evaluated at x^k and not at y^{k+1}. This small difference is essential! In particular, this new method does not have the same convergence guarantees as Algorithm 8.1 and can converge with a slower rate. The algorithmic scheme (8.4) is known as *inertial gradient method* or *Heavy-ball method* and was introduced by Polyak (1964).

Although inertial methods do not enjoy the same theoretical advantages as accelerated algorithms, they are the still popular in applications – especially in nonconvex problems when f is nonconvex. This is partly due to the following reasons:

- The inertial term $\beta_k(x^k - x^{k-1})$ contains information about the past iterates and can be used to leverage the gradient step and improve the gradient direction. We have added "momentum".

- The update (8.4) has a precise interpretation: it corresponds to an explicit finite differences discretization of the so-called *Heavy-ball with friction dynamical system*:

$$\ddot{x}(t) + \gamma\dot{x}(t) + \nabla f(x(t)) = 0.$$

"*Instead of walking along the path of slow gradient steps, we roll a heavy ball along the landscape of the optimization problem. This reduces oscillations and potentially accelerates the method*".

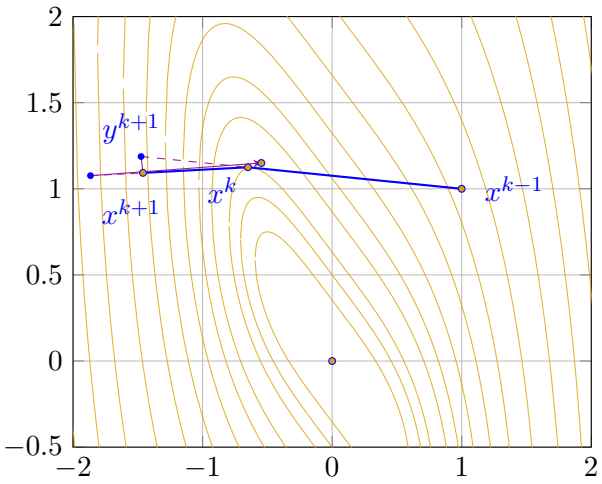


Figure 8.2: Illustration of the inertial gradient method. The plot shows the level sets of the convex mapping $f(x) = x_1^4 + x_1^2 + x_2^2 + 2x_1x_2 + 0.25$ and two exemplary steps of Algorithm 8.2 using $\beta = 0.5$ and $\alpha_k = 0.1$.

The full method is presented below in Algorithm 8.2. A visualization of the extrapolation scheme and the inertial step is given in Figure 8.2.

Algorithm 8.2: An Inertial Gradient Method

Initialization: Choose an initial point $x^0 \in \mathbb{R}^n$ and set $x^{-1} = x^0$.
for $k = 0, 1, 2, \ldots$ **do**
  Select an extrapolation parameter β_k and compute $y^{k+1} = x^k + \beta_k(x^k - x^{k-1})$.
  Select a step size $\alpha_k > 0$ and set $x^{k+1} = y^{k+1} - \alpha_k \nabla f(x^k)$.

As in the accelerated gradient method, the parameters β_k and step sizes α_k need to be chosen carefully to guarantee convergence. If the Lipschitz constant L is unknown, we can utilize the following iterative scheme: select the minimal $L_k \in \{L_{k-1}, \sigma^{-1}L_{k-1}, \sigma^{-2}L_{k-1}, \ldots\}$ and a corresponding $\alpha_k < 2(1-\beta)/L_k$ satisfying:

$$f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^\top (x^{k+1} - x^k) + \frac{L_k}{2}\|x^{k+1} - x^k\|^2.$$

If ∇f is Lipschitz continuous, this procedure will stop after finitely many reductions of L_k. We only present a specialized convergence theorem that establishes global convergence in the general nonconvex case.

Theorem 8.2: Convergence of the Inertial Gradient Method

Let $f \in C_L^{1,1}(\mathbb{R}^n)$ be given and let $(x^k)_k$ be the sequence generated by the inertial gradient method Algorithm 8.2 utilizing the following extrapolation step size strategy:

$$\beta_k \equiv \bar{\beta} \in (0,1), \quad \alpha_k \equiv \bar{\alpha} < 2(1-\bar{\beta})/L.$$

Then, we either have $f(x^k) \to -\infty$ or $(f(x^k))_k$ converges to a finite value and it holds that $\nabla f(x^k) \to 0$ as $k \to \infty$.

Proof. Let us set $\delta := \frac{1}{\bar{\alpha}} - \frac{L}{2} - \frac{\bar{\beta}}{2\bar{\alpha}} > \frac{\bar{\beta}}{2\bar{\alpha}}$ an define the mapping $H(x,y) = f(x) + \delta\|x-y\|^2$. By the descent lemma, it follows

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \nabla f(x^k)^\top (x^{k+1} - x^k) + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \frac{1}{\bar{\alpha}}\langle x^{k+1} - y^{k+1}, x^{k+1} - x^k\rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2. \\ &= f(x^k) - \left[\frac{1}{\bar{\alpha}} - \frac{L}{2}\right]\|x^{k+1} - x^k\|^2 + \frac{\bar{\beta}}{\bar{\alpha}}\langle x^k - x^{k-1}, x^{k+1} - x^k\rangle \\ &\leq f(x^k) - \left[\frac{2-\bar{\beta}}{2\bar{\alpha}} - \frac{L}{2}\right]\|x^{k+1} - x^k\|^2 + \frac{\bar{\beta}}{2\bar{\alpha}}\|x^k - x^{k-1}\|^2, \end{aligned}$$

where we applied the estimate $a^\top b \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$ for all $a, b \in \mathbb{R}^n$. Using the definition of H, we then obtain

$$H(x^{k+1}, x^k) \leq H(x^k, x^{k-1}) - \left[\delta - \frac{\bar{\beta}}{2\bar{\alpha}}\right]\|x^k - x^{k-1}\|^2.$$

Consequently, the sequence $(H(x^k, x^{k-1}))_k$ is non-increasing and converges to some $\xi \in \mathbb{R} \cup \{-\infty\}$. In the case $\xi = -\infty$, we clearly have $f(x^k) \to -\infty$. Otherwise, summing the latter estimate, we can infer

$$H(x^0, x^{-1}) - \xi = \lim_{k\to\infty} \sum_{i=0}^{k-1} H(x^i, x^{i-1}) - H(x^{i+1}, x^i) \geq \left[\delta - \frac{\bar{\beta}}{2\bar{\alpha}}\right] \sum_{i=0}^{\infty} \|x^i - x^{i-1}\|^2.$$

This shows $\|x^k - x^{k-1}\| \to 0$ and hence, it follows $\lim_{k\to\infty} f(x^k) = \lim_{k\to\infty} H(x^k, x^{k-1}) = \xi \in \mathbb{R}$. Moreover, we have

$$\|\nabla f(x^k)\| = \frac{1}{\bar{\alpha}}\|x^{k+1} - y^{k+1}\| \leq \frac{1}{\bar{\alpha}}\|x^{k+1} - x^k\| + \frac{\bar{\beta}}{\bar{\alpha}}\|x^k - x^{k-1}\| \to 0$$

as $k \to \infty$. This finishes the proof of Theorem 8.2. ■

9. Constrained Optimization

In this part of the lecture, we consider general nonlinear programs of the form

$$\min_{x\in\mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad g(x) \leq 0, \quad h(x) = 0, \tag{9.1}$$

where the functions $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$, and $h : \mathbb{R}^n \to \mathbb{R}^p$ are assumed to be continuously differentiable. We start by repeating several definitions and terminologies that were introduced during our first lecture.

Definition 9.1: Feasible Set and Active Constraints
The set

$$X := \{x \in \mathbb{R}^n : g(x) \leq 0,\, h(x) = 0\}, \tag{9.2}$$

is called **feasible set** of the problem (9.1). A point $x \in \mathbb{R}^n$ is called **feasible** if we have $x \in X$. For a feasible point $x \in X$, we define the **index set of the active constraints** $\mathcal{A}(x)$ and the **index set of the inactive constraints** $\mathcal{I}(x)$ as follows:

$$\begin{aligned}\mathcal{A}(x) &:= \{i \in \{1,...,m\} : g_i(x) = 0\},\\ \mathcal{I}(x) &:= \{1,...,m\}\backslash\mathcal{A}(x) = \{i \in \{1,...,m\} : g_i(x) < 0\}.\end{aligned}$$

We refer to Definition 2.1 for a definition of local and global minima of constrained optimization problems. We now briefly mention related subclasses of problem (9.1).

- In the case $m = 0$, we do not have any inequality constraints and problem (9.1) is called an *equality constrained optimization problem*.
- If all functions are (affine)-linear, i.e., $f(x) = c^\top x$, $g(x) = Ax - b$, and $h(x) = Bx - d$, then we obtain a *linear optimization problem*.
- If f is a quadratic function, i.e.,

  $$f(x) = \frac{1}{2}x^\top Cx + c^\top x, \quad c \in \mathbb{R}^n, \quad C \in \mathbb{R}^{n\times n} \text{ symmetric}$$

  and if g and h are affine-linear, then problem (9.1) is *quadratic optimization problem*.
- If the functions f and g_i, $i = 1,...,m$ are convex and if h is affine-linear, then problem (9.1) is a *convex optimization problem*. (Compare this with our results and definitions in section 4).

9.1. First-Order Necessary Conditions

In this section, we develop first-order optimality conditions for the constrained minimization problem (9.1).

Let $x^* \in X$ be a local solution of (9.1). We are now trying to derive conditions that x^* necessarily has to satisfy. In contrast to the unconstrained setting, we also need to take the

geometry of the set X into account. In particular, we have

$$f(x^* + td) \geq f(x^*) \quad \text{for all } \; x^* + td \; \text{ such that } \; x^* + td \in X$$

and td is sufficiently small. This motivates the definition of *set of feasible directions* or the so-called *radial cone* of X at x:

$$\mathcal{R}_X(x) := \{d \in \mathbb{R}^n : \exists \; t_* > 0 \text{ such that } \; x + td \in X \; \text{ for all } \; t \in [0, t_*]\}.$$

Unfortunately, the radial cone can often be too small as the following example illustrates.

Example 9.2. Consider the spherical constraints $X := \{x : \|x\| = 1\}$. Then, for every $x \in X$, we have $d \in \mathcal{R}_X(x)$ if and only if there exists $t_* > 0$ such that

$$1 = \|x + td\|^2 = 1 + 2tx^\top d + t^2\|d\|^2 \quad \iff \quad t(2x^\top d + t\|d\|^2) = 0$$

for all $t \in [0, t_*]$. This can only hold in the case $d = 0$, i.e., it follows $\mathcal{R}_X(x) = \{0\}$.

Next, we introduce a different set of directions, the so-called tangent cone.

Definition 9.3: Tangent Cone

Let $X \subset \mathbb{R}^n$ be a nonempty set. The **tangent cone** (or **Bouligand cone**) of X at a point $x \in X$ is given by

$$T_X(x) := \{d \in \mathbb{R}^n : \exists \; (\eta_k)_k \in \mathbb{R}_{++}, (x^k)_k \in X \; \text{ such that } \; x^k \to x, \;\; \eta_k(x^k - x) \to d\}.$$

Remark 9.4. Let us note that a set $K \subset \mathbb{R}^n$ is called **cone** if $\lambda x \in K$ for all $\lambda > 0$ and $x \in K$. If X is nonempty, then the tangent cone $T_X(x)$ can be shown to be a closed and nonempty cone. Moreover, if X is additionally convex, then $T_X(x)$ can also be represented as follows:

$$T_X(x) = \mathrm{cl}(\{d \in \mathbb{R}^n : \exists \; \eta > 0, \;\; y \in X \; \text{ such that } \; d = \eta(y - x)\}),$$

where $\mathrm{cl}(K)$ denotes the closure of a set K.

Example 9.5. Let us reconsider the spherical constraints $X := \{x : \|x\| = 1\}$. Then, for every $x \in X$, we claim that $T_X(x) = \{d \in \mathbb{R}^n : x^\top d = 0\}$. Let $d \in T_X(x)$ and let $(x^k)_k \subset X$ and $(\eta_k)_k \subset \mathbb{R}_{++}$ be given with $x^k \to x$ and $d^k := \eta_k(x^k - x) \to d$. It follows

$$x^\top d = \lim_{k\to\infty} \eta_k(x^\top x^k - \|x\|^2) = \lim_{k\to\infty} -\frac{\eta_k}{2}\|x^k - x\|^2 = \lim_{k\to\infty} \|d^k\|\|x^k - x\| = 0.$$

This implies $T_X(x) \subseteq \{d : x^\top d = 0\}$. On the other hand, if d satisfies $x^\top d = 0$, then define

$$x^k := \frac{x + t_k d}{\|x + t_k d\|} = \frac{1}{\|x + t_k d\|}x + \frac{t_k}{\|x + t_k d\|}d, \quad (t_k)_k \subset \mathbb{R}_{++}, \quad t_k \to 0.$$

Then, we have $x^k \in X$ for all k and $x^k \to x$ as $k \to \infty$. Setting $\eta_k = t_k^{-1}$, we obtain

$$\begin{aligned}\eta_k(x^k - x) &= \frac{1 - \|x + t_k d\|}{t_k \|x + t_k d\|} x + \frac{d}{\|x + t_k d\|} = \frac{1 - (\|x\|^2 + 2t_k x^\top d + t_k^2 \|d\|^2)}{t_k \|x + t_k d\|(1 + \|x + t_k d\|)} x + \frac{d}{\|x + t_k d\|} \\ &= \frac{-t_k \|d\|^2}{\|x + t_k d\|(1 + \|x + t_k d\|)} x + \frac{d}{\|x + t_k d\|} \to d.\end{aligned}$$

This shows $\{d : x^\top d = 0\} \subseteq T_X(x)$ and verifies our claim.

We can now formulate a first necessary optimality condition.

Theorem 9.6: First-Order Optimality via Tangent Cones

Let x^* be a local solution of problem (9.1). Then, it holds that $x^* \in X$ and

$$\nabla f(x^*)^\top d \geq 0 \quad \forall\, d \in T_X(x^*).$$

Proof. The feasibility of x^* is obvious. Let $d \in T_X(x^*)$ be given. Then, using the representation in Definition 9.3, we have

$$X \ni x^k \to x^*, \quad \eta_k > 0, \quad d^k := \eta_k(x^k - x^*) \to d.$$

In the case $d = 0$, the condition $\nabla f(x^*)^\top d \geq 0$ is clearly satisfied. Hence, let us assume $d \neq 0$. Since x^* is a local minimum of f on X, we obtain $f(x^k) \geq f(x^*)$ for all k sufficiently large and $x^k \neq x^*$ (since $d \neq 0$). Applying Taylor's theorem, this yields

$$\begin{aligned}0 \leq \eta_k(f(x^k) - f(x^*)) &= \eta_k \nabla f(x^*)^\top (x^k - x^*) + \eta_k o(\|x^k - x^*\|) \\ &= \nabla f(x^*)^\top d^k + \|d^k\| \cdot \frac{o(\|x^k - x^*\|)}{\|x^k - x^*\|} \to \nabla f(x^*)^\top d\end{aligned}$$

which finishes the proof. ■

This type of necessary optimality condition is very general but also cumbersome to apply (as we require a full characterization of the tangent cone $T_X(x)$ for all $x \in X$ and as was demonstrated in the derivations in Example 9.5). A criterion that is solely based on f, g, and h and their derivatives would be preferable. Using a Taylor expansion, we can locally approximate the feasible set X around some $x \in X$ via linearization.

Definition 9.7: Linearized Tangent Cone

The set

$$T_\ell(g, h, x) := \{d \in \mathbb{R}^n : \nabla g_i(x)^\top d \leq 0 \;\; \forall\, i \in \mathcal{A}(x), \;\; \nabla h(x)^\top d = 0\}$$

is called the **linearized tangent cone** at $x \in X$ (given the representation (9.2) of X).

Remark 9.8. The linearized tangent cone can also be interpreted as follows: If we linearize all constraints in the definition of the feasible set X at a point $\bar{x} \in X$, then we obtain the

polyhedral set

$$X_\ell(\bar{x}) = \{x : g(\bar{x}) + \nabla g(\bar{x})^\top (x - \bar{x}) \leq 0, \;\; h(\bar{x}) + \nabla h(\bar{x})^\top (x - \bar{x}) = 0\}.$$

It is now possible to show $T_{X_\ell(\bar{x})}(\bar{x}) = T_\ell(g, h, \bar{x})$ (exercise).

We further notice that the tangent cone $T_X(\bar{x})$ only depends on the set X while $T_\ell(g, h, \bar{x})$ depends on the specific representation of X utilizing the constraint functions g and h. This representation does not need to be unique!

While the condition $d \in T_X(x)$ is generally hard to verify, it is fairly easy to check if a direction d is contained in the linearized tangent cone $T_\ell(g, h, \bar{x})$ or not. In order to derive simpler optimality conditions that can also be used algorithmically, our idea is substitute the more complicated cone $T_X(x)$ by its linearized and simpler version $T_\ell(g, h, x)$. We first show that such a step does not yield weaker optimality conditions in Theorem 9.6, since $T_X(x)$ is always contained in $T_\ell(g, h, x)$.

Lemma 9.9
For all $x \in X$ it holds that $T_X(x) \subseteq T_\ell(g, h, x)$.

Proof. Let us set $d = \lim_{k\to\infty} d^k$ where $d^k = \eta_k(x^k - x)$, $\eta_k > 0$ and $x^k \in X$ with $x^k \to x$. Then, applying a Taylor expansion, we obtain

$$\begin{aligned} 0 \geq \eta_k(g_i(x^k) - g_i(x)) &= \nabla g_i(x)^\top d^k + \eta_k o(\|x^k - x\|) \quad \forall\, i \in \mathcal{A}(x) \\ 0 = \eta_k(h_j(x^k) - h_j(x)) &= \nabla h_j(x)^\top d^k + \eta_k o(\|x^k - x\|) \quad \forall\, j \in \{1, ..., p\} \end{aligned}$$

which, by taking the limit $k \to \infty$, implies $\nabla g_i(x)^\top d \leq 0$ for all $i \in \mathcal{A}(x)$ and $\nabla h(x)^\top d = 0$. As consequence, every $d \in T_X(x)$ satisfies $d \in T_\ell(g, h, x)$. ■

The reverse inclusion generally does not hold (see exercises).

Definition 9.10: Abadie Constraint Qualification
The condition

(ACQ) $$T_\ell(g, h, x) = T_X(x)$$

is called **Abadie Constraint Qualification** (ACQ) for $x \in X$.

Under the (ACQ) and using Theorem 9.6, we immediately obtain the following corollary.

Corollary 9.11: First-Order Optimality under the ACQ
Let x^* be a local solution of problem (9.1) and assume that the Abadie Constraint Qualification (ACQ) does hold at x^*. Then, we have $x^* \in X$ and

$$\nabla f(x^*)^\top d \geq 0 \quad \forall\, d \in T_\ell(g, h, x^*).$$

It is possible to slightly weaken the (ACQ) by considering polar cones. Let $K \subset \mathbb{R}^n$ be a nonempty cone. Then, the *polar cone* of K is defined via

$$K^\circ := \{v \in \mathbb{R}^n : v^\top d \leq 0 \ \ \forall\, d \in K\}.$$

The conditions in Theorem 9.6 and Corollary 9.11 are then equivalent to $-\nabla f(x^*) \in T_X(x^*)^\circ$ and $-\nabla f(x^*) \in T_\ell(g,h,x^*)^\circ$. Hence, Corollary 9.11 is still valid under the weaker condition $T_X(x^*)^\circ = T_\ell(g,h,x^*)^\circ$.

Definition 9.12: Guignard Constraint Qualification

The condition

$$\text{(GCQ)} \qquad T_X(x^*)^\circ = T_\ell(g,h,x^*)^\circ$$

is called **Guignard Constraint Qualification** (GCQ) for $x \in X$.

The next corollary is an immediate consequence of our last considerations.

Corollary 9.13: First-Order Optimality under the GCQ

Let x^* be a local solution of problem (9.1) and assume that the Guignard Constraint Qualification (GCQ) does hold at x^*. Then, we have $x^* \in X$ and

$$\nabla f(x^*)^\top d \geq 0 \quad \forall\, d \in T_\ell(g,h,x^*).$$

We will see that the conditions (ACQ) and (GCQ) are generally only violated in rare cases and constructed examples. Furthermore, in the next section, we will discuss and study simpler constraint qualifications that imply the (GCQ).

Definition 9.14: Constraint Qualification

Let $x \in X$ be given. A condition that implies the (GCQ) is called **Constraint Qualification** (CQ) for x.

In particular, the (ACQ) is a constraint qualification. The simple structure of the linearized tangent cone $T_\ell(g,h,x)$ allows to apply the following important result for linear inequalities.

Lemma 9.15: Farkas' Lemma

Let $A \in \mathbb{R}^{n\times m}$, $B \in \mathbb{R}^{n\times p}$, and $c \in \mathbb{R}^n$ be given. Then, the following two statements are equivalent:

(i) For all $d \in \mathbb{R}^n$ with $A^\top d \leq 0$ and $B^\top d = 0$ it follows $c^\top d \leq 0$.

(ii) There exist $u \in \mathbb{R}^m$, $u \geq 0$, and $v \in \mathbb{R}^p$ such that $c = Au + Bv$.

This results remains valid in the cases $m = 0$ ($A = 0$) and $p = 0$ ($B = 0$).

We will prove this result at a later point.

Remark 9.16. Farkas' Lemma can also be reinterpreted. The first condition states that the objective function of the minimization problem

$$\min_{d\in\mathbb{R}^n} \; -c^\top d \quad \text{s.t.} \quad A^\top d \leq 0, \quad B^\top d = 0$$

is zero. By Farkas' Lemma this is equivalent to saying that the optimization problem

$$\max_{u\in\mathbb{R}^m, v\in\mathbb{R}^p} \; 0 \quad \text{s.t.} \quad u \geq 0, \quad c = Au + Bv$$

has a nonempty feasible set. (In the case of maximization, the optimal value would be $-\infty$ if the feasible set is empty). Consequently, the optimal objective function of those two problem coincides and is zero. This is a very first form of *duality* that we will explore in more detail later.

We can now derive the standard first-order optimality conditions for the constrained problem (9.1).

Theorem 9.17: Karush-Kuhn-Tucker (KKT) Conditions

Let x^* be a local solution of (9.1) and assume that a Constraint Qualification holds at x^*. Then, the **Karush-Kuhn-Tucker conditions** (KKT conditions) are satisfied:

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

(i) $\nabla f(x^*) + \nabla g(x^*)\lambda^* + \nabla h(x^*)\mu^* = 0$ (*multiplier rule*).

(ii) $h(x^*) = 0$

(iii) $\lambda^* \geq 0, \quad g(x^*) \leq 0, \quad (\lambda^*)^\top g(x^*) = 0$ (*complementarity conditions*).

Proof. Let x^* be a local solution of problem (9.1) at which a Constraint Qualification is satisfied. By Corollary 9.13, we then have

$$-\nabla f(x^*)^\top d \leq 0 \quad \forall\, d \in \mathbb{R}^n \quad \text{with} \quad \nabla g_i(x^*)^\top d \leq 0, \;\; i \in \mathcal{A}(x^*), \;\; \nabla h(x^*)^\top d = 0.$$

Setting $c = -\nabla f(x^*)$, $A = \nabla g_{\mathcal{A}(x^*)}(x^*) = (\nabla g_i(x^*))_{i\in\mathcal{A}(x^*)}$ and $B = \nabla h(x^*)$, Farkas' Lemma guarantees the existence of $u \in \mathbb{R}^{|\mathcal{A}(x^*)|}$ and $v \in \mathbb{R}^p$ such that

$$u \geq 0 \quad \text{and} \quad c = Au + Bv.$$

We can now set $\lambda^* \in \mathbb{R}^m$, $\lambda^*_{\mathcal{A}(x^*)} = u$, $\lambda^*_{\mathcal{I}(x^*)} = 0$, and $\mu^* = v$. This implies the multiplier rule stated in (i). The choice of λ^* also implies that the complementarity condition in (iii) is satisfied. ∎

The complementarity conditions can also be expressed slightly differently. We summarize some of our observations in the next remark.

Remark 9.18. The complementarity condition can be substituted by one of the following reformulations:

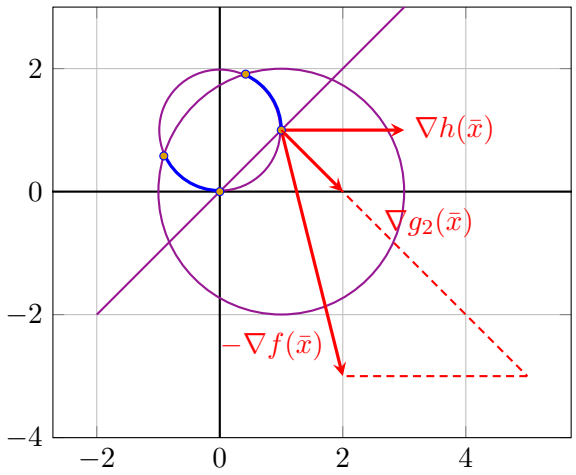


Figure 9.1: Visualization of the multiplier rule and the KKT conditions for the optimization problem in Example 9.20. The feasible set is shown using a blue color.

(iii′) $\lambda_i^* \geq 0, \quad g_i(x^*) \leq 0, \quad \lambda_i^* g_i(x^*) = 0 \quad$ for all $i = 1, ..., m.$

(iii″) $g(x^*) \leq 0, \quad \lambda_i^* \geq 0 \quad$ for all $i \in \mathcal{A}(x^*), \quad \lambda_i^* = 0 \quad$ for all $i \in \mathcal{I}(x^*).$

The complementarity condition ensures that at least one of the numbers λ_i^* and $g_i(x^*)$ is zero. As a consequence, the Lagrange multipliers associated with the inactive inequality constraints need to vanish. We continue with several definitions.

Definition 9.19: KKT-Points and the Lagrange Function

If the triple $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ satisfies the KKT conditions, then we call x^* a **KKT-point** of problem (9.1) and (x^*, λ^*, μ^*) a **KKT-triple** of (9.1). Furthermore, the function $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$,

$$L(x, \lambda, \mu) := f(x) + \lambda^\top g(x) + \mu^\top h(x)$$

is called **Lagrange function** for problem (9.1).

Using the Lagrange function, we can express the multiplier condition (i) in Theorem 9.17 compactly as $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$. In addition, if (x^*, λ^*, μ^*) is a KKT-triple for (9.1), we say that the *strict complementarity condition* is satisfied if

$$\lambda_i^* > 0 \quad \forall\, i \in \mathcal{A}(x^*).$$

We continue with an example to illustrate the application of the KKT-conditions.

Example 9.20. Consider the problem

$$\min_x \; x_1^3 x_2 - 2x_1^2 + 3x_2 \quad \text{s.t.} \quad g(x) \leq 0, \quad h(x) = 0,$$

where $g_1(x) = (x_1 - 1)^2 + x_2^2 - 4$

$$g_2(x) = x_1 - x_2, \quad h(x) = x_1^2 + (x_2 - 1)^2 - 1.$$

It holds that $\nabla f(x) = (3x_1^2 x_2 - 4x_1, x_1^3 + 3)^\top$ and

$$\nabla g_1(x) = (2(x_1 - 1), 2x_2)^\top, \quad \nabla g_2(x) = (1, -1)^\top, \quad \nabla h(x) = (2x_1, 2(x_2 - 1))^\top.$$

At $\bar{x} = (1, 1)^\top$, we have $g_1(\bar{x}) - 3 < 0$, $g_2(\bar{x}) = 0$, and $h(\bar{x}) = 0$. Hence, we obtain $\mathcal{A}(\bar{x}) = \{2\}$ and

$$\nabla f(\bar{x}) = (-1, 4)^\top, \quad \nabla g_2(\bar{x}) = (1, -1)^\top, \quad \nabla h(\bar{x}) = (2, 0)^\top.$$

Due to $\mathcal{I}(\bar{x}) = \{1\}$, we can set $\lambda_1 = 0$ and the KKT-conditions are satisfied if we can find $\lambda_2 \geq 0$ and $\mu \in \mathbb{R}$ such that $-\nabla f(\bar{x}) = \nabla g_2(\bar{x})\lambda_2 + \nabla h(\bar{x})\mu$. This is true for $\lambda_2 = 4$ and $\mu = 1.5$. Since the complementarity conditions are clearly satisfied, $\bar{x}$ is a KKT point. The multiplier rule is visualized in Figure 9.1.

9.2. Constraint Qualifications

In this section, we derive the most important constraint qualifications. We start with one of the simplest CQs.

