Numerical Mathematics I

Peter Philip*

Lecture Notes

Originally Created for the Class of Winter Semester 2008/2009 at LMU Munich,
Revised and Extended for Several Subsequent Classes

February 25, 2020

Contents

1 Introduction and Motivation .................................................. 5

2 Tools: Landau Symbols, Norms, Condition .............................. 9
   2.1 Landau Symbols ................................................................ 9
   2.2 Operator Norms and Natural Matrix Norms ....................... 15
   2.3 Condition of a Problem ...................................................... 20

3 Interpolation ........................................................................... 32
   3.1 Motivation ........................................................................ 32
   3.2 Polynomial Interpolation .................................................... 33
      3.2.1 Polynomials and Polynomial Functions ...................... 33
      3.2.2 Existence and Uniqueness ......................................... 35
      3.2.3 Hermite Interpolation ............................................... 37
      3.2.4 Divided Differences and Newton’s Interpolation Formula .. 44
      3.2.5 Convergence and Error Estimates ............................... 52
      3.2.6 Chebyshev Polynomials ............................................. 54
   3.3 Spline Interpolation .......................................................... 57

*E-Mail: philip@math.lmu.de
# CONTENTS

3.3.1 Introduction, Definition of Splines .................................. 57  
3.3.2 Linear Splines ....................................................... 59  
3.3.3 Cubic Splines ....................................................... 60  

4 Numerical Integration .................................................... 68  
4.1 Introduction ............................................................. 68  
4.2 Quadrature Rules Based on Interpolating Polynomials .............. 73  
4.3 Newton-Cotes Formulas ................................................ 76  
4.3.1 Definition, Weights, Degree of Accuracy ....................... 76  
4.3.2 Rectangle Rules \( n = 0 \) ........................................... 79  
4.3.3 Trapezoidal Rule \( n = 1 \) ......................................... 80  
4.3.4 Simpson’s Rule \( n = 2 \) ........................................... 80  
4.3.5 Higher Order Newton-Cotes Formulas ......................... 81  
4.4 Convergence of Quadrature Rules ................................... 82  
4.5 Composite Newton-Cotes Quadrature Rules ....................... 84  
4.5.1 Introduction, Convergence ........................................ 84  
4.5.2 Composite Rectangle Rules \( n = 0 \) ............................ 86  
4.5.3 Composite Trapezoidal Rules \( n = 1 \) ......................... 87  
4.5.4 Composite Simpson’s Rules \( n = 2 \) ........................... 88  
4.6 Gaussian Quadrature ................................................... 88  
4.6.1 Introduction ......................................................... 88  
4.6.2 Orthogonal Polynomials .......................................... 89  
4.6.3 Gaussian Quadrature Rules ..................................... 92  

5 Numerical Solution of Linear Systems .................................. 96  
5.1 Background and Setting .............................................. 96  
5.2 Pivot Strategies for Gaussian Elimination ......................... 97  
5.3 Cholesky Decomposition .............................................. 101  
5.4 QR Decomposition ..................................................... 103  
5.4.1 Definition and Motivation ....................................... 103  
5.4.2 QR Decomposition via Gram-Schmidt Orthogonalization ...... 105  
5.4.3 QR Decomposition via Householder Reflections ............. 109  

6 Iterative Methods, Solution of Nonlinear Equations ................ 115
CONTENTS

6.1 Motivation: Fixed Points and Zeros .................................................. 115
6.2 Newton’s Method ................................................................................... 115
   6.2.1 Newton’s Method in One Dimension ............................................. 116
   6.2.2 Newton’s Method in Several Dimensions ..................................... 119

7 Eigenvalues ........................................................................................................... 123
   7.1 Introductory Remarks .............................................................................. 123
   7.2 Estimates, Localization ......................................................................... 125
   7.3 Power Method .......................................................................................... 129
   7.4 QR Method .............................................................................................. 134

8 Minimization ............................................................................................................. 140
   8.1 Motivation, Setting .................................................................................. 140
   8.2 Local Versus Global and the Role of Convexity .................................... 141
   8.3 Gradient-Based Descent Methods ............................................................ 146
   8.4 Conjugate Gradient Method ..................................................................... 154

A Representations of Real Numbers, Rounding Errors .............................................. 158
   A.1 $b$-Adic Expansions of Real Numbers .................................................. 158
   A.2 Floating-Point Numbers Arithmetic, Rounding ....................................... 160
   A.3 Rounding Errors ....................................................................................... 166

B Operator Norms and Natural Matrix Norms .......................................................... 170

C Linear Algebra ..................................................................................................... 172
   C.1 Positive Semidefinite Matrices ................................................................. 172

D Integration ............................................................................................................... 174
   D.1 $\mathbb{K}^m$-Valued Integration ................................................................. 174
   D.2 Continuity ................................................................................................. 176

E Divided Differences ................................................................................................. 176

F DFT and Trigonometric Interpolation ..................................................................... 180
   F.1 Discrete Fourier Transform (DFT) ........................................................... 180
   F.2 Approximation of Periodic Functions ....................................................... 183
CONTENTS

F.3 Trigonometric Interpolation ............................... 187
F.4 Fast Fourier Transform (FFT) ............................ 191

G The Weierstrass Approximation Theorem .............. 197

H Baire Category and the Uniform Boundedness Principle 199

I Matrix Decomposition .................................. 202
  I.1 Cholesky Decomposition of Positive Semidefinite Matrices 202
  I.2 Multiplication with Householder Matrix over $\mathbb{R}$ .... 206

J Newton’s Method ........................................ 207

K Eigenvalues ........................................... 209
  K.1 Continuous Dependence on Matrix Coefficients .......... 209
  K.2 Transformation to Hessenberg Form ..................... 211
  K.3 QR Method for Matrices in Hessenberg Form .......... 214

L Minimization ........................................ 217
  L.1 Golden Section Search .............................. 217
  L.2 Nelder-Mead Method .............................. 222
1 Introduction and Motivation

The central motivation of Numerical Mathematics is to provide constructive and effective methods (so-called algorithms, see Def. 1.1 below) that reliably compute solutions (or sufficiently accurate approximations of solutions) to classes of mathematical problems. Moreover, such methods should also be efficient, i.e. one would like the algorithm to be as quick as possible while one would also like it to use as little memory as possible. Frequently, both goals can not be achieved simultaneously: For example, one might decide to recompute intermediate results (which needs more time) to avoid storing them (which would require more memory) or vice versa. One typically also has a trade-off between accuracy and requirements for memory and execution time, where higher accuracy means use of more memory and longer execution times.

Thus, one of the main tasks of Numerical Mathematics consists of proving that a given method is constructive, effective, and reliable. That a method is constructive, effective, and reliable means that, given certain hypotheses, it is guaranteed to converge to the solution. This means, it either finds the solution in a finite number of steps, or, more typically, given a desired error bound, within a finite number of steps, it approximates the true solution such that the error is less than the given bound. Proving error estimates is another main task of Numerical Mathematics and so is proving bounds on an algorithm’s complexity, i.e. bounds on its use of memory (i.e. data) and run time (i.e. number of steps). Moreover, in addition to being convergent, for a method to be useful, it is of crucial importance that is also stable in the sense that a small perturbation of the input data does not destroy the convergence and results in, at most, a small increase of the error. This is of the essence as, for most applied problems, the input data will not be exact, and most algorithms are subject to round-off errors.

Instead of a method, we will usually speak of an algorithm, by which we mean a “useful” method. To give a mathematically precise definition of the notion algorithm is beyond the scope of this lecture (it would require an unjustifiably long detour into the field of logic), but the following definition will be sufficient for our purposes.

Definition 1.1. An algorithm is a finite sequence of instructions for the solution of a class of problems. Each instruction must be representable by a finite number of symbols. Moreover, an algorithm must be guaranteed to terminate after a finite number of steps.

Remark 1.2. Even though we here require an algorithm to terminate after a finite number of steps, in the literature, one sometimes omits this part from the definition. The question if a given method can be guaranteed to terminate after a finite number of steps is often tremendously difficult (sometimes even impossible) to answer.

Example 1.3. Let \( a, a_0 \in \mathbb{R}^+ \) and consider the sequence \((x_n)_{n \in \mathbb{N}_0}\) defined recursively by

\[
x_0 := a_0, \quad x_{n+1} := \frac{1}{2} \left( x_n + \frac{a}{x_n} \right) \quad \text{for each } n \in \mathbb{N}_0.
\]  

It is an exercise to show that, for each \( a, a_0 \in \mathbb{R}^+ \), this sequence is well-defined (i.e. \( x_n > 0 \) for each \( n \in \mathbb{N}_0 \)) and converges to \( \sqrt{a} \) (this is Newton’s method for the computation
of the zero of the function $f : \mathbb{R}^+ \to \mathbb{R}$, $f(x) := x^2 - a$, cf. Ex. 6.4(b)). The $x_n$ can be computed using the following finite sequence of instructions:

1: $x = a_0$ % store the number $a_0$ in the variable $x$
2: $x = (x + a/x)/2$ % compute $(x + a/x)/2$ and replace the 
   % contents of the variable $x$ with the computed value \hfill (1.2)
3: goto 2 % continue with instruction 2

Even though the contents of the variable $x$ will converge to $\sqrt{a}$, (1.2) does not constitute an algorithm in the sense of Def. 1.1 since it does not terminate. To guarantee termination and to make the method into an algorithm, one might introduce the following modification:

1: $\epsilon = 10^{-10} \times a$ % store the number $10^{-10}a$ in the variable $\epsilon$
2: $x = a_0$ % store the number $a_0$ in the variable $x$
3: $y = x$ % copy the contents of the variable $x$ to 
   % the variable $y$ to save the value for later use
4: $x = (x + a/x)/2$ % compute $(x + a/x)/2$ and replace the 
   % contents of the variable $x$ with the computed value \hfill (1.3)
5: if $|x - y| > \epsilon$
   then goto 3 % if $|x - y| > \epsilon$, then continue with instruction 3
   else quit % if $|x - y| \leq \epsilon$, then terminate the method

Now the convergence of the sequence guarantees that the method terminates within finitely many steps.

Another problem with regard to algorithms, that we already touched on in Ex. 1.3, is the implicit requirement of Def. 1.1 for an algorithm to be well-defined. That means, for every initial condition, given a number $n \in \mathbb{N}$, the method has either terminated after $m \leq n$ steps, or it provides a (feasible!) instruction to carry out step number $n + 1$. Such methods are called complete. Methods that can run into situations, where they have not reached their intended termination point, but can not carry out any further instruction, are called incomplete. Algorithms must be complete! We illustrate the issue in the next example:

**Example 1.4.** Let $a \in \mathbb{R} \setminus \{2\}$ and $N \in \mathbb{N}$. Define the following finite sequence of instructions:

1: $n = 1$; \hspace{0.5em} $x = a$
2: $x = 1/(2 - x)$; \hspace{0.5em} $n = n + 1$
3: if $n \leq N$
   then goto 2
   else quit \hfill (1.4)
Consider what occurs for $N = 10$ and $a = \frac{5}{4}$. The successive values contained in
the variable $x$ are $\frac{5}{4}, \frac{4}{3}, \frac{3}{2}, 2$. At this stage $n = 4 \leq N$, i.e. instruction 3 tells the
method to continue with instruction 2. However, the denominator has become 0, and
the instruction has become meaningless. The following modification makes the method
complete and, thereby, an algorithm:

\begin{equation}
\begin{aligned}
1 : & \quad n = 1; \quad x = a \\
2 : & \quad \text{if } x \neq 2 \\
& \quad \quad \text{then } x = 1/(2 - x); \quad n = n + 1 \\
& \quad \quad \text{else } x = -5; \quad n = n + 1 \\
3 : & \quad \text{if } n \leq N \\
& \quad \quad \text{then goto 2} \\
& \quad \quad \text{else quit}
\end{aligned}
\end{equation}

We can only expect to find stable algorithms if the underlying problem is sufficiently
benign. This leads to the following definition:

**Definition 1.5.** We say that a mathematical problem is **well-posed** if, and only if, its
solutions enjoy the three benign properties of **existence**, **uniqueness**, and **continuity** with
respect to the input data. More precisely, given admissible input data, the problem
must have a unique solution (output), thereby providing a map between the set of
admissible input data and (a superset of) the set of possible solutions. This map must
be continuous with respect to suitable norms or metrics on the respective sets (small
changes of the input data must only cause small changes in the solution). A problem
which is not well-posed is called **ill-posed**.

We can thus add to the important tasks of Numerical Mathematics mentioned earlier
the additional important tasks of investigating a problem’s well-posedness. Then, once
well-posedness is established, the task is to provide a stable algorithm for its solution.

**Example 1.6. (a)** The problem “find a minimum of a given polynomial function $p : \mathbb{R} \rightarrow \mathbb{R}$” is inherently ill-posed: Depending on $p$, the problem has no solution
(e.g. $p(x) = x$), a unique solution (e.g. $p(x) = x^2$), finitely many solutions (e.g.
$p(x) = x^2(x - 1)^2(x + 2)^2$) or infinitely many solutions (e.g. $p(x) = 1$).

(b) Frequently, one can transform an ill-posed problem into a well-posed problem, by
choosing an appropriate setting: Consider the problem “find a zero of $f(x) = \lambda x^2 + c$”. If, for example, one admits $a, c \in \mathbb{R}$ and looks for solutions in $\mathbb{R}$, then the
problem is ill-posed as one has no solutions for $ac > 0$ and no solutions for $a = 0,
c \neq 0$. Even for $ac < 0$, the problem is not well-posed as the solution is not always
unique. However, in this case, one can make the problem well-posed by considering
solutions in $\mathbb{R}^2$. The correspondence between the input and the solution (sometimes referred to as the solution operator) is then given by the continuous map

$$S : \{(a, c) \in \mathbb{R}^2 : ac < 0\} \to \mathbb{R}^2, \quad S(a, c) := \left(\sqrt{-\frac{c}{a}}, -\sqrt{-\frac{c}{a}}\right). \quad (1.6a)$$

The problem is also well-posed when just requiring $a \neq 0$, but admitting complex solutions. The continuous solution operator is then given by

$$S : \{(a, c) \in \mathbb{R}^2 : a \neq 0\} \to \mathbb{C}^2,$$

$$S(a, c) := \begin{cases} 
\left(\sqrt{|c| / |a|}, -\sqrt{|c| / |a|}\right) & \text{for } ac \leq 0, \\
(i \sqrt{|c| / |a|}, -i \sqrt{|c| / |a|}) & \text{for } ac > 0.
\end{cases} \quad (1.6b)$$

(c) The problem “determine if $x \in \mathbb{R}$ is positive” might seem simple at first glance, however it is ill-posed, as it is equivalent to computing the values of the function

$$S : \mathbb{R} \to \{0, 1\}, \quad S(x) := \begin{cases} 
1 & \text{for } x > 0, \\
0 & \text{for } x \leq 0,
\end{cases} \quad (1.7)$$

which is discontinuous at 0.

As stated before, the analysis and control of errors is of central interest. Errors occur due to several causes:

(1) **Modeling Errors**: A mathematical model can only approximate the physical situation in the best of cases. Often models have to be further simplified in order to compute solutions and to make them accessible to mathematical analysis.

(2) **Data Errors**: Typically, there are errors in the input data. Input data often result from measurements of physical experiments or from calculations that are potentially subject to every type of error in the present list.

(3) **Blunders**: For example, logical errors and implementation errors.

One should always be aware that errors of the types just listed will or can be present. However, in the context of Numerical Mathematics, one focuses mostly on the following error types:

(4) **Truncation Errors**: Such errors occur when replacing an infinite process (e.g. an infinite series) by a finite process (e.g. a finite summation).

(5) **Round-Off Errors**: Errors occurring when discarding digits needed for the exact representation of a (e.g. real or rational) number.
In an increasing manner, the functioning of our society relies on the use of numerical algorithms. In consequence, avoiding and controlling numerical errors is vital. Several examples of major disasters caused by numerical errors can be found on the following web page of D.N. Arnold at the University of Minnesota:
http://www.ima.umn.edu/~arnold/disasters/
For a much more comprehensive list of numerical and related errors that had significant consequences, see the web page of T. Huckle at TU Munich:
http://www5.in.tum.de/~huckle/bugse.html

2 Tools: Landau Symbols, Norms, Condition

Notation 2.1. We will write $\mathbb{K}$ in situations, where we allow $\mathbb{K}$ to be $\mathbb{R}$ or $\mathbb{C}$.

2.1 Landau Symbols

When calculating errors (and also when calculating the complexity of algorithms), one is frequently not so much interested in the exact value of the error (or the computing time and size of an algorithm), but only in the order of magnitude and in the asymptotics. The Landau symbols $O$ (big $O$) and $o$ (small $o$) are a notation in support of these facts. Here is the precise definition:

Definition 2.2. Let $(X, d)$ be a metric space and let $(Y, \| \cdot \|), (Z, \| \cdot \|)$ be normed vector spaces over $\mathbb{K}$. Let $D \subseteq X$ and consider functions $f : D \rightarrow Y$, $g : D \rightarrow Z$, where we assume $g(x) \neq 0$ for each $x \in D$. Moreover, let $x_0 \in X$ be a cluster point of $D$ (note that $x_0$ does not have to be in $D$, and $f$ and $g$ do not have to be defined in $x_0$).

(a) $f$ is called of order big $O$ of $g$ or just big $O$ of $g$ (denoted by $f(x) = O(g(x))$) for $x \to x_0$ if, and only if,
$$\limsup_{x \to x_0} \frac{\|f(x)\|}{\|g(x)\|} < \infty$$
(i.e. there exists $M \in \mathbb{R}_0^+$ such that, for each sequence $(x_n)_{n \in \mathbb{N}}$ in $D$ with the property $\lim_{n \to \infty} x_n = x_0$, the sequence $(\frac{\|f(x_n)\|}{\|g(x_n)\|})_{n \in \mathbb{N}}$ in $\mathbb{R}_0^+$ is bounded from above by $M$).

(b) $f$ is called of order small $o$ of $g$ or just small $o$ of $g$ (denoted by $f(x) = o(g(x))$) for $x \to x_0$ if, and only if,
$$\lim_{x \to x_0} \frac{\|f(x)\|}{\|g(x)\|} = 0$$
(i.e., for each sequence $(x_n)_{n \in \mathbb{N}}$ in $D$ with the property $\lim_{n \to \infty} x_n = x_0$, we have $\lim_{n \to \infty} \frac{\|f(x_n)\|}{\|g(x_n)\|} = 0$).

As mentioned before, $O$ and $o$ are known as Landau symbols.
Remark 2.3. In applications, the space $X$ in Def. 2.2 is often $\mathbb{R} := \mathbb{R} \cup \{-\infty, \infty\}$, where convergence is defined in the usual way. This usual notion of convergence is, indeed, given by a suitable metric, where, as a metric space, $\mathbb{R}$ is then homeomorphic to $[-1, 1]$, where

$$\phi : \mathbb{R} \rightarrow [-1, 1], \quad \phi(x) := \begin{cases} 
-1 & \text{for } x = -\infty, \\
\frac{x}{|x|+1} & \text{for } x \in \mathbb{R}, \\
1 & \text{for } x = \infty, 
\end{cases}$$

provides a homeomorphism (cf. [Phi17, Sec. A], in particular, [Phi17, Rem. A.5]).

Proposition 2.4. Consider the setting of Def. 2.2. In addition, define the functions

$$f_0, g_0 : D \rightarrow \mathbb{R}, \quad f_0(x) := \|f(x)\|, \quad g_0(x) := \|g(x)\|.$$ 

(a) The following statements are equivalent:

(i) $f(x) = O(g(x))$ for $x \rightarrow x_0$.

(ii) $f_0(x) = O(g_0(x))$ for $x \rightarrow x_0$.

(iii) There exist $C, \delta > 0$ such that

$$\frac{\|f(x)\|}{\|g(x)\|} \leq C \quad \text{for each } x \in D \setminus \{x_0\} \text{ with } d(x, x_0) < \delta. \quad (2.3)$$

(b) The following statements are equivalent:

(i) $f(x) = o(g(x))$ for $x \rightarrow x_0$.

(ii) $f_0(x) = o(g_0(x))$ for $x \rightarrow x_0$.

(iii) For every $C > 0$, there exists $\delta > 0$ such that (2.3) holds.

Proof. Note that

$$\forall x \in D \quad \frac{\|f(x)\|}{\|g(x)\|} = \frac{f_0(x)}{g_0(x)}. \quad (2.4)$$

(a): The equivalence between (i) and (ii) is immediate from (2.4). Now suppose (i) holds, and let $M := \limsup_{x \rightarrow x_0} \frac{\|f(x)\|}{\|g(x)\|}$, $C := 1 + M$. If there were no $\delta > 0$ such that (2.3) holds, then, for each $\delta_n := 1/n$, $n \in \mathbb{N}$, there were some $x_n \in D \setminus \{x_0\}$ with $d(x, x_0) < \delta_n$ and $y_n := \frac{\|f(x_n)\|}{\|g(x_n)\|} > C = 1 + M$, implying $\limsup_{n \rightarrow \infty} y_n \geq 1 + M > M$, in contradiction to the definition of $M$. Thus, (iii) must hold. Conversely, suppose (iii) holds, i.e. there exist $C, \delta > 0$ such that (2.3) is valid. Then, for each sequence $(x_n)_{n \in \mathbb{N}}$ in $D \setminus \{x_0\}$ such that $\lim_{n \rightarrow \infty} x_n = x_0$, the sequence $(y_n)_{n \in \mathbb{N}}$ with $y_n := \frac{\|f(x_n)\|}{\|g(x_n)\|}$ can not have a cluster point larger than $C$ (only finitely many of the $x_n$ are not in $B_\delta(x_0) := \{x \in D : d(x, x_0) < \delta\}$, which is the neighborhood of $x_0$ determined by $\delta$). Thus $\limsup_{n \rightarrow \infty} y_n \leq C$, thereby implying (2.1) and (i).

(b): Again, the equivalence between (i) and (ii) is immediate from (2.4). The equivalence between (i) and (iii), we know from Analysis 2, cf. [Phi16b, Cor. 2.9].
Due to the respective equivalences in Prop. 2.4, in the literature, the Landau symbols are often only defined for real-valued functions.

**Proposition 2.5.** Consider the setting of Def. 2.2. Suppose \( \| \cdot \|_Y \) and \( \| \cdot \|_Z \) are equivalent norms on \( Y \) and \( Z \), respectively. Then \( f(x) = O(g(x)) \) (resp. \( f(x) = o(g(x)) \)) for \( x \to x_0 \) with respect to the original norms if, and only if, \( f(x) = O(g(x)) \) (resp. \( f(x) = o(g(x)) \)) for \( x \to x_0 \) with respect to the new norms \( \| \cdot \|_Y \) and \( \| \cdot \|_Z \).