Theorem 9.21: Concavity and Linearity

The following condition is a constraint qualification for $x \in X$:

$$g_i \quad \text{is concave}, \quad i \in \mathcal{A}(x), \quad h \quad \text{is affine-linear}.$$

Proof. Suppose that the condition in Theorem 9.21 holds at $x \in X$ and let $d \in T_\ell(g, h, x)$ be arbitrary. We now define $(t_k)_k$ via $t_k := \tau/k$. If τ is sufficiently small, then due to the continuity of g_i, it follows

$$g_i(x + t_k d) \leq 0 \quad \forall\, i \in \mathcal{I}(x), \quad k \geq 1.$$

Moreover, using the concavity of g_i and the definition of $T_\ell(g, h, x)$, we obtain

$$g_i(x + t_k d) \leq g_i(x) + t_k \nabla g_i(x)^\top d = t_k \nabla g_i(x)^\top d \leq 0.$$

Similarly, due to the linearity of h, we have

$$h(x + t_k d) = h(x) + t_k \nabla h(x)^\top d = t_k \nabla h(x)^\top d = 0.$$

This shows $x^k := x + t_k d \in X$ for all $k \geq 1$ and setting $\eta_k = 1/t_k$, this implies $d \in T_X(x)$. ■

We continue with one of the most important constraint qualifications.

Definition 9.22: Mangasarian-Fromovitz Constraint Qualification

The **Mangasarian-Fromovitz Constraint Qualification (MFCQ)** holds at $x \in X$ if

(i) $\nabla h(x)$ has full column rank or h is affine-linear.

(ii) There exists $d \in \mathbb{R}^n$ such that

$$\nabla g_i(x)^\top d < 0, \quad i \in \mathcal{A}(x), \quad \nabla h(x)^\top d = 0.$$

In the case $m = 0$ or $\mathcal{A}(x) = \emptyset$, then the condition (ii) can be omitted. In the case $p = 0$, the conditions (i) and $\nabla h(x)^\top d = 0$ are not required.

Theorem 9.23

Let the (MFCQ) be satisfied at $x \in X$, then the (ACQ) holds at x, i.e., the Mangasarian-Fromovitz Constraint Qualification is a constraint qualification for x.

Proof. As before, the easiest solution would be to consider $x^k := x + t_k v$ for some $v \in T_\ell(g, h, x)$ and a sequence $(t_k)_k$ with $t_k \to 0$. However, such a construction might violate the active inequality constraints and the equality constraints. Hence, we first try to find a feasible path $\bar{x}(t) \in X$, t sufficiently small, along the direction d satisfying condition (ii) in the formulation of the (MFCQ).

The path should satisfy the following properties: there is $\varepsilon > 0$ such that $\bar{x} : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ is continuously differentiable and we have $\bar{x}(t) \in X$ for all $t \in [0, \varepsilon)$, $\bar{x}(0) = x$, and $\bar{x}'(0) = d$.

Case 1: $\nabla h(x)$ has full column rank. Let us define the auxiliary mapping $H : \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}^p$ via

$$H(t, u) := h(x + td + \nabla h(x)u).$$

The nonlinear equation $H(t, u) = 0$ has the solution $(\bar{t}, \bar{u}) = (0, 0)$ and the partial derivative

$$\frac{\partial}{\partial u} H(0, 0) = \nabla h(x)^\top \nabla h(x)$$

is invertible (even positive definite) due to the linear independence of the vectors $\nabla h_j(x)$, $j = 1, ..., p$. We can now apply the implicit function theorem, which ensures the existence of $\varepsilon > 0$ and a C^1-function $u : (-\varepsilon, \varepsilon) \to \mathbb{R}^p$ such that $u(0) = 0$, $H(t, u(t)) = h(x+td+\nabla h(x)u(t)) = 0$ and

$$u'(t) = -\left[\frac{\partial}{\partial u} H(t, u(t))\right]^{-1} \frac{\partial}{\partial t} H(t, u(t))$$

for all $t \in (-\varepsilon, \varepsilon)$. Hence, we obtain

$$u'(0) = -\left[\frac{\partial}{\partial u} H(0, 0)\right]^{-1} \frac{\partial}{\partial t} H(0, 0) = -[\nabla h(x)^\top \nabla h(x)]^{-1} \nabla h(x)^\top d = 0.$$

Case 2: h is affine-linear. Since h is affine-linear, it follows $h(x+td) = h(x)+t\nabla h(x)^\top d = 0$ for all $t \in \mathbb{R}$. Thus, in this case, we can just set $u \equiv 0$.

Combining the two cases, we can define $\bar{x}(t) = x + td + \nabla h(x)u(t)$ for all $t \in (-\varepsilon, \varepsilon)$. By

reducing ε if necessary, the path $\bar{x}$ has all the desired properties. Obviously, we have $\bar{x} \in C^1$, $\bar{x}(0) = x$, $\bar{x}'(0) = d$, and $h(\bar{x}(t)) = 0$ for all $t \in (-\varepsilon, \varepsilon)$. Moreover, using the continuity of g, we have $g_i(\bar{x}(t)) < 0$ for all $i \in \mathcal{I}(x)$ and $|t|$ sufficiently small. For the active indices $i \in \mathcal{A}(x)$, it follows $g_i(\bar{x}(0)) = g_i(x) = 0$ and

$$\frac{\mathrm{d}}{\mathrm{d}t} g_i(\bar{x}(t))\Big|_{t=0} = \nabla g_i(\bar{x}(0))^\top \bar{x}'(0) = \nabla g_i(x)^\top d < 0.$$

Hence, we have $g_i(\bar{x}(t)) < 0$ for all $t > 0$ sufficiently small.

By Lemma 9.9, it suffices to show $T_\ell(g, h, x) \subseteq T_X(x)$. Consequently, let $v \in T_\ell(g, h, x)$ be arbitrary. Setting $v_\delta := v + \delta d$, we obtain

$$\nabla g_i(x)^\top v_\delta = \nabla g_i(x)^\top v + \delta \nabla g_i(x)^\top d < 0, \quad \forall\, i \in \mathcal{A}(x), \quad \nabla h(x)^\top v_\delta = 0$$

for all $\delta > 0$. This implies that v_δ satisfies condition (ii) of the (MFCQ) for all $\delta > 0$ and according to the first part of our proof there exist $\varepsilon = \varepsilon_\delta > 0$ and a C^1-path $\bar{x}_\delta : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ such that $\bar{x}_\delta(t) \in X$ for all $t \in [0, \varepsilon)$, $\bar{x}_\delta(0) = x$, and $\bar{x}_\delta'(0) = v_\delta$. For any arbitrary sequence $(t_k)_k$ with $t_k \downarrow 0$, we now define $x^k := \bar{x}_\delta(t_k)$. This yields $x^k \in X$ for all k sufficiently large, $x^k \to x$, and

$$v_\delta = \bar{x}_\delta'(0) = \lim_{k\to\infty} \frac{\bar{x}_\delta(t_k) - x}{t_k} = \lim_{k\to\infty} \frac{x^k - x}{t_k} \in T_X(x).$$

Since the tangent cone $T_X(x)$ is closed, this implies $v = \lim_{\delta \downarrow 0} v_\delta \in T_X(x)$. ■

Remark 9.24. The implicit function theorem used in the proof of Theorem 9.23 is a standard tool to study solvability properties of nonlinear equations. Let $G : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be a continuously differentiable function and let $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ be a point with $G(\bar{x}, \bar{y}) = 0$. Our goal is to express all y with $G(x, y) = 0$ locally as a function g of x such that $G(x, g(x)) = 0$. Suppose that the Jacobian

$$D_y G(\bar{x}, \bar{y}) = \left[\frac{\partial G_i}{\partial y_j}(\bar{x}, \bar{y})\right] \in \mathbb{R}^{m\times m}$$

is invertible. Then, the implicit function theorem guarantees the existence of an open set $U \subset \mathbb{R}^n$, $\bar{x} \in U$, and a continuously differentiable function $g : U \to \mathbb{R}^m$ such that $g(\bar{x}) = \bar{y}$ and $G(x, g(x)) = 0$ for all $x \in U$. Moreover, it holds that

$$Dg(x) = -\left[\frac{\partial G_i}{\partial y_j}(x, g(x))\right]^{-1} \left[\frac{\partial G_i}{\partial x_j}(x, g(x))\right] = -D_y G(x, g(x))^{-1} D_x G(x, g(x)).$$

We illustrate the application of the implicit function theorem with an example. Consider the equation

$$G(x, y, z) = xy + xz\log(yz) - 1 = 0.$$

A solution of this nonlinear equation is given by $(1, 1, 1)$ and we now want to implicitly write z as a function $g(x, y)$ for (x, y) near $(1, 1)$. Due to $\frac{\partial}{\partial z} G(x, y, z) = x\log(yz) + x$ and $\frac{\partial}{\partial z} G(1, 1, 1) = 1$, the implicit function theorem is applicable. Hence, there exist an open set

$U \in \mathbb{R}^2$, $(1,1) \in U$, and a function $g : U \to \mathbb{R}$ such that $g(1,1) = 1$ and $G(x,y,g(x,y)) = 0$ for all $(x,y) \in U$. We can also calculate the derivative of g:

$$\begin{aligned}\nabla g(x,y) &= -\left[\tfrac{\partial}{\partial x}G(x,y,g(x,y)) \quad \tfrac{\partial}{\partial y}G(x,y,g(x,y))\right]^\top \left[\tfrac{\partial}{\partial z}G(x,y,g(x,y))\right]^{-1} \\ &= -\frac{1}{x\log(yg(x,y)) + x}\begin{bmatrix} y + g(x,y)\log(yg(x,y)) \\ x + \frac{x}{y}\cdot g(x,y)\end{bmatrix}.\end{aligned}$$

Specifically, at $x = y = 1$, this yields $\nabla g(1,1) = (-1,-2)^\top$.

We often work with a slightly stronger version of the Mangasarian-Fromovitz constraint qualification – the so-called regularity condition.

Definition 9.25: Regularity and the LICQ

The point $x \in X$ is called **regular** if the columns of the matrix

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

are linearly independent. Furthermore, we say that the **Linear Independence Constraint Qualification (LICQ)** is satisfied at a regular point $x \in X$.

Lemma 9.26

Let the (LICQ) hold at $x \in X$, then the (MFCQ) is satisfied, i.e., the Linear Independence Constraint Qualification is a constraint qualification for x.

Proof. Obviously, $\nabla h(x)$ has full column rank. Since the matrix

$$\begin{pmatrix} \nabla g_{\mathcal{A}(x)}(x)^\top \\ \nabla h(x)^\top \end{pmatrix} \in \mathbb{R}^{(|\mathcal{A}(x)|+p)\times n}$$

has full row rank, we can add more rows to obtain a non-singular square matrix $H(x) \in \mathbb{R}^{n\times n}$. Consequently, the linear system of equations

$$H(x)d = \begin{pmatrix} -\mathbb{1} \\ 0 \end{pmatrix}, \quad \mathbb{1} \in \mathbb{R}^{|\mathcal{A}(x)|}$$

has a solution which satisfies the requirements of the (MFCQ). ■

9.3. Karush-Kuhn-Tucker Conditions and Convexity

In the following, we suppose that the optimization problem (9.1) is convex, i.e., the functions f, g_i, $i = 1,...,m$ are convex and h is an affine-linear mapping. Similar to the unconstrained setting, the KKT-conditions share a remarkable property: if the problem (9.1) is convex, then the KKT-conditions ensure local optimality.

Theorem 9.27: KKT-Conditions and Convexity

Suppose that the problem (9.1) is convex. Then, every local solution of (9.1) is a global solution. Let $x^* \in X$ be a global solution and assume that a constraint qualification holds at x^*, then the KKT-conditions are satisfied. Conversely, if x^* is a KKT-point, then x^* is a global solution of problem (9.1).

Proof. The first two claims directly follow from Theorem 4.19 and Theorem 9.17. Let x^* be a KKT-point with multiplier λ^* and μ^* and let $x \in X$ be arbitrary. Setting $d = x - x^*$, the convexity of g_i and the complementarity conditions imply

$$\lambda_i^* \cdot \nabla g_i(x^*)^\top d \leq \lambda_i^*(g_i(x) - g_i(x^*)) = \lambda_i^* g_i(x) \leq 0.$$

Furthermore, since h is affine-linear, we have $\nabla h(x^*)\top d = h(x) - h(x^*) = 0$. We now apply the convexity of f and the main condition:

$$f(x) - f(x^*) \geq \nabla f(x^*)^\top d = -(\lambda^*)^\top \nabla g(x^*)d - (\mu^*)^\top \nabla h(x^*)^\top d = -(\lambda^*)^\top \nabla g(x^*)d \geq 0.$$

Since $x \in X$ is arbitrary, this verifies that x^* is a global solution of (9.1). ■

9.4. Second-Order Optimality Conditions

In this section, we want to derive second-order optimality conditions that include the curvature information of the objective function and of the feasible set and constraints. Similar to the discussion in the unconstrained case, we can expect that certain definiteness properties need to be satisfied. We will see that in constrained problems, these conditions are formulated using the Hessian of the Lagrange function instead of the Hessian of f.

In order to motivate the second-order conditions, let us first assume that $x^* \in X$ is a feasible point satisfying

$$\nabla f(x^*)^\top d > 0 \quad \forall\, d \in T_\ell(g, h, x^*)\backslash\{0\}. \tag{9.3}$$

In particular, this implies that x^* is a KKT-point and there are Lagrange multiplier $\lambda^* \geq 0$ and μ^* such that (x^*, λ^*, μ^*) is a KKT-triple. We now argue that x^* is a strict local minimum. Suppose that this claim is wrong. Then there exists a sequence $(x^k)_k$, $x^k \in X$ for all k, with $x^k \to x$ and $f(x^k) \leq f(x^*)$ for all k. Defining $t_k := \|x^k - x\|$ and $d^k = t_k^{-1}(x^k - x)$, it follows $\|d^k\| = 1$ for all k. Hence, $(d^k)_k$ is bounded and there exists a subsequence $(d^{k_\ell})_\ell$ and $d \in \mathbb{R}^n$, $\|d\| = 1$ such that $d^{k_\ell} \to d$ as $\ell \to \infty$. In the following and without loss of generality, we assume that the full sequence $(d^k)_k$ converges to d (to simplify the derivation and notation). Using a Taylor expansion, we obtain

$$0 \geq g_i(x^k) = g_i(x^* + t_k d^k) - g_i(x^*) = t_k \nabla g_i(x^*)^\top d^k + o(t_k) \quad \forall\, i \in \mathcal{A}(x^*)$$

and

$$0 = h(x^k) = h(x^* + t_k d^k) - h(x^*) = t_k \nabla h(x^*)^\top d^k + o(t_k)$$

as $k \to \infty$. Dividing the last two results by t_k and taking the limit $k \to \infty$, we can infer $d \in T_\ell(g,h,x^*)\backslash\{0\}$. Similary, we now have

$$0 \geq f(x^k) - f(x^*) = f(x^* + t_k d^k) - f(x^*) = t_k \nabla f(x^*)^\top d^k + o(t_k)$$

and by taking the limit, it follows $\nabla f(x^*)^\top d \leq 0$ which is a contradiction to (9.3).

Condition (9.3) is a first-order sufficient optimality condition, which often might not be satisfied in practice. However, our derivation illustrates that second-order and curvature information is only required for directions $d \in T_\ell(g,h,x^*)$ with $\nabla f(x^*)^\top d = 0$. This motivates the definition of the so-called critical cone.

Definition 9.28: Critical Cone
Let $x \in X$ be given. The set

$$\mathcal{C}(x) := \{d \in T_\ell(g,h,x) : \nabla f(x)^\top d = 0\}$$

is called the **critical cone** at x.

We can now formulate the sufficient second-order optimality conditions.

Theorem 9.29: Second-Order Sufficient Optimality Conditions (SOSC)
Let x^* be a KKT-point with Lagrange multiplier $\lambda^* \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^p$ and suppose that the condition

$$d^\top \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) d > 0 \quad \forall\, d \in \mathcal{C}(x^*)\backslash\{0\}$$

is satisfied. Then, x^* is strict local minimizer of problem (9.1).

Proof. The proof is similar to our previous derivation. Specifically, assuming that x^* is not a strict local minimum, we can construct $(x^k)_k \subset X$ and $(d^k)_k \subset \mathbb{R}^n$ such that $x^k \to x^*$, $d^k \to d$, and $\|d\| = 1$. As before it follows

$$\nabla g_i(x^*)^\top d \leq 0, \quad \forall\, i \in \mathcal{A}(x^*), \quad \nabla h(x^*)^\top d = 0, \quad \nabla f(x^*)^\top d \leq 0.$$

Since x^* is a KKT-point, it holds that $\nabla f(x^*)^\top d \geq 0$ for all $d \in T_\ell(g,h,x^*)$ (see derivation of Theorem 9.17) and thus, we can infer $\nabla f(x^*)^\top d = 0$ and $d \in \mathcal{C}(x^*)\backslash\{0\}$. Next, we have

$$L(x^k, \lambda^*, \mu^*) = f(x^k) + (\lambda^*)^\top g(x^k) \leq f(x^k) \leq f(x^*) = L(x^*, \lambda^*, \mu^*)$$

and a second-order Taylor expansion and $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$ yield

$$0 \geq L(x^k, \lambda^*, \mu^*) - L(x^*, \lambda^*, \mu^*) = \frac{1}{2} t_k^2 (d^k)^\top \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) d^k + o(t_k^2).$$

Dividing both sides by t_k^2 and taking the limit $k \to \infty$, we obtain a contradiction to the condition in Theorem 9.29. ■

We next present an example that shows that the sufficient second-order optimality conditions can hold at a (global) solution x^* while the Hessian of f is indefinite on the cone

$\mathcal{C}(x^*)$. Since the Hessian of f typically can only represent the curvature information of the objective function f and not of the constraints, this is not surprising. On the other hand, the Hessian of the Lagrangian $\nabla^2_{xx}L(x^*,\lambda^*,\mu^*)$ contains this information in suitable way.

Example 9.30. Consider the optimization problem

$$\min_{x\in\mathbb{R}^2} \; f(x) = -x_1^2 + 2x_2 \quad \text{s.t.} \quad g(x) = x_1^2 - x_2 \leq 0. \tag{9.4}$$

Then, we have

$$\nabla f(x) = \begin{pmatrix} -2x_1 \\ 2 \end{pmatrix}, \;\; \nabla^2 f(x) = \begin{pmatrix} -2 & 0 \\ 0 & 0 \end{pmatrix}, \;\; \nabla g(x) = \begin{pmatrix} 2x_1 \\ -1 \end{pmatrix}, \;\; \nabla^2 g(x) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}.$$

We first want to determine all KKT-points of this problem. By the complementarity conditions, if $x_1^2 < x_2$, then we have $\lambda = 0$ and the main condition reduces to $\nabla f(x) = 0$. Since there is no feasible x satisfying this condition, we must have $x_1^2 = x_2$ and $\mathcal{A}(x) = \{1\}$. We obtain

$$\begin{pmatrix} -2x_1 \\ 2 \end{pmatrix} + \lambda \begin{pmatrix} 2x_1 \\ -1 \end{pmatrix} = 0, \quad x_2 = x_1^2,$$

and hence, the point $x_1^* = x_2^* = 0$ and $\lambda^* = 2$ is the only KKT-point of this problem. The critical cone is given by

$$\mathcal{C}(x^*) = \{d : 2d_2 = 0, \; -d_2 \leq 0\} = \{d \in \mathbb{R}^2 : d_2 = 0\}$$

and thus, we have

$$d^\top \nabla^2_{xx} L(x^*,\lambda^*) d = d^\top \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} d = 2d_1^2 > 0 \quad \forall \; d \in \mathcal{C}(x^*)\backslash\{0\}.$$

This shows that the sufficient second-order optimality conditions are satisfied although the Hessian $\nabla^2 f(x^*)$ is negative semidefinite. Moreover, using $x_1^2 \leq x_2$, we have $f(x) = -x_1^2 + 2x_2 \geq x_1^2 \geq 0 = f(x^*)$ which implies that x^* is the unique global solution of problem (9.4).

Under the (LICQ) we can establish the following second-order necessary optimality conditions.

Theorem 9.31: Second-Order Necessary Optimality Conditions (SONC)

Let f, g, and h be twice continuously differentiable and suppose that x^* is a local solution of problem (9.1) at which the (LICQ) is satisfied. Then there is a unique pair of Lagrange multiplier $(\lambda^*,\mu^*) \in \mathbb{R}^m \times \mathbb{R}^p$ such that

(i) $\nabla_x L(x^*,\lambda^*,\mu^*) = 0$ (*multiplier rule*).

(ii) $h(x^*) = 0$

(iii) $\lambda^* \geq 0, \quad g(x^*) \leq 0, \quad (\lambda^*)^\top g(x^*) = 0$ (*complementarity conditions*)

and we further have $d^\top \nabla^2_{xx} L(x^*,\lambda^*,\mu^*) d \geq 0$ for all $d \in \mathcal{C}(x^*)$.

Proof. Since the (LICQ) holds at x^*, we can apply Theorem 9.17 to show (i)–(iii). The uniqueness of the Lagrange multiplier is a consequence of the (LICQ) and is shown in one of the exercises. To prove the last statement, let $d \in \mathcal{C}(x^*)$ be arbitrary and let us define

$$H : \mathbb{R}^n \to \mathbb{R}^{(|\mathcal{A}(x^*)|+p)\times n}, \quad H(x) := \begin{bmatrix} \nabla g_{\mathcal{A}(x^*)}(x)^\top \\ \nabla h(x)^\top \end{bmatrix}.$$

The (LICQ) implies that the matrix $H(x^*)$ has full row rank $q := |\mathcal{A}(x^*)| + p$. As in the proof of Lemma 9.26, the full row rank of $H(x^*)$ also implies that there exists a matrix $Z \in \mathbb{R}^{n\times(n-q)}$ such that the matrix

$$\begin{bmatrix} H(x^*) \\ Z^\top \end{bmatrix} \in \mathbb{R}^{n\times n}$$

is invertible. We now construct the mapping $G : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ via

$$G(t,x) := \begin{bmatrix} g_{\mathcal{A}(x^*)}(x) - t \cdot \nabla g_{\mathcal{A}(x^*)}(x^*)^\top d \\ h(x) \\ Z^\top(x - x^* - td) \end{bmatrix}.$$

We have $G(0,x^*) = 0$ and $D_x G(0,x^*) = [\nabla g_{\mathcal{A}(x^*)}(x^*)\ \nabla h(x^*)\ Z]^\top = [H(x^*)^\top\ Z]^\top$. Consequently, $D_x G(0,x^*)$ is invertible and the implicit function is applicable. In particular, there exist an open interval $I \supset \{0\}$ and a C^2-function $x : I \to \mathbb{R}^n$ with $x(0) = x^*$, $G(t,x(t)) = 0$ for all $t \in I$, and

$$x'(t) = -D_x G(t,x(t))^{-1} D_t G(t,x(t)) \quad \forall\, t \in I.$$

For $t = 0$, the last expression simplifies to:

$$\begin{aligned} x'(0) &= -D_x G(0,x^*)^{-1} D_t G(0,x^*) \\ &= -\begin{bmatrix} H(x^*) \\ Z^\top \end{bmatrix}^{-1} \begin{bmatrix} -\nabla g_{\mathcal{A}(x^*)}(x^*)^\top d \\ 0 \\ -Z^\top d \end{bmatrix} = \begin{bmatrix} H(x^*) \\ Z^\top \end{bmatrix}^{-1} \begin{bmatrix} \nabla g_{\mathcal{A}(x^*)}(x^*)^\top \\ \nabla h(x^*)^\top \\ Z^\top \end{bmatrix} d = d, \end{aligned}$$

where we used $0 = \nabla h(x^*)^\top d$ ($d \in \mathcal{C}(x^*) \subset T_\ell(g,h,x^*)$). Due to $G(t,x(t)) = 0$ for all $t \in I$, we further obtain

$$h(x(t)) = 0 \quad \text{and} \quad g_{\mathcal{A}(x^*)}(x(t)) = t \cdot \nabla g_{\mathcal{A}(x^*)}(x^*)^\top d \le 0$$

for all $t \in I \cap \mathbb{R}_+$. Utilizing $g_{\mathcal{I}(x^*)}(x^*) < 0$, the continuity of g and $x(t)$, and $x(0) = x^*$, we also have $g_{\mathcal{I}(x^*)}(x(t)) < 0$ for all t sufficiently small. In summary, there exists $\tau > 0$ such that $x(t) \in X$ for all $t \in [0,\tau]$. Thus, by the local optimality of x^* we can infer $f(x(t)) \ge f(x^*)$ for all $t \ge 0$ sufficiently small. Moreover, we have

$$\begin{aligned} L(x(t),\lambda^*,\mu^*) &= f(x(t)) + (\lambda^*)^\top g(x(t)) + (\mu^*)^\top h(x(t)) \\ &= f(x(t)) + \sum_{i\in\mathcal{A}(x^*)} \lambda_i^* \cdot g_i(x(t)) = f(x(t)) + t \cdot \sum_{i\in\mathcal{A}(x^*)} \lambda_i^* \cdot \nabla g_i(x^*)^\top d \end{aligned}$$

and the KKT-conditions and $d \in \mathcal{C}(x^*)$ imply

$$\begin{aligned} 0 &= \nabla_x L(x^*, \lambda^*, \mu^*)^\top d \\ &= \nabla f(x^*)^\top d + (\lambda^*)^\top \nabla g(x^*)^\top d + (\mu^*)^\top \nabla h(x^*)^\top d = \sum_{i \in \mathcal{A}(x^*)} \lambda_i^* \cdot \nabla g_i(x^*)^\top d. \end{aligned}$$

Combing the last derivations, this yields $L(x(t), \lambda^*, \mu^*) = f(x(t))$ (for all $t \in [0, \tau]$). By the complementarity conditions, we also have $L(x^*, \lambda^*, \mu^*) = f(x^*)$. Utilizing $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$ and a second order Taylor expansion, we obtain

$$\begin{aligned} 0 \leq f(x(t)) - f(x^*) &= L(x(t), \lambda^*, \mu^*) - L(x^*, \lambda^*, \mu^*) \\ &= \frac{1}{2}(x(t) - x^*)^\top \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*)(x(t) - x^*) + o(\|x(t) - x^*\|^2) \end{aligned}$$

for all $t \geq 0$ sufficiently small. Furthermore, by the properties of the path $t \mapsto x(t)$, it holds that $x(t) = x(0) + x'(0) \cdot t + o(t) = x^* + t \cdot d + o(t)$ for t sufficiently small. In particular, this implies $x(t) - x^* = t \cdot d + o(t)$, $o(\|x(t) - x^*\|^2) = o(t^2)$, and

$$0 \leq \frac{t^2}{2} \cdot d^\top \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) d + o(t^2).$$

Dividing both sides by t^2 and taking the limit $t \downarrow 0$, it follows $d^\top \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) d \geq 0$. Since $d \in \mathcal{C}(x^*)$ was chosen arbitrarily, this finishes the proof. ■

Using duality theory (which we will discuss in the next section), the necessary conditions in Theorem 9.31 can also be further strengthened. Suppose that x^* is a local solution of problem (9.1) at which the (MFCQ) is satisfied and set

$$\mathcal{M}(x^*) := \{(\lambda, \mu) \in \mathbb{R}^m_+ \times \mathbb{R}^p : \text{the triple } (x^*, \lambda, \mu) \text{ satisfies the KKT-conditions for (9.1)}\}.$$

Then, the following more general second-order necessary condition can be shown to hold:

$$\max_{(\lambda,\mu) \in \mathcal{M}(x^*)} d^\top \nabla_{xx}^2 L(x^*, \lambda, \mu) d \geq 0 \quad \forall\, d \in \mathcal{C}(x^*).$$

References. This section mainly follows [9, Section 16]. The proof of Theorem 9.23 is taken from [7]. The proof of Theorem 9.31 is taken from [8, Lemma 12.2 and Theorem 12.5]. We also refer to [5, Chapter 4] and [2, Chapter 11].

10. Duality Theory

We consider the nonlinear program

$$\min_{x\in\mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad g(x) \leq 0, \quad h(x) = 0 \tag{10.1}$$

with the associated Lagrange function $L(x,\lambda,\mu) = f(x) + \lambda^\top g(x) + \mu^\top h(x)$.

10.1. The Dual Problem

We want to derive a *dual problem* for (10.1) that can be equivalent to problem (10.1) in certain situations but which guarantees to give a lower bound for the optimal value of (10.1). The construction of the dual problem is based on the following observation:

$$p(x) := \sup_{\lambda\in\mathbb{R}^m_+,\,\mu\in\mathbb{R}^p} L(x,\lambda,\mu) = \begin{cases} f(x) & \text{if } x \in X, \\ +\infty & \text{otherwise.} \end{cases}$$

Hence, our *primal problem* (10.1) is equivalent to

$$\min_{x\in\mathbb{R}^n} \sup_{\lambda\in\mathbb{R}^m_+,\,\mu\in\mathbb{R}^p} L(x,\lambda,\mu). \tag{10.2}$$

The dual problem results from this problem formulation by exchanging "min" and "sup".