**Proof.** Let \( C_Y, C_Z \in \mathbb{R}^+ \) be such that

\[
\forall y \in Y \quad \|y\|_Y \leq C_Y \|y\|, \quad \forall z \in Z \quad C_Z \|z\| \leq \|z\|_Z.
\]

Suppose \( f(x) = O(g(x)) \) for \( x \to x_0 \) with respect to the original norms. Then

\[
\limsup_{x \to x_0} \frac{\|f(x)\|_Y}{\|g(x)\|_Z} \leq \frac{C_Y}{C_Z} \limsup_{x \to x_0} \frac{\|f(x)\|}{\|g(x)\|} < \infty,
\]

showing \( f(x) = O(g(x)) \) for \( x \to x_0 \) with respect to the new norms. Suppose \( f(x) = o(g(x)) \) for \( x \to x_0 \) with respect to the original norms. Then

\[
\lim_{x \to x_0} \frac{\|f(x)\|_Y}{\|g(x)\|_Z} = \frac{C_Y}{C_Z} \lim_{x \to x_0} \frac{\|f(x)\|}{\|g(x)\|} = 0,
\]

showing \( f(x) = o(g(x)) \) for \( x \to x_0 \) with respect to the new norms. The converse implications now also follow, since one can switch the roles of the old norms and the new norms. \( \blacksquare \)

**Example 2.6.** (a) Consider a polynomial function on \( \mathbb{R} \), i.e. \( (a_0, \ldots, a_n) \in \mathbb{R}^{n+1}, n \in \mathbb{N}_0 \), and

\[
P : \mathbb{R} \rightarrow \mathbb{R}, \quad P(x) := \sum_{i=0}^{n} a_i x^i, \quad a_n \neq 0.
\]

(2.5)

Then, for \( p \in \mathbb{R} \),

\[
P(x) = O(x^p) \quad \text{for} \quad x \to \infty \quad \Leftrightarrow \quad p \geq n, \quad \text{(2.6a)}
\]

\[
P(x) = o(x^p) \quad \text{for} \quad x \to \infty \quad \Leftrightarrow \quad p > n : \quad \text{(2.6b)}
\]

For each \( x \neq 0 \):

\[
\frac{P(x)}{x^p} = \sum_{i=0}^{n} a_i x^{i-p}.
\]

(2.7)

Since

\[
\lim_{x \to \infty} x^{i-p} = \begin{cases} 
\infty & \text{for } i - p > 0, \\
1 & \text{for } i = p, \\
0 & \text{for } i - p < 0,
\end{cases}
\]

(2.8)

(2.7) implies (2.6). Thus, in particular, for \( x \to \infty \), each constant function is big \( O \) of 1: \( a_0 = O(1) \).
Recall the notion of differentiability: If $G$ is an open subset of $\mathbb{R}^n$, $n \in \mathbb{N}$, $\xi \in G$, then $f : G \rightarrow \mathbb{R}^m$, $m \in \mathbb{N}$, is called differentiable in $\xi$ if, and only if, there exists a linear map $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and another (not necessarily linear) map $r : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that
\[
f(\xi + h) - f(\xi) = L(h) + r(h)
\] for each $h \in \mathbb{R}^n$ with sufficiently small $\|h\|_2$, and
\[
\lim_{h \rightarrow 0} \frac{r(h)}{\|h\|_2} = 0.
\]
(2.9b)

Now, using the Landau symbol $o$, (2.9b) can be equivalently expressed as
\[
r(h) = o(h) \quad \text{for } h \rightarrow 0.
\]
(2.9c)

In the literature, one also finds statements of the form $f(x) = h(x) + o(g(x))$, which are not covered by Def. 2.2 since, according to Def. 2.2, $o(g(x))$ does not denote an element of a set, i.e. adding $o(g(x))$ to (or otherwise combining it with) an element of a set does not immediately make any sense. To give a meaning to such expressions is the purpose of the following definition:

**Definition 2.7.** As in Def. 2.2, let $(X, d)$ be a metric space, let $(Y, \| \cdot \|)$, $(Z, \| \cdot \|)$ be normed vector spaces over $K$, let $D \subseteq X$ and consider functions $f : D \rightarrow Y$, $g : D \rightarrow Z$, where $g(x) \neq 0$ for each $x \in D$. Moreover, let $x_0 \in X$ be a cluster point of $D$. Now let $(V, \| \cdot \|)$ be another normed vector space and let $S \subseteq V$ be a subset\(^1\) and, for each $x \in D$, let $F_x : S \rightarrow Y$ be invertible on $f(D) \subseteq Y$. Define
\[
f(x) = F_x(O(g(x))) \quad \text{for } x \rightarrow x_0 \quad :\Leftrightarrow \quad F_x^{-1}(f(x)) = O(g(x)) \quad \text{for } x \rightarrow x_0,
\]
\[
f(x) = F_x(o(g(x))) \quad \text{for } x \rightarrow x_0 \quad :\Leftrightarrow \quad F_x^{-1}(f(x)) = o(g(x)) \quad \text{for } x \rightarrow x_0.
\]

**Example 2.8.** (a) In the situation of Ex. 2.6(b), we can now state that $f$ is differentiable in $\xi$ if, and only if, there exists a linear map $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that
\[
f(\xi + h) - f(\xi) = L(h) + o(h) \quad \text{for } h \rightarrow 0:
\]
(2.10)

Indeed, for each $h \in \mathbb{R}^n$, the map
\[
F_h : \mathbb{R}^n \rightarrow \mathbb{R}^m, \quad F_h(s) := L(h) + s,
\]
is invertible and, by Def. 2.7,
\[
f(\xi + h) - f(\xi) = L(h) + o(h) = F_h(o(h)) \quad \text{for } h \rightarrow 0
\]
is equivalent to
\[
r(h) := f(\xi + h) - f(\xi) - L(h) = F_h^{-1}(f(\xi + h) - f(\xi)) = o(h),
\]
which is (2.9c).

---

\(^1\) Even though, when using this kind of notation, one typically thinks of $F_x$ as operating on $g(x)$, as $O(x)$ and $o(x)$ are merely notations and not actual mathematical objects, it is not even necessary for $S$ to contain $g(D)$.\]
(b) Using the notation of Def. 2.7, we claim
\[
\sin x = x \ln(O(1)) \quad \text{for } x \to 0 : \tag{2.11}
\]
Indeed, for each \( x \in \mathbb{R} \setminus \{0\} \), the map
\[
F_x : \mathbb{R}^+ \to \mathbb{R}, \quad F_x(s) := x \ln s,
\]
is invertible with \( F_x^{-1} : \mathbb{R} \to \mathbb{R}^+ \), \( F_x^{-1}(y) := \exp(\frac{y}{x}) \) and, by Def. 2.7, (2.11) is equivalent to
\[
F_x^{-1}(\sin x) = e^{\sin x} = O(1),
\]
which holds, due to \( \lim_{x \to 0} \exp(\sin x) = e^{1} = e \).

(c) Let \( I \subseteq \mathbb{R} \) be an open interval and \( a, x \in I, x \neq a \). If \( m \in \mathbb{N}_0 \) and \( f \in C^{m+1}(I) \), then
\[
f(x) = T_m(x, a) + R_m(x, a), \tag{2.12a}
\]
where
\[
T_m(x, a) := \sum_{k=0}^{m} \frac{f^{(k)}(a)}{k!} (x - a)^k \tag{2.12b}
\]
is the \( m \)th Taylor polynomial and
\[
R_m(x, a) := \frac{f^{(m+1)}(\theta)}{(m + 1)!} (x - a)^{m+1} \quad \text{with some suitable } \theta \in ]x, a[ \tag{2.12c}
\]
is the Lagrange form of the remainder term. Since the continuous function \( f^{(m+1)} \) is bounded on each compact interval \( [a, y], y \in I \), (2.12) imply
\[
f(x) = T_m(x, a) + O((x - a)^{m+1})
\]
for \( x \to x_0 \) for each \( x_0 \in I \).

**Proposition 2.9.** As in Def. 2.2, let \((X, d)\) be a metric space, let \((Y, \| \cdot \|), (Z, \| \cdot \|)\) be normed vector spaces over \( \mathbb{K} \), let \( D \subseteq X \) and consider functions \( f : D \to Y, g : D \to Z \), where \( g(x) \neq 0 \) for each \( x \in D \). Moreover, let \( x_0 \in X \) be a cluster point of \( D \).

(a) Suppose \( \|f\| \leq \|g\| \) in some neighborhood of \( x_0 \), i.e. there exists \( \epsilon > 0 \) such that
\[
\|f(x)\| \leq \|g(x)\| \quad \text{for each } x \in D \setminus \{x_0\} \text{ with } d(x, x_0) < \epsilon.
\]
Then \( f(x) = O(g(x)) \) for \( x \to x_0 \).

(b) \( f(x) = O(f(x)) \) for \( x \to x_0 \), but not \( f(x) = o(f(x)) \) for \( x \to x_0 \) (assuming \( f \neq 0 \)).

(c) \( f(x) = o(g(x)) \) for \( x \to x_0 \) implies \( f(x) = O(g(x)) \) for \( x \to x_0 \).

**Proof.** Exercise. \( \blacksquare \)
Proposition 2.10. As in Def. 2.2, let \((X, d)\) be a metric space, let \((Y, \| \cdot \|), (Z, \| \cdot \|)\) be normed vector spaces over \(\mathbb{K}\), let \(D \subseteq X\) and consider functions \(f : D \rightarrow Y, g : D \rightarrow Z\), where \(g(x) \neq 0\) for each \(x \in D\). Moreover, let \(x_0 \in X\) be a cluster point of \(D\). In addition, for use in some of the following assertions, we introduce functions \(f_1 : D \rightarrow Y_1, f_2 : D \rightarrow Y_2, g_1 : D \rightarrow Z_1, g_2 : D \rightarrow Z_2\), where \((Y_1, \| \cdot \|), (Y_2, \| \cdot \|), (Z_1, \| \cdot \|), (Z_2, \| \cdot \|)\) are normed vector spaces over \(\mathbb{K}\). Assume \(g_1(x) \neq 0\) and \(g_2(x) \neq 0\) for each \(x \in D\).

(a) If \(f(x) = O(g(x))\) (resp. \(f(x) = o(g(x))\)) for \(x \rightarrow x_0\), and if \(\|f_1\| \leq \|f\|\) and \(\|g\| \leq \|g_1\|\) in some neighborhood of \(x_0\), then \(f_1(x) = O(g_1(x))\) (resp. \(f_1(x) = o(g_1(x))\)) for \(x \rightarrow x_0\).