Definition 10.1: The Dual Problem
The problem

$$\sup_{\lambda\in\mathbb{R}^m_+,\,\mu\in\mathbb{R}^p} \inf_{x\in\mathbb{R}^n} \; L(x,\lambda,\mu) \tag{10.3}$$

is called the **dual problem** of (10.1) or (10.2). The function $d(\lambda,\mu) := \inf_{x\in\mathbb{R}^n} L(x,\lambda,\mu)$ is the **dual (objective) function** and $p(x) := \sup_{\lambda\in\mathbb{R}^m_+,\,\mu\in\mathbb{R}^p} L(x,\lambda,\mu)$ is called the **primal objective function**.

Remark 10.2. For every fixed $x \in \mathbb{R}^n$ the mapping $(\lambda,\mu) \mapsto L(x,\lambda,\mu)$ is a linear function. Hence, the dual function $d(\lambda,\mu)$ is concave since it is the infimum of linear functions. (This basically follows from Lemma 4.12). Consequently, the dual problem is a maximization problem of a concave function on a convex set – this is equivalent to a convex minimization problem!

Example 10.3. Let us consider the quadratic optimization problem

$$\min_x \; \frac{1}{2}x^\top Q x + c^\top x \quad \text{s.t.} \quad Ax \leq b, \quad Cx = d,$$

where $\mathbb{R}^{n\times n} \ni Q \succ 0$, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $C \in \mathbb{R}^{p\times n}$, $d \in \mathbb{R}^p$ are given. In the following, we calculate the dual of this problem. The associated Lagrange function is given

by $L(x,\lambda,\mu) := \frac{1}{2}x^\top Qx + c^\top x + \lambda^\top(Ax-b) + \mu^\top(Cx-d)$. Since $x \mapsto L(x,\lambda,\mu)$ is convex, we can derive the dual objective function via discussing the first-order optimality condition

$$0 = \nabla_x L(x,\lambda,\mu) = Qx + c + A^\top\lambda + C^\top\mu \quad \iff \quad x = -Q^{-1}[c + A^\top\lambda + C^\top\mu].$$

Consequently, we have

$$d(\lambda,\mu) = \inf_x \; L(x,\lambda,\mu) = -\frac{1}{2}[c + A^\top\lambda + C^\top\mu]^\top Q^{-1}[c + A^\top\lambda + C^\top\mu] - b^\top\lambda - d^\top\mu.$$

and the dual problem is given by:

$$\max_{\lambda,\,\mu} \; -\frac{1}{2}[c + A^\top\lambda + C^\top\mu]^\top Q^{-1}[c + A^\top\lambda + C^\top\mu] - b^\top\lambda - d^\top\mu \quad \text{s.t.} \quad \lambda \geq 0.$$

10.2. Weak Duality and Saddle Points of the Lagrange Function

We now show that the dual problem (10.3) always yields a lower bound to the optimal value of the primal problem (10.2).

Theorem 10.4: Weak Duality Theorem

Suppose that $\bar{x}$ and $(\bar{\lambda},\bar{\mu})$ are feasible points of the primal and dual problem, respectively. Then, it holds that

$$p(\bar{x}) = f(\bar{x}) \geq d(\bar{\lambda},\bar{\mu}).$$

Proof. Due to $\bar{\lambda} \geq 0$, $g(\bar{x}) \leq 0$, and $h(\bar{x}) = 0$, we have

$$d(\bar{\lambda},\bar{\mu}) = \inf_{x\in\mathbb{R}^n} L(x,\bar{\lambda},\bar{\mu}) \leq L(\bar{x},\bar{\lambda},\bar{\mu}) = f(\bar{x}) + \bar{\lambda}^\top g(\bar{x}) + \bar{\mu}^\top h(\bar{x}) \leq f(\bar{x}) = p(\bar{x})$$

as desired. ■

We will now show that the optimal objective function values of the primal and dual problem are equal, if the Lagrangian possesses a saddle point.

Definition 10.5: Saddle Points of the Lagrangian

A point $(x^*,\lambda^*,\mu^*) \in \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^p$ is called **saddle point** of the Lagrange function if

$$L(x^*,\lambda,\mu) \leq L(x^*,\lambda^*,\mu^*) \leq L(x,\lambda^*,\mu^*) \quad \forall\, x \in \mathbb{R}^n,\; \lambda \in \mathbb{R}^m_+,\; \mu \in \mathbb{R}^p.$$

Theorem 10.6: Saddle Points and Global Optimality

The following statements are equivalent:

(i) (x^*,λ^*,μ^*) is a saddle point of the Lagrange function.

(ii) x^* is a global minimizer of (10.1), (λ^*,μ^*) is a global maximizer of the dual problem (10.3) and we have $f(x^*) = d(\lambda^*,\mu^*)$.

Notice that for every function $L : \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^p \to \mathbb{R}$, it holds that

$$\sup_{\lambda\in\mathbb{R}^m_+,\,\mu\in\mathbb{R}^p} \inf_{x\in\mathbb{R}^n} L(x,\lambda,\mu) \leq \inf_{x\in\mathbb{R}^n} \sup_{\lambda\in\mathbb{R}^m_+,\,\mu\in\mathbb{R}^p} L(x,\lambda,\mu). \tag{10.4}$$

Indeed for any $y \in \mathbb{R}^n$, we have $\sup_{\lambda\in\mathbb{R}^m_+,\,\mu\in\mathbb{R}^p} \inf_{x\in\mathbb{R}^n} L(x,\lambda,\mu) \leq \sup_{\lambda\in\mathbb{R}^m_+,\,\mu\in\mathbb{R}^p} L(y,\lambda,\mu)$. Taking the infimum with respect to y verifies our general claim. We now turn to the proof of Theorem 10.6.

Proof. We first show (i) $\implies$ (ii). Using (10.4), we obtain

$$\begin{aligned} L(x^*,\lambda^*,\mu^*) &= \inf_{x\in\mathbb{R}^n} L(x,\lambda^*,\mu^*) \leq \sup_{\lambda\in\mathbb{R}^m_+,\,\mu\in\mathbb{R}^p} \inf_{x\in\mathbb{R}^n} L(x,\lambda,\mu) \\ &\leq \inf_{x\in\mathbb{R}^n} \sup_{\lambda\in\mathbb{R}^m_+,\,\mu\in\mathbb{R}^p} L(x,\lambda,\mu) \leq \sup_{\lambda\in\mathbb{R}^m_+,\,\mu\in\mathbb{R}^p} L(x^*,\lambda,\mu) = L(x^*,\lambda^*,\mu^*). \end{aligned}$$

Hence, it follows

$$L(x^*,\lambda^*,\mu^*) = \inf_{x\in\mathbb{R}^n} L(x,\lambda^*,\mu^*) = d(\lambda^*,\mu^*) = \sup_{\lambda\in\mathbb{R}^m_+,\,\mu\in\mathbb{R}^p} L(x^*,\lambda,\mu) = p(x^*) < \infty.$$

This shows that x^* is feasible and we have $f(x^*) = p(x^*)$. Moreover, due to $d(\lambda^*,\mu^*) = p(x^*)$, the global optimality of x^* and (λ^*,μ^*) is a consequence of the weak duality theorem. In order to verify "(ii) $\implies$ (i)", we note

$$\begin{aligned} L(x^*,\lambda^*,\mu^*) \leq f(x^*) = p(x^*) &= \sup_{\lambda\in\mathbb{R}^m_+,\,\mu\in\mathbb{R}^p} L(x^*,\lambda,\mu) \\ &= d(\lambda^*,\mu^*) = \inf_{x\in\mathbb{R}^n} L(x,\lambda^*,\mu^*) \leq L(x^*,\lambda^*,\mu^*). \end{aligned}$$

This establishes the saddle point property. ■

10.3. Strong Duality

Motivated by the results in Theorem 10.6, we now want to discuss more general conditions and situations that guarantee zero gap between the optimal objective function values of the primal and dual problem. We first start with linear programs and separations theorems.

10.3.1. Separation Results and Strong Duality in Linear Programming

Let us first consider the linear program

$$\min_x \; c^\top x \quad \text{s.t.} \quad Ax \leq b, \quad Cx = d, \tag{10.5}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $C \in \mathbb{R}^{p\times n}$, $d \in \mathbb{R}^p$ are given. The dual function can be calculated as follows:

$$d(\lambda,\mu) = \inf_x \; c^\top x + \lambda^\top(Ax-b) + \mu^\top(Cx-d) = \inf_x \; (c + A^\top\lambda + C^\top\mu)^\top x - b^\top\lambda - d^\top\mu$$
$$= \begin{cases} -b^\top\lambda - d^\top\mu & \text{if } c = -A^\top\lambda - C^\top\mu, \\ -\infty & \text{otherwise.} \end{cases}$$

Consequently, the dual of the linear program (10.5) is given by

$$\max_{\lambda\in\mathbb{R}^m,\,\mu\in\mathbb{R}^p} -b^\top\lambda - d^\top\mu \quad \text{s.t.} \quad \lambda \geq 0, \quad c = -A^\top\lambda - C^\top\mu. \tag{10.6}$$

We can now state the following duality result for linear programs.

Theorem 10.7: Strong Duality in Linear Programming

Consider the linear program (10.5) and its dual (10.6). The following statements are equivalent:

(i) The primal problem (10.5) has a solution.

(ii) The dual problem (10.6) has a solution.

(iii) The feasible set of the dual problem is nonempty and the dual objective function is bounded from above on the feasible set.

(iv) The feasible set of the primal problem is nonempty and the primal objective function is bounded from below on the feasible set.

In addition, any of the latter conditions implies that there is no duality gap between the primal and dual problem, i.e., every solution x^* of (10.5) and every solution (λ^*,μ^*) of the dual problem (10.6) satisfy $p(x^*) = c^\top x^* = d(\lambda^*,\mu^*)$.

This theorem can be shown by applying Farkas' Lemma and will be covered in the exercises. Before continuing with more general duality results, we first turn to a proof of Farkas' Lemma.

We will see that separation theorems play a major role in the derivation of Farkas' Lemma and strong duality results. Next, we state two of such fundamental hyperplane separation properties.

Theorem 10.8: Hahn-Banach Separation Theorem

Let $C \subset \mathbb{R}^n$ be a nonempty, closed, and convex set and let $y \notin C$ be given. Then there exist $p \in \mathbb{R}^n\backslash\{0\}$ and $\alpha \in \mathbb{R}$ such that

$$p^\top y - \alpha > 0 \quad \text{and} \quad p^\top x - \alpha \leq 0 \quad \forall\, x \in C.$$

Additionally, if C is a cone, then we can choose $\alpha = 0$.

Proof. As shown in Assignment A2.1, the projection $\mathcal{P}_C(y)$ satisfies the first-order opti-

mality condition

$$(\mathcal{P}_C(y) - y)^\top (x - \mathcal{P}_C(y)) \geq 0 \quad \forall\, x \in C.$$

Consequently, setting $p := y - \mathcal{P}_C(y) \neq 0$ and $\alpha := p^\top \mathcal{P}_C(y)$, we have $p^\top x \leq \alpha$ for all $x \in C$ and $p^\top y = p^\top p + p^\top \mathcal{P}_C(y) > \alpha$. In addition, if C is a cone, then we have $0 \in C$ (this follows from the closedness of C) and it follows $\alpha \geq p^\top 0 = 0$. Moreover, due to $\lambda x \in C$ for all $x \in C$ and $\lambda > 0$, we have $\lambda p^\top x \leq \alpha$. Taking the limit $\lambda \to \infty$, this implies $p^\top x \leq 0$. Hence, we have

$$p^\top x \leq 0 \quad \forall\, x \in C \quad \text{and} \quad p^\top y > \alpha \geq 0,$$

which finishes the proof. ■

Theorem 10.9: Supporting Hyperplane Theorem

Let $C \subset \mathbb{R}^n$ be a nonempty convex set and let $y \notin C$ be given. Then there exists a vector $p \in \mathbb{R}^n \backslash \{0\}$ such that

$$p^\top x \leq p^\top y \quad \forall\, x \in C.$$

Proof. The proof relies on the following properties of convex sets:

- $\text{int}(C) = \text{int}(\text{cl } C)$.
- The closure cl C is again a convex set.

Since $y \notin C$, we have $y \notin \text{int}(C) = \text{int}(\text{cl } C)$. Therefore, there exists a sequence $(y^k)_k$ satisfying $y^k \notin \text{cl } C$ and $y^k \to y$. Utilizing the convexity of cl C, Theorem 10.8 is applicable and there exists $p^k \neq 0$ such that

$$(p^k)^\top x < (p^k)^\top y^k \quad \forall\, x \in \text{cl } C, \quad \forall\, k.$$

Defining $q^k := p^k / \|p^k\|$, we have $\|q^k\| = 1$ and without loss of generality, we may assume that $(q^k)_k$ converges to some q with $\|q\| = 1$. Taking the limit $k \to \infty$ and dividing both sides by $\|p^k\|$ in the last inequality, we therefore obtain

$$q^\top x \leq q^\top y \quad \forall\, x \in \text{cl } C.$$

This readily establishes the result noticing $C \subset \text{cl } C$. ■

We now present a proof of Farkas' Lemma – Lemma 9.15.

Proof. Suppose that the statement (ii) in Lemma 9.15 does not hold. Then, we have

$$c \notin K := \{x \in \mathbb{R}^n : x = Au + Bv,\, u \in \mathbb{R}^m_+,\, v \in \mathbb{R}^p\}.$$

We first show that K is a convex closed cone. In fact, let $x, y \in K$ and $\lambda \in [0, 1]$ be arbitrary and consider $u_x, u_y \in \mathbb{R}^m_+$ and $v_x, v_y \in \mathbb{R}^p$ with $x = Au_x + Bv_x$ and $y = Au_y + Bv_y$. Then, it follows

$$\lambda x + (1 - \lambda) y = A[\lambda u_x + (1 - \lambda) u_y] + B[\lambda v_x + (1 - \lambda) v_y] \in K.$$

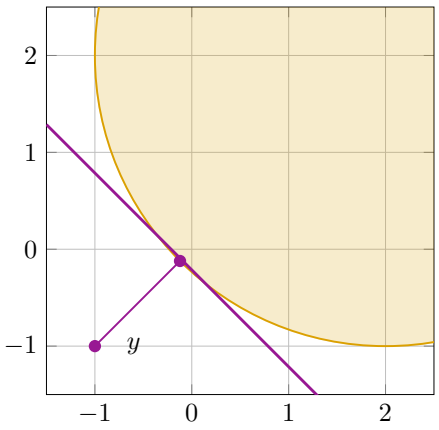


Figure 10.1: Illustration of the Hahn-Banach separation theorem Theorem 10.8: separating $y = (-1,-1)^\top$ from the ball $\{x : (x_1-2)^2 + (x_2-2)^2 \leq 9\}$.

This establishes convexity of K and similarly, we can verify that K is a cone. We show closedness via an inductive argument. Let us first note that without loss of generality we can assume that K has the simpler form

$$K = K_m := \{x \in \mathbb{R}^n : x = Au,\, u \in \mathbb{R}^m_+\} = \left\{\sum\nolimits_{i=1}^m u_i a_i : u \in \mathbb{R}^m_+\right\},$$

where a_i denotes the i-th column vector of A. We now perform an induction with respect to the dimension m of $A \in \mathbb{R}^{n\times m}$. In the case $m = 1$, we have $K_1 = \{ua_1 : u \geq 0\}$. This set is obviously closed. Suppose now that the sets $K_1, ..., K_{m-1}$ are closed for arbitrary choices of the matrix $A \in \mathbb{R}^{n\times j}$, $j = 1, ..., m-1$. Let $A \in \mathbb{R}^{n\times m}$ be arbitrary and consider a sequence $(x^k)_k \subset K_m$ with $x^k \to x$. Then there exists $u^k \in \mathbb{R}^m_+$ such that $x^k = Au^k$. If A has full column rank, we have

$$u^k = (A^\top A)^{-1} A^\top x^k$$

and hence, $(u^k)_k$ converges to $u = (A^\top A)^{-1}A^\top x$. Due to $u^k \geq 0$, we can infer $u \geq 0$ and it holds that $x = \lim_{k\to\infty} x^k = \lim_{k\to\infty} Au^k = Au \in K$.

If the columns of A are linearly dependent, there exists $w \in \mathbb{R}^n\backslash\{0\}$ such that $Aw = 0$. Let γ_k be the smallest number such that the vector $\tilde{u}^k = u^k + \gamma_k w$ possesses a component that equals zero. (It follows $\gamma_k = 0$ if and only if u^k already has a zero component). Let i_k further denote the index of such a component. The sequence $(i_k)_k \subset \{1, ..., m\}$ then has to attain one value $i \in \{1, ..., m\}$ infinitely many times. Let us set $L := \{k : i_k = i\}$. Then we have $\tilde{u}^k_i = 0$ for all $k \in L$ and setting

$$\bar{A} := (a_1, ..., a_{i-1}, a_{i+1}, ..., a_m), \quad \bar{u}^k = (\tilde{u}^k_1, ..., \tilde{u}^k_{i-1}, \tilde{u}^k_{i+1}, ...\tilde{u}^k_m)^\top,$$

it follows $(x^k)_{k\in L} \subset K' := \{x : x = \bar{A}u,\, u \in \mathbb{R}^{m-1}_+\}$. More specifically, we have

$$x^k = Au^k = Au^k + \gamma_k Aw = A\tilde{u}^k = \bar{A}\bar{u}^k, \quad k \in L.$$

By construction, we obtain $K' \subset K_m$ and applying the induction hypothesis, we can infer

that the set K' is closed. Thus, it follows $x \in K' \subset K_m$ which finishes the induction.

Due to $c \notin K$, we can now utilize the Hahn-Banach separation theorem, i.e., there exists $p \in \mathbb{R}^n \backslash \{0\}$ such that $p^\top c > 0$ and $p^\top x \leq 0$ for all $x \in K$. Setting $u = e_i$ and $v = 0$ it is easy to see that the i-th column a_i is contained in K, i.e., we have $a_i \in K$. Hence, it follows $p^\top a_i \leq 0$ for all i. Similarly, setting $u = 0$ and $v = \pm e_j$, we have $\pm b_j \in K$ where b_j denotes the j-th column of the matrix B. This yields $p^\top b_j = 0$ and together, we obtain

$$p^\top c > 0, \quad A^\top p \leq 0, \quad B^\top p = 0.$$

Consequently, statement (i) in Lemma 9.15 is not satisfied. ■

10.3.2. The General Case

We now consider the general optimization problem

$$\min_{x \in \mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad g(x) \leq 0, \quad Ax \leq b, \quad Cx = d, \tag{10.7}$$

where $A \in \mathbb{R}^{q \times n}$, $b \in \mathbb{R}^q$, and $C \in \mathbb{R}^{p \times n}$, $d \in \mathbb{R}^p$ are given as before. The feasible set of problem (10.7) is given by $X := \{x \in \mathbb{R}^n : g(x) \leq 0, Ax \leq b, Cx = d\}$ and the associated dual function of (10.7) is defined via

$$d(\lambda, \nu, \mu) := \inf_{x \in \mathbb{R}^n} \; f(x) + \lambda^\top g(x) + \nu^\top (Ax - b) + \mu^\top (Cx - d). \tag{10.8}$$

Let x^* and $(\lambda^*, \nu^*, \mu^*)$ be feasible global solutions of the primal and dual problem, respectively. In the following, we study a condition that ensures that the so-called *duality gap* $p(x^*) - d(\lambda^*, \nu^*, \mu^*) = f(x^*) - d(\lambda^*, \nu^*, \mu^*)$ is zero. We will work with *Slater's condition*.

Definition 10.10: Slater's Condition

We say that **Slater's condition** is satisfied for problem (10.7) if there exists a point $\bar{x} \in \mathbb{R}^n$ such that $g_i(\bar{x}) < 0$ for all $i = 1, ..., m$ and $A\bar{x} \leq b$, and $C\bar{x} = d$.

Theorem 10.11: Strong Duality Theorem

Suppose that problem (10.7) is feasible with finite optimal value $f^* = \inf_{x \in X} f(x)$. In addition, assume that $f, g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, ..., m$, are convex functions and that Slater's condition holds for (10.7). Then, the dual problem $\sup_{(\lambda,\nu,\mu) \in \mathbb{R}^m_+ \times \mathbb{R}^q_+ \times \mathbb{R}^p} d(\lambda, \nu, \mu)$ has a global solution $(\lambda^*, \nu^*, \mu^*) \in \mathbb{R}^m_+ \times \mathbb{R}^q_+ \times \mathbb{R}^p$ and strong duality holds, i.e., $f^* = d(\lambda^*, \nu^*, \mu^*)$.

We only present a proof for the special case without the additional linear constraints, i.e., we consider the simpler problem $\min_x f(x)$ subject to the constraints $g(x) \leq 0$.

Proof. We define the $K := \{(u, v) \in \mathbb{R}^m \times \mathbb{R} : \exists\, x \in \mathbb{R}^n$ such that $g(x) \leq u$, $f(x) \leq v\}$. The set K is obviously nonempty due to $(g(x), f(x)) \in K$ for all $x \in \mathbb{R}^n$. We now want to verify that K is convex. Let (u_1, v_1), $(u_2, v_2) \in K$ and $\lambda \in [0, 1]$ be arbitrary. Then there are x_i with $g(x_i) \leq u_i$ and $f(x_i) \leq v_i$ for $i = 1, 2$. Setting $x_\lambda := \lambda x_1 + (1 - \lambda) x_2$ and using

the convexity of g_j and f, we obtain

$$f(x_\lambda) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \leq \lambda v_1 + (1-\lambda) v_2$$
$$g(x_\lambda) \leq \lambda g(x_1) + (1-\lambda) g(x_2) \leq \lambda u_1 + (1-\lambda) u_2.$$

This shows $(\lambda u_1 + (1-\lambda)u_2, \lambda v_1 + (1-\lambda)v_2) \in K$ and verifies that K is convex. Furthermore, the point $(0, f^*)$ does not lie in the interior of K. Otherwise there would exist $\varepsilon > 0$ such that $(0, f^* - \varepsilon) \in K$. But then, by the definition of the set K there exists x with $g(x) \leq 0$ and $f(x) \leq f^* - \varepsilon$. Due to $f^* = \inf_{x \in X} f(x)$ this is a contradiction. Consequently, the separation result in Theorem 10.9 is applicable and there exists $(\bar{u}, \bar{v}) \in \mathbb{R}^m \times \mathbb{R} \backslash \{0\}$ such that

$$\bar{v} \cdot f^* \leq \bar{u}^\top u + \bar{v} \cdot v \quad \forall\, (u, v) \in K. \tag{10.9}$$

By definition of K we have $(u + s, v) \in K$ and $(u, v + t) \in K$ for all $s, t \geq 0$ and $(u, v) \in K$. Taking the limits $s_i \to \infty$ and $t \to \infty$, this implies $\bar{u}, \bar{v} \geq 0$. We will now show $\bar{v} > 0$.

Let us assume $\bar{v} = 0$. In this case, condition (10.9) reduces to $0 \leq \bar{u}^\top u$ for all $(u, v) \in K$. Specifically let us choose the point $(g(\bar{x}), f(\bar{x})) \in K$ (notice that $\bar{x}$ was the point satisfying Slater's condition). Using $\bar{u} \geq 0$, $\bar{u} \neq 0$, we obtain the contradiction

$$0 \leq \sum_{i=1}^{m} \bar{u}_i g_i(\bar{x}) < 0.$$

Thus, we have $\bar{v} > 0$. Defining $\lambda^* = \bar{u}/\bar{v}$ and using $(g(x), f(x)) \in K$, this yields

$$f^* \leq (\lambda^*)^\top g(x) + f(x) = L(x, \lambda^*) \quad \forall\, x \in \mathbb{R}^n.$$

Taking the infimum with respect to x, we then can infer $f^* \leq \inf_{x \in \mathbb{R}^n} L(x, \lambda^*) = d(\lambda^*)$. Applying the weak duality theorem, it follows $d(\lambda^*) \leq f^*$ which shows that the duality gap is zero and strong duality holds. Furthermore, the weak duality theorem also implies $d(\lambda^*) \geq f^* = \inf_{x \in X} f(x) \geq d(\lambda)$ for all $\lambda \in \mathbb{R}^m_+$. This verifies that λ^* is a global solution of the dual problem and establishes the existence result. ■

The more general duality result presented in Theorem 10.11 requires a more specialized separation theorem for polyhedral sets, see, e.g., [4, Proposition 1.5.7]. A proof can then be obtained by combining the techniques in [5, Proposition 6.2.3 and Proposition 6.3.2]. Notice that here we assumed convexity of f and g_i, $i = 1, ..., m$ on the whole $\mathbb{R}^n$.

10.3.3. Application: Sparse Approximation

We consider the nonsmooth optimization problem

$$\min_{x \in \mathbb{R}^n} \; \|Ax - b\|_1 + \frac{\alpha}{2}\|x\|^2,$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $\alpha > 0$ are given. Introducing an auxiliary variable $y \in \mathbb{R}^m$, this problem can be rewritten as follows

$$\min_{x,y} \ \|y\|_1 + \frac{\alpha}{2}\|x\|^2 \quad \text{s.t.} \quad Ax - y = b.$$

Since the point $(0, b)$ is feasible, Slater's condition is obviously satisfied and we can apply the strong duality theorem. It holds that

$$\begin{aligned} L(x, y, \mu) &= \|y\|_1 + \frac{\alpha}{2}\|x\|^2 + \mu^\top(Ax - y - b), \\ d(\mu) &= \inf_{x,y} \ L(x, y, \mu) = \inf_y \inf_x \ \|y\|_1 + \frac{\alpha}{2}\|x\|^2 + \mu^\top(Ax - y - b). \end{aligned}$$

Due to $\nabla_x L(x, y, \mu) = \alpha x + A^\top\mu$ and $\nabla^2_{xx} L(x, y, \mu) = \alpha I$, the minimum with respect to x is attained at $x(\mu) = -\alpha^{-1}A^\top\mu$ and it follows

$$\inf_x \ L(x, y, \mu) = L(x(\mu), y, \mu) = \|y\|_1 - \frac{1}{2\alpha}\|A^\top\mu\|^2 - \mu^\top(y + b).$$

Next, suppose there exists i with $|\mu_i| > 1$. Then selecting $y(t) = t\,\mathrm{sgn}(\mu_i)e_i$, we obtain:

$$\|y(t)\|_1 - \mu^\top y(t) = |t| - t|\mu_i| \to -\infty \quad \text{as} \quad t \to \infty.$$

In the case $\|\mu\|_\infty \leq 1$, we have

$$\|y\|_1 - \mu^\top y = \sum_{i=1}^m (1 - \mathrm{sgn}(y_i)\mu_i)|y_i| \geq 0.$$

We can achieve the optimal value 0 by selecting y as follows:

$$y_i > 0 \ \text{ if } \ \mu_i = 1, \quad y_i < 0 \ \text{ if } \ \mu_i = -1, \quad y_i = 0 \ \text{ otherwise.}$$

Together, this now yields

$$d(\mu) = \inf_{y,x} \ L(x, y, \mu) = \begin{cases} -\frac{1}{2\alpha}\|A^\top\mu\|^2 - \mu^\top b & \text{if } \|\mu\|_\infty \leq 1, \\ -\infty & \text{if } \|\mu\|_\infty > 1 \end{cases}$$

and the associated dual problem is given by

$$\sup_\mu \ -\frac{1}{2\alpha}\|A^\top\mu\|^2 - \mu^\top b \quad \text{s.t.} \quad \mu_i \in [-1, 1] \quad \forall\, i.$$

This problem is a concave, quadratic problem with simple box constraints. Since the constraints define a compact set, the dual problem always possesses a solution $\mu^* \in \mathbb{R}^m$.

A visualization of the discussed sparse approximation problem and its dual can be found in Figure 10.2.

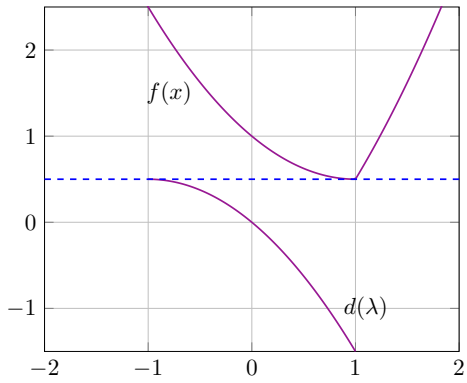


Figure 10.2: Illustration of strong duality and of the connection between the primal and dual problem in subsection 10.3.3. Visualization in $\mathbb{R}$ with $b = \alpha = 1$ and $A = 1$.

10.4. Fenchel Duality

Inspired by the sparse approximation problem, we now want to consider a special class of optimization problems of the form

$$\min_{x\in\mathbb{R}^n} \; f(x) + g(Ax - b), \tag{10.10}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}$ are convex functions and $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are given.