(b) If \(\alpha \in \mathbb{K} \setminus \{0\}\), then for \(x \rightarrow x_0\):

\[
\begin{align*}
\alpha f(x) &= O(g(x)) & \Rightarrow & & f(x) &= O(g(x)), \\
\alpha f(x) &= o(g(x)) & \Rightarrow & & f(x) &= o(g(x)), \\
f(x) &= O(\alpha g(x)) & \Rightarrow & & f(x) &= O(g(x)), \\
f(x) &= o(\alpha g(x)) & \Rightarrow & & f(x) &= o(g(x)).
\end{align*}
\]

(c) For \(x \rightarrow x_0\):

\[
\begin{align*}
f(x) &= O(g_1(x)) & \text{and} & & g_1(x) &= O(g_2(x)) & \Rightarrow & & f(x) &= O(g_2(x)), \\
f(x) &= o(g_1(x)) & \text{and} & & g_1(x) &= o(g_2(x)) & \Rightarrow & & f(x) &= o(g_2(x)).
\end{align*}
\]

(d) Assume \((Y_1, \| \cdot \|) = (Y_2, \| \cdot \|)\). Then, for \(x \rightarrow x_0\):

\[
\begin{align*}
f_1(x) &= O(g(x)) & \text{and} & & f_2(x) &= O(g(x)) & \Rightarrow & & f_1(x) + f_2(x) &= O(g(x)), \\
f_1(x) &= o(g(x)) & \text{and} & & f_2(x) &= o(g(x)) & \Rightarrow & & f_1(x) + f_2(x) &= o(g(x)).
\end{align*}
\]

(e) For \(x \rightarrow x_0\):

\[
\begin{align*}
f_1(x) &= O(g_1(x)) & \text{and} & & f_2(x) &= O(g_2(x)) & \Rightarrow & & \|f_1(x)\| \|f_2(x)\| &= O(\|g_1(x)\| \|g_2(x)\|), \\
f_1(x) &= o(g_1(x)) & \text{and} & & f_2(x) &= o(g_2(x)) & \Rightarrow & & \|f_1(x)\| \|f_2(x)\| &= o(\|g_1(x)\| \|g_2(x)\|).
\end{align*}
\]

(f) For \(x \rightarrow x_0\):

\[
\begin{align*}
f(x) &= O(\|g_1(x)\| \|g_2(x)\|) & \Rightarrow & & \|f(x)\| = O(g_2(x)), \\
f(x) &= o(\|g_1(x)\| \|g_2(x)\|) & \Rightarrow & & \|f(x)\| = o(g_2(x)).
\end{align*}
\]

Proof. Exercise.
2.2 Operator Norms and Natural Matrix Norms

When considering maps between normed vector spaces, the corresponding norms provide a natural measure for error analyses. A special class of norms of importance to us are so-called operator norms defined on linear maps between normed vector spaces (cf. Def. 2.12 below).

**Definition 2.11.** Let \( A : X \rightarrow Y \) be a linear map between two normed vector spaces \( (X, \| \cdot \|) \) and \( (Y, \| \cdot \|) \) over \( K \). Then \( A \) is called bounded if, and only if, \( A \) maps bounded sets to bounded sets, i.e. if, and only if, \( A(B) \) is a bounded subset of \( Y \) for each bounded \( B \subseteq X \). The vector space of all bounded linear maps between \( X \) and \( Y \) is denoted by \( \mathcal{L}(X,Y) \).

**Definition 2.12.** Let \( A : X \rightarrow Y \) be a linear map between two normed vector spaces \( (X, \| \cdot \|) \) and \( (Y, \| \cdot \|) \) over \( K \). The number

\[
\| A \| := \sup \left\{ \frac{\| Ax \|}{\| x \|} : x \in X, \ x \neq 0 \right\}
\]

\[= \sup \left\{ \| Ax \| : x \in X, \ \| x \| = 1 \right\} \in [0, \infty] \tag{2.13}
\]

is called the operator norm of \( A \) induced by the norms on \( X \) and \( Y \), respectively (strictly speaking, the term operator norm is only justified if the value is finite, but it is often convenient to use the term in the generalized way defined here).

In the special case, where \( X = K^n \), \( Y = K^m \), and \( A \) is given via a real \( m \times n \) matrix, the operator norm is also called a natural matrix norm or an induced matrix norm.

**Remark 2.13.** According to Th. B.1 of the Appendix, for a linear map \( A : X \rightarrow Y \) between two normed vector spaces \( X \) and \( Y \) over \( K \), the following statements are equivalent:

(a) \( A \) is bounded.

(b) \( \| A \| < \infty \).

(c) \( A \) is Lipschitz continuous.

(d) \( A \) is continuous.

(e) There is \( x_0 \in X \) such that \( A \) is continuous at \( x_0 \).

For linear maps between finite-dimensional spaces, the above equivalent properties always hold: Each linear map \( A : K^n \rightarrow K^m \), \( (n, m) \in \mathbb{N}^2 \), is continuous (cf. [Phi16b, Ex. 2.16]). In particular, each linear map \( A : K^n \rightarrow K^m \), has all the above equivalent properties.

**Remark 2.14.** According to Th. B.2 of the Appendix, for two normed vector spaces \( X \) and \( Y \) over \( K \), the operator norm does, indeed, constitute a norm on the vector space of bounded linear maps \( \mathcal{L}(X,Y) \), and, moreover, if \( A \in \mathcal{L}(X,Y) \), then \( \| A \| \) is the smallest Lipschitz constant for \( A \).
Lemma 2.15. If $\text{Id} : X \to X$, $\text{Id}(x) := x$, is the identity map on a normed vector space $X$ over $K$, then $\|\text{Id}\| = 1$ (in particular, the operator norm of a unit matrix is always 1). Caveat: In principle, one can consider two different norms on $X$ simultaneously, and then the operator norm of the identity can differ from 1.

Proof. If $\|x\| = 1$, then $\|\text{Id}(x)\| = \|x\| = 1$. □

Lemma 2.16. Let $X, Y, Z$ be normed vector spaces over $K$ and consider linear maps $A \in \mathcal{L}(X, Y)$, $B \in \mathcal{L}(Y, Z)$. Then

$$\|BA\| \leq \|B\| \|A\|. \quad (2.14)$$

Proof. Let $x \in X$ with $\|x\| = 1$. If $Ax = 0$, then $\|B(A(x))\| = 0 \leq \|B\| \|A\|$. If $Ax \neq 0$, then one estimates

$$\|B(Ax)\| = \|Ax\| \left\| B \left( \frac{Ax}{\|Ax\|} \right) \right\| \leq \|A\| \|B\|,$$

thereby establishing the case. □

Notation 2.17. For each $m, n \in \mathbb{N}$, let $\mathcal{M}(m, n, K) \cong K^{mn}$ denote the vector space over $K$ of $m \times n$ matrices $(a_{kl}) : (k,l) \in \{1, \ldots, m\} \times \{1, \ldots, n\}$ over $K$. Moreover, we define $\mathcal{M}(n, K) := \mathcal{M}(n, n, K)$.

Example 2.18. Let $m, n \in \mathbb{N}$ and let $A : K^n \to K^m$ be the $K$-linear map given by $(a_{kl}) \in \mathcal{M}(m, n, K)$ (with respect to the standard bases of $K^n$ and $K^m$, respectively).

(a) The norm defined by

$$\|A\|_\infty := \max \left\{ \sum_{k=1}^{n} |a_{kl}| : k \in \{1, \ldots, m\} \right\}$$

is called the row sum norm of $A$. It is an exercise to show that $\|A\|_\infty$ is the operator norm induced if $K^n$ and $K^m$ are endowed with the $\infty$-norm.

(b) The norm defined by

$$\|A\|_1 := \max \left\{ \sum_{k=1}^{m} |a_{kl}| : l \in \{1, \ldots, n\} \right\}$$

is called the column sum norm of $A$. It is an exercise to show that $\|A\|_1$ is the operator norm induced if $K^n$ and $K^m$ are endowed with the 1-norm.

(c) The operator norm on $A$ induced if $K^n$ and $K^m$ are endowed with the 2-norm (i.e. the Euclidean norm) is called the spectral norm of $A$ and is denoted $\|A\|_2$. Unfortunately, there is no formula as simple as the ones in (a) and (b) for the computation of $\|A\|_2$. We will see in Th. 2.24 below that the value of $\|A\|_2$ is related to the eigenvalues of $A^*A$. 
Caveat 2.19. Let \( m, n \in \mathbb{N} \) and \( p \in [1, \infty] \). Let \( A : \mathbb{K}^n \rightarrow \mathbb{K}^m \) be the \( \mathbb{K} \)-linear map given by \( (a_{kl}) \in \mathcal{M}(m, n, \mathbb{K}) \) (with respect to the standard bases of \( \mathbb{K}^n \) and \( \mathbb{K}^m \), respectively). It is common to denote the operator norm on \( A \in \mathcal{L}(\mathbb{K}^n, \mathbb{K}^m) \) induced by the \( p \)-norms on \( \mathbb{K}^n \) and \( \mathbb{K}^m \), respectively, by \( \|A\|_p \) (as we did in Ex. 2.18 for \( p = \infty, 1, 2 \)). However, for \( n, m > 1 \), the operator norm \( \|\cdot\|_p \) is not(!) the \( p \)-norm on \( \mathbb{K}^{mn} \) (the 2-norm on \( \mathbb{K}^{mn} \) is known as the Frobenius norm or the Hilbert-Schmidt norm on \( \mathbb{K}^{mn} \)). As it turns out, for \( n > 1 \), the \( p \)-norm on \( \mathbb{K}^{n^2} \) is not an operator norm induced by any norm on \( \mathbb{K}^n \) at all: Let

\[
\|A\|_{p,\text{nonop}} := \begin{cases} 
(\sum_{k=1}^{n} \sum_{l=1}^{n} |a_{kl}|^p)^{1/p} & \text{for } p < \infty, \\
\max\{|a_{kl}| : k, l \in \{1, \ldots, n\}\} & \text{for } p = \infty,
\end{cases}
\]

denote the \( p \)-norm on \( \mathbb{K}^{n^2} \). If \( p \in [1, \infty] \), then

\[
\|\text{Id}\|_{p,\text{nonop}} = \sqrt[n]{n} \neq 1 \quad \text{for } n > 1,
\]

showing that \( \|\cdot\|_{p,\text{nonop}} \) does not satisfy Lem. 2.15. If \( a_{kl} := 1 \) for each \( k, l \in \{1, \ldots, n\} \), then

\[n = \|AA\|_{\infty,\text{nonop}} > \|A\|_{\infty,\text{nonop}} \cdot \|A\|_{\infty,\text{nonop}} = 1 \cdot 1 = 1 \quad \text{for } n > 1,
\]

showing that \( \|\cdot\|_{\infty,\text{nonop}} \) does not satisfy Lem. 2.16. For \( p \in [1, \infty] \), one might think that rescaling \( \|\cdot\|_{p,\text{nonop}} \) might make it into an operator norm: The modified norm

\[
\|\cdot\|'_{p,\text{nonop}} : \mathbb{K}^{n^2} \rightarrow \mathbb{R}_+^+, \quad \|A\|'_{p,\text{nonop}} := \frac{\|A\|_{p,\text{nonop}}}{\sqrt[n]{n}},
\]

does satisfy Lem. 2.15. However, if

\[
a_{kl} := \begin{cases} 
1 & \text{for } k = l = 1, \\
0 & \text{otherwise},
\end{cases}
\]

then \( AA = A \) and

\[
\frac{1}{\sqrt[n]{n}} = \|AA\|'_{p,\text{nonop}} > \|A\|'_{p,\text{nonop}} \cdot \|A\|'_{p,\text{nonop}} = \frac{1}{\sqrt[n]{n}} \cdot \frac{1}{\sqrt[n]{n}} \quad \text{for } n > 1,
\]

i.e. the modified norm does not satisfy Lem. 2.16.

In the rest of this section, we will further investigate the spectral norm defined in Ex. 2.18(c). In preparation, we recall the following notions and notation:

Notation 2.20. Let \( m, n \in \mathbb{N} \). If \( A = (a_{kl})_{(k,l) \in \{1,\ldots,m\} \times \{1,\ldots,n\}} \in \mathcal{M}(m, n, \mathbb{K}) \), then

\[
\overline{A} := (\overline{a}_{kl})_{(k,l) \in \{1,\ldots,m\} \times \{1,\ldots,n\}} \in \mathcal{M}(m, n, \mathbb{K}),
\]

\[
A^t := (a_{kl})_{(l,k) \in \{1,\ldots,n\} \times \{1,\ldots,m\}} \in \mathcal{M}(n, m, \mathbb{K}),
\]

\[
A^* := (\overline{A})^t \in \mathcal{M}(n, m, \mathbb{K}),
\]

where \( \overline{A} \) is called the (complex) conjugate of \( A \), \( A^t \) is the transpose of \( A \), and \( A^* \) is the conjugate transpose, also called the adjoint, of \( A \) (thus, if \( \mathbb{K} = \mathbb{R} \), then \( \overline{A} = A \) and \( A^* = A^t \)).
**Definition 2.21.** Let \( n \in \mathbb{N} \) and let \( A = (a_{kl}) \in \mathcal{M}(n, \mathbb{K}) \).

(a) \( A \) is called **symmetric** if, and only if, \( A = A^t \), i.e. if, and only if, \( a_{kl} = a_{lk} \) for each \((k, l) \in \{1, \ldots, n\}^2\).

(b) \( A \) is called **Hermitian** if, and only if, \( A = A^* \), i.e. if, and only if, \( a_{kl} = \overline{a}_{kl} \) for each \((k, l) \in \{1, \ldots, n\}^2\).

(c) \( A \) is called **positive semidefinite** if, and only if, \( x^*Ax \in \mathbb{R}_0^+ \) for each \( x \in \mathbb{K}^n \).

(d) \( A \) is called **positive definite** if, and only if, \( A \) is positive semidefinite and \( x^*Ax = 0 \iff x = 0 \), i.e. if, and only if, \( x^*Ax > 0 \) for each \( 0 \neq x \in \mathbb{K}^n \).

**Lemma 2.22.** Let \( m, n \in \mathbb{N} \) and \( A \in \mathcal{M}(m, n, \mathbb{K}) \). Then both \( A^*A \in \mathcal{M}(n, \mathbb{K}) \) and \( AA^* \in \mathcal{M}(m, \mathbb{K}) \) are **Hermitian** and positive semidefinite.

**Proof.** Since \((A^*)^* = A\), it suffices to consider \( A^*A \). That \( A^*A \) is Hermitian due to

\[
(A^*A)^* = A^*(A^*)^* = A^*A.
\]

Moreover, if \( x \in \mathbb{K}^n \), then \( x^*A^*Ax = (Ax)^*(Ax) = \|Ax\|^2_2 \in \mathbb{R}_0^+ \), showing \( A^*A \) to be positive semidefinite. \(\blacksquare\)

**Definition 2.23.** We define, for each \( n \in \mathbb{N} \) and each \( A \in \mathcal{M}(n, \mathbb{C}) \),

\[
r(A) := \max \{ |\lambda| : \lambda \in \mathbb{C} \text{ and } \lambda \text{ is eigenvalue of } A \}.
\]

The number \( r(A) \) is called the **spectral radius** of \( A \).

**Theorem 2.24.** Let \( m, n \in \mathbb{N} \). If \( A \in \mathcal{M}(m, n, \mathbb{K}) \), then its spectral norm \( \|A\|_2 \) (cf. Ex. 2.18(c)) satisfies

\[
\|A\|_2 = \sqrt{r(A^*A)}.
\]

(2.16a)

If \( m = n \) and \( A \) is Hermitian (i.e. \( A^* = A \)), then \( \|A\|_2 \) is also given by the following simpler formula:

\[
\|A\|_2 = r(A).
\]

(2.16b)

**Proof.** According to Lem. 2.22, \( A^*A \) is a Hermitian and positive semidefinite \( n \times n \) matrix. Then there exist (cf. Th. C.1 and Th. C.2) real nonnegative (not necessarily distinct) eigenvalues \( \lambda_1, \ldots, \lambda_n \geq 0 \) and a corresponding ordered orthonormal basis \((v_1, \ldots, v_n)\) of \( \mathbb{K}^n \), consisting of eigenvectors satisfying, in particular,

\[
A^*Av_k = \lambda_kv_k \quad \text{for each } k \in \{1, \ldots, n\}.
\]

As \((v_1, \ldots, v_n)\) is a basis of \( \mathbb{K}^n \), for each \( x \in \mathbb{K}^n \), there are numbers \( x_1, \ldots, x_n \in \mathbb{K} \) such that \( x = \sum_{k=1}^n x_kv_k \). Thus, one computes

\[
\|Ax\|^2_2 = (Ax) \cdot (Ax) = x^*A^*Ax = \left( \sum_{k=1}^n x_kv_k \right)^* \left( \sum_{k=1}^n x_k\overline{\lambda}_k v_k \right)
\]

\[
= \sum_{k=1}^n \lambda_k |x_k|^2 \leq r(A^*A) \sum_{k=1}^n |x_k|^2 = r(A^*A)\|x\|^2_2,
\]

(2.17)
proving \( \|A\|_2 \leq \sqrt{r(A^*A)} \). To verify the remaining inequality, let \( \lambda_j := r(A^*A) \) be the largest of the nonnegative \( \lambda_k \), and let \( v := v_j \) be the corresponding eigenvector from the orthonormal basis. Then \( \|v\|_2 = 1 \) and choosing \( x = v = v_j \) (i.e. \( x_k = \delta_{jk} \)) in (2.17) yields \( \|Av\|_2^2 = r(A^*A) \), thereby proving \( \|A\|_2 \geq \sqrt{r(A^*A)} \) and completing the proof of (2.16a). It remains to consider the case where \( A = A^* \). Then \( A^*A = A^2 \) and, since

\[
Av = \lambda v \implies A^2v = \lambda^2v \implies r(A^2) = r(A)^2,
\]

we have

\[
\|A\|_2 = \sqrt{r(A^*A)} = \sqrt{r(A^2)} = r(A),
\]

thereby establishing the case. \( \blacksquare \)

**Caveat 2.25.** In general, it is *not* admissible to use the simpler formula (2.16b) for non-Hermitian \( n \times n \) matrices: For example, for \( A := \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \), one has

\[
A^*A = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix},
\]

such that \( 1 = r(A) \neq \sqrt{2} = \sqrt{r(A^*A)} = \|A\|_2 \).

**Lemma 2.26.** Let \( n \in \mathbb{N} \). Then

\( \|y\|_2 = \max \{ |v \cdot y| : v \in \mathbb{K}^n \text{ and } \|v\|_2 = 1 \} \) for each \( y \in \mathbb{K}^n \). \( (2.18) \)

**Proof.** Let \( v \in \mathbb{K}^n \) such that \( \|v\|_2 = 1 \). One estimates, using the Cauchy-Schwarz inequality [Phi16b, (1.41)],

\[
|v \cdot y| \leq \|v\|_2 \|y\|_2 = \|y\|_2.
\]

(2.19a)

Note that (2.18) is trivially true for \( y = 0 \). If \( y \neq 0 \), then letting \( w := y/\|y\|_2 \) yields \( \|w\|_2 = 1 \) and

\[
|w \cdot y| = \frac{|y \cdot y| \|y\|_2}{\|y\|_2} = \|y\|_2.
\]

(2.19b)

Together, (2.19a) and (2.19b) establish (2.18). \( \blacksquare \)

**Proposition 2.27.** Let \( m, n \in \mathbb{N} \). If \( A \in \mathcal{M}(m,n,\mathbb{K}) \), then \( \|A\|_2 = \|A^*\|_2 \) and, in particular, \( r(A^*A) = r(AA^*) \). This allows one to use the simpler (smaller) matrix of \( A^*A \) and \( AA^* \) to compute \( \|A\|_2 \) (see Ex. 2.29 below).

**Proof.** We know from Linear Algebra (cf. [Phi19b, Th. 10.16(a)] and [Phi19b, Th. 10.18(a)]) that

\[
\forall x \in \mathbb{K}^n \quad \forall v \in \mathbb{K}^m \quad v \cdot Ax = (A^*v) \cdot x.
\]

Thus, one calculates

\[
\|A\|_2 = \max \{ \|Ax\|_2 : x \in \mathbb{K}^n \text{ and } \|x\|_2 = 1 \} \overset{(2.18)}{=} \max \{ \|v \cdot Ax\|_2 : v \in \mathbb{K}^m, \ x \in \mathbb{K}^n, \text{ and } \|v\|_2 = \|x\|_2 = 1 \} = \max \{ \|(A^*v) \cdot x\|_2 : v \in \mathbb{K}^m, \ x \in \mathbb{K}^n, \text{ and } \|v\|_2 = \|x\|_2 = 1 \} \overset{(2.18)}{=} \max \{ \|A^*v\|_2 : v \in \mathbb{K}^m \text{ and } \|v\|_2 = 1 \} = \|A^*\|_2,
\]
proving the proposition.

**Remark 2.28.** One can actually show more than we did in Prop. 2.27: The eigenvalues (including multiplicities) of $A^*A$ and $AA^*$ are always identical, except that the larger of the two matrices (if any) can have additional eigenvalues of value 0 – the square roots of the nonzero (i.e. positive) eigenvalues of $A^*A$ and $AA^*$ are the so-called *singular values* of $A$ and $A^*$ (see, e.g., [HB09, Def. and Th. 12.1]).

**Example 2.29.** Consider $A := \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \in \mathbb{M}(2, 4, \mathbb{K})$. One obtains $AA^* = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, $A^*A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$. So one would tend to compute the eigenvalues using $AA^*$. One finds $\lambda_1 = 1$ and $\lambda_2 = 3$. Thus, $r(AA^*) = 3$ and $\|A\|_2 = \sqrt{3}$.

### 2.3 Condition of a Problem

We are now in a position to apply our acquired knowledge of operator norms to error analysis. The general setting is the following: The input is given in some normed vector space $X$ (e.g. $\mathbb{K}^n$) and the output is likewise a vector lying in some normed vector space $Y$ (e.g. $\mathbb{K}^m$). The solution operator, i.e. the map $f$ between input and output, is some function $f$ defined on an open subset of $X$. The goal in this section is to study the behavior of the output given small changes (errors) of the input.

**Definition 2.30.** Let $(X, \| \cdot \|)$ and $(Y, \| \cdot \|)$ be normed vector spaces over $\mathbb{K}$, $U \subseteq X$ open, and $f : U \rightarrow Y$. Fix $x \in U \setminus \{0\}$ such that $f(x) \neq 0$ and $\delta > 0$ such that $B_\delta(x) = \{ \hat{x} \in X : \|\hat{x} - x\| < \delta \} \subseteq U$. We call (the problem represented by the solution operator) $f$ *well-conditioned* in $B_\delta(x)$ if, and only if, there exists $K \geq 0$ such that

$$\frac{\|f(\hat{x}) - f(x)\|}{\|f(x)\|} \leq K \frac{\|\hat{x} - x\|}{\|x\|} \quad \text{for every } \hat{x} \in B_\delta(x). \quad (2.20)$$

If $f$ is well-conditioned in $B_\delta(x)$, then we define $K(\delta) := K(f, x, \delta)$ to be the smallest number $K \in \mathbb{R}_0^+$ such that (2.20) is valid. If there exists $\delta > 0$ such that $f$ is well-conditioned in $B_\delta(x)$, then $f$ is called *well-conditioned* at $x$.