We now introduce the so-called *convex conjugate* of the function f:

$$f^* : \mathbb{R}^n \to (-\infty, +\infty], \quad f^*(x) := \sup_{y\in\mathbb{R}^n} \; y^\top x - f(y).$$

Since the mapping $x \mapsto y^\top x - f(y)$ is linear in x for every fixed $y \in \mathbb{R}^n$, Lemma 4.12 implies that f^* is convex (independent of the convexity of f) on the set of points where it is well-defined. We can now utilize the convex conjugate to determine the dual of (10.10) explicitly. As before, we first introduce an auxiliary variable $y \in \mathbb{R}^m$ to rewrite problem (10.10) as follows:

$$\min_{x,y} \; f(x) + g(y) \quad \text{s.t.} \quad Ax - y = b.$$

The associated Lagrange function is then given by $L(x, y, \mu) = f(x) + g(y) + \mu^\top(Ax - y - b)$ and the dual function satisfies

$$\begin{aligned} d(\mu) = \inf_{x,y} \; L(x, y, \mu) &= \inf_x \; [x^\top A^\top \mu + f(x)] + \inf_y \; [-y^\top \mu + g(y)] - \mu^\top b \\ &= -\sup_x \; [x^\top(-A^\top \mu) - f(x)] - \sup_y \; [y^\top \mu - g(y)] - \mu^\top b \\ &= -\mu^\top b - f^*(-A^\top \mu) - g^*(\mu). \end{aligned}$$

Hence, the dual problem of (10.10) can be expressed as follows:

$$\max_{\mu\in\mathbb{R}^m} \; -\mu^\top b - f^*(-A^\top \mu) - g^*(\mu). \tag{10.11}$$

If the primal problem (10.10) has a finite optimal value, then the strong duality theorem Theorem 10.11 is applicable (notice that the point $(0, -b)$ is feasible for the auxiliary problem) and we have:

$$\min_{x\in\mathbb{R}^n} \; f(x) + g(Ax - b) = \max_{\mu\in\mathbb{R}^m} \; -\mu^\top b - f^*(-A^\top \mu) - g^*(\mu).$$

This result and the alternative duality concept using convex conjugates is known as *Fenchel duality*. It is also possible to derive a direct connection between primal and dual solutions using Fenchel duality.

Theorem 10.12: Primal and Dual Solutions via Fenchel Duality

Let us consider the primal and dual problem (10.10) and (10.11), respectively. Then the following statements are true:

- Let x^* be a solution of the primal problem (10.10) and suppose that the function g is differentiable. Then, $\mu^* = \nabla g(Ax^* - b)$ is a dual solution of (10.11).
- Let μ^* be a solution of the dual problem $\sup_{\mu\in\mathbb{R}^m} \; d(\mu)$ and let f^* be differentiable. Then, $x^* = \nabla f^*(-A^\top \mu^*)$ is a primal solution.

We will not present a proof of this result here, but refer to section 12 and [1, Theorem 19.1 and Proposition 19.3] for further details.

An alternative way to recover primal and dual solutions is to find x^* or μ^* such that the pair $(x^*, Ax^* - b, \mu^*)$ is a saddle point of the Lagrangian $L(x, y, \mu) = f(x) + g(y) + \mu^\top (Ax - y - b)$.

References. The first sections closely follow [9, Section 16.5 and Chapter 17]. The proof of Theorem 10.11 is presented in [5, Proposition 6.3.1]. See also [5, Chapter 6] and [2, Chapter 12] for more and different duality results.

11. Algorithms for Constrained Optimization Problems

In this section, we study and develop optimization algorithms for general nonlinear programs of the form

$$\min_{x\in\mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad g(x) \leq 0, \quad h(x) = 0. \tag{11.1}$$

We will start with the so-called penalty method for constrained problems.

11.1. The Penalty Method

The penalty method is a classical optimization approach in nonlinear programming. The principle idea of the penalty method is to solve a general nonlinear program via considering a sequence of unconstrained optimization problems. These unconstrained problems are built by adding *penalty terms* for the constraints to the objective function. The penalty terms are then balanced by positive parameters – the so-called *penalty parameters*. We obtain penalty subproblems of the form

$$\min_{x\in\mathbb{R}^n} \; f(x) + \alpha p(x),$$

where $\alpha > 0$ is the penalty parameter and $p : \mathbb{R}^n \to \mathbb{R}$ is a penalty function satisfying

- $p(x) = 0$ for all $x \in X$ and $p(x) > 0$ on $\mathbb{R}^n \backslash X$.

The larger the penalty parameter α, the better we will approximate the initial constrained problem. Unfortunately, the numerical performance of the penalty subproblems becomes worse the larger the penalty parameter is chosen. Hence, we typically generate a sequence of penalty problems corresponding to a monotonically growing sequence of penalty parameters $(\alpha_k)_k$. The solution x^k of the k-th subproblem will then be used as initial point for the next problem.

The quadratic penalty method for (11.1) utilizes the *quadratic penalty function*

$$\begin{aligned} P_\alpha(x) &:= f(x) + \frac{\alpha}{2}\sum_{i=1}^{m}(\max\{0, g_i(x)\})^2 + \frac{\alpha}{2}\sum_{j=1}^{p} h_j(x)^2 \\ &= f(x) + \frac{\alpha}{2}\|(g(x))_+\|^2 + \frac{\alpha}{2}\|h(x)\|^2. \end{aligned}$$

Here, for $v \in \mathbb{R}^n$ we use the notation $(v_+)_i = \max\{0, v_i\}$ for all i. The scalar $\alpha > 0$ is the penalty parameter. We have introduced the two different penalty functions $p_{\mathtt{i}}(g_i(x))$, $p_{\mathtt{i}}(t) = (t)_+^2 = (\max\{0, t\})^2$ and $p_{\mathtt{e}}(h_j(x))$, $p_{\mathtt{e}}(t) = t^2$ for the inequality and equality constraints, respectively. Since the mappings $p_{\mathtt{i}}$ and $p_{\mathtt{e}}$ are continuously differentiable, the quadratic penalty function P_α is continuously differentiable as well, as long as the underlying problem is C^1. Specifically, we have

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

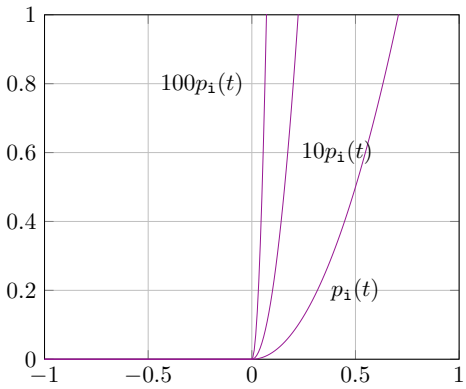


Figure 11.1: Illustration of the penalty term $t \mapsto \frac{\alpha}{2} p_{\mathtt{i}}(t)$ for $\alpha \in \{2, 20, 200\}$.

and consequently, it follows

$$P_\alpha(x) = f(x) \quad \text{and} \quad \nabla P_\alpha(x) = \nabla f(x) \quad \forall\, x \in X.$$

This implies that the penalty terms have slope zero when leaving the feasible set and hence, the penalization of the constraints does not immediately take effect. In particular, a feasible point $x \in X$ can only be a stationary point of P_α in the case $\nabla f(x) = 0$. Since solutions and KKT-points of the underlying constrained problem typically do not satisfy this condition, the penalty subproblem often returns infeasible points as solutions.

The full penalty method is summarized in Algorithm 11.1.

11.1.1. Convergence Analysis of the Penalty Method

We now present a basic convergence analysis of the quadratic penalty method.

Theorem 11.1: Convergence of the Penalty Method

Suppose that f, g, and h are continuous functions and let the feasible set X be nonempty. Let the sequence of penalty parameters $(\alpha_k)_k \subset (0, \infty)$ be monotonically increasing and diverging to $+\infty$. Furthermore, let $(x^k)_k$ be generated by Algorithm 11.1, then we have:

(i) The sequence $(P_{\alpha_k}(x^k))_k$ is monotonically increasing.

(ii) The sequence $(\|(g(x^k))_+\|^2 + \|h(x^k)\|^2)_k$ is monotonically decreasing.

(iii) The sequence $(f(x^k))_k$ is monotonically increasing.

(iv) It holds that $\lim_{k\to\infty}(g(x^k))_+ = 0$ and $\lim_{k\to\infty} h(x^k) = 0$.

(v) Every accumulation point of $(x^k)_k$ is a global solution of problem (11.1).

Algorithm 11.1: The Quadratic Penalty Method

1 Initialization: Choose an initial point $x^{-1} \in \mathbb{R}^n$ and a penalty parameter $\alpha_0 > 0$.
for $k = 0, 1, 2, \ldots$ **do**
2 Calculate the global solution x^k of the penalty problem $\min_{x \in \mathbb{R}^n} P_{\alpha_k}(x)$. (Here, we often utilize x^{k-1} as initial point).
3 Terminate if $x^k \in X$. Otherwise select $\alpha_{k+1} > \alpha_k$.

Proof. We set $p(x) := \frac{1}{2}(\|(g(x))_+\|^2 + \|h(x)\|^2)$. Using the optimality of x^k and $\alpha_{k+1} > \alpha_k$, we have

$$\begin{aligned} P_{\alpha_k}(x^k) \leq P_{\alpha_k}(x^{k+1}) &= f(x^{k+1}) + \alpha_k p(x^{k+1}) \\ &\leq f(x^{k+1}) + \alpha_{k+1} p(x^{k+1}) = P_{\alpha_{k+1}}(x^{k+1}). \end{aligned}$$

This shows part (i). Next, summing the estimates $P_{\alpha_k}(x^k) \leq P_{\alpha_k}(x^{k+1})$ and $P_{\alpha_{k+1}}(x^{k+1}) \leq P_{\alpha_{k+1}}(x^k)$, we obtain

$$\alpha_k p(x^k) + \alpha_{k+1} p(x^{k+1}) \leq \alpha_k p(x^{k+1}) + \alpha_{k+1} p(x^k).$$

Due to $\alpha_{k+1} > \alpha_k$, this yields $p(x^k) \geq p(x^{k+1})$. Using this estimate, it further follows

$$0 \leq P_{\alpha_k}(x^{k+1}) - P_{\alpha_k}(x^k) = f(x^{k+1}) - f(x^k) + \alpha_k(p(x^{k+1}) - p(x^k)) \leq f(x^{k+1}) - f(x^k).$$

This establishes part (ii) and (iii). We continue with a verification of part (iv). In particular, we show $p(x^k) \to 0$. Due to $X \neq \emptyset$ there exists $y \in X$ and hence, we have $P_{\alpha_k}(x^k) \leq P_{\alpha_k}(y) = f(y)$. Utilizing part (iii), we obtain:

$$f(y) \geq P_{\alpha_k}(x^k) = f(x^k) + \alpha_k p(x^k) \geq f(x^0) + \alpha_k p(x^k).$$

Thus, due to $\alpha_k \to \infty$, this implies $p(x^k) \to 0$. Let x^* now be an accumulation point of $(x^k)_k$. Then, by the continuity of $(g)_+$ and h and part (iv), we have $x^* \in X$. Let $(x^{k_\ell})_\ell$ be a subsequence converging to x^*. Then, for all $x \in X$ and $\ell \in \mathbb{N}$, it follows

$$f(x^{k_\ell}) \leq P_{\alpha_{k_\ell}}(x^{k_\ell}) \leq P_{\alpha_{k_\ell}}(x) = f(x).$$

This shows $f(x^*) = \lim_{\ell \to \infty} f(x^{k_\ell}) \leq f(x)$ for all $x \in X$ which finishes the proof. ■

We assumed that Algorithm 11.1 generates an infinite sequence of iterates $(x^k)_k$. If this is not possible, then either one of the subproblems is not solvable or we have found an iterate $x^k \in X$ and terminate in step 3 of the algorithm. This makes sense, since x^k is a global solution of the underlying problem (11.1) in this case. Indeed, due to $x^k \in X$ we have

$$f(x) = P_{\alpha_k}(x) \geq P_{\alpha_k}(x^k) = f(x^k) \quad \forall\, x \in X.$$

Suppose that f, g, and h are continuously differentiable. Since each iterate x^k necessarily is

a stationary point of P_{α_k}, we have

$$\begin{aligned} 0 = \nabla P_{\alpha_k}(x^k) &= \nabla f(x^k) + \sum_{i=1}^{m} \alpha_k (g_i(x^k))_+ \nabla g_i(x^k) + \sum_{j=1}^{p} \alpha_k h_j(x^k) \nabla h_j(x^k) \\ &= \nabla f(x^k) + \nabla g(x^k)\lambda^k + \nabla h(x^k)\mu^k, \end{aligned}$$

where

$$\lambda_i^k := \alpha_k \max\{0, g_i(x^k)\} \quad \text{and} \quad \mu_j^k := \alpha_k h_j(x^k). \tag{11.2}$$

Consequently, if there exist subsequences $x^{k_\ell} \to x^*$, $\lambda^{k_\ell} \to \lambda^*$, and $\mu^{k_\ell} \to \mu^*$, then we have $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$ and (x^*, λ^*, μ^*) is a KKT-triple of (11.1). We make this statement more precise in the next theorem.

Theorem 11.2: Penalty Method and KKT-Conditions

Let f, g, and h be continuously differentiable and suppose that the feasible set X is nonempty. Let $(\alpha_k)_k \subset (0, \infty)$ be monotonically increasing and diverging to $+\infty$. Furthermore, let $(x^k)_k$ be generated by Algorithm 11.1 and define $(\lambda^k)_k$ and $(\mu^k)_k$ as in (11.2), then we have:

(i) Let $(x^{k_\ell}, \lambda^{k_\ell}, \mu^{k_\ell})_\ell$ be a subsequence of $(x^k, \lambda^k, \mu^k)_k$ converging to (x^*, λ^*, μ^*). Then x^* is a global solution of (11.1) and (x^*, λ^*, μ^*) is a KKT-triple.

(ii) Let x^* be accumulation point of $(x^k)_k$ and let $(x^{k_\ell})_\ell$ converge to x^*. Furthermore, suppose that x^* is regular. Then, $(x^{k_\ell}, \lambda^{k_\ell}, \mu^{k_\ell})_\ell$ converges to a KKT-triple of (11.1) and x^* is a global solution of (11.1).

Proof. Applying Theorem 11.1, x^* is a global solution of problem (11.1). Following our previous discussion, we can take the limit

$$\nabla_x L(x^*, \lambda^*, \mu^*) = \lim_{\ell \to \infty} \nabla_x L(x^{k_\ell}, \lambda^{k_\ell}, \mu^{k_\ell}) = 0.$$

Since we already know that x^* is feasible, we only need to verify the complementarity conditions. By definition, we have $\lambda^k \geq 0$ for all k and thus, it follows $\lambda^* \geq 0$. Next, for all $i \in \mathcal{I}(x^*)$, we can infer $g_i(x^{k_\ell}) < 0$ for all i and all ℓ sufficiently large. Consequently, it holds that

$$\lambda_i^{k_\ell} = \alpha_{k_\ell} \max\{0, g_i(x^{k_\ell})\} = 0 \quad \forall\, i \in \mathcal{I}(x^*)$$

and ℓ sufficiently large. This implies $\lambda_i^* = 0$ for all $i \in \mathcal{I}(x^*)$ and thus, the complementarity conditions are satisfied.

We now verify part (ii). As in part (i), we can infer that x^* is a global solution of (11.1). Hence, we only need to show convergence of $(x^{k_\ell}, \lambda^{k_\ell}, \mu^{k_\ell})_\ell$. As before, we have $g_i(x^{k_\ell}) < 0$ for all $i \in \mathcal{I}(x^*)$ and ℓ sufficiently large. This implies $\lambda_i^{k_\ell} = 0$, $i \in \mathcal{I}(x^*)$ and it follows

$$\lambda^*_{\mathcal{I}(x^*)} = \lim_{\ell \to \infty} \lambda^{k_\ell}_{\mathcal{I}(x^*)} = 0.$$

By assumption, the matrix $H_* = (\nabla g_{\mathcal{A}(x^*)}(x^*), \nabla h(x^*))$ has full column rank and thus, $H_*^\top H_*$ is invertible and positive definite. Consequently, utilizing the continuity of ∇g and ∇h and Lemma 6.2, the matrix $H_{k_\ell}^\top H_{k_\ell}$, $H_{k_\ell} := (\nabla g_{\mathcal{A}(x^*)}(x^{k_\ell}), \nabla h(x^{k_\ell}))$ is positive definite and invertible for all ℓ sufficiently large. But then we obtain

$$0 = H_{k_\ell}^\top \nabla_x L(x^{k_\ell}, \lambda^{k_\ell}, \mu^{k_\ell}) = H_{k_\ell}^\top \nabla f(x^{k_\ell}) + H_{k_\ell}^\top H_{k_\ell} \begin{pmatrix} \lambda^{k_\ell}_{\mathcal{A}(x^*)} \\ \mu^{k_\ell} \end{pmatrix}$$

and

$$\begin{pmatrix} \lambda^{k_\ell}_{\mathcal{A}(x^*)} \\ \mu^{k_\ell} \end{pmatrix} = -(H_{k_\ell}^\top H_{k_\ell})^{-1} H_{k_\ell}^\top \nabla f(x^{k_\ell}) \to -(H_*^\top H_*)^{-1} H_*^\top \nabla f(x^*).$$

(Here, we utilized the continuity of the matrix inversion $M \mapsto M^{-1}$). This establishes convergence of the subsequence $(x^{k_\ell}, \lambda^{k_\ell}, \mu^{k_\ell})_\ell$. ■

Our theoretical results demonstrate that the sequence of penalty parameters has to converge to infinity in the quadratic penalty method to guarantee convergence and feasibility (unless the approach terminates after finitely many steps which is unlikely in practice). Unfortunately, this will result in an increasingly bad condition number.

In order to explain this phenomenon, we consider an equality constrained problem with affine-linear constraints $h(x) = Ax - b$, $A \in \mathbb{R}^{p\times n}$ and $b \in \mathbb{R}^p$. Then, we have

$$\nabla^2 P_\alpha(x) = \nabla^2 f(x) + \alpha \nabla h(x) \nabla h(x)^\top + \alpha \sum_{j=1}^{p} h_j(x) \nabla^2 h_j(x) = \nabla^2 f(x) + \alpha A^\top A.$$

For every v with $Av \neq 0$, we then obtain

$$v^\top \nabla^2 P_\alpha(x) v = v^\top \nabla^2 f(x) v + \alpha \|Av\|^2 = \mathcal{O}(\alpha) \quad \alpha \to \infty.$$

On the other hand, for every $w \neq 0$ with $Aw = 0$ (if such an w does not exist, then A is invertible and X only consists of a single point), we have

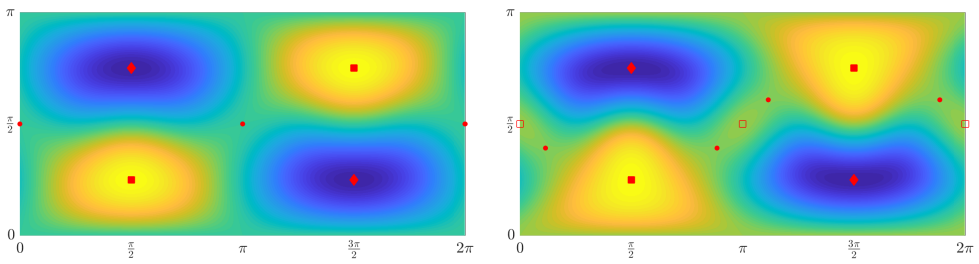
$$w^\top \nabla^2 P_\alpha(x) w = w^\top \nabla^2 f(x) w = \mathcal{O}(1) \quad \alpha \to \infty.$$

Thus, the condition number of $\nabla^2 P_\alpha(x)$ asymptotically behaves like $\mathcal{O}(\alpha)$ for $\alpha \to \infty$. As we have seen, this will affect the performance of gradient-based approaches and it will lead to smaller regions of fast local convergence when utilizing Newton-type methods.

11.1.2. Numerical Experiment: Bose-Einstein Condensates

We now briefly illustrate and discuss the numerical performance of the quadratic penalty method. We consider the nonconvex quartic-quadratic problem:

$$\min_{x\in\mathbb{R}^n} \; f(x) = \frac{1}{2} x^\top A x + \beta \|x\|_4^4 \quad \text{s.t.} \quad \|x\| = 1. \tag{11.3}$$



(a) $\beta = 0.25$ (b) $\beta = 0.75$

(c) $\beta = 1.25$ (d) $\beta = 3.25$

Figure 11.2: Plot of the landscape of the objective function (11.3) for fixed A and different values of β. The red point marker depicts the location of saddle points. Local and global minima are indicated by non-filled and filled diamond markers. The location of local and global maxima is marked by non-filled and filled squares.

Due to the spherical constraints this problem is highly nonconvex. The so-called Bose-Einstein condensation (BEC) problem is an important application and example that can be expressed using the optimization model (11.3) and that has attracted great interests in electronic structure calculations and in the atomic, molecule and optical physics community. In Figure 11.2, we illustrate different landscapes of the mapping f in $\mathbb{R}^3$ and when the parameter β changes. We use spherical coordinates to plot f on the sphere and we consider the choice

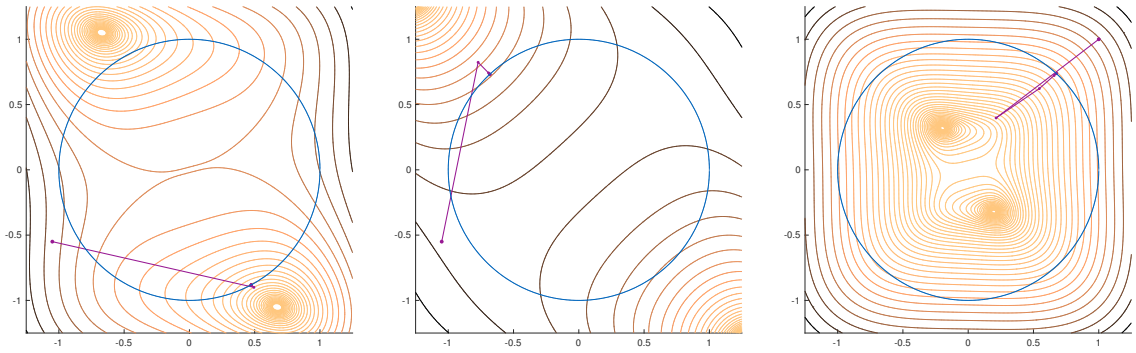
$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad \beta \in \{0.25, 0.75, 1.25, 3.25\}.$$

Figure 11.2 demonstrates that the landscape of f varies a lot when the interaction coefficient β changes. Specifically, the number of stationary points and local minima increases from 6 to 26 and from 2 to 8.

In the numerical test, we set $A = -aa^\top$ where $a =$ `randn(n,1)` is randomly generated and we select $\beta \in \{0.1, 5\}$. We use a globalized Newton method to solve the resulting penalty subproblems:

$$\min_x \; P_{\alpha_k}(x) = f(x) + \frac{\alpha_k}{2}(\|x\|^2 - 1)^2.$$

We start with $\alpha_0 = 1$ and increase α_k by a factor of 10 after each successful penalty step (other parameters are standard: $s = 1$, $\gamma = p = 0.1$, $\sigma = 0.5$, $\beta_1 = \beta_2 = 10^{-6}$). The method



(a) $\beta = 0.1$ (b) $\beta = 0.1$ (c) $\beta = 5$

Figure 11.3: Plot of the solution path of the quadratic penalty method with different coefficients β and random $A = -aa^\top$.

typically requires only few iterations to converge, i.e., in the case $n = 1000$ and $x^0 = \mathbb{1}$, the quadratic penalty method usually converges within 6–20 iterations. Since the iterates are re-used as initial points for the next iteration, the globalized Newton method also converges quickly within 3–8 iterations.

In Figure 11.3, the solution path of the penalty method is shown in $\mathbb{R}^2$ for a few different settings.

11.2. An Augmented-Lagrange Method

We now want to derive a different penalty approach that enjoys better properties and also works for bounded penalty parameters. Let us consider the equality constrained optimization problem

$$\min_{x\in\mathbb{R}^n} \; f(x) \quad \text{s.t.} \quad h(x) = 0, \tag{11.4}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^p$ are continuously differentiable functions. Let $L(x, \mu) = f(x) + \mu^\top h(x)$ denote the Lagrangian associated with problem (11.4).

The *augmented-Lagrange method* is an iterative method which, at each iteration, computes the minimizer $x^k \in \mathbb{R}^n$ of the unconstrained problem

$$\min_{x\in\mathbb{R}^n} \; L_{\alpha_k}(x, \mu^k) := L(x, \mu^k) + \frac{\alpha_k}{2}\|h(x)\|_2^2, \tag{11.5}$$

where the multiplier $\mu^k \in \mathbb{R}^p$ and the penalty parameter $\alpha_k > 0$ are chosen appropriately in each step k.

We first consider the case where $\mu^k = \mu \in \mathbb{R}^p$ is fixed for all $k \in \mathbb{N}$ and the sequence $(\alpha_k)_k$ is monotonically increasing with $\alpha_k \to \infty$. Furthermore, let us suppose that each iterate x^k is a global solution of subproblem (11.5) with $\mu^k = \mu$ i.e., $x^k \in \arg\min_{x\in\mathbb{R}^n} \; L_{\alpha_k}(x, \mu)$.

We can interpret this procedure as applying the quadratic penalty method to the following

problem:

$$\min_{x\in\mathbb{R}^n} \; L(x,\mu) \quad \text{s.t.} \quad h(x)=0. \tag{11.6}$$

Let x^* be an arbitrary accumulation point of the sequence (x^k) and let $(x^{k_\ell})_\ell$ be a corresponding subsequence of $(x^k)_k$ converging to x^*. Then it holds that:

- The point x^* is a global solution of problem (11.4).
- Assume that the LICQ is satisfied at x^* and define $\mu^k_{\text{app}} := \mu + \alpha_k h(x^k)$. Then the limit $\mu^* := \lim_{\ell\to\infty} \mu^{k_\ell}_{\text{app}}$ exists and (x^*,μ^*) is a KKT pair of (11.4).

Proof. Since L_{α_k} is exactly the quadratic penalty function associated with problem (11.6), the described method can be interpreted as an application of the quadratic penalty method to problem (11.6). Thus, since $(\alpha_k)_k$ is monotonically increasing and diverges to ∞ and x^* is an accumulation point of $(x^k)_k$, Theorem 11.1 implies that x^* is a global solution of problem (11.6). Moreover, we have

$$\begin{aligned}
& x^* \text{ is a global solution of (11.6)} \\
\iff \quad & h(x^*)=0 \quad \wedge \quad L(x^*,\mu)\le L(x,\mu) \quad \forall\, x\in\mathbb{R}^n \text{ with } h(x)=0 \\
\iff \quad & h(x^*)=0 \quad \wedge \quad f(x^*)=L(x^*,\mu)\le L(x,\mu)=f(x) \quad \forall\, x\in\mathbb{R}^n \text{ with } h(x)=0 \\
\iff \quad & x^* \text{ is a global solution of (11.4)}.
\end{aligned}$$

This shows that x^* is a solution of problem (11.4). To verify the second statement, we can apply Theorem 11.2. In particalur, the term $\mu^{k_\ell}_{\text{app}} - \mu = \alpha_{k_\ell} h(x^{k_\ell})$ converges to a multiplier $\bar\mu := \mu^* - \mu$ of (11.6). This also implies that $(\mu^{k_\ell}_{\text{app}})_\ell$ converges to some μ^*. Next, we have

$$\begin{aligned}
& (x^*,\mu^*-\mu) \text{ is a KKT pair of (11.6)} \\
\iff \quad & h(x^*)=0 \quad \wedge \quad \nabla_x L(x^*,\mu) + \nabla h(x^*)(\mu^*-\mu)=0 \\
\iff \quad & h(x^*)=0 \quad \wedge \quad \nabla_x L(x^*,\mu^*)=0 \\
\iff \quad & (x^*,\mu^*) \text{ is a KKT pair of (11.4)},
\end{aligned}$$

which finishes the proof. ■

Next, we consider an inexact version of the augmented-Lagrange method, where the subproblem (11.5) is only solved approximately and the multiplier μ is no longer fixed. For a given iterate x^k and a penalty parameter α_k we use the update formula:

$$\mu^{k+1} = \mu^k + \alpha_k \nabla_\mu L_{\alpha_k}(x^k,\mu^k) = \mu^k + \alpha_k h(x^k).$$

The different steps of the method are shown and summarized in Algorithm 11.2.

If Algorithm 11.2 terminates at iteration k in step 3, then we have $h(x^k)=0$ and

$$0 = \nabla_x L_{\alpha_k}(x^k,\mu^k) = \nabla_x L(x^k,\mu^k) + \alpha_k \nabla h(x^k) h(x^k) = \nabla_x L(x^k,\mu^k).$$

This immediately implies that (x^k,μ^k) is a KKT-pair of problem (11.4). We now present a

Algorithm 11.2: An Inexact Augmented-Lagrange Method

1 Initialization: Choose an initial point $x^{-1} \in \mathbb{R}^n$, $\mu^0 \in \mathbb{R}^p$ and a penalty parameter $\alpha_0 > 0$.
for $k = 0, 1, 2, \ldots$ **do**
2 Choose $\varepsilon_k \geq 0$ and compute the new iterate x^k as an inexact solution of the subproblem (11.5), i.e., calculate x^k such that:

$$\|\nabla_x L_{\alpha_k}(x^k, \mu^k)\| \leq \varepsilon_k.$$

3 STOP, if $\nabla_x L_{\alpha_k}(x^k, \mu^k) = 0$ and $h(x^k) = 0$.
4 Set $\mu^{k+1} = \mu^k + \alpha_k \nabla_\mu L_{\alpha_k}(x^k, \mu^k)$ and choose $\alpha_{k+1} > 0$.

basic convergence result similar to Theorem 11.2 for the augmented-Lagrange method.