**Definition and Remark 2.31.** We remain in the situation of Def. 2.30 and notice that $0 < \alpha \leq \beta \leq \delta$ implies $B_\alpha(x) \subseteq B_\beta(x) \subseteq B_\delta(x)$, and, thus $0 \leq K(\alpha) \leq K(\beta) \leq K(\delta)$. In consequence, the following definition makes sense if $f$ is well-conditioned at $x$:

$$k_{rel} := k_{rel}(f, x) := \lim_{\alpha \rightarrow 0} K(\alpha). \quad (2.21)$$

The number $k_{rel} \in \mathbb{R}_0^+$ is called the *relative condition* of (the problem represented by the solution operator) $f$ at $x$. The number $k_{rel}$ provides a measure for how much a relative
error in $x$ might get inflated by the application of $f$, i.e. the smaller $k_{\text{rel}}$, the more well-behaved is the problem in the vicinity of $x$. What value for $k_{\text{rel}}$ is still acceptable in terms of numerical stability depends on the situation: While $k_{\text{rel}} \approx 1$ generally indicates stability, even $k_{\text{rel}} \approx 100$ might be acceptable in some situations.

**Remark 2.32.** Let us compare the newly introduced notion of Def. and Rem. 2.31 of $f$ being well-conditioned with some other regularity notions for $f$. For example, if $f$ is $L$-Lipschitz in $B_\delta(x)$, $x \in X$, $\delta > 0$, then

$$\forall \hat{x} \in B_\delta(x) \quad \|f(\hat{x}) - f(x)\| \leq L \|\hat{x} - x\|,$$

i.e. $f$ is well-conditioned in $B_\delta(x)$ and (2.20) holds with $K := \frac{L \|x\|}{\|f(x)\|}$. The converse is not true: There are examples, where $f$ is well-conditioned in some $B_\delta(x)$ without being Lipschitz continuous on $B_\delta(x)$ (exercise). On the other hand, (2.20) does imply continuity in $x$, and, once again, the converse is not true (another exercise). In particular, a problem can be well-posed in the sense of Def. 1.5 without being well-conditioned. On the other hand, if $f$ is everywhere well-conditioned, then it is everywhere continuous, and, thus, well-posed. In that sense well-conditionedness is stronger than well-posedness. However, one should also note that we defined well-conditionedness as a *local* property, whereas well-posedness was defined as a *global* property.

As an example, we consider the problem of solving the linear system $Ax = b$ with an invertible $n \times n$ matrix $A \in \mathcal{M}(n, \mathbb{K})$. Before studying the problem systematically using the notion of well-conditionedness, let us look at a particular example that illustrates a typical instability that can occur:

**Example 2.33.** Consider the linear system

$$Ax = b$$

for the unknown $x \in \mathbb{R}^4$ with

$$A := \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}, \quad b := \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

One finds that the solution is

$$x := \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

If instead of $b$, we are given the following perturbed $\tilde{b}$,

$$\tilde{b} := \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix},$$
where the absolute error is 0.1 and relative error is even smaller, then the corresponding solution is
\[ \tilde{x} := \begin{pmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{pmatrix}, \]
in no way similar to the solution to \( b \). In particular, the absolute and relative errors have been hugely amplified.

One might suspect that the issue behind the instability lies in \( A \) being “almost” singular such that applying \( A^{-1} \) is similar to dividing by a small number. However, that is not the case, as we have
\[ A^{-1} = \begin{pmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{pmatrix}. \]
We will see in Ex. 2.42 below that the actual reason behind the instability is the range of eigenvalues of \( A \), in particular, the relation between the largest and the smallest eigenvalue. One has:
\[ \lambda_1 \approx 0.010, \quad \lambda_2 \approx 0.843, \quad \lambda_3 \approx 3.86, \quad \lambda_4 \approx 30.3, \quad \frac{\lambda_4}{\lambda_1} \approx 2984. \]

**Proposition 2.34.** Let \( m, n \in \mathbb{N} \) and consider the normed vector spaces \((\mathbb{K}^n, \| \cdot \|)\) and \((\mathbb{K}^m, \| \cdot \|)\). Let \( A : \mathbb{K}^n \to \mathbb{K}^m \) be \( \mathbb{K} \)-linear. Moreover, let \( x \in \mathbb{K}^n \) be such that \( x \neq 0 \) and \( Ax \neq 0 \). Then the following statements hold true:

(a) \( A \) is well-conditioned at \( x \) and
\[ k_{rel}(A, x) = \frac{\| A \|}{\| Ax \|}. \]
where \( \| A \| \) denotes the corresponding operator norm of \( A \).

(b) If \( n = m \) and \( A \) is invertible, then
\[ k_{rel}(A, x) \leq \frac{\| A \|}{\| A^{-1} \|}, \]
where \( \| A \| \) and \( \| A^{-1} \| \) denote the corresponding operator norms of \( A \) and \( A^{-1} \), respectively.

**Proof.** (a): For each \( \tilde{x} \in \mathbb{K}^n \), we estimate
\[ \frac{\| A\tilde{x} - Ax \|}{\| Ax \|} \leq \frac{\| A \|}{\| Ax \|} \frac{\| \tilde{x} - x \|}{\| x \|} = \frac{\| A \|}{\| Ax \|} \frac{\| \tilde{x} - x \|}{\| x \|}, \]
showing \( k_{rel}(A, x) \leq \| A \| \frac{\| \tilde{x} \|}{\| Ax \|} \). It still remains to prove the inverse inequality. To that end, choose \( v \in \mathbb{K}^n \) such that \( \| v \| = 1 \) and \( \| Av \| = \| A \| \) (such a vector \( v \) must exist due
to the fact that the continuous function \( z \mapsto \|Az\| \) must attain its max on the compact unit sphere \( S_1(0) = \{ z \in \mathbb{K}^n : \|z\| = 1 \} \) – note that this argument does not work in infinite-dimensional spaces, where \( S_1(0) \) is not compact). Consider \( \tilde{x} := x + \alpha v \). Then

\[
\frac{\|A\tilde{x} - Ax\|}{\|Ax\|} = \alpha \frac{\|Av\|}{\|Ax\|} = \|A\| \frac{\|\tilde{x} - x\|}{\|x\|}.
\]

As this equality is independent of \( \alpha \), for \( \alpha \to 0 \), we obtain \( k_{\text{rel}}(A, x) = \|A\| \), as desired.

(b): If \( n = m \) and \( A \) is invertible, then (a) implies

\[
k_{\text{rel}}(A, x) = \|A\| \|x\| = \|A\| \|A^{-1}Ax\| \leq \|A\| \|A^{-1}\| \|Ax\| = \|A\| \|A^{-1}\|,
\]

thereby proving (b).

Example 2.35. Let \( A \in \mathcal{M}(n, \mathbb{K}) \) be an invertible \( n \times n \) matrix, \( n \in \mathbb{N} \) and consider the problem of solving the linear system \( Ax = b, b \in \mathbb{K}^n \). Then the solution operator is

\[
A^{-1} : \mathbb{K}^n \to \mathbb{K}^n, \quad b \mapsto A^{-1}b.
\]

Fixing some norm on \( \mathbb{K}^n \) and using the induced natural matrix norm, we obtain from Prop. 2.34(b) that, for each \( 0 \neq b \in \mathbb{K}^n \):

\[
k_{\text{rel}}(A^{-1}, b) \leq \|A\| \|A^{-1}\|. \tag{2.22}
\]

Definition and Remark 2.36. In view of Prop. 2.34(b) and (2.22), we define the condition number \( \kappa(A) \) (also just called the condition) of an invertible \( A \in \mathcal{M}(n, \mathbb{K}) \), \( n \in \mathbb{N} \), by

\[
\kappa(A) := \|A\| \|A^{-1}\|, \tag{2.23}
\]

where \( \| \cdot \| \) denotes a natural matrix norm induced by some norm on \( \mathbb{K}^n \). Then one immediately gets from (2.22) that

\[
k_{\text{rel}}(A^{-1}, b) \leq \kappa(A) \quad \text{for each } b \in \mathbb{K}^n \setminus \{0\}. \tag{2.24}
\]

The condition number clearly depends on the underlying natural matrix norm. If the natural matrix norm is the spectral norm, then one calls the condition number the spectral condition and one also writes \( \kappa_2 \) instead of \( \kappa \).

Notation 2.37. For each \( A \in \mathcal{M}(n, \mathbb{C}) \), \( n \in \mathbb{N} \), let

\[
\sigma(A) := \{ \lambda \in \mathbb{C} : \lambda \text{ is eigenvalue of } A \},
\]

\[
|\sigma(A)| := \{ |\lambda| : \lambda \in \mathbb{C} \text{ is eigenvalue of } A \},
\]

denote the set of eigenvalues of \( A \) (also known as the spectrum of \( A \)) and the set of absolute values of eigenvalues of \( A \), respectively.
Lemma 2.38. For the spectral condition of an invertible matrix $A \in \mathcal{M}(n, \mathbb{K})$, $n \in \mathbb{N}$, one obtains

$$
\kappa_2(A) = \sqrt{\frac{\max \sigma(A^*A)}{\min \sigma(A^*A)}}
$$

(2.25)

(recall that all eigenvalues of $A^*A$ are real and positive). Moreover, if $A$ is Hermitian, then (2.25) simplifies to

$$
\kappa_2(A) = \frac{\max |\sigma(A)|}{\min |\sigma(A)|}
$$

(2.26)

(note that $\min |\sigma(A)| > 0$, as $A$ is invertible).

Proof. Exercise. ■

Theorem 2.39. Let $A \in \mathcal{M}(n, \mathbb{K})$ be invertible, $n \in \mathbb{N}$. Assume that $x, b, \Delta x, \Delta b \in \mathbb{K}^n$ satisfy

$$
Ax = b, \quad A(x + \Delta x) = b + \Delta b,
$$

(2.27)
i.e. $\Delta b$ can be seen as a perturbation of the input and $\Delta x$ can be seen as the resulting perturbation of the output. One then has the following estimates for the absolute and relative errors:

$$
\|\Delta x\| \leq \|A^{-1}\| \|\Delta b\|, \quad \frac{\|\Delta x\|}{\|x\|} \leq \kappa(A) \frac{\|\Delta b\|}{\|b\|},
$$

(2.28)

where $x, b \neq 0$ is assumed for the second estimate (as before, it does not matter which norm on $\mathbb{K}^n$ one uses, as long as one uses the induced operator norm for the matrix).