Lemma 11.3: Convergence of the Augmented-Lagrange Method

Let the sequences $(x^k)_k$, $(\mu^k)_k$ be generated by Algorithm 11.2 and let $x^* \in \mathbb{R}^n$ be a regular point. Suppose that $(x^k)_k$ converges to x^*. Furthermore, assume that there exists $\bar{\alpha}$ with $\alpha_k \geq \bar{\alpha}$ for all $k \in \mathbb{N}$ and $\varepsilon_k \to 0$ for $k \to \infty$. Then $(x^k, \mu^k)_k$ converges to a KKT-pair of (11.4).

Proof. Since the LICQ is satisfied at x^* and $x^k \to x^*$ for $k \to \infty$, we can infer that $\nabla h(x^k)$ has full column rank for all k sufficiently large. It holds that

$$\begin{aligned}\nabla_x L_{\alpha_k}(x^k, \mu^k) = \nabla_x L(x^k, \mu^k) + \alpha_k \nabla h(x^k) h(x^k) &= \nabla f(x^k) + \nabla h(x^k)(\mu^k + \alpha_k h(x^k)) \\ &= \nabla f(x^k) + \nabla h(x^k)\mu^{k+1}.\end{aligned}$$

Consequently, for k sufficiently large, we obtain:

$$\mu^{k+1} = (\nabla h(x^k)^\top \nabla h(x^k))^{-1} \nabla h(x^k)^\top (\nabla_x L_{\alpha_k}(x^k, \mu^k) - \nabla f(x^k)).$$

Now, from $\varepsilon_k \to 0$ and step 2 of Algorithm 11.2, it follows

$$\nabla_x L_{\alpha_k}(x^k, \mu^k) \to 0 \quad \text{for} \quad k \to \infty.$$

Since ∇f and ∇h are continuous, the latter observation implies

$$\mu^{k+1} \to -(\nabla h(x^*)^\top \nabla h(x^*))^{-1} \nabla h(x^*)^\top \nabla f(x^*) =: \mu^* \quad \text{as} \quad k \to \infty.$$

Moreover, it follows

$$\bar{\alpha}\|h(x^k)\| \leq \alpha_k \|h(x^k)\| = \|\mu^{k+1} - \mu^k\| \to 0 \quad \text{as} \quad k \to \infty$$

and

$$\nabla_x L(x^k, \mu^k) = \nabla L_{\alpha_k}(x^k, \mu^k) - \nabla h(x^k)(\alpha_k h(x^k)) \to 0 \quad k \to \infty.$$

Together, this establishes $(x^k, \mu^k) \to (x^*, \mu^*)$ and

$$\nabla_x L(x^*, \mu^*) = 0, \qquad h(x^*) = 0.$$

Thus, (x^*, μ^*) is a KKT point of (11.4), as desired. ■

Lemma 11.3 requires convergence of $(x^k)_k$ as prerequisite which is typically a rather strong assumption. Next, we give a different motivation and justification of the augmented-Lagrange method presented in Algorithm 11.2. Specifically, we show that when we have knowledge of the exact multiplier μ^*, a solution x^* of (11.4) is also a strict local minimizer of $L_\alpha(x, \mu^*)$ for all α sufficiently large. This suggests that we can obtain a good estimate of x^* by minimizing the augmented Lagrangian $L_\alpha(x, \mu)$ if μ is a reasonable estimate of μ^*. Different from the quadratic penalty method this property does not require α to diverge to infinity which is one of the main advantages of the augmented-Lagrange method.

Lemma 11.4: The Augmented Lagrangian and Local Solutions

Let (x^*, μ^*) be a KKT-pair of problem (11.4) and suppose that the second-order sufficient conditions are satisfied at (x^*, μ^*). Then there exists $\hat{\alpha} > 0$ such that x^* is a strict local minimum of the augmented-Lagrange function $x \mapsto L_\alpha(x, \mu^*)$ for all $\alpha \geq \hat{\alpha}$.

Proof. First, we show the following auxiliary result: Let $Q = Q^\top \in \mathbb{R}^{n\times n}$ and $H \in \mathbb{R}^{p\times n}$ be given. Suppose that for all $d \in \operatorname{null} H \setminus \{0\}$ it holds that $d^\top Q d > 0$. Then for all α sufficiently large the matrix $Q + \alpha H^\top H$ is positive definite.

Let us first prove this additional result. Assume that for every $k \in \mathbb{N}$ there exists $d^k \in \mathbb{R}^n$ with $\|d^k\|_2 = 1$ such that

$$0 \geq (d^k)^\top (Q + kH^\top H) d^k = (d^k)^\top Q d^k + k\|Hd^k\|_2^2. \tag{11.7}$$

Since the set $S := \{x \in \mathbb{R}^n : \|x\| = 1\} \supset (d^k)_{k\in\mathbb{N}}$ is compact, there exists a convergent subsequence $(d^{k_\ell})_\ell \to \bar{d} \in S$. Hence, inequality (11.7) implies

$$\begin{aligned}\|H\bar{d}\|^2 = \lim_{\ell\to\infty} \|Hd^{k_\ell}\|^2 &\leq \lim_{\ell\to\infty} -\frac{1}{k_\ell}(d^{k_\ell})^\top Q d^{k_\ell} \\ &\leq \lim_{\ell\to\infty} \frac{\max\{|\lambda_{\min}(Q)|, |\lambda_{\max}(Q)|\}}{k_\ell}\|d^{k_\ell}\|_2^2 = 0,\end{aligned}$$

where $\lambda_{\min}(Q)$ and $\lambda_{\max}(Q)$ denote the minimum and maximum eigenvalue of Q. Thus, we have $H\bar{d} = 0$, $\|\bar{d}\| = 1$, and $\bar{d} \in \operatorname{null} H \setminus \{0\}$. Now, from (11.7), it follows

$$(\bar{d})^\top Q \bar{d} = \lim_{\ell\to\infty} (d^{k_\ell})^\top (Q) d^{k_\ell} \leq \lim_{\ell\to\infty} -k_\ell \|Hd^{k_\ell}\|^2 \leq 0.$$

This contradicts the assumption $d^\top Q d > 0$ for all $d \in \operatorname{null} H \setminus \{0\}$. Consequently, $Q + \alpha H^\top H$ is positive definite for α sufficiently large.

Let us now continue with the proof of Lemma 11.4. Obviously, x^* is also a local solution of

problem (11.4) in this situation. Since (x^*, μ^*) is a KKT pair of (11.4), we have

$$h(x^*) = 0 \quad \text{and} \quad \nabla_x L_\alpha(x^*, \mu^*) = \nabla_x L(x^*, \mu^*) + \alpha \nabla h(x^*) h(x^*) = 0.$$

Thus, it holds that:

$$\begin{aligned}\nabla^2_{xx} L_\alpha(x^*, \mu^*) &= \nabla^2_{xx} L(x^*, \mu^*) + \alpha \sum_{i=1}^{p} h_i(x^*) \cdot \nabla^2 h_i(x^*) + \alpha \nabla h(x^*) \nabla h(x^*)^\top \\ &= \nabla^2_{xx} L(x^*, \mu^*) + \alpha \nabla h(x^*) \nabla h(x^*)^\top.\end{aligned}$$

Moreover, the second-order sufficient conditions for (11.4) imply

$$d^\top \nabla^2_{xx} L(x^*, \mu^*) d > 0 \quad \forall\, d \in \mathcal{C}(x^*) \setminus \{0\} = \text{null } \nabla h(x^*)^\top \setminus \{0\}.$$

Hence, setting $Q = \nabla^2_{xx} L(x^*, \mu^*)$ and $H = \nabla h(x^*)^\top$, the auxiliary result is applicable and there exists $\hat{\alpha} > 0$ such that $\nabla^2_{xx} L_\alpha(x^*, \mu^*)$ is positive definite for all $\alpha \geq \hat{\alpha}$. Consequently, the second-order sufficient conditions for the unconstrained problem $\min_x L_\alpha(x, \mu^*)$ are satisfied and x^* is a strictly local solution of $\min_x L_\alpha(x, \mu^*)$ for all $\alpha \geq \hat{\alpha}$. ■

We finally state several convergence properties of the augmented-Lagrange method when μ does not coincide with the optimal multiplier μ^*.

Theorem 11.5: General Convergence Properties

Suppose that the conditions in Lemma 11.4 are satisfied at x^* and μ^* and let $\hat{\alpha} > 0$ be given as specified in Lemma 11.4. Let us further assume that the LICQ holds at x^*. Then there exist positive scalars δ, ε and M such that following claims hold:

(i) For all μ^k and α_k with $\|\mu^k - \mu^*\| \leq \alpha_k \delta$ and $\alpha_k \geq \hat{\alpha}$, the problem

$$\min_x \; L_{\alpha_k}(x, \mu^k) \quad \text{s.t.} \quad \|x - x^*\| \leq \varepsilon$$

has a unique solution x^k. Moreover, we have

$$\|x^k - x^*\| \leq \frac{M}{\alpha_k} \|\mu^k - \mu^*\| \quad \text{and} \quad \|\mu^{k+1} - \mu^*\| \leq \frac{M}{\alpha_k} \|\mu^k - \mu^*\|,$$

where the multiplier μ^{k+1} is updated as specified in Algorithm 11.2.

(ii) For all μ^k and α_k satisfying the condition stated in (i), the matrix $\nabla^2_{xx} L_{\alpha_k}(x^k, \mu^k)$ is positive definite and the gradients $\nabla h_i(x^k)$, $i = 1, ..., p$ are linearly independent.

This theorem illustrates several main features of the augmented-Lagrange approach. The estimate in part (i) indicates that x^k will be close to x^* if μ^k is accurate or if the penalty parameter α_k is large. Hence, this observation gives us two ways of improving the accuracy of the iterates x^k. (In the quadratic penalty method we only have the option to increase α_k to achieve higher accuracy). The theorem also shows that the multiplier $(\lambda_k)_k$ converge q-

Algorithm 11.3: The Local Lagrange-Newton Method

Initialization: Choose an initial point $x^0 \in \mathbb{R}^n$ and $\mu^0 \in \mathbb{R}^p$.
for $k = 0, 1, 2, \ldots$ **do**
  STOP, if $\nabla_x L(x^k, \mu^k) = 0$ and $h(x^k) = 0$. ((x^k, μ^k) is a KKT-pair).
  Compute $d^k = (d_x^k, d_\mu^k)$ by solving the Newton equation (11.9).
  Set $x^{k+1} = x^k + d_x^k$ and $\mu^{k+1} = \mu^k + d_\mu^k$.

linearly if α_k is sufficiently large. Thus, in this situation, we can expect r-linear convergence of the iterates $(x^k)_k$. A proof of Theorem 11.5 can be found in [3, Proposition 2.4].

11.3. Lagrange Newton-Methods

Next, we present a family of approaches that utilizes and exploits higher-order information to achieve faster rates of convergence. The principle idea of Lagrange Newton-methods is to interpret the KKT-conditions as a system of nonlinear equations to apply Newton's method to solve this system.

11.3.1. The Lagrange Newton-Method for Equality Constraints

We continue to discuss the equality constrained problem (11.4). Suppose that x^* is a local solution of (11.4) and let a CQ be satisfied at x^* (h is affine-linear or $\nabla h(x^*)$ has full column rank). Then the KKT-conditions are satisfied and there exists $\mu^* \in \mathbb{R}^p$ such that

$$\nabla_x L(x^*, \mu^*) = 0 \quad \text{and} \quad h(x^*) = 0.$$

This defines a nonlinear system of equations with $n+p$ unknowns and $n+p$ equations. Consequently, we can apply Newton's method to determine a solution of the nonlinear equation $F : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n \times \mathbb{R}^p$,

$$F(x, \mu) := \begin{pmatrix} \nabla_x L(x, \mu) \\ h(x) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{11.8}$$

We now assume that f and h are twice continuously differentiable. Let (x^k, μ^k) denote the current iterate, then Newton's step d^k for (11.8) is given by $DF(x^k, \mu^k)d^k = -F(x^k, \mu^k)$. In particular, we can calculate

$$DF(x, \mu) = \begin{pmatrix} \nabla_{xx}^2 L(x, \mu) & \nabla_{x\mu}^2 L(x, \mu) \\ \nabla h(x)^\top & 0 \end{pmatrix} = \begin{pmatrix} \nabla_{xx}^2 L(x, \mu) & \nabla h(x) \\ \nabla h(x)^\top & 0 \end{pmatrix}$$

and the complete Newton update simplifies to:

$$\begin{pmatrix} \nabla_{xx}^2 L(x^k, \mu^k) & \nabla h(x^k) \\ \nabla h(x^k)^\top & 0 \end{pmatrix} \begin{pmatrix} d_x^k \\ d_\mu^k \end{pmatrix} = \begin{pmatrix} -\nabla_x L(x^k, \mu^k) \\ -h(x^k) \end{pmatrix}, \quad \begin{pmatrix} x^{k+1} \\ \mu^{k+1} \end{pmatrix} = \begin{pmatrix} x^k \\ \mu^k \end{pmatrix} + \begin{pmatrix} d_x^k \\ d_\mu^k \end{pmatrix}. \tag{11.9}$$

Our discussion motivates the local Lagrange-Newton method Algorithm 11.3.

Since this approach does not contain any globalization strategy, we can only expect it to work in a neighborhood of a suitable KKT-pair. The local convergence analysis basically coincides with the one in section 6. As before, the invertibility of the matrix $DF(x,\mu)$ will play a key ingredient to ensure convergence. We will now connect invertibility of $DF(x,\mu)$ to our second order optimality conditions.

Lemma 11.6: Invertibility via Second-Order Condition

Let $f:\mathbb{R}^n\to\mathbb{R}$ and $h:\mathbb{R}^n\to\mathbb{R}^p$ be twice continuously differentiable and let $x\in\mathbb{R}^n$ and $\mu\in\mathbb{R}^p$ be arbitrary. Suppose that $\nabla h(x)$ has full column rank and that we have

$$d^\top\nabla^2_{xx}L(x,\mu)d>0\quad\forall\, d\in\mathbb{R}^n\backslash\{0\}\quad\text{with}\quad\nabla h(x)^\top d=0.$$

Then the matrix

$$DF(x,\mu)=\begin{pmatrix}\nabla^2_{xx}L(x,\mu) & \nabla h(x)\\ \nabla h(x)^\top & 0\end{pmatrix}$$

is invertible.

Proof. Let v,w be given with $DF(x,\mu)(v^\top,w^\top)^\top=0$. We need to show that this implies $v=0$ and $w=0$. From the second block equation, we can infer $\nabla h(x)^\top v=0$. Multiplying the first block equation with $v^\top$, we obtain

$$0=v^\top\nabla^2_{xx}L(x,\mu)v+v^\top\nabla h(x)w=v^\top\nabla^2_{xx}L(x,\mu)v.$$

Due to $\nabla h(x)^\top v=0$, the second-order condition stated in the lemma is applicable and it follows $v=0$. But then, we have $\nabla h(x)w=0$. Since $\nabla h(x)$ has full column rank, this implies $w=0$. ■

We can formulate the following convergence result:

Theorem 11.7: Local Convergence of the Lagrange-Newton Method

Let f and h be twice continuously differentiable and let (x^*,μ^*) be a KKT-pair satisfying the sufficient second-order optimality conditions:

$$d^\top\nabla^2_{xx}L(x^*,\mu^*)d>0\quad\forall\, d\in\mathbb{R}^n\backslash\{0\}\quad\text{with}\quad\nabla h(x^*)^\top d=0.$$

Furthermore, suppose that the LICQ holds at x^*. Then there exists $\varepsilon>0$ such that for all initial points $(x^0,\mu^0)\in B_\varepsilon(x^*,\mu^*)$, Algorithm 11.3 either terminates after finitely many iterations with $(x^k,\mu^k)=(x^0,\mu^0)$ or it generates a sequence $(x^k,\mu^k)_k$ that converges q-superlinearly to (x^*,μ^*):

$$\|(x^{k+1}-x^*,\mu^{k+1}-\mu^*)\|=o(\|(x^k-x^*,\mu^k-\mu^*)\|)\quad k\to\infty.$$

If the Hessians $\nabla^2 f$ and $\nabla^2 h_i$ are Lipschitz continuous on $B_\varepsilon(x^*)$, then the rate of convergence is q-quadratic.

Proof. Thanks to Lemma 11.6, the matrices $DF(x,\mu)$ are boundedly invertible close to

Algorithm 11.4: The Globalized Lagrange-Newton Method

Initialization: Choose an initial point $x^0 \in \mathbb{R}^n$ and $\mu^0 \in \mathbb{R}^p$, $\beta > 0$, and $p > 2$.
for $k = 0, 1, 2, \ldots$ **do**
1. STOP, if $\|\nabla g(x^k, \mu^k)\| = 0$.
2. Compute $d^k = (d_x^k, d_\mu^k)$ by solving the Lagrange Newton-equation (11.9). If this is not possible or if
$$\nabla g(x^k, \mu^k)^\top d^k > -\beta \|d^k\|^p,$$
then set $d^k = -\nabla g(x^k, \mu^k)$.
3. Perform backtracking on g to calculate a step size α_k. Set $x^{k+1} = x^k + \alpha_k d_x^k$ and $\mu^{k+1} = \mu^k + \alpha_k d_\mu^k$.

(x^*, μ^*). We can then mimic the proof of Theorem 6.1 to show local convergence. (See also Appendix B for further details). For q-quadratic convergence, we only need to verify that DF is Lipschitz continuous on $B_\varepsilon(x^*, \mu^*)$. We have

$$\|DF(x,\mu) - DF(x',\mu')\| \leq \|\nabla^2_{xx} L(x,\mu) - \nabla^2_{xx} L(x',\mu')\| + 2\|\nabla h(x) - \nabla h(x')\|.$$

Since ∇h is continuously differentiable, the mapping $x \mapsto \nabla h(x)$ is locally Lipschitz continuous. Moreover, we obtain

$$\|\nabla^2_{xx} L(x,\mu) - \nabla^2_{xx} L(x',\mu')\| \leq \|\nabla^2 f(x) - \nabla^2 f(x')\| + \sum_{j=1}^{p} \|\mu_j \nabla^2 h_j(x) - \mu_j' \nabla^2 h_j(x')\|$$
$$\|\mu_j \nabla^2 h_j(x) - \mu_j' \nabla^2 h_j(x')\| \leq |\mu_j| \|\nabla^2 h_j(x) - \nabla^2 h_j(x')\| + |\mu_j - \mu_j'| \|\nabla^2 h_j(x')\|.$$

This establishes local Lipschitz continuity of the derivative mapping DF. ■

11.3.2. Globalizing the Lagrange Newton-Method

In order to globalize the Lagrange-Newton method, we consider the following *merit function problem*:

$$\min_{x,\mu} \; g(x,\mu) = \frac{1}{2}\|F(x,\mu)\|^2. \tag{11.10}$$

Clearly, (x^*, μ^*) is a KKT-pair of problem (11.4) if and only if $g(x^*, \mu^*) = 0$, i.e., if (x^*, μ^*) is a global solution of (11.10). Moreover, using $\nabla g(x,\mu) = DF(x,\mu)^\top F(x,\mu)$, it follows that every stationary point (x^*, μ^*) of g with $\nabla g(x^*, \mu^*) = 0$ needs to satisfy $F(x^*, \mu^*) = 0$ if $DF(x^*, \mu^*)$ is invertible. Finally, let d be a Lagrange Newton-direction with $DF(x,\mu)d = -F(x,\mu)$, then we have:

$$\nabla g(x,\mu)^\top d = F(x,\mu)^\top DF(x,\mu) d = -\|F(x,\mu)\|^2 < 0. \tag{11.11}$$

Hence, the Lagrange Newton-direction is always a descent direction of the merit function g. This shows that there is a very natural connection between finding a solution of the nonlinear system $F(x,\mu)=0$ and minimizing the merit function g. The full method is summarized in Algorithm 11.4 and we now present a global-local result for Algorithm 11.4.

Theorem 11.8: Global-Local Result for the Globalized LN-Method

Let $f:\mathbb{R}^n\to\mathbb{R}$ and $h:\mathbb{R}^n\to\mathbb{R}^p$ be twice continuously differentiable and let $(x^k,\mu^k)_k$ be generated by Algorithm 11.4. Then, every accumulation point of $(x^k,\mu^k)_k$ is a stationary point of g.

Additionally, if we have $\gamma\in(0,\frac{1}{2})$, $s=1$, and if there exists an accumulation point (x^*,μ^*) of $(x^k,\mu^k)_k$ such that x^* is regular and the second-order sufficient conditions

$$d^\top\nabla^2_{xx}L(x^*,\mu^*)d>0\quad\forall\, d\in\mathbb{R}^n\backslash\{0\}\quad\text{with}\quad\nabla h(x^*)^\top d=0,$$

are satisfied, then it follows:

(i) The point (x^*,μ^*) is a KKT-pair and x^* is a strict local minimum of problem (11.4).

(ii) The whole sequence $(x^k,\mu^k)_k$ converges to (x^*,μ^*).

(iii) The algorithm turns into the pure Lagrange Newton-method, i.e., there is $K\in\mathbb{N}$ such that for all $k\geq K$ we have $\alpha_k=1$ and the LN-direction is accepted. Moreover, $(x^k,\mu^k)_k$ converges q-superlinearly to (x^*,μ^*). If $\nabla^2 f$ and $\nabla^2 h_i$, $i=1,...,p$ are Lipschitz continuous in a neighborhood of x^*, the rate of convergence is q-quadratic.

Proof. Let us set $\zeta^k=(x^k,\mu^k)$. Using the Armijo condition, we have

$$g(\zeta^{k+1})-g(\zeta^k)\leq\gamma\alpha_k\cdot\nabla g(\zeta^k)^\top d^k\begin{cases}=-\gamma\alpha_k\|\nabla g(\zeta^k)\|^2 & \text{if a gradient step is performed}\\ =-\gamma\alpha_k\|F(\zeta^k)\|^2 & \text{if a LN-step is performed.}\end{cases}$$

This shows that the sequence $(g(\zeta^k))_k$ is monotonically decreasing and thus, it converges to some $\xi\in\mathbb{R}\cup\{-\infty\}$. As usual, if $\zeta^*=(x^*,\mu^*)$ is an accumulation point of $(\zeta^k)_k$, then we have $g(\zeta^k)\to\xi=g(\zeta^*)\in\mathbb{R}$. Summing the last estimate, this establishes

$$-\sum_{k=0}^{\infty}\alpha_k\nabla g(\zeta^k)^\top d^k<\infty.$$

Let $(\zeta^{k_\ell})_\ell$ be a subsequence of $(\zeta^k)_k$ converging to ζ^*. Then, by the continuity of DF, there exists $C>0$ such that $\|\nabla g(\zeta^{k_\ell})\|\leq\|DF(\zeta^{k_\ell})\|\|F(\zeta^{k_\ell})\|\leq C\|F(\zeta^{k_\ell})\|$. Combining this observation with (11.11), we can mimic the proof of Theorem 5.11 or Theorem 6.4 to show that $\nabla g(\zeta^{k_\ell})\to\nabla g(\zeta^*)=0$.

We now continue to verify part (i)–(iii) of Theorem 11.8. Lemma 11.6 implies that the matrix $DF(x^*,\mu^*)=DF(\zeta^*)$ is invertible. Thus, using $0=\nabla g(\zeta^*)=DF(\zeta^*)F(\zeta^*)$, we can infer $F(\zeta^*)=0$ and $\zeta^*=(x^*,\mu^*)$ is a KKT-point of problem (11.4). Utilizing the second-order sufficient conditions, see Theorem 9.29, x^* also has to be a strict local minimum of (11.4). As shown in Lemma B.1, the invertibility of $DF(\zeta^*)$ implies that $DF(\zeta)$ is boundedly

invertible in a neighborhood of ζ^*, i.e., there exist $\varepsilon_1 > 0$ and $C_F > 0$ such that

$$\|DF(\zeta)^{-1}\| \leq C_F \quad \forall\, \zeta \in B_{\varepsilon_1}(\zeta^*).$$

Let $(\zeta^{k_\ell})_\ell$ be an arbitrary subsequence of $(\zeta^k)_k$ converging to ζ^*, then it follows:

$$\|\zeta^{k_\ell+1} - \zeta^{k_\ell}\| = \alpha_{k_\ell}\|d^{k_\ell}\| \leq \begin{cases} \|DF(\zeta^k)^{-1}F(\zeta^k)\| \leq C_F^2\|\nabla g(\zeta^k)\| & \text{if } d^{k_\ell} \text{ is a LN-step} \\ \|\nabla g(\zeta^k)\| & \text{if } d^{k_\ell} \text{ is a gradient step} \end{cases}$$
$$\to 0 \quad \ell \to 0.$$

By Lemma B.2, the point ζ^* is an isolated solution of the equation $F(\zeta) = 0$ and there exists $\varepsilon_2 \in (0, \varepsilon_1]$ and η such that $\|F(\zeta)\| \geq \eta\|\zeta - \zeta^*\|$ for all $\zeta \in B_{\varepsilon_2}(\zeta^*)$. But this implies $\|\nabla g(\zeta)\| \geq (\eta/C_F)\|\zeta - \zeta^*\|$ for all $\zeta \in B_{\varepsilon_2}(\zeta^*)$ and thus, ζ^* is also an isolated solution of the equation $\nabla g(\zeta) = 0$. Since every accumulation point of $(\zeta^k)_k$ is a stationary point of the mapping g, this establishes that ζ^* is an isolated accumulation point of the sequence $(\zeta^k)_k$. Consequently, by Lemma 5.20, the whole sequence $(\zeta^k)_k$ has to converge to ζ^*.

Let us set $s^k = -DF(\zeta^k)^{-1}F(\zeta^k)$. By the continuity of F, we have $F(\zeta^k) \to 0$ and hence, it follows $\beta\|s^k\|^{p-2} \leq C_F^{p-2}\beta\|F(\zeta^k)\|^{p-2} \leq C_F^{-2}$ for all k sufficiently large. This implies

$$\begin{aligned} -\nabla g(\zeta^k)^\top s^k &= (s^k)^\top DF(\zeta^k)^2 s^k \\ &= \|DF(\zeta^k)s^k\|^2 \geq \|s^k\|^2 \cdot \min_{s\neq 0} \frac{\|DF(\zeta^k)s\|^2}{\|s\|^2} \geq C_F^{-2}\|s^k\|^2 \geq \beta\|s^k\|^p. \end{aligned}$$

for all k sufficiently, and consequently, the full Lagrange Newton-step $d^k = s^k$ will eventually always be accepted. Next, in order to show transition to fast local convergence, we need to utilize a smoothness property of the mapping g. Notice that formally, the derivatives of g are given by

$$\nabla g(\zeta) = \sum_{i=1}^{n+p} F_i(\zeta)\nabla F_i(\zeta), \quad \nabla^2 g(\zeta) = \sum_{i=1}^{n+p} \nabla F_i(\zeta)\nabla F_i(\zeta)^\top + F_i(\zeta)\nabla^2 F_i(\zeta).$$

However, this requires third-order differentiability of f and h. Nonetheless, since $F(\zeta^*) = 0$ vanishes at ζ^*, it can be shown that g is twice differentiable at ζ^* (more specifically, the gradient ∇g is strictly differentiable at ζ^*) with $\nabla^2 g(\zeta^*) = DF(\zeta^*)^\top DF(\zeta^*) = DF(\zeta^*)^2$ and Taylor's theorem is applicable:

$$\begin{aligned} g(\zeta^k + s^k) &= g(\zeta^*) + \nabla g(\zeta^*)^\top(\zeta^k + s^k - \zeta^*) + \frac{1}{2}(\zeta^k + s^k - \zeta^*)^\top \nabla^2 g(\zeta^*)(\zeta^k + s^k - \zeta^*) \\ &\quad + o(\|\zeta^k + s^k - \zeta^*\|^2), \\ g(\zeta^k) &= g(\zeta^*) + \nabla g(\zeta^*)^\top(\zeta^k - \zeta^*) + \frac{1}{2}(\zeta^k - \zeta^*)^\top \nabla^2 g(\zeta^*)(\zeta^k - \zeta^*) + o(\|\zeta^k - \zeta^*\|^2) \end{aligned}$$

as $k \to \infty$. Moreover, by the continuity of DF and differentiability of F, we have

$$\begin{aligned}\|\zeta^k + s^k - \zeta^*\| &\leq \|DF(\zeta^k)^{-1}\|\|F(\zeta^k) - DF(\zeta^k)(\zeta^k - \zeta^*)\| \\ &\leq C_F[\|F(\zeta^k) - F(\zeta^*) - DF(\zeta^*)(\zeta^k - \zeta^*)\| + \|DF(\zeta^k) - DF(\zeta^*)\|\|\zeta^k - \zeta^*\|] \\ &= o(\|\zeta^k - \zeta^*\|) = o(\|F(\zeta^k)\|).\end{aligned}$$

In the last step, we used the estimate $\|F(\zeta)\| \geq \eta\|\zeta - \zeta^*\|$ for $\zeta \in B_{\varepsilon_2}(\zeta^*)$. Consequently, it follows

$$\begin{aligned}&g(\zeta^k + s^k) - g(\zeta^k) - \gamma\nabla g(\zeta^k)^\top s^k \\ &\quad = -\left[\gamma - \frac{1}{2}\right]\nabla g(\zeta^k)^\top s^k + \frac{1}{2}\|DF(\zeta^k)s^k\|^2 - \frac{1}{2}\|DF(\zeta^*)(\zeta^k - \zeta^*)\|^2 + o(\|F(\zeta^k)\|^2) \\ &\quad = \left[\gamma - \frac{1}{2}\right]\|F(\zeta^k)\|^2 + \frac{1}{2}[\|DF(\zeta^k)(\zeta^k - \zeta^*)\|^2 - \|DF(\zeta^*)(\zeta^k - \zeta^*)\|^2] + o(\|F(\zeta^k)\|^2) \\ &\quad \leq \left[\gamma - \frac{1}{2}\right]\|F(\zeta^k)\|^2 + \frac{1}{2}\|DF(\zeta^k)^2 - DF(\zeta^*)^2\|\|\zeta^k - \zeta^*\|^2 + o(\|F(\zeta^k)\|^2).\end{aligned}$$

Hence, due to $\gamma < \frac{1}{2}$ and $o(\|\zeta^k - \zeta^*\|) = o(\|F(\zeta^k)\|)$, we can infer that the full step $\zeta^k + s^k$ with $\alpha_k = 1$ satisfies the Armijo condition for all k sufficiently large. Altogether, we have $d^k = s^k$ for all k sufficiently large and the algorithm eventually turns into a pure Lagrange Newton-method. Q-superlinear and q-quadratic convergence follow from Theorem 11.7. ■

11.3.3. Extensions to Inequality Constraints

It possible to extend the Lagrange Newton-method to general nonlinear programs by using so-called *nonlinear complementarity problem* functions (NCP-functions). Here, a function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is a NCP-function if we have

$$\varphi(a, b) = 0 \quad \iff \quad a \geq 0, \quad b \geq 0, \quad ab = 0.$$

NCP-functions allow to transform the complementarity conditions

$$g_i(x) \leq 0, \quad \lambda_i \geq 0, \quad \lambda_i g_i(x) = 0, \quad i = 1, ..., m,$$

into a set of nonlinear equations:

$$\varphi(-g_i(x), \lambda_i) = 0 \quad i = 1, ..., m.$$

The most common NCP-functions are $\varphi(a, b) = \min\{a, b\}$ and the *Fischer-Burmeister function* $\varphi_{\text{FB}}(a, b) := \sqrt{a^2 + b^2} - a - b$. Hence, setting

$$\Phi(x, \lambda) := (\varphi_{\text{FB}}(-g_1(x), \lambda_1), ..., \varphi_{\text{FB}}(-g_m(x), \lambda_m))^\top,$$

we can again rewrite the KKT-conditions as a system of nonlinear equations. More specifically, the tupel (x^*, λ^*, μ^*) is a KKT-triple if and only if

$$F(x^*, \lambda^*, \mu^*) = 0 \quad \text{where} \quad F(x, \lambda, \mu) = \begin{pmatrix} \nabla_x L(x, \lambda, \mu) \\ h(x) \\ \Phi(x, \lambda) \end{pmatrix}. \tag{11.12}$$

We continue with several remarks:

- We can again use $g(x, \lambda, \mu) := \frac{1}{2}\|F(x, \lambda, \mu)\|^2$ as merit function for globalization.
- We can now generally mimic the Lagrange Newton-method for equality constrained problems and apply Newton's method to solve (11.12).
- However, notice that the Fischer-Burmeister function $\varphi_{\texttt{FB}}$ is not differentiable at $(0, 0)$. This is particularly relevant if the generated iterates converge to a KKT-triple at which the strict complementarity condition is violated. Utilizing generalized derivatives, we can still perform (nonsmooth) Lagrange Newton-type steps – this will lead to so-called *semismooth Newton steps*.

11.3.4. Sequential Quadratic Programming

We can also reinterpret the Lagrange Newton-update for equality constrained problems. In particular, consider the quadratic problem:

$$\min_s \ \nabla f(x^k)^\top s + \frac{1}{2} s^\top B_k s \quad \text{s.t.} \quad h(x^k) + \nabla h(x^k)^\top s = 0,$$

where $B_k = \nabla^2_{xx} L(x^k, \mu^k)$.

Since the constraints are affine-linear, every local solution s^k of this quadratic program satisfies the KKT-conditons, i.e., there exists μ^k_{qp} with

$$\nabla f(x^k) + B_k s^k + \nabla h(x^k)\mu^k_{qp} = 0, \quad h(x^k) + \nabla h(x^k)^\top s^k = 0.$$

Consequently, setting $d^k_x = s^k$ and $d^k_\mu = \mu^k_{qp} - \mu^k$, we obtain:

$$\begin{aligned} B_k d^k_x + \nabla h(x^k) d^k_\mu &= -\nabla f(x^k) - \nabla h(x^k)\mu^k, \\ \nabla h(x^k)^\top d^k_x &= -h(x^k). \end{aligned}$$

Hence, performing a Lagrange Newton-step is equivalent to the following procedure:

- Calculate a KKT-pair (s^k, μ^k_{qp}) of the quadratic problem:

$$\min_s \ \nabla f(x^k)^\top s + \frac{1}{2} s^\top \nabla^2_{xx} L(x^k, \mu^k) s \quad \text{s.t.} \quad h(x^k) + \nabla h(x^k)^\top s = 0.$$

- Set $x^{k+1} = x^k + s^k$ and $\mu^{k+1} = \mu^k_{qp}$.

This approach defines one step of the *local SQP method* for equality constrained problems. It is also possible to incorporate inequality constraints. In this case the SQP subproblem is given by

$$\begin{aligned} \min_s \quad & \nabla f(x^k)^\top s + \frac{1}{2} s^\top \nabla^2_{xx} L(x^k, \lambda_k, \mu^k) s \\ \text{s.t.} \quad & g(x^k) + \nabla g(x^k)^\top s \leq 0 \\ & h(x^k) + \nabla h(x^k)^\top s = 0, \end{aligned} \tag{11.13}$$

i.e., we adjust the Hessian of the Lagrangian and add a linearized version of the inequality constraints $g(x) \leq 0$ to the quadratic problem. We continue with two final comments.

- Globalization mechanisms for the SQP-method typically are based on the ℓ_1-*penalty function* $P^1_\alpha(x) = f(x) + \alpha(\|(g(x))_+\|_1 + \|h(x)\|_1)$.
- The general quadratic subproblem (11.13) can be solved by specialized optimization algorithms, such as the *active set method*.

References. We note that subsection 11.1, subsection 11.3.1, and subsection 11.3.4 closely follow [9, Chapter 18 and 19]. Let us also refer to [5, Chapter 5] and [8, Chapter 17 and 18] for a similar coverage. More details and additional results on the augmented-Lagrange approaches presented in subsection 11.2 can be found in [5, Section 5.2] and [3].

12. Projected and Proximal Gradient Method

So far many of our theoretical results and optimization methods are based on the fundamental assumption that the underlying minimization problem is sufficiently smooth. In this section, we will investigate a class of nonsmooth optimization problems that can be expressed as

$$\min_x \ f(x) + \varphi(x) \quad \text{s.t.} \quad x \in \mathbb{R}^n, \tag{12.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a general smooth (possibly nonconvex) and $\varphi : \mathbb{R}^n \to (-\infty, \infty]$ is a convex but possibly not differentiable mapping. As we have already seen, problems of this form arise frequently in applications and are often referred to as *composite-type* or *convex composite* optimization problems. Typically, the mapping f is used as a loss model that depends on given data and φ corresponds to a specific *regularization* that promotes a certain structure. A very popular choice for φ is the weighted ℓ_1-norm

$$\varphi(x) = \mu\|x\|_1$$

which promotes *sparse* solutions, i.e., solutions $x^* \in \mathbb{R}^n$ where many components vanish and are zero. In the introductory part of the lecture we have already discussed and introduced several exemplary applications that fit the framework (12.1). For instance, possible applications are support vector machines, recommender systems, machine and deep learning problems, dictionary learning, image reconstruction problems and problems in computer vision.

We first briefly discuss problem with convex constraints and the projected gradient method which allows us to complement the algorithms discussed in the last section.

12.1. The Projected Gradient Method

12.1.1. Optimization Problems with Convex Constraints

We first consider constrained minimization problems of the form

$$\min_x \ f(x) \quad \text{s.t.} \quad x \in C, \tag{12.2}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function and $C \subset \mathbb{R}^n$ is a convex (and often closed) set. We can derive the following optimality condition. We now restate several results that were already shown in the exercises.

Theorem 12.1: First-Order Optimality

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on an open set that contains the convex set $C \subset \mathbb{R}^n$ and let $x^* \in C$ be a local minimizer of (12.2). Then, it holds that

$$\nabla f(x^*)^\top (x - x^*) \geq 0, \quad \forall\, x \in C. \tag{12.3}$$

Additionally, if f is convex, then $x^* \in C$ is global minimizer of (12.2) if and only if the condition (12.3) is satisfied.

As before we will call a point $x^* \in C$ with

$$\nabla f(x^*)^\top (x - x^*) \geq 0 \quad \forall\, x \in C,$$

a *stationary point* of problem (12.2). We continue with an important special case and application of Theorem 12.1.

Theorem 12.2: Projection Theorem

Let $C \subset \mathbb{R}^n$ be a closed and convex set.

(i) For every $y \in \mathbb{R}^n$, the constrained optimization problem

$$\min_x \ \frac{1}{2}\|x - y\|^2 \quad \text{s.t.} \quad x \in C,$$

has a unique global solution $x^* \in C$. This minimizer is called the **(Euclidean) projection** of y onto C and we write $x^* = \mathcal{P}_C(y)$.

(ii) A point x^* is the projection of y onto C, i.e., $x^* = \mathcal{P}_C(y)$, if and only if

$$(x^* - y)^\top (x - x^*) \geq 0, \quad \forall\, x \in C.$$

(iii) The mapping $\mathcal{P}_C : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with constant $L = 1$ and satisfies

$$\|\mathcal{P}_C(y) - \mathcal{P}_C(z)\|^2 \leq (\mathcal{P}_C(y) - \mathcal{P}_C(z))^\top (y - z), \quad \forall\, y, z \in \mathbb{R}^n.$$

(iv) Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on an open set containing C. Then, x^* is a stationary point of (12.2) if and only if

$$x^* - \mathcal{P}_C(x^* - \lambda \nabla f(x^*)) = 0 \quad \text{for any } \lambda > 0.$$

Depending on the structure of the set C, it might be hard to compute the projection $\mathcal{P}_C(y)$ since we basically need to solve a constrained optimization problem for each input $y \in \mathbb{R}^n$. However, in certain situations and applications it is possible to calculate $\mathcal{P}_C(y)$ explicitly and to derive closed-form solutions.

Example 12.3. Let $C = [a, b]$ with $a, b \in \mathbb{R}^n$ and $a \leq b$. Then, it holds that

$$\mathcal{P}_{[a,b]}(x) = \max\{\min\{x, b\}, a\}.$$

Example 12.4. Let $C = \{x \in \mathbb{R}^n : \|x\| \leq r\}$ for some $r > 0$. Then, it holds that

$$\mathcal{P}_C(x) = \frac{rx}{\max\{r, \|x\|\}} = \begin{cases} x & \text{if } \|x\| \leq r, \\ r \cdot \frac{x}{\|x\|} & \text{if } \|x\| \geq r. \end{cases}$$

Proof. In the case $\|x\| \leq r$, the projection obviously reduces to $\mathcal{P}_C(x) = x$. Let us now

assume $\|x\| > r$. Then, by the Cauchy-Schwarz inequality, it follows

$$\min_{y\in C} \frac{1}{2}\|x-y\|^2 = \min_{y\in C} \frac{1}{2}\|x\|^2 - x^\top y + \frac{1}{2}\|y\|^2 \geq \min_{y\in C} \frac{1}{2}(\|x\| - \|y\|)^2 = \frac{1}{2}(\|x\| - r)^2$$

and equality only holds if there exists $y \in C$ with $\|y\| = r$ and $y = \alpha x$ for some $\alpha \in \mathbb{R}$. However using the condition $\|y\| = r$, we can infer $\alpha = r/\|x\|$ and thus, $y = rx/\|x\|$. ■

Example 12.5. Let $C = \mathbb{S}^n_+$ be the set of symmetric, positive semidefinite matrices and let $X \in \mathbb{R}^{n\times n}$ be a symmetric matrix with eigenvalue decomposition $X = Q\,\mathrm{diag}(\lambda)Q^\top$, $\lambda \in \mathbb{R}^n$. Then, we have

$$\mathcal{P}_{\mathbb{S}^n_+}(X) := \operatorname*{arg\,min}_{Y\in\mathbb{S}^n_+} \frac{1}{2}\|X-Y\|_F^2 = Q\,\mathrm{diag}(\max\{0,\lambda\})Q^\top.$$

Proof. The trace of the product of two matrices is invariant under permutation, i.e., for $A, B \in \mathbb{R}^{n\times n}$ we have $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. Together with $QQ^\top = I$, this yields

$$\begin{aligned}\|X-Y\|_F^2 &= \mathrm{tr}((X-Y)^\top(X-Y)) = \mathrm{tr}(QQ^\top(X-Y)^\top QQ^\top(X-Y)) \\ &= \mathrm{tr}([Q^\top(X-Y)Q]^\top[Q^\top(X-Y)Q]) = \|\,\mathrm{diag}(\lambda) - Q^\top YQ\|_F^2.\end{aligned}$$

This term is only minimized if $Q^\top YQ$ is also a diagonal matrix, i.e., we have $Q^\top YQ = \mathrm{diag}(y)$. The constraint $Y \in \mathbb{S}^n_+$ then implies $y \geq 0$ and we obtain

$$\min_{Y\in\mathbb{S}^n_+} \frac{1}{2}\|X-Y\|_F^2 = \min_{y\geq 0} \frac{1}{2}\|\,\mathrm{diag}(\lambda) - \mathrm{diag}(y)\|_F^2 = \min_{y\geq 0} \frac{1}{2}\|\lambda - y\|^2.$$

By Example 12.3, the unique minimizer of the latter problem is given by the projection $y = \mathcal{P}_{\mathbb{R}^n_+}(\lambda) = \max\{0,\lambda\}$ and hence, it follows $Y = Q\,\mathrm{diag}(\max\{0,\lambda\})Q^\top$. ■

12.1.2. The Projected Gradient Method

We now investigate a projection based, first order approach that is designed to solve the constrained problem (12.2). Here, we assume that the set $C \subseteq \mathbb{R}^n$ is nonempty, convex, and closed and f is continuously differentiable on an open neighborhood of C. As in the standard gradient method our basic idea is to perform a gradient descent step of the form

$$x^k - \lambda_k \nabla f(x^k),$$

where $\lambda_k > 0$ is a step size. However, setting $x^{k+1} = x^k - \lambda_k\nabla f(x^k)$ might generate iterates that are not feasible, i.e., we might have $x^{k+1} \notin C$ and the method might converge to an infeasible point. In order to resolve this feasibility issue, we project the step $x^k - \lambda_k\nabla f(x^k)$ back onto C and perform the update

$$x^{k+1} = \mathcal{P}_C(x^k - \lambda_k\nabla f(x^k)). \tag{12.4}$$

Algorithm 12.1: A Projected Gradient Method

Initialization: Choose an initial point $x^0 \in C$ and $\sigma, \gamma \in (0,1)$.
for $k = 0, 1, 2, \ldots$ **do**
  Select $\lambda_k > 0$ and compute $\nabla f(x^k)$ and the new direction $d^k = -F_{\lambda_k}(x^k)$.
  If $\|d^k\| \leq \lambda_k \varepsilon$, then STOP and x^k is the output.
  Choose a maximal step size $\alpha_k \in \{1, \sigma, \sigma^2, \ldots\}$ that satisfies the Armijo condition

$$f(x^k + \alpha_k d^k) - f(x^k) \leq \gamma \alpha_k \cdot \nabla f(x^k)^\top d^k.$$

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

In principle, the iterative scheme (12.4) can be used directly to solve (12.2). However, this has one drawback. Currently, it is not clear how to choose $\lambda_k > 0$. If λ_k is determined by a line search-type procedure then each adjustment of λ_k requires to reevaluate the projection $\mathcal{P}_C(x^k - \lambda_k \nabla f(x^k))$. This can be expensive, if the set C and the projection $\mathcal{P}_C$ are complicated.

We will now discuss a (very similar) algorithm that only needs to compute one single projection onto C in each iteration. For given $\lambda > 0$, we define the mapping $F_\lambda : \mathbb{R}^n \to \mathbb{R}^n$, $F_\lambda(x) := x - \mathcal{P}_C(x - \lambda \nabla f(x))$. We can show the following result.

Lemma 12.6: Generating Descent Directions

Let C be a nonempty, convex, and closed set and let $x \in C$ and $\lambda > 0$ be given. If x is not a stationary point of problem (12.2), then the direction $d := -F_\lambda(x)$ is a descent direction and it holds that

$$\nabla f(x)^\top d \leq -\lambda^{-1} \|d\|^2 < 0.$$

In addition, if the iterate $x^k \in C$ is feasible, then the convexity of C implies

$$x^k + \alpha d^k = x^k - \alpha F_{\lambda_k}(x^k) = (1-\alpha)x^k + \alpha \mathcal{P}(x^k - \lambda_k \nabla f(x^k)) \in C$$

for all $\alpha \in [0,1]$, i.e. the step $x^k + \alpha d^k$ maintains feasibility for all $\alpha \in [0,1]$! The idea of the projected gradient method is to utilize d^k as a descent direction – with fixed $\lambda_k > 0$ – and to determine a suitable step size α_k for the step $x^k + \alpha_k d^k$ via backtracking. The full algorithm is presented in Algorithm 12.1.

Theorem 12.7: Global Convergence of the Projected Gradient Method

Let $C \subset \mathbb{R}^n$ be a nonempty, convex, and closed set and let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable on an open set that contains C. Let $(x^k)_k$ be generated by the projected gradient method and assume that $(\lambda_k)_k$ is bounded, i.e., there exist $\lambda_m, \lambda_M > 0$ such that

$$0 < \lambda_m \leq \lambda_k \leq \lambda_M \quad \forall\, k.$$

Then, every accumulation point of $(x^k)_k$ is a stationary point of problem (12.2).

This convergence result was already analyzed in Assignment A2.1. Additionally, if ∇f is Lipschitz continuous with constant L, we can show:

- If λ_k satisfies $\lambda_k \leq \frac{2(1-\gamma)}{L}$, then the step size $\alpha_k = 1$ is accepted in step 4 of Algorithm 12.1.
- Furthermore, if f is also strongly convex, then the problem (12.2) has a unique global solution $x^* \in C$ and $(x^k)_k$ converges q-linearly to x^*. Similar to the basic gradient method, the rate again depends on the strong convexity parameter μ and the Lipschitz constant L and convergence can be slow if L/μ is large.

The first comment implies that $\lambda_k \approx \frac{2}{L}$ or $\lambda_k \approx \frac{1}{L}$ are good choices of the parameter λ_k. In practice, if the Lipschitz constant is not known, λ_k can be constructed to estimate L within the iterative process.

The projected gradient is often only practicable if the projection $\mathcal{P}_C$ is simple and has a closed-form expression. Fortunately, for many important examples and applications such closed-form formulae exist and can be utilized. Other algorithmic approaches might be more suitable if calculating $\mathcal{P}_C$ is too expensive.

12.2. The Proximal Gradient Method

We now return to our initial problem

$$\min_x \; f(x) + \varphi(x), \tag{12.5}$$

where $\varphi : \mathbb{R}^n \to (-\infty, \infty]$ is a convex but potentially nonsmooth function. We first discuss a connection between this problem and the constrained optimization problem (12.2). Let $C \subset \mathbb{R}^n$ be nonempty, convex, and closed and let us define the so-called *indicator function*

$$\iota_C : \mathbb{R}^n \to (-\infty, \infty], \quad \iota_C(x) := \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{if } x \notin C. \end{cases}$$

Then, we obtain

$$\min_{x \in C} \; f(x) \quad \equiv \quad \min_{x \in \mathbb{R}^n} \; f(x) + \iota_C(x)$$

and hence, these two problems share the same structure. The function ι_C is a real-extended valued function, i.e., it is allowed to take the value "$+\infty$". Many definitions and properties can be transferred to real-extended valued functions and only need to hold on the so-called *effective domain* $\operatorname{dom} \iota_C := \{x \in \mathbb{R}^n : \iota_C(x) < \infty\} = C$. In particular, the convexity of C implies that ι_C is a convex function.

- The principle idea of the proximal gradient method is to replace the indicator function ι_C with a general convex function $\varphi : \mathbb{R}^n \to \mathbb{R}$ (or $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$). This also leads to generalized projections – the so-called *proximity operator*.

Using this connection, we will see that the results in the previous section can the treated as special cases of the more general proximal framework and of the proximal gradient method. In the next subsection, we collect several elementary results on nonsmooth convex analysis that can be useful in the calculations.

12.2.1. Convex Analysis: Revisited

In order to allow a uniform treatment of constrained and unconstrained nonsmooth optimization problems, we will work with functions that can take the values $+\infty$ and $-\infty$. Let us introduce the following sets of extended real numbers:

$$(-\infty, \infty] := \mathbb{R} \cup \{+\infty\}, \quad [-\infty, \infty] := \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}.$$

Notice that the usual rules of calculus are still applicable for these extended sets. However, the following expressions will remain undefined:

$$+\infty + (-\infty), \qquad 0 \cdot (+\infty), \qquad +\infty / + \infty.$$

Next, we introduce several definitions and notions for extended real-valued functions.

- The domain of an extended real-valued function $\varphi : \mathbb{R}^n \to [-\infty, \infty]$ is given by

  $$\text{dom } \varphi := \{x \in \mathbb{R}^n : \varphi(x) < \infty\}.$$

  The function φ is called *proper* if dom $\varphi \neq \emptyset$ and $\varphi(x) \neq -\infty$ for all $x \in \mathbb{R}^n$.

- The *epigraph* of φ is defined as:

  $$\text{epi } \varphi := \{(x, t) \in \mathbb{R}^n \times \mathbb{R} : \varphi(x) \leq t\} \subset \mathbb{R}^n \times \mathbb{R}.$$

- The function φ is called *lower semicontinuous* if for all $x \in \mathbb{R}^n$ we have

  $$\liminf_{k \to \infty} \ f(x^k) \geq f(x), \quad \forall\ (x^k)_k \subset \mathbb{R}^n, \quad x^k \to x.$$

- The mapping φ is called *convex* if

  $$\varphi(\lambda x + (1-\lambda)y) \leq \lambda \varphi(x) + (1-\lambda)\varphi(y), \quad \forall\ x, y \in \text{dom } \varphi, \quad \lambda \in [0, 1].$$

  (Similar definitions for strict and strong convexity).

Many properties and results for functions and optimization problems can be generalized to the extended real-valued case. We now will collect and present several of those results that are helpful in the context of nonsmooth minimization.

Proposition 12.8: Characterizing Lower Semicontinuity

Let $\varphi : \mathbb{R}^n \to (-\infty, \infty]$ be given. Then, the following statements are equivalent:

(i) The function φ is lower semicontinuous.

(ii) The epigraph epi φ is closed in $\mathbb{R}^n \times \mathbb{R}$.

(iii) For every $\alpha \in \mathbb{R}$, the level set $L_{\leq \alpha} := \{x \in \mathbb{R}^n : \varphi(x) \leq \alpha\}$ is closed in $\mathbb{R}^n$.

A proper function $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ is again called *coercive* if $\lim_{\|x\|\to\infty} \varphi(x) = \infty$. Using a generalized version of the Weierstraß theorem, we can establish the following existence result for lower semicontinuous functions.

Proposition 12.9: Global Solutions and Coercivity

Let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ be proper, lower semicontinuous, and coercive and let $C \subset \mathbb{R}^n$ be a nonempty, closed set with $C \cap \operatorname{dom} \varphi \neq \emptyset$. Then φ attains its minimal value over C.

Next, we give an equivalent characterization of the continuity of a convex function.

Proposition 12.10: Convexity and Continuity

Let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ be a convex and proper function. The mapping φ is continuous at $x \in \operatorname{dom} \varphi$ if and only if $x \in \operatorname{int}(\operatorname{dom} \varphi)$. Furthermore, in that case, φ is also locally Lipschitz continuous near x and on the whole set $\operatorname{int}(\operatorname{dom} \varphi)$.

Proposition 12.10 implies that a real-valued function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is continuous and locally Lipschitz continuous on $\mathbb{R}^n$. In the general extended real-valued case, lack of continuity is a direct consequence of potential jumps when leaving $\operatorname{dom} \varphi$.

If φ is not differentiable, the so-called *subdifferential* of φ is used as an alternative to the derivative of φ. This notion can be motivated as follows: in the differentiable case, we have

$$\varphi(y) - \varphi(x) \geq \nabla\varphi(x)^\top (y - x), \quad \forall\, y \in \mathbb{R}^n,$$

i.e., the tangent $y \mapsto \varphi(x) + \nabla\varphi(x)^\top (y - x)$ supports the function φ at x from below. In the nonsmooth case, many such supporting functions might exist and the subdifferential of φ is defined as the collection of the *subgradients* of these functions.

Definition 12.11: The Convex Subdifferential

Let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ and $x \in \operatorname{dom} \varphi$ be given. The **subdifferential** of φ is the set

$$\partial\varphi(x) := \{g \in \mathbb{R}^n : \varphi(y) - \varphi(x) \geq g^\top (y - x),\ \forall\, y \in \mathbb{R}^n\}.$$

The function φ is called **subdifferentiable** at x if $\partial\varphi(x) \neq \emptyset$. The elements $g \in \partial\varphi(x)$ are called **subgradients** of φ at x.

In Figure 12.1, the construction of the subdifferential is shown for $\varphi(x) = |x|$. In the following, we list several important properties of the convex subdifferential.

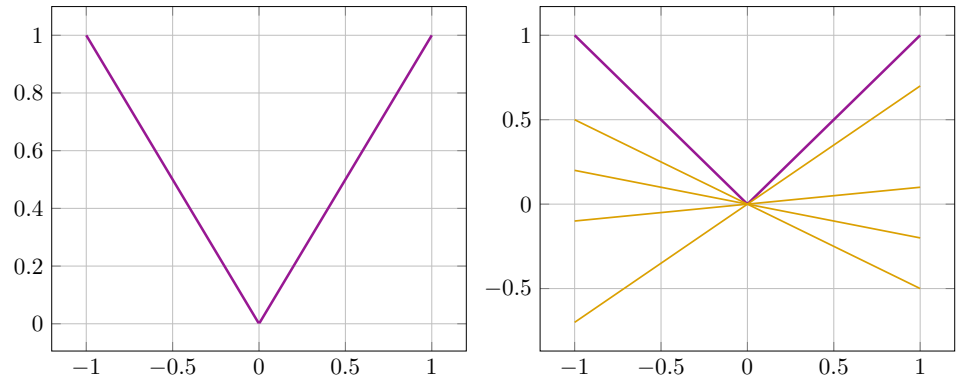
Proposition 12.12: Subdifferentiability and Continuity

Let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ be convex, proper and let $x \in \operatorname{dom} \varphi$ be given. Then, it holds:

(i) φ is continuous at x if and only if $\partial\varphi(x)$ is nonempty and bounded.

(ii) If φ is continuous at x, then there exists $\varepsilon > 0$ such that $\partial\varphi(B_\varepsilon(x))$ is bounded.

Next, we present a connection between the subdifferentiability of φ and its convex conju-



(a) Plot of $\varphi(x) = |x|$ (b) Supporting tangents at $x = 0$

Figure 12.1: Illustration: The subdifferential of $\varphi(x) = |x|$. The absolute value is differentiable for $x \neq 0$ and we have $\partial\varphi(x) = \{+1\}$ if $x > 0$ and $\partial\varphi(x) = \{-1\}$ if $x < 0$. In the case $x = 0$, we obtain $\partial\varphi(0) = [-1, 1]$.

gate φ^*. Recall that the *convex conjugate* or *Fenchel conjugate* of φ is defined as

$$\varphi^* : \mathbb{R}^n \to [-\infty, +\infty], \quad \varphi^*(x) := \sup_{y \in \mathbb{R}^n} \; y^\top x - \varphi(y).$$

The convex conjugate φ^* of a mapping $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ is always a convex and lower semicontinuous function.