Proof. Since $A$ is linear, (2.27) implies $A(\Delta x) = \Delta b$, i.e. $\Delta x = A^{-1}(\Delta b)$, which already yields the first estimate in (2.28). For $x, b \neq 0$, the first estimate together with $Ax = b$ easily implies the second:

$$
\frac{\|\Delta x\|}{\|x\|} \leq \|A^{-1}\| \frac{\|\Delta b\|}{\|b\|} \frac{\|Ax\|}{\|x\|},
$$

thereby establishing the case. ■

Example 2.40. Suppose we want to find out how much we can perturb $b$ in the problem

$$
Ax = b, \quad A := \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}, \quad b := \begin{pmatrix} 1 \\ 4 \end{pmatrix},
$$

if the resulting relative error $e_r$ in $x$ is to be less than $10^{-2}$ with respect to the $\infty$-norm. From (2.28), we know

$$
e_r \leq \kappa(A) \frac{\|\Delta b\|_\infty}{\|b\|_\infty} = \kappa(A) \frac{\|\Delta b\|_\infty}{4}.
$$

Moreover, since $A^{-1} = \begin{pmatrix} -1 & 2 \\ 1 & -1 \end{pmatrix}$, from (2.23) and Ex. 2.18(a), we obtain

$$
\kappa(A) = \|A\|_\infty \|A^{-1}\|_\infty = 3 \cdot 3 = 9.
$$

Thus, if the perturbation $\Delta b$ satisfies $\|\Delta b\|_\infty < 4/900 \approx 0.0044$, then $e_r < \frac{9}{4} \cdot \frac{4}{900} = 10^{-2}$. 
Remark 2.41. Note that (2.28) is a much stronger and more useful statement than the $k_{rel}(A^{-1}, b) \leq \kappa(A)$ of (2.22). The relative condition only provides a bound in the limit of smaller and smaller neighborhoods of $b$ and without providing any information on how small the neighborhood actually has to be (one can estimate the amplification of the error in $x$ provided that the error in $b$ is very small). On the other hand, (2.28) provides an effective control of the absolute and relative errors without any restrictions with regard to the size of the error in $b$ (it holds for each $\Delta b$).

Example 2.42. We come back to the instability observed in Ex. 2.33 and investigate its actual cause. Suppose $A \in \mathcal{M}(n, \mathbb{C})$ is an arbitrary Hermitian invertible matrix, $\lambda_{\min}, \lambda_{\max} \in \sigma(A)$ such that $|\lambda_{\min}| = \min |\sigma(A)|$, $|\lambda_{\max}| = \max |\sigma(A)|$. Let $v_{\min}$ and $v_{\max}$ be eigenvectors for $\lambda_{\min}$ and $\lambda_{\max}$, respectively, satisfying $\|v_{\min}\| = \|v_{\max}\| = 1$. The most unstable behavior of $Ax = b$ occurs if $b = v_{\max}$ and $b$ is perturbed in the direction of $v_{\min}$, i.e. $\tilde{b} := b + \epsilon v_{\min}, \epsilon > 0$. The solution to $Ax = b$ is $x = \lambda_{\max}^{-1}v_{\max}$, whereas the solution to $Ax = \tilde{b}$ is

$$\tilde{x} = A^{-1}(v_{\max} + \epsilon v_{\min}) = \lambda_{\max}^{-1}v_{\max} + \epsilon\lambda_{\min}^{-1}v_{\min}.$$ 

Thus, the resulting relative error in the solution is

$$\frac{\|\tilde{x} - x\|}{\|x\|} = \epsilon \frac{|\lambda_{\max}|}{|\lambda_{\min}|},$$

while the relative error in the input was merely

$$\frac{\|\tilde{b} - b\|}{\|b\|} = \epsilon.$$

This shows that, in the worst case, the relative error in $b$ can be amplified by a factor of $\kappa_2(A) = |\lambda_{\max}|/|\lambda_{\min}|$, and that is exactly what had occurred in Ex. 2.33.

Next, we study a generalization of the problem from (2.27). Now, in addition to a perturbation of the right-hand side $b$, we also allow a perturbation of the matrix $A$ by a matrix $\Delta A$:

$$Ax = b, \quad (A + \Delta A)(x + \Delta x) = b + \Delta b. \tag{2.29}$$

We would now like to control $\Delta x$ in terms of $\Delta b$ and $\Delta A$.

Remark 2.43. Let $n \in \mathbb{N}$. The set $\text{GL}_n(\mathbb{K})$ of invertible $n \times n$ matrices over $\mathbb{K}$ is open in $\mathcal{M}(n, \mathbb{K}) \cong \mathbb{K}^{n^2}$ (where all norms are equivalent) – this follows, for example, from the determinant being a continuous map (even a polynomial, cf. [Phi19b, Rem. 4.33]) from $\mathbb{K}^{n^2}$ into $\mathbb{K}$, which implies that $\text{GL}_n(\mathbb{K}) = \text{det}^{-1}(\mathbb{K} \setminus \{0\})$ must be open. Thus, if $A$ in (2.29) is invertible and $\|\Delta A\|$ is sufficiently small, then $A + \Delta A$ must also be invertible.

Lemma 2.44. Let $n \in \mathbb{N}$, consider some norm on $\mathbb{K}^n$, and the induced natural matrix norm on $\mathcal{M}(n, \mathbb{K})$. If $B \in \mathcal{M}(n, \mathbb{K})$ is such that $\|B\| < 1$, then $\text{Id} + B$ is invertible and

$$\|(\text{Id} + B)^{-1}\| \leq \frac{1}{1 - \|B\|}. \tag{2.30}$$
Proof. For each $x \in \mathbb{K}^n$:
\[
\|(\text{Id} + B)x\| \geq \|x\| - \|Bx\| \geq \|x\| - \|B\| \|x\| = (1 - \|B\|) \|x\|,
\]
showing that $x \neq 0$ implies $(\text{Id} + B)x \neq 0$, i.e. $\text{Id} + B$ is invertible. Fixing $y \in \mathbb{K}^n$ and applying (2.31) with $x := (\text{Id} + B)^{-1}y$ yields
\[
\|(\text{Id} + B)^{-1}y\| \leq \frac{1}{1 - \|B\|} \|y\|,
\]
thereby proving (2.30) and concluding the proof of the lemma.

\[\blacksquare\]

Lemma 2.45. Let $n \in \mathbb{N}$, consider some norm on $\mathbb{K}^n$ and the induced natural matrix norm on $\mathcal{M}(n, \mathbb{K})$. If $A, \Delta A \in \mathcal{M}(n, \mathbb{K})$, $A$ is invertible, and $\|\Delta A\| < \|A^{-1}\|^{-1}$, then $A + \Delta A$ is invertible and
\[
\|(A + \Delta A)^{-1}\| \leq \frac{1}{\|A^{-1}\|^{-1} - \|\Delta A\|},
\]
Moreover,
\[
\|\Delta A\| \leq \frac{1}{2\|A^{-1}\|} \implies \|(A + \Delta A)^{-1} - A^{-1}\| \leq C\|\Delta A\|,
\]
where the constant can be chosen as $C := 2\|A^{-1}\|^2$.

Proof. We can write $A + \Delta A = A(\text{Id} + A^{-1}\Delta A)$. As $\|A^{-1}\Delta A\| \leq \|A^{-1}\|\|\Delta A\| < 1$, Lem. 2.44 yields that $\text{Id} + A^{-1}\Delta A$ is invertible. Since $A$ is also invertible, so is $A + \Delta A$. Moreover,
\[
\|(A + \Delta A)^{-1}\| = \|(\text{Id} + A^{-1}\Delta A)^{-1}A^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\Delta A\|} \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\Delta A\|},
\]
proving (2.32). The identity
\[
(A + \Delta A)^{-1} - A^{-1} = -(A + \Delta A)^{-1}(\Delta A)A^{-1}
\]
holds, as it is equivalent to
\[
\text{Id} - A^{-1}(A + \Delta A) = -(\Delta A)A^{-1}
\]
and
\[
A - A - \Delta A = -\Delta A.
\]
For $\|\Delta A\| \leq \frac{1}{2\|A^{-1}\|}$, (2.35) yields
\[
\|(A + \Delta A)^{-1} - A^{-1}\| \leq \|(A + \Delta A)^{-1}\|\|\Delta A\|\|A^{-1}\| \leq \frac{\|A^{-1}\|\|\Delta A\|\|A^{-1}\|}{1 - \|A^{-1}\|\|\Delta A\|} \leq 2\|A^{-1}\|^2\|\Delta A\|,
\]
proving (2.33).
Theorem 2.46. As in Lem. 2.45, let \( n \in \mathbb{N} \), consider some norm on \( \mathbb{K}^n \) with the induced natural matrix norm on \( \mathcal{M}(n, \mathbb{K}) \) and let \( A, \Delta A \in \mathcal{M}(n, \mathbb{K}) \) such \( A \) is invertible and \( \| \Delta A \| < \| A^{-1} \|^{-1} \). Moreover, assume \( b, x, \Delta b, \Delta x \in \mathbb{K}^n \) satisfy
\[
Ax = b, \quad (A + \Delta A)(x + \Delta x) = b + \Delta b.
\]
Then, for \( b, x \neq 0 \), one has the following bound for the relative error:
\[
\frac{\| \Delta x \|}{\| x \|} \leq \frac{\kappa(A)}{1 - \| A^{-1} \| \| \Delta A \|} \left( \frac{\| \Delta A \|}{\| A \|} + \frac{\| \Delta b \|}{\| b \|} \right).
\]

Proof. The equations (2.36) imply
\[
(A + \Delta A)(\Delta x) = \Delta b - (\Delta A)x.
\]
In consequence, from (2.32) and \( \| b \| \leq \| A \| \| x \| \), one obtains
\[
\frac{\| \Delta x \|}{\| x \|} \overset{(2.38)}{=} \frac{\| (A + \Delta A)^{-1}(\Delta b - (\Delta A)x) \|}{\| x \|} \overset{(2.32)}{\leq} \frac{1}{\| A^{-1} \| - \| \Delta A \|} \left( \frac{\| \Delta A \|}{\| A \|} + \frac{\| \Delta b \|}{\| b \|} \right) = \frac{\kappa(A)}{1 - \| A^{-1} \| \| \Delta A \|} \left( \frac{\| \Delta A \|}{\| A \|} + \frac{\| \Delta b \|}{\| b \|} \right),
\]
thereby establishing the case.

For nonlinear problems, the following result can be useful:

Theorem 2.47. Let \( m, n \in \mathbb{N} \), let \( U \subseteq \mathbb{R}^n \) be open, and assume that \( f : U \to \mathbb{R}^m \) is continuously differentiable: \( f \in C^1(U, \mathbb{R}^m) \). If \( 0 \neq x \in U \) and \( f(x) \neq 0 \), then \( f \) is well-conditioned at \( x \) and
\[
k_{rel} = k_{rel}(f, x) = \| Df(x) \| \frac{\| x \|}{\| f(x) \|},
\]
where \( Df(x) : \mathbb{R}^n \to \mathbb{R}^m \) is the (total) derivative of \( f \) in \( x \) (which is a linear map represented by the Jacobian \( J_f(x) \)). In (2.39), any norm on \( \mathbb{R}^n \) and \( \mathbb{R}^m \) will work as long as \( \| Df(x) \| \) is the corresponding induced operator norm. While the present result is only formulated and proved for \( \mathbb{R} \)-differentiable functions, see Rem. 2.48 and Ex. 2.49(c) below to learn how it can also be applied to \( \mathbb{C} \)-differentiable functions.