Proposition 12.13: Subdifferentiability and Convex Conjugates

Suppose that $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ is convex, proper, and lower semicontinuous and let $x, g \in \mathbb{R}^n$ be arbitrary. Then, the following statements are equivalent:

(i) $g \in \partial\varphi(x)$. (ii) $x \in \partial\varphi^*(g)$. (iii) $\varphi(x) + \varphi^*(s) = x^\top g$.

Next, we present a chain rule for the subdifferential of a composition of convex functions.

Proposition 12.14: Chain Rule for Subdifferentials

Let the two functions $f : \mathbb{R}^n \to (-\infty, +\infty]$ and $\varphi : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, proper, and lower semicontinuous and let $A \in \mathbb{R}^{m \times n}$ be an arbitrary matrix. Furthermore, let us set $\psi(x) := f(x) + \varphi(Ax)$ and suppose that $A(\text{dom } f) \cap \text{dom } \varphi \neq \emptyset$. Then, it holds

$$\partial f(x) + A^\top \partial\varphi(Ax) \subseteq \partial\psi(x), \quad \forall\, x \in \mathbb{R}^n. \tag{12.6}$$

In addition, if the regularity condition $0 \in \text{int}(A(\text{dom } f) - \text{dom } \varphi)$ is satisfied, then it follows

$$\partial\psi(x) = \partial f(x) + A^\top \partial\varphi(Ax), \quad \forall\, x \in \mathbb{R}^n. \tag{12.7}$$

The regularity condition is obviously satisfied if dom $\varphi = \mathbb{R}^n$, i.e., when φ is real-valued.

We now present a connection between classical derivatives and subgradients.

Proposition 12.15: Subdifferentiability and Differentiability

Let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ be a convex and proper mapping and let $x \in \operatorname{dom} \varphi$ be given.

(i) Suppose that φ is (Fréchet) differentiable at x. Then, we have $\partial\varphi(x) = \{\nabla\varphi(x)\}$.

(ii) Suppose we have $x \in \operatorname{int}(\operatorname{dom} \varphi)$ and $\partial\varphi(x)$ consists of a single element g. Then φ is differentiable at x with $\nabla\varphi(x) = g$.

We conclude this subsection with some basic and explicit examples.

Example 12.16. Consider the convex function $\varphi : \mathbb{R} \to (-\infty, +\infty]$,

$$\varphi(x) := \begin{cases} -\sqrt{x} & x \geq 0, \\ \infty & \text{otherwise.} \end{cases}$$

This function is not subdifferentiable at $x = 0$. Assume, by contradiction, that there exists $g \in \mathbb{R}$ such that $\varphi(y) - \varphi(0) \geq g(y - 0)$ for all $y \geq 0$. This is equivalent to the condition

$$-\sqrt{y} \geq gy \quad \forall\ y \geq 0.$$

Setting $y = 1$, this implies $g \leq -1$ and for $y = 1/(2g^2)$ we obtain $-\sqrt{1/(2g^2)} \geq 1/(2g)$ which is equivalent to $\sqrt{2} \leq 1$. This is a contradiction and hence, φ can not be subdifferentiable.

Example 12.17. Let $C \subset \mathbb{R}^n$ be a convex set. The indicator function ι_C of C is subdifferentiable at $x \in \mathbb{R}^n$ if and only if $x \in C$. Consequently, the subdifferential of ι_C is given by

$$\partial\iota_C(x) = \begin{cases} \{g \in \mathbb{R}^n : g^\top(y - x) \leq 0,\ \forall\ y \in C\} = N_C(x) & \text{if } x \in C, \\ \emptyset & \text{otherwise.} \end{cases}$$

12.2.2. First-Order Optimality and the Proximity Operator

As in the last section, we first characterize local solutions and stationary points of (12.5). Throughout this section, we assume that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ is convex, lower semicontinuous, and proper.

Lemma 12.18: First-Order Necessary Conditions

Let $x^* \in \mathbb{R}^n$ be a local minimizer of (12.5). Then, it holds that

$$\nabla f(x^*)^\top(x - x^*) + \varphi(x) - \varphi(x^*) \geq 0, \quad \forall\ x \in \mathbb{R}^n. \tag{12.8}$$

Additionally, if f is convex, then x^* is global minimizer of (12.1) if and only if the condition (12.8) is satisfied.

Proof. Let $x^* \in \operatorname{dom} \varphi$ first be a local minimum of the objective function $\psi = f + \varphi$. Due to the convexity of φ, we have $\varphi(\lambda x + (1 - \lambda)x^*) \leq \lambda\varphi(x) + (1 - \lambda)\varphi(x^*)$ for all $\lambda \in [0, 1]$

and every $x \in \mathbb{R}^n$. (This is obviously true for $x \notin \operatorname{dom} \varphi$). Thus, it follows

$$\begin{aligned} 0 &\leq \lim_{\lambda \to 0} \frac{f(x^* + \lambda(x - x^*)) - f(x^*) + \varphi(\lambda x + (1-\lambda)x^*) - \varphi(x^*)}{\lambda} \\ &\leq \nabla f(x^*)^\top (x - x^*) + \varphi(x) - \varphi(x^*). \end{aligned}$$

On the other hand, if f is convex, the condition (12.8) and Theorem 4.14 imply

$$0 \leq \nabla f(x^*)^\top (x - x^*) + \varphi(x) - \varphi(x^*) \leq \psi(x) - \psi(x^*) \quad \forall\, x \in \mathbb{R}^n.$$

Hence, x^* is a global solution of the problem $\min_{x \in \mathbb{R}^n} f(x) + \varphi(x)$. ■

Again we will call a point x^* satisfying the inequality (12.8) a *stationary point* of (12.5). Let us further notice that the optimality condition (12.8) is equivalent to

$$-\nabla f(x^*) \in \partial \varphi(x^*).$$

We now generalize Theorem 12.2 and formulate a similar variant for general convex functions. The resulting operation generalizes the notion of a projection and was initially introduced by Moreau in 1965. More specifically, we now consider functions of the form

$$\operatorname{env}_{\lambda\varphi} : \mathbb{R}^n \to \mathbb{R}, \quad \operatorname{env}_{\lambda\varphi}(x) := \min_y \ \varphi(y) + \frac{1}{2\lambda}\|x - y\|^2$$

where $\lambda > 0$ is a parameter. The mapping $\operatorname{env}_{\lambda\varphi}$ is the so-called *Moreau envelope* of φ.

Theorem 12.19: The Proximity Operator

Let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ be a convex, lower semicontinuous, and proper function.

(i) For every $x \in \mathbb{R}^n$ and every $\lambda > 0$, the optimization problem

$$\min_y \ \varphi(y) + \frac{1}{2\lambda}\|x - y\|^2$$

has a unique global solution $x^* \in \operatorname{dom} \varphi$. This minimizer is called the **proximity operator** of φ at x and we write $x^* = \operatorname{prox}_{\lambda\varphi}(x)$.

(ii) The mapping $\operatorname{prox}_{\lambda\varphi} : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with $L = 1$ and satisfies

$$\|\operatorname{prox}_{\lambda\varphi}(x) - \operatorname{prox}_{\lambda\varphi}(z)\|^2 \leq (\operatorname{prox}_{\lambda\varphi}(x) - \operatorname{prox}_{\lambda\varphi}(z))^\top (x - z), \quad \forall\, x, z \in \mathbb{R}^n.$$

(iii) Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Then, x^* is a stationary point of problem (12.5) if and only if

$$x^* - \operatorname{prox}_{\lambda\varphi}(x^* - \lambda \nabla f(x^*)) = 0 \quad \text{for any } \lambda > 0.$$

Proof. Since the mapping φ is proper there exists $\bar{x} \in \operatorname{dom} \varphi$ and thus, the level set $\bar{L} := \{y \in \mathbb{R}^n : \varphi(y) + \frac{1}{2\lambda}\|x - y\|^2 \leq \varphi(\bar{x}) + \frac{1}{2\lambda}\|x - \bar{x}\|^2\}$ is nonempty and closed. Furthermore,

due to $\bar{L} \subset \operatorname{dom} \varphi$, we can invoke the convexity of φ to infer that $\bar{L}$ is convex and that $y \mapsto \varphi(y) + \frac{1}{2\lambda}\|x-y\|^2$ is strongly convex on $\bar{L}$. Applying Assignment A1.6, this establishes compactness of the level set $\bar{L}$ and the generalized Weierstraß theorem guarantees existence of a global solution of $\min_y \varphi(y) + \frac{1}{2\lambda}\|x-y\|^2$. Uniqueness is a consequence of the strong convexity.

For $x, z \in \mathbb{R}^n$ the optimality condition (12.8) implies

$$\left[\begin{array}{l} \varphi(\operatorname{prox}_{\lambda\varphi}(z)) - \varphi(\operatorname{prox}_{\lambda\varphi}(x)) \geq \lambda^{-1}\langle x - \operatorname{prox}_{\lambda\varphi}(x), \operatorname{prox}_{\lambda\varphi}(z) - \operatorname{prox}_{\lambda\varphi}(x)\rangle, \\ \varphi(\operatorname{prox}_{\lambda\varphi}(x)) - \varphi(\operatorname{prox}_{\lambda\varphi}(z)) \geq \lambda^{-1}\langle z - \operatorname{prox}_{\lambda\varphi}(z), \operatorname{prox}_{\lambda\varphi}(x) - \operatorname{prox}_{\lambda\varphi}(z)\rangle. \end{array}\right.$$

Adding those two inequalities, we obtain

$$\|\operatorname{prox}_{\lambda\varphi}(x) - \operatorname{prox}_{\lambda\varphi}(z)\|^2 + \langle z - x, \operatorname{prox}_{\lambda\varphi}(x) - \operatorname{prox}_{\lambda\varphi}(z)\rangle \leq 0.$$

Again by (12.8), we have

$$\begin{aligned} x^* = \operatorname{prox}_{\lambda\varphi}(x^* - \lambda\nabla f(x^*)) & \\ \iff \quad & x^* \text{ is a solution of } \min_y \ \varphi(y) + \frac{1}{2\lambda}\|(x^* - \lambda\nabla f(x^*)) - y\|^2 \\ \iff \quad & \lambda^{-1}(x^* - (x^* - \lambda\nabla f(x^*))^\top (x - x^*) + \varphi(x) - \varphi(x^*) \geq 0 \quad \forall\, x \in \mathbb{R}^n \\ \iff \quad & \nabla f(x^*)^\top (x - x^*) + \varphi(x) - \varphi(x^*) \geq 0 \quad \forall\, x \in \mathbb{R}^n. \end{aligned}$$

This verifies statement (iii) of Theorem 12.19. ■

As before, it is possible to derive explicit representations of the proximity operator for many interesting examples and applications.

Example 12.20. Let us consider the weighted ℓ_1-norm $\varphi(x) := \mu\|x\|_1$, $\mu > 0$. The proximity operator of φ is also known as *shrinkage operator* or *soft thresholding* and it satisfies

$$\operatorname{prox}_{\lambda\varphi}(x) = x - \mathcal{P}_{[-\lambda\mu,\lambda\mu]}(x), \quad [\operatorname{prox}_{\lambda\varphi}(x)]_i = \begin{cases} x_i + \lambda\mu & \text{if } x_i < -\lambda\mu, \\ 0 & \text{if } x_i \in [-\lambda\mu, \lambda\mu], \\ x_i - \lambda\mu & \text{if } x_i > \lambda\mu. \end{cases}$$

Proof. We need to solve the optimization problem

$$\min_y \ \mu\|y\|_1 + \frac{1}{2\lambda}\|x-y\|^2 = \min_y \ \sum_{i=1}^n \left[\mu|y_i| + \frac{1}{2\lambda}(x_i - y_i)^2\right].$$

Since the problem is separable, it is enough to determine the solution of the one dimensional problem $\min_t \mu|t| + \frac{1}{2\lambda}(x_i - t)^2$. We then have $[\operatorname{prox}_{\lambda\varphi}(x)]_i = \operatorname{prox}_{\lambda\mu|\cdot|}(x_i) = t$ for all i. The

associated optimality conditions are given by

$$0 \in \mu\partial|\cdot|(t) + \frac{1}{\lambda}(t - x_i) \quad \Longleftrightarrow \quad x_i \in t + \begin{cases} \{\lambda\mu\} & \text{if } t > 0, \\ [-\lambda\mu, \lambda\mu] & \text{if } t = 0, \\ \{-\lambda\mu\} & \text{if } t < 0. \end{cases}$$

Rearranging these conditions yields

$$t = \begin{cases} x_i - \lambda\mu & \text{if } x_i > \lambda\mu, \\ 0 & \text{if } x_i \in [-\lambda\mu, \lambda\mu], \\ x_i + \lambda\mu & \text{if } x_i < -\lambda\mu, \end{cases} \quad = \quad x_i - \mathcal{P}_{[-\lambda\mu,\lambda\mu]}(x_i),$$

which finishes the proof. ■

Example 12.21. In our second example, let us consider a weighted ℓ_2-norm $\varphi(x) := \mu\|x\|_2$, $\mu > 0$. The proximity operator of φ is then given by

$$\operatorname{prox}_{\lambda\varphi}(x) = x - \mathcal{P}_{B_{\lambda\mu}(0)}(x) = \frac{x}{\|x\|} \cdot \max\{\|x\| - \lambda\mu, 0\}.$$

Proof. The associated optimality conditions defining the proximity operator $y = \operatorname{prox}_{\lambda\varphi}(x)$ are given by

$$0 \in \mu\partial\|\cdot\|(y) + \frac{1}{\lambda}(y - x) \quad \Longleftrightarrow \quad x \in y + \begin{cases} \lambda\mu \cdot \frac{y}{\|y\|} & \text{if } y \neq 0, \\ B_{\lambda\mu}(0) & \text{if } y = 0. \end{cases}$$

Taking the norm in the case $y \neq 0$, we obtain $\|x\| = (1 + \lambda\mu/\|y\|)\|y\| = \|y\| + \lambda\mu > 0$ and $\|y\| = \|x\| - \lambda\mu$. This allows to rearrange the terms and it follows $y = (1 + \lambda\mu/\|y\|)^{-1}x = \frac{\|y\|}{\|y\|+\lambda\mu}x = x - \lambda\mu\frac{x}{\|x\|}$ if $\|x\| > \lambda\mu$. In the case $\|x\| \leq \lambda\mu$, we obtain $y = 0$. ■

12.2.3. The Proximal Gradient Method

As before, the principal idea of the proximal gradient method is to generate a descent-type direction based on the proximity operator of φ. Let us define

$$F_\lambda(x) := x - \operatorname{prox}_{\lambda\varphi}(x - \lambda\nabla f(x)).$$

Then we can show the following result.

Lemma 12.22: Nonsmooth Descent Condition
Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ be a convex, proper, and lower semicontinuous mapping. Let $x \in \mathbb{R}^n$ and $\lambda > 0$ be given and set $d := -F_\lambda(x)$. Then, it holds that

$$\Delta := \nabla f(x)^\top d + \varphi(x+d) - \varphi(x) \leq -\frac{1}{\lambda}\|d\|^2.$$

In addition, suppose that $x \in \operatorname{dom} \varphi$ is not a stationary point of (12.5) and let $\gamma \in (0,1)$ be arbitrary. Then, setting $\psi = f + \varphi$, there exists $\bar{\alpha} > 0$ such that

$$\psi(x + \alpha d) - \psi(x) \leq \gamma\alpha \cdot \Delta \quad \forall\, \alpha \in [0, \bar{\alpha}].$$

Proof. By the definition of the proximity operator, we have

$$\begin{aligned}
\varphi(x+d) - \varphi(x) &= \varphi(\operatorname{prox}_{\lambda\varphi}(x - \lambda\nabla f(x))) - \varphi(x) \\
&\leq \lambda^{-1}\langle \operatorname{prox}_{\lambda\varphi}(x - \lambda\nabla f(x)) - (x - \lambda\nabla f(x)), x - \operatorname{prox}_{\lambda\varphi}(x - \lambda\nabla f(x))\rangle \\
&= -\nabla f(x)^\top d - \frac{1}{\lambda}\|d\|^2.
\end{aligned}$$

This implies $\Delta \leq -\frac{1}{\lambda}\|d\|^2$. We continue to show the second part. Using a Taylor expansion and the convexity of φ, it holds that

$$\begin{aligned}
\psi(x+\alpha d) - \psi(x) &= f(x+\alpha d) - f(x) + \varphi((1-\alpha)x + \alpha\operatorname{prox}_{\lambda\varphi}(x - \lambda\nabla f(x))) - \varphi(x) \\
&\leq \alpha\nabla f(x)^\top d + o(\alpha) + \alpha\varphi(x+d) - \alpha\varphi(x) \\
&= \alpha\Delta + o(\alpha) = \gamma\alpha\cdot\Delta + [(1-\gamma)\alpha\cdot\Delta + o(\alpha)].
\end{aligned}$$

Since x is not a stationary point, we have $d \neq 0$ and $\Delta < 0$ and thus, the last term in brackets is negative for all α sufficiently small. This finishes the proof of Lemma 12.22. ■

Lemma 12.22 basically tells us that we can use the direction $d^k = -F_{\lambda_k}(x^k)$ as a descent direction and perform an Armijo-type line search technique to determine a suitable step size α_k. These different strategies define the basic proximal gradient method which is presented in Algorithm 12.2.

The convergence results for the proximal gradient method are basically identical to the standard and projected gradient method.

Algorithm 12.2: A Proximal Gradient Method

1 Initialization: Choose an initial point $x^0 \in \mathbb{R}^n$ and $\sigma, \gamma \in (0,1)$.
for $k = 0, 1, 2, \ldots$ **do**
2 Select $\lambda_k > 0$ and compute $\nabla f(x^k)$ and the new direction $d^k = -F_{\lambda_k}(x^k)$.
3 If $\|d^k\| \leq \lambda_k \varepsilon$, then STOP and x^k is the output.
4 Choose a maximal step size $\alpha_k \in \{1, \sigma, \sigma^2, \ldots\} \subset (0,1]$ that satisfies the Armijo-type condition

$$\psi(x^k + \alpha_k d^k) - \psi(x^k) \leq \gamma \alpha_k \cdot \Delta^k.$$

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

Theorem 12.23: Global Convergence of the Proximal Gradient Method

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ be a convex, proper, and lower semicontinuous mapping. Let $(x^k)_k$ be generated by Algorithm 12.2 and assume that $(\lambda_k)_k$ is bounded, i.e., there exist $\lambda_m, \lambda_M > 0$ such that

$$0 < \lambda_m \leq \lambda_k \leq \lambda_M \quad \forall\, k.$$

Then, every accumulation point of $(x^k)_k$ is a stationary point of problem (12.5).

If ∇f is Lipschitz continuous with constant L, we again have:

- If λ_k satisfies $\lambda_k \leq \frac{2(1-\gamma)}{L}$, then the step size $\alpha_k = 1$ is accepted in step 4 of Algorithm 12.2 and the method reduces to

$$x^{k+1} = \operatorname{prox}_{\lambda_k \varphi}(x^k - \lambda_k \nabla f(x^k)). \tag{12.9}$$

- Q-linear convergence can be established if f is additionally strongly convex.

The update in (12.9) can also be expressed differently by using the definition of the proximity operator:

$$\begin{aligned} x^{k+1} &= \operatorname{prox}_{\lambda_k \varphi}(x^k - \lambda_k \nabla f(x^k)) \\ &= \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \ \varphi(y) + \frac{1}{2\lambda_k} \|x^k - \lambda_k \nabla f(x^k) - y\|^2 \\ &= \operatorname*{arg\,min}_{y \in \mathbb{R}^n} \ \varphi(y) + \nabla f(x^k)^\top (y - x^k) + \frac{1}{2\lambda_k} \|y - x^k\|^2. \end{aligned}$$

Here, we dropped the term $\frac{\lambda_k}{2}\|\nabla f(x^k)\|^2$ in the last step, since it does not depend on y and we are only interested in the global minimizer "arg min". Hence, the principle idea of the proximal gradient method can be interpreted as follows:

- We build a simpler model of the objective function $\psi = \varphi + f$ by keeping the nonsmooth function φ and by using a quadratic approximation $f(y) \approx f(x^k) + \nabla f(x^k)^\top (y - x^k) + \frac{1}{2\lambda_k}(y - x^k)^\top (y - x^k)$ for the smooth function f.

- The global minimizer of this model is then used to define the next iterate x^{k+1}.
- This immediately motivates possible extensions of the form

$$x^{k+1} = \operatorname*{arg\,min}_{y\in\mathbb{R}^n} \ \varphi(y) + \nabla f(x^k)^\top (y - x^k) + \frac{1}{2}(y - x^k)^\top B_k (y - x^k), \tag{12.10}$$

where $B_k \in \mathbb{S}^n_{++}$ is a symmetric, positive definite matrix that can either be the Hessian $\nabla^2 f(x^k)$ or a suitable approximation.

In contrast to the simple proximal gradient step (12.9) such updates typically no longer have closed-form expressions and a subproblem needs to be solved to calculate x^{k+1}. Methods that are based on the formulation (12.10) are called *proximal Newton methods*.

12.2.4. Proximal Calculus

Similar to the projected gradient method, the proximal gradient method assumes that the proximity operator $\operatorname{prox}_{\lambda\varphi}$ either has a closed-form expression or can be computed cheaply. We have already discussed several important examples and applications where explicit formulae for the proximity operator are available and additional examples are provided on the document "Projections and Proximity Operators".

In this subsection, we will discuss several computational rules for proximity operators.

Lemma 12.24: Translation and Scaling

Let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ be a convex, lower semicontinuous, and proper function and let $\lambda > 0$ and $x \in \mathbb{R}^n$ be given. We have:

(i) Let us define $g(\cdot) := \varphi(\cdot - b)$, $b \in \mathbb{R}^n$. Then, it follows $\operatorname{prox}_{\lambda g}(x) = b + \operatorname{prox}_{\lambda\varphi}(x - b)$.

(ii) Let us define $h(\cdot) := \varphi(\cdot/\beta)$, $\beta \neq 0$. Then, it follows $\operatorname{prox}_{\lambda h}(x) = \beta \cdot \operatorname{prox}_{\lambda\varphi/\beta^2}(x/\beta)$.

Proof. We start with the proof of the first part. Let us set $p = \operatorname{prox}_{\lambda g}(x)$. Then, by the definition of the proximity operator and by a change of variable, we obtain

$$\begin{aligned} p &= \operatorname*{arg\,min}_{y\in\mathbb{R}^n} \ \varphi(y - b) + \frac{1}{2\lambda}\|y - x\|^2 \\ &= b + \operatorname*{arg\,min}_{w = y - b \in \mathbb{R}^n} \ \varphi(w) + \frac{1}{2\lambda}\|w - (x - b)\|^2 = b + \operatorname{prox}_{\lambda\varphi}(x - b). \end{aligned}$$

Similarly for $p = \operatorname{prox}_{\lambda h}(x)$, we have

$$\begin{aligned} p = \operatorname*{arg\,min}_{y} \ \varphi(y/\beta) + \frac{1}{2\lambda}\|y - x\|^2 &= \left[\operatorname*{arg\,min}_{w = y\beta} \ \varphi(w) + \frac{1}{2\lambda}\|\beta w - x\|^2\right] \cdot \beta \\ &= \beta \cdot \operatorname{prox}_{\lambda\varphi/\beta^2}(x/\beta) \end{aligned}$$

which finishes the proof of this lemma. ■

Lemma 12.25: Composition with a Linear Operator

Let $\varphi : \mathbb{R}^m \to (-\infty, +\infty]$ be a convex, proper, and lower semicontinuous function and let $x \in \mathbb{R}^n$, $\lambda > 0$ and $A \in \mathbb{R}^{m\times n}$ be given. Suppose that the matrix A satisfies $AA^\top = I$. Then, for $g(\cdot) := \varphi(A\cdot)$, it follows

$$\operatorname{prox}_{\lambda g}(x) = x - A^\top(Ax - \operatorname{prox}_{\lambda\varphi}(Ax)).$$

Proof. Let us set $y = \operatorname{prox}_{\lambda g}(x)$. Then, by the optimality condition

$$\begin{aligned} 0 \in \partial g(y) + \frac{1}{\lambda}(y - x) \quad &\iff \quad 0 \in A^\top \partial\varphi(Ay) + \frac{1}{\lambda}(y - x) \\ &\iff \quad 0 \in \partial\varphi(Ay) + \frac{1}{\lambda}(Ay - Ax) \\ &\iff \quad Ay = \operatorname{prox}_{\lambda\varphi}(Ax). \end{aligned}$$

This shows that there exists $v \in \partial\varphi(Ay)$ such that $y = x - \lambda A^\top v$ and we have $Ay = Ax - \lambda v$, i.e., $v = (Ax - Ay)/\lambda$. Together, this implies $y = x - A^\top(Ax - \operatorname{prox}_{\lambda\varphi}(Ax))$. ■

Lemma 12.25 can be seen as special but important chain rule. In particular, the conditions in Lemma 12.25 are satisfied if A is an orthogonal matrix with $m = n$. Finally, we consider the case where the function φ is *separable*.

Lemma 12.26: Separable Functions

Let $\varphi_i : \mathbb{R} \to (-\infty, +\infty]$, $i = 1, ..., n$, be a family of convex, lower semicontinuous, and proper functions and let us set $\varphi(x) = \sum_{i=1}^n \varphi_i(x_i)$. Then, it holds that

$$[\operatorname{prox}_{\lambda\varphi}(x)]_i = \operatorname{prox}_{\lambda\varphi_i}(x_i) \quad \forall\, i, \quad \lambda > 0.$$

We continue with an important example.

Example 12.27. Let us consider the optimization problem

$$\min_x \; f(x) \quad \text{s.t.} \quad \|Ax - b\|_2 \le \sigma,$$

where $A \in \mathbb{R}^{m\times n}$ satisfies $AA^\top = I$ and $b \in \mathbb{R}^m$, $\sigma \ge 0$ are given. Let us define the feasible set $C := \{x : \|Ax - b\| \le \sigma\}$. In this example, we want to compute the projection operator $\mathcal{P}_C(x)$. Introducing $B_\sigma(0) := \{y \in \mathbb{R}^m : \|y\| \le \sigma\}$ and using $b = AA^\top b$, we obtain

$$\iota_C(x) = \iota_{B_\sigma(0)}(Ax - b) = \iota_{B_\sigma(0)}(A(x - A^\top b)).$$

Hence, applying Lemma 12.25 and Lemma 12.24, it follows

$$\begin{aligned}\mathcal{P}_C(x) &= \operatorname{prox}_{(\iota_{B_\sigma(0)}\circ A)(\cdot - A^\top b)}(x) = A^\top b + \operatorname{prox}_{\iota_{B_\sigma(0)}\circ A}(x - A^\top b)\\ &= A^\top b + x - A^\top b - A^\top(Ax - AA^\top b - \mathcal{P}_{B_\sigma(0)}(A(x - A^\top b)))\\ &= (I - A^\top A)x + A^\top(b + \mathcal{P}_{B_\sigma(0)}(Ax - b)).\end{aligned}$$

Partial permutation matrices are typical type of matrices that fulfill the condition $AA^\top = I$. Let $e_1, ...e_n \in \mathbb{R}^n$ denote the n different unit vectors in $\mathbb{R}^n$. Then, a partial permutation matrix A is constructed by selecting m different unit vectors e_i ($m \leq n$) and by setting the rows of A as $e_i^\top$. Thus, A selects certain components of x which should be close to b. As we have seen such operations appear frequently in image reconstruction problems, where A selects pixels of the "image" x and b contains information of the original image. In this case, σ can be interpreted as a given noise level.

Theorem 12.28: Moreau's Decomposition Principle

Let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ be a convex, proper, and lower semicontinuous function and let $\lambda > 0$ be arbitrary. Then, for all $x \in \mathbb{R}^n$, it holds that

$$x = \operatorname{prox}_{\lambda\varphi}(x) + \lambda \cdot \operatorname{prox}_{\lambda^{-1}\varphi^*}(\lambda^{-1}x), \tag{12.11}$$

where φ^* denotes the convex conjugate of φ.

Proof. We have $\lambda^{-1}p = \operatorname{prox}_{\lambda^{-1}\varphi^*}(\lambda^{-1}x)$ if and only if $0 \in \partial\varphi^*(\lambda^{-1}p) + p - x$. Using Proposition 12.13 this is further equivalent to $\lambda^{-1}p \in \partial\varphi(x - p)$. Hence, we have

$$0 \in \partial\varphi(x - p) + \frac{1}{\lambda}(x - p - x) \quad \Longleftrightarrow \quad x - p = \operatorname{prox}_{\lambda\varphi}(x).$$

This finishes the proof of Theorem 12.28. ■

12.2.5. The Accelerated Proximal Gradient Method

It is possible to combine the acceleration techniques discussed in section 8 and the proximal gradient method under the assumption that the smooth function $f : \mathbb{R}^n \to \mathbb{R}$ is additionally convex.

In principle, the acceleration mechanism is identical to the one presented in section 8. We first perform an extrapolation step

$$y^{k+1} = x^k + \beta_k(x^k - x^{k-1}), \qquad \beta_k > 0$$

to approximate and extrapolate the next iterate. Afterwards we compute a proximal gradient step based on the predicted information y^{k+1}. The full accelerated proximal gradient method is shown in Algorithm 12.3.

Algorithm 12.3: The Accelerated Proximal Gradient Method

Initialization: Choose an initial point $x^0 \in \mathbb{R}^n$ and set $x^{-1} = x^0$, $t_{-1} = t_0 = 1$.
for $k = 0, 1, 2, \ldots$ **do**
  Select an extrapolation parameter β_k and compute $y^{k+1} = x^k + \beta_k(x^k - x^{k-1})$.
  Select a step size $\lambda_k > 0$ and set $x^{k+1} = \operatorname{prox}_{\lambda_k \varphi}(y^{k+1} - \lambda_k \nabla f(y^{k+1}))$.

As before we are specifically interested in extrapolation parameters of the form:

$$\beta_k = \frac{\theta_k(1-\theta_{k-1})}{\theta_{k-1}} = \frac{\theta_k}{\theta_{k-1}} - \theta_k \quad \text{and} \quad \theta_{-1} = \theta_0 = 1. \tag{12.12}$$

The following convergence result can be shown:

Theorem 12.29: Convergence of the Accelerated Proximal Gradient Method

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex mapping satisfying $f \in C_L^{1,1}(\mathbb{R}^n)$ and let $\varphi : \mathbb{R}^n \to (-\infty, +\infty]$ be convex, lower semicontinuous, and proper. Let us further assume that the solution set $\mathcal{X}^*$ of the problem $\min_x \psi(x)$ is nonempty. Let $(x^k)_k$, $(\lambda_k)_k$, and $(\beta_k)_k$ be generated by Algorithm 12.3 with $\lambda_k = \bar{\lambda} \in (0, \frac{1}{L}]$,

$$\beta_k \text{ satisfies (12.12)} \quad \text{and we have} \quad 0 \leq \theta_k \leq \frac{2}{k+2}, \quad \frac{1-\theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}$$

for all $k \in \mathbb{N}$. Then, for every $x^* \in \mathcal{X}^*$, it follows

$$\psi(x^k) - \psi(x^*) \leq \frac{2\|x^0 - x^*\|^2}{\bar{\lambda}(k+1)^2} \quad \forall\, k \in \mathbb{N}.$$

We continue with several remarks:

- In the case $\beta_k = 0$, $k \in \mathbb{N}$, we can only guarantee the following complexity bound:

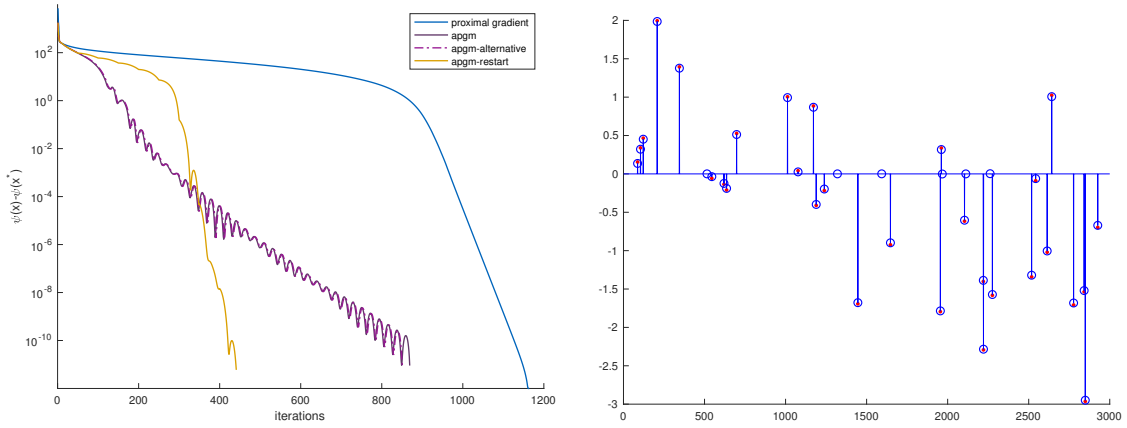
  $$\psi(x^k) - \psi(x^*) \leq \frac{\|x^0 - x^*\|^2}{2\bar{\lambda}k},$$

  i.e., the proximal gradient method approaches the optimal value $\psi(x^*)$ with the much slower rate $\mathcal{O}(k^{-1})$ whereas the accelerated proximal gradient method converges with rate $\mathcal{O}(k^{-2})$. This is basically identical to our observations in the smooth case $\varphi \equiv 0$.

- As in the smooth case, we can utilize the choices $\beta_k = \frac{k-2}{k+1}$ or $\theta_k = t_k^{-1}$ and

  $$t_k = \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2}, \quad \beta_k = \frac{t_{k-1} - 1}{t_k}, \quad t_{-1} = t_0 = 1.$$

  These choices satisfy the requirements in Theorem 12.29.



(a) Plot of convergence (b) Sparsity pattern

Figure 12.2: Example 12.30: Convergence of the sequence $(\psi(x^k))$ with respect to number of iterations. The result of the proximal gradient method is shown using blue color. The behavior of the accelerated PGM without and with restart is depicted using red and green color. The yellow plot corresponds to the accelerated PGM with $\beta_k = \frac{k-2}{k+1}$. In Figure b) the true sparsity pattern of the signal x^* is shown in red. The blue signal depicts the reconstruction x^k obtained by APGM (with restart).

- If the Lipschitz constant L is unknown, then the step size λ_k can be determined by the following line search procedure: Choose $\eta \in (0,1)$;
  - Set $\lambda_k = \lambda_{k-1}$ and $x^{k+1} = \operatorname{prox}_{\lambda_k \varphi}(y^{k+1} - \lambda_k \nabla f(y^{k+1}))$.
  - `while` $f(x^{k+1}) > f(y^{k+1}) + \nabla f(y^{k+1})^\top (x^{k+1} - y^{k+1}) + \frac{1}{2\lambda_k}\|x^{k+1} - y^{k+1}\|^2$ `do`:
    Set $\lambda_k = \eta \lambda_k$ and compute $x^{k+1} = \operatorname{prox}_{\lambda_k \varphi}(y^{k+1} - \lambda_k \nabla f(y^{k+1}))$.

We now continue with several examples to illustrate the potential acceleration of APGM and the differences in convergence.

Example 12.30 (Sparse Recovery). In this example, we discuss the ℓ_1-optimization problem

$$\min_x \ \psi(x) = \frac{1}{2}\|Ax - b\|^2 + \mu\|x\|_1 \tag{12.13}$$

The data is generated as follows:

- Set `m = 300`, `n = 3000`, `s = 30` and create a mask `mask = randperm(n,s)`. We then generate a sparse signal $x^* \in \mathbb{R}^n$ via

$$x^* = \texttt{zeros(n,1)}, \quad x^*(\texttt{mask}) = \texttt{randn(s,1)}.$$

  (x^* has only 30 nonzero components that are chosen randomly).

- We choose `A = randn(m,n)` and generate the partial measurement b via `b = A*`x^* `+ 0.01*randn(m,1)`. The goal is to reconstruct x^* from the much smaller measurements b via solving the problem (12.13).
- The Lipschitz constant of ∇f is given by $L = \lambda_{\max}(A^\top A)$. All other parameter are set as usual.

We compare the basic proximal gradient method (with quasi-Armijo linesearch and $\gamma = 0.1$, $s = 1$, $\sigma = 0.5$) with three different variants of the accelerated proximal gradient method (basic: $\beta_k = (t_{k-1} - 1)t_k^{-1}$, alternative: $\beta_k = \frac{k-2}{k+1}$, basic with restart $t_{k-1} = t_k = 1$ after 50 iterations). The tolerance is set to `tol` $= 10^{-8}$. Setting $\mu = 5$, we obtain the following results:

```
-- gradient method:  [ITER; OBJ]: [1154; 1.28e+02]; TIME: 1.45 sec
-- apgm:  [ITER; OBJ]: [ 753; 1.28e+02]; TIME: 0.97 sec
-- apgm (alter.  beta):  [ITER; OBJ]: [ 755; 1.28e+02]; TIME: 0.92 sec
-- apgm (restart):  [ITER; OBJ]: [ 393; 1.28e+02]; TIME: 0.49 sec
```

The variants of APGM converge significantly faster than the proximal gradient method. The (heuristic) restart strategy can enhance the convergence even further. (This is not always the case: if we select A via `A = 10*rand(m,n)-5`, APGM with restart converges slower. However, increasing the restarting parameter to 500 iterations can again yield faster convergence). The convergence results are also illustrated in Figure 12.2.

13. Alternating Direction Method of Multiplier

In this section, we discuss an algorithm that is related to the Augmented Lagrangian method introduced in subsection 11.2 and that is well-suited for optimization problems with special structures involving among others separability and large sums of component functions. The algorithm uses alternate minimization to decouple sets of variables that are coupled within the augmented Lagrangian, and is known as the *alternating direction method of multipliers* or ADMM.

We will now formulate the ADMM and discuss several applications as well as its convergence properties. The starting point are minimization problems of the form

$$\min_{x\in\mathbb{R}^n} \; f(x) + g(Ax), \tag{13.1}$$

where A is an $m \times n$ matrix and $f : \mathbb{R}^n \to (-\infty, +\infty]$ and $g : \mathbb{R}^m \to (-\infty, +\infty]$ are convex, lower semicontinuous, and proper functions. We convert this problem to the following equivalent constrained minimization problem

$$\min \; f(x) + g(y) \quad x \in \mathbb{R}^n, \quad y \in \mathbb{R}^m, \quad Ax - y = 0, \tag{13.2}$$

and we introduce the associated augmented Lagrangian function

$$L_\sigma(x, y, \lambda) = f(x) + g(y) + \lambda^\top (Ax - y) + \frac{\sigma}{2}\|Ax - y\|^2.$$

The ADMM, given the current iterates $(x^k, y^k, \lambda^k) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$, generates a new iterate $(x^{k+1}, y^{k+1}, \lambda^{k+1})$ by first minimizing the augmented Lagrangian with respect to x, then with respect to y, and finally performing a multiplier update:

$$\begin{aligned} x^{k+1} &\in \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \; L_\sigma(x, y^k, \lambda^k) \\ y^{k+1} &\in \operatorname*{arg\,min}_{y\in\mathbb{R}^m} \; L_\sigma(x^{k+1}, y, \lambda^k) \\ \lambda^{k+1} &= \lambda^k + \gamma\sigma(Ax^{k+1} - y^{k+1}), \end{aligned}$$

where $\gamma > 0$. Utilizing the notion of the proximity operator, the minimization with respect to y can also be written more compactly

$$y^{k+1} = \operatorname*{arg\,min}_{y\in\mathbb{R}^m} \; g(y) + (\lambda^k)^\top (Ax^{k+1} - y) + \frac{\sigma}{2}\|Ax^{k+1} - y\|^2 = \operatorname{prox}_{g/\sigma}(Ax^{k+1} + \sigma^{-1}\lambda^k).$$

The penalty parameter σ is typically kept constant in ADMM. The important advantage that the ADMM may offer over the augmented Lagrangian method is that it involves separate minimization with respect to x and with respect to y. Thus, the complications resulting from the coupling of x and y in the penalty term $\|Ax - y\|^2$ are eliminated.

13.1. Applications: Sparse Recovery and TV-Models

Before presenting convergence results and a more refined version of the algorithm, we consider several applications and examples.

Example 13.1 (Sparse Recovery). We reconsider the ℓ_1-optimization problem

$$\min_x \ \frac{1}{2}\|Ax-b\|^2 + \mu\|x\|_1,$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, and $\mu > 0$ are given. Introducing the auxiliary variable $y = x$, this problem is equivalent to

$$\min_{x,y} \ \frac{1}{2}\|Ax-b\|^2 + \mu\|y\|_1 \quad \text{s.t.} \quad x - y = 0.$$

The ADMM iteration then takes the form

$$\left[\begin{array}{l} x^{k+1} \in \arg\min_{x\in\mathbb{R}^n} \ \frac{1}{2}\|Ax-b\|^2 + (\lambda^k)^\top(x-y^k) + \frac{\sigma}{2}\|x-y^k\|^2 \\ y^{k+1} = \operatorname{prox}_{\mu\sigma^{-1}\|\cdot\|_1}(x^{k+1} + \sigma^{-1}\lambda^k) \\ \lambda^{k+1} = \lambda^k + \gamma\sigma(x^{k+1} - y^{k+1}). \end{array}\right.$$

The x-step can be further simplified. Specifically, rearranging the associated optimality condition $A^\top(Ax-b) + \lambda^k + \sigma(x-y^k) = 0$, we obtain

$$x^{k+1} = [A^\top A + \sigma I]^{-1}(A^\top b - \lambda^k + \sigma y^k).$$

Moreover, using the Sherman-Morrison-Woodbury formula

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

the matrix inversion can also be calculated via $[A^\top A + \sigma I]^{-1} = \frac{1}{\sigma}I - \frac{1}{\sigma}A^\top(\sigma I + AA^\top)^{-1}A$. If $m \ll n$, then the $m \times m$ matrix $\sigma I + AA^\top$ has a much smaller dimension and its inverse can be either computed explicitly or via a Cholesky decomposition.

Example 13.2 (Total Variation Minimization). Next, let us consider the total variation inpainting problem

$$\min_x \ \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} \|D_{(i,j)}x\|_2 \quad \text{s.t.} \quad Ax = b,$$

where $A \in R^{s\times mn}$ is a given inpainting mask, $b \in \mathbb{R}^s$ stores undamaged parts of the ground truth image and $D_{(i,j)} \in \mathbb{R}^{2\times mn}$ represents the image gradient at pixel (i,j) for the vectorized image $x \in \mathbb{R}^{mn}$. Defining $C := \{x \in \mathbb{R}^{mn} : Ax = b\}$, this problem can be reformulated as follows:

$$\min_{x,y_{ij}} \ \sum_{i=1}^{m-1}\sum_{j=1}^{n-1} \|y_{ij}\|_2 + \iota_C(x) \quad \text{s.t.} \quad D_{(i,j)}x - y_{ij} = 0 \quad \forall\, i,j.$$

The ADMM steps for this formulation are given by:

$$
\begin{aligned}
x^{k+1} &\in \arg\min_x \ \iota_C(x) + \sum_{i=1}^{m-1}\sum_{j=1}^{n-1}(\lambda_{ij}^k)^\top(D_{(i,j)}x - y_{ij}^k) + \frac{\sigma}{2}\|D_{(i,j)}x - y_{ij}^k\|^2 \\
y_{ij}^{k+1} &= \operatorname{prox}_{\|\cdot\|_2/\sigma}(D_{(i,j)}x^{k+1} + \sigma^{-1}\lambda_{ij}^k) \\
\lambda_{ij}^{k+1} &= \lambda_{ij}^k + \gamma\sigma(D_{(i,j)}x^{k+1} + \sigma^{-1}\lambda_{ij}^k)
\end{aligned}
$$

for all $i = 1, ..., m-1$, $j = 1, ..., n-1$. Notice that the first step corresponds to a quadratic program with linear constraints which might be time-consuming to solve.

13.2. Semi-Proximal Alternating Direction Method of Multiplier

We now analyze a more general method than ADMM in which a quadratic proximity term is added to the objective function in the two minimization problems defined in ADMM. We suppose that there are two given positive semidefinite matrices $S \in \mathbb{S}_+^n$, $T \in \mathbb{S}_+^m$ and we set $\|x\|_S^2 = x^\top Sx$ and $\|y\|_T^2 = y^\top Ty$. The so-called semi-proximal ADMM is presented in Algorithm 13.1.

Algorithm 13.1: sp-ADMM

1 Initialization: Choose an initial points $x^0 \in \mathbb{R}^n$, $y^0, \lambda^0 \in \mathbb{R}^m$, and $\sigma, \gamma \in (0,1)$.
for $k = 0, 1, 2, ...$ **do**
Perform the following updates:
2 $x^{k+1} \in \arg\min_{x\in\mathbb{R}^n} \ f(x) + \frac{\sigma}{2}\|Ax - y^k + \sigma^{-1}\lambda^k\|^2 + \frac{1}{2}\|x - x^k\|_S^2$.
3 $y^{k+1} \in \arg\min_{y\in\mathbb{R}^m} \ g(y) + \frac{\sigma}{2}\|Ax^{k+1} - y + \sigma^{-1}\lambda^k\|^2 + \frac{1}{2}\|y - y^k\|_T^2$.
4 $\lambda^{k+1} = \lambda^k + \gamma\sigma(Ax^{k+1} - y^{k+1})$.

One important observation and motivation for considering sp-ADMM is that by using the additional proximity terms, the minimization in step 2 and 3 can be considerably simplified by choosing $S = \tau I - \sigma A^\top A$ with $\tau \geq \sigma\lambda_{\max}(A^\top A)$ and $T = 0$ (other choices are possible in more general settings). Then, we obviously have $S \in \mathbb{S}_n^+$ and the x-step can be simplified as follows:

$$
\begin{aligned}
&f(x) + \frac{\sigma}{2}\|Ax - y^k + \sigma^{-1}\lambda^k\|^2 + \frac{1}{2}\|x - x^k\|_S^2 \\
&\quad = f(x) + \frac{\sigma}{2}\|A(x - x^k)\|^2 + \sigma\langle A(x - x^k), Ax^k - y^k + \sigma^{-1}\lambda^k\rangle + \frac{1}{2}\|x - x^k\|_S^2 + \text{cons.} \\
&\quad = f(x) + \langle A(x - x^k), \sigma(Ax^k - y^k) + \lambda^k\rangle + \frac{\tau}{2}\|x - x^k\|^2 + \text{cons.} \\
&\quad = f(x) + \frac{\tau}{2}\|x - x^k + \frac{\sigma}{\tau}A^\top[Ax^k - y^k + \sigma^{-1}\lambda^k]\|^2 + \text{cons.}
\end{aligned}
$$

Thus, step 2 and 3 can be expressed explicitly via

$$x^{k+1} = \mathrm{prox}_{f/\tau}(x^k - \frac{\sigma}{\tau} A^\top [Ax^k - y^k + \sigma^{-1}\lambda^k])$$
$$y^{k+1} = \mathrm{prox}_{g/\sigma}(Ax^{k+1} + \sigma^{-1}\lambda^k).$$

This special version of ADMM is called semi-proximal linearized ADMM. In the updates in step 2 and 3, we actually linearize the original quadratic terms and add a quadratic proximity term. We now present a streamlined convergence result for Algorithm 13.1.

Theorem 13.3: Global Convergence of sp-ADMM

Suppose that $f : \mathbb{R}^n \to (-\infty, +\infty]$ and $g : \mathbb{R}^m \to (-\infty, +\infty]$ are convex, lower semicontinuous, and proper functions and that there exists

$$\hat{x} \in \mathrm{ri}(\mathrm{dom}\ f), \quad \hat{y} \in \mathrm{ri}(\mathrm{dom}\ g) \quad \text{such that} \quad A\hat{x} = \hat{y}.$$

We further assume $\sigma > 0$, $\gamma \in (0, (1+\sqrt{5})/2)$,

$$T \succeq 0, \quad S + \sigma A^\top A \succ 0,$$

and that the solution set of problem (13.1) (or (13.2)) is nonempty. Then, the sequence $(x^k, y^k, \lambda^k)_k$ generated by Algorithm 13.1 is well-defined and converges to a saddle-point (x^*, y^*, λ^*) of the Lagrange function $L(x, y, \lambda) = f(x) + g(y) + \lambda^\top(Ax - y)$.

A. MATLAB Code for section 5

Listing 1: Gradient Method for Quadratic Problems

```
function [x,obj] = gm_quadratic(A,b,x0,eps)

% === INPUT ==========
% A      the positive definite matrix associated with the objective function
% b      a column vector associated with the linear part of the objective
%        function
% x0     initial point
% eps    tolerance parameter
% === OUTPUT =========
% x      an optimal solution of min 0.5*x^T A x + b^T x
% obj    the optimal function value up to a tolerance

x      = x0;
iter   = 0;
g      = A*x + b;
ng     = norm(g);

fprintf(1,'- - - gradient method for quadratic programs; n = %g\n',length(b));
fprintf(1,'ITER ; OBJ.VAL ; G.NORM ; STEP.SIZE\n');

% main loop
while ng > eps && iter < 10000
 iter   = iter + 1;
 alpha  = ng^2 / (g'*A*g);
 x      = x - alpha*g;
 g      = A*x + b;
 ng     = norm(g);
 obj    = 0.5*x'*A*x + b'*x;

 fprintf(1,'[%4i] ; %2.6f ; %2.6f ; %1.2f\n',iter,obj,ng,alpha);
end
```

Listing 2: Gradient Method with Backtracking

```
function [x,obj] = gm_armijo(f,x0,opts)

% === INPUT ==========
% f      a struct for the function f: f.obj(x) returns f(x); f.grad(x)
%        returns the gradient of f at x
% x0     initial point
% opts   a struct for the options:
%          - .maxit  maximum number iterations
%.         - .tol    tolerance
%          - .gamma  line search parameter
%          - .s      line search parameter
% === OUTPUT =========
% x      an optimal solution of min f(x)
```

```
% obj    the optimal function value up to a tolerance

x      = x0;       iter   = 0;
f_old  = f.obj(x); alpha  = 0;

fprintf(1,'- - - gradient method with backtracking;\n');
fprintf(1,'ITER ; OBJ.VAL ; G.NORM ; STEP.SIZE\n');

% main loop
while iter < opts.maxit
 iter     = iter + 1;   x_old   = x;
 g        = f.grad(x);  ng      = norm(g);

 fprintf(1,'[%4i] ; %2.6f ; %2.6f ; %1.2f\n',iter-1,f_old,ng,alpha);

 if ng <= opts.tol
  break;
 end

 alpha      = opts.s;   x         = x_old - alpha*g;
 f_new      = f.obj(x); a_counter = 1;

 while f_new - f_old > - alpha*opts.gamma*ng^2 && a_counter <= 10
  alpha = alpha/2;  x         = x_old - alpha*g;
  f_new = f.obj(x); a_counter = a_counter + 1;
 end

 f_old = f_new;
end
```

B. Newton's Method for General Nonlinear Equations

In this section, we briefly discuss Newton's method to solve nonlinear equations of the form

$$F(x) = 0,$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$ is a continuously differentiable function. Similar to Newton's method for optimization problems, the key idea is to linearize $F(x^k + d)$ and to solve the resulting linear system of equations to obtain the next step. Specifically, we consider the iterative scheme:

- Solve the linear system $DF(x^k)d = -F(x^k)$ and perform the update $x^{k+1} = x^k + d$.

The full method is shown in Algorithm B.1.

The convergence of this general Newton method is based on a generalization of the invertibility result Lemma 6.2.

Algorithm B.1: Pure Newton's Method for Nonlinear Equations

Initialization: Select an initial point $x^0 \in \mathbb{R}^n$.

for $k = 0, 1, ...$ **do**

Compute the Newton direction d^k which is the solution of the linear system

$$DF(x^k)d^k = -F(x^k).$$

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

If $\|F(x^{k+1})\| \leq \varepsilon$, then STOP and x^{k+1} is the output.

Lemma B.1: Banach Lemma

The set $\mathcal{W} \subset \mathbb{R}^{n\times n}$ of invertible matrices is open and the mapping $\mathcal{W} \ni M \mapsto M^{-1} \in \mathbb{R}^{n\times n}$ is continuous. More specifically, for all $A \in \mathcal{W}$ and $B \in \mathbb{R}^{n\times n}$ with $\|A^{-1}B\| < 1$ (this particularly holds if $\|A^{-1}\|\|B\| < 1$), the matrix $A + B$ is invertible and we have

$$\|(A+B)^{-1}\| \leq \frac{\|A^{-1}\|}{1-\|A^{-1}B\|} \quad \text{and} \quad \|(A+B)^{-1} - A^{-1}\| \leq \frac{\|A^{-1}\|\|A^{-1}B\|}{1-\|A^{-1}B\|}.$$

Proof. Let $A \in \mathcal{W}$ and $B \in \mathbb{R}^{n\times n}$ with $\|A^{-1}B\| < 1$ be arbitrary and let us define $M := -A^{-1}B$. Then the Neumann series $S = \sum_{k=0}^{\infty} M$ (we set $M^0 = I$) converges. In particular, setting $S_n := \sum_{k=0}^{n} M^k$, we have

$$\|S - S_n\| \leq \sum_{k=n+1}^{\infty} \|M\|^k = \|M\|^{n+1} \sum_{k=0}^{\infty} \|M\|^k = \frac{\|M\|^{n+1}}{1-\|M\|} \to 0$$

as $n \to \infty$. Moreover, it follows

$$S_n(I-M) = (I-M)S_n = (I-M)\sum_{k=0}^{n} M^k = I - M^{n+1}.$$

Taking the limit $n \to \infty$, this yields $S(I-M) = (I-M)S = I$ which implies $I - M \in \mathcal{W}$ and $(I-M)^{-1} = S$. Due to $A+B = A(I-M)$ we have $A+B \in \mathcal{W}$ and $(A+B)^{-1} = SA^{-1}$. Finally, we obtain

$$\begin{aligned}
\|(A+B)^{-1}\| &\leq \|A^{-1}\|\|S\| \leq \|A^{-1}\| \sum_{k=0}^{\infty} \|M\|^k = \frac{\|A^{-1}\|}{1-\|M\|}, \\
\|(A+B)^{-1} - A^{-1}\| &\leq \|SA^{-1} - A^{-1}\| \left\| \sum_{k=0}^{\infty} M^k A^{-1} - A^{-1} \right\| \\
&\leq \|A^{-1}\| \sum_{k=1}^{\infty} \|M\|^k \leq \frac{\|A^{-1}\|\|M\|}{1-\|M\|}.
\end{aligned}$$

For $\|B\| \to 0$, the term $\|M\|$ converges to 0, which shows continuity of matrix inversion. ■

We can also establish the following variant of Lemma 5.19.

Lemma B.2: Isolated Solution
Suppose that $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable in a neighborhood of a solution $\bar{x}$ with $F(\bar{x}) = 0$ and assume that the derivative $DF(\bar{x})$ is invertible. Then there exists $\varepsilon, \eta > 0$ such that

$$\|F(x)\| \geq \eta \|x - \bar{x}\| \quad \forall\, x \in B_\varepsilon(\bar{x}).$$

In particular, $\bar{x}$ is an isolated solution of the mapping F.

The proof is identical to the proof of Lemma 5.19 exchanging ∇f and F.

Theorem B.3: Local Convergence of the General Newton Method
Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable function and let $x^* \in \mathbb{R}^n$ be a point satisfying $F(x^*) = 0$ and assume that the Jacobian $DF(x^*)$ is invertible. Then there are $\varepsilon > 0$ and $C > 0$ such that

- The point x^* is the only zero of F in $B_\varepsilon(x^*)$.
- We have $\|DF(x)^{-1}\| \leq C$ for all $x \in B_\varepsilon(x^*)$.
- For every $x^0 \in B_\varepsilon(x^*)$, Algorithm B.1 either terminates with $x^k = x^*$ or it generates a sequence $(x^k)_k \subset B_\varepsilon(x^*)$ that converges q-superlinearly to x^*.
- In addition, if DF is Lipschitz continuous on $B_\delta(x^*)$ with constant L, then the rate of convergence is q-quadratic (if the algorithm does not terminate after finitely many steps) and it holds that:

$$\|x^{k+1} - x^*\| \leq \frac{CL}{2} \|x^k - x^*\|^2 \quad \forall\, k.$$

The proof is based on Lemma B.1 and Lemma B.2. The statements can then be shown by mimicking the steps of the proof of Theorem 6.1.

References

[1] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, New York, 2011.

[2] A. Beck, *Introduction to nonlinear optimization*, vol. 19 of MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2014. Theory, algorithms, and applications with MATLAB.

[3] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*, Computer Science and Applied Mathematics, Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1982.

[4] ——, *Convex optimization theory*, Athena Scientific, Nashua, NH, 2009.

[5] ——, *Nonlinear programming*, Athena Scientific Optimization and Computation Series, Athena Scientific, Belmont, MA, third ed., 2016.

[6] R. Burachik, L. M. G. n. Drummond, A. N. Iusem, and B. F. Svaiter, *Full convergence of the steepest descent method with inexact line searches*, Optimization, 32 (1995), pp. 137–146.

[7] J. V. Burke, *Numerical optimization.* Course Notes, AMath/Math 516, University of Washington, Spring Term, 2012.

[8] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer Series in Operations Research and Financial Engineering, Springer, New York, second ed., 2006.

[9] M. Ulbrich and S. Ulbrich, *Nichtlineare Optimierung*, Birkhäuser, Basel, 2012.