Proof. Choose \( \delta > 0 \) sufficiently small such that \( B_\delta(x) \subseteq U \). Since \( f \in C^1(U, \mathbb{R}^m) \), we can apply Taylor’s theorem. Recall that its lowest order version with the remainder term in integral form states that, for each \( h \in B_\delta(0) \) (which implies \( x + h \in B_\delta(x) \)), the following holds (cf. [Phi16b, Th. 4.44], which can be applied to the components \( f_1, \ldots, f_m \) of \( f \)):
\[
f(x + h) = f(x) + \int_0^1 Df(x + th)(h) \, dt.
\]

\( \text{ } \)
If \( \tilde{x} \in B_\delta(x) \), then we can let \( h := \tilde{x} - x \), which permits to restate (2.40) in the form

\[
f(\tilde{x}) - f(x) = \int_0^1 Df ((1 - t)x + t\tilde{x}) \, \tilde{x} - x \) \, dt .
\]  

(2.41)

This allows to estimate the norm as follows:

\[
\left\| f(\tilde{x}) - f(x) \right\| = \left\| \int_0^1 Df ((1 - t)x + t\tilde{x}) \, \tilde{x} - x \) \, dt \right\| \leq \int_0^1 \left\| Df ((1 - t)x + t\tilde{x}) \right\| \left\| \tilde{x} - x \right\| dt \leq \sup \left\{ \left\| Df(y) \right\| : y \in B_\delta(x) \right\} \left\| \tilde{x} - x \right\| = S(\delta) \left\| \tilde{x} - x \right\|
\]  

(2.42)

with \( S(\delta) := \sup \left\{ \left\| Df(y) \right\| : y \in B_\delta(x) \right\} \). This implies

\[
\frac{\left\| f(\tilde{x}) - f(x) \right\|}{\left\| f(x) \right\|} \leq \frac{\left\| x \right\|}{\left\| f(x) \right\|} S(\delta) \frac{\left\| \tilde{x} - x \right\|}{\left\| x \right\|}
\]

and, therefore,

\[
K(\delta) \leq \frac{\left\| x \right\|}{\left\| f(x) \right\|} S(\delta).
\]  

(2.43)

The hypothesis that \( f \) is continuously differentiable implies \( \lim_{y \to x} Df(y) = Df(x) \), which implies \( \lim_{y \to x} \left\| Df(y) \right\| = \left\| Df(x) \right\| \) (continuity of the norm, cf. [Phi16b, Ex. 2.6(a)]), which, in turn, implies \( \lim_{\delta \to 0} S(\delta) = \left\| Df(x) \right\| \). Since, also, \( \lim_{\delta \to 0} K(\delta) = k_{rel} \), (2.43) yields

\[
k_{rel} \leq \left\| Df(x) \right\| \frac{\left\| x \right\|}{\left\| f(x) \right\|}.
\]  

(2.44)

It still remains to prove the inverse inequality. To that end, choose \( v \in \mathbb{R}^n \) such that \( \left\| v \right\| = 1 \) and

\[
\left\| Df(x)(v) \right\| = \left\| Df(x) \right\|
\]  

(2.45)

(such a vector \( v \) must exist due to the fact that the continuous function \( y \mapsto \left\| Df(x)(y) \right\| \) must attain its max on the compact unit sphere \( S_1(0) \) – as in the proof of Prop. 2.34(a), this argument does not extend to infinite-dimensional spaces, where \( S_1(0) \) is not compact). For \( 0 < \epsilon < \delta \) consider \( \tilde{x} := x + \epsilon v \). Then \( \left\| \tilde{x} - x \right\| = \epsilon < \delta \), i.e. \( \tilde{x} \in B_\delta(x) \). In particular, \( \tilde{x} \) is admissible in (2.41), which provides

\[
f(\tilde{x}) - f(x) = \epsilon \int_0^1 Df ((1 - t)x + t\tilde{x}) (v) \, dt
\]

\[
= \epsilon Df(x)(v) + \epsilon \int_0^1 \left( Df ((1 - t)x + t\tilde{x}) - Df(x) \right)(v) \, dt .
\]  

(2.46)
Once again using \( \| \tilde{x} - x \| = \epsilon \) as well as the triangle inequality in the form \( \| a + b \| \geq \| a \| - \| b \| \), we obtain from (2.46) and (2.45):

\[
\left\| \frac{f(\tilde{x}) - f(x)}{f(x)} \right\| \geq \frac{\| x \|}{\| f(x) \|} \left( \| Df(x) \| - \int_0^1 \| Df((1 - t)x + t\tilde{x}) - Df(x) \| \, dt \right) \frac{\| \tilde{x} - x \|}{\| x \|}.
\]

(2.47)

Thus,

\[
K(\epsilon) \geq \frac{\| x \|}{\| f(x) \|} \left( \| Df(x) \| - T(\epsilon) \right),
\]

where

\[
T(\epsilon) := \sup \left\{ \int_0^1 \| Df((1 - t)x + t\tilde{x}) - Df(x) \| \, dt : \tilde{x} \in B_\epsilon(x) \right\}.
\]

(2.49)

Since \( Df \) is continuous in \( x \), for each \( \alpha > 0 \), there is \( \epsilon > 0 \) such that \( \| y - x \| \leq \epsilon \) implies \( \| Df(y) - Df(x) \| < \alpha \). Thus, since \( \| (1 - t)x + t\tilde{x} - x \| = t \| \tilde{x} - x \| \leq \epsilon \) for \( \tilde{x} \in B_\epsilon(x) \) and \( t \in [0, 1] \), one obtains

\[
T(\epsilon) \leq \int_0^1 \alpha \, dt = \alpha,
\]

implying

\[
\lim_{\epsilon \to 0} T(\epsilon) = 0.
\]

In particular, we can take limits in (2.48) to get

\[
k_{\text{rel}} \geq \frac{\| x \|}{\| f(x) \|} \| Df(x) \|.\]

(2.50)

Finally, (2.50) together with (2.44) completes the proof of (2.39).

\[\square\]

**Remark 2.48.** While Th. 2.47 was formulated and proved for (continuously) \( \mathbb{R} \)-differentiable functions, it can also be used for \( \mathbb{C} \)-differentiable (i.e. for holomorphic) functions: Recall that \( \mathbb{C} \)-differentiability is actually a much stronger condition than continuous \( \mathbb{R} \)-differentiability: A function that is \( \mathbb{C} \)-differentiable is, in particular, \( \mathbb{R} \)-differentiable and even \( \mathbb{C}^\infty \) (see [Phi16b, Def. 4.18] and [Phi16b, Rem. 4.19(c)]). As a concrete example, we apply Th. 2.47 to complex multiplication in Ex. 2.49(c) below. For more information regarding the relation between norms on \( \mathbb{C}^n \) and norms on \( \mathbb{R}^{2n} \), also see [Phi16b, Sec. D].

**Example 2.49. (a)** Let us investigate the condition of the problem of real multiplication, by considering

\[
f : \mathbb{R}^2 \to \mathbb{R}, \quad f(x, y) := xy, \quad Df(x, y) = (y, x).
\]

(2.51)

Using the Euclidean norm \( \| \cdot \|_2 \) on \( \mathbb{R}^2 \), one obtains, for each \( (x, y) \in \mathbb{R}^2 \) such that \( xy \neq 0 \),

\[
k_{\text{rel}}(x, y) := k_{\text{rel}}(f, (x, y)) = \| Df(x, y) \|_2 \frac{\| (x, y) \|_2}{\| f(x, y) \|} = \frac{x^2 + y^2}{|xy|} = \frac{|x|}{|y|} + \frac{|y|}{|x|}.
\]

(2.52)
Thus, we see that the relative condition explodes if $|x| \gg |y|$ or $|x| \ll |y|$. To show $f$ is well-conditioned in $B_3(x, y)$ for each $\delta > 0$, let $(\tilde{x}, \tilde{y}) \in B_3(x, y)$ and estimate

$$|f(x, y) - f(\tilde{x}, \tilde{y})| = |xy - \tilde{x}\tilde{y}| \leq |xy - \tilde{y}\tilde{x}| + |x\tilde{y} - \tilde{x}\tilde{y}| = |x||y - \tilde{y}| + |\tilde{y}||x - \tilde{x}|$$

$$\leq \max \left\{ \| (x, y) \|_2, \| (\tilde{x}, \tilde{y}) \|_2 \right\} \| (x, y) - (\tilde{x}, \tilde{y}) \|_1$$

$$\leq C(\delta + \| (x, y) \|_2) \| (x, y) - (\tilde{x}, \tilde{y}) \|_2$$

with suitable $C \in \mathbb{R}^+$. Thus, $f$ is well-conditioned in $B_3(x, y)$. In consequence, we see that multiplication is numerically stable if, and only if, $|x|$ and $|y|$ are roughly of the same order of magnitude.

(b) Analogous to (a), we now investigate real division, i.e.

$$g : \mathbb{R} \times (\mathbb{R} \setminus \{0\}) \to \mathbb{R}, \quad g(x, y) := \frac{x}{y}, \quad Dg(x, y) = \left( \frac{1}{y}, -\frac{x}{y^2} \right). \quad (2.53)$$

Once again using the Euclidean norm on $\mathbb{R}^2$, one obtains from (2.39) that, for each $(x, y) \in \mathbb{R}^2$ such that $xy \neq 0$,

$$k_{\text{rel}}(x, y) := k_{\text{rel}}(g, (x, y)) = \| Dg(x, y) \|_2 \| (x, y) \|_2 = \frac{|y|}{|x|} \sqrt{x^2 + y^2} \sqrt{1 + \frac{x^2}{y^2}}$$

$$= \frac{1}{|x|} \sqrt{x^2 + y^2} \sqrt{1 + \frac{x^2}{y^2}} = \frac{x^2 + y^2}{|x||y|} = \frac{|x|}{|y|} + \frac{|y|}{|x|}, \quad (2.54)$$

i.e. the relative condition is the same as for multiplication. In particular, division also becomes unstable if $|x| \gg |y|$ or $|x| \ll |y|$, while $k_{\text{rel}}(x, y)$ remains small if $|x|$ and $|y|$ are of the same order of magnitude. However, we now again investigate if $g$ is well-conditioned in $B_3(x, y)$, $\delta > 0$, and here we find an important difference between division and multiplication. For $\delta \geq |y|$, it is easy to see that $g$ is not well-conditioned in $B_3(x, y)$: For each $n \in \mathbb{N}$, let $z_n := (x, \frac{y}{n})$. Then $\| (x, y) - z_n \|_2 = (1 - \frac{1}{n})|y| < |y| \leq \delta$, but

$$\left| \frac{x}{y} - \frac{x}{y^n} \right| = (n - 1) \frac{|x|}{|y|} \to \infty \quad \text{for } n \to \infty.$$
for some suitable $C \in \mathbb{R}^+$, showing that $g$ is well-conditioned in $B_\delta(x,y)$ for $0 < \delta < |y|$. The significance of the present example lies in its demonstration that the knowledge of $k_{\text{rel}}(x,y)$ alone is not enough to determine the numerical stability of $g$ at $(x,y)$: Even though $k_{\text{rel}}(x,y)$ can remain bounded for $y \to 0$, the problem is that the neighborhood $B_\delta(x,y)$, where division is well-conditioned, becomes arbitrarily small. Thus, without an effective bound (from below) on $B_\delta(x,y)$, the knowledge of $k_{\text{rel}}(x,y)$ can be completely useless and even misleading.

(c) We now consider the problem of complex multiplication, i.e.

$$ f : \mathbb{C}^2 \rightarrow \mathbb{C}, \quad f(z,w) := zw. \quad (2.55) $$

We check that $f$ is $\mathbb{C}$-differentiable with $Df(z,w) = (w,z)$: Indeed,

$$ \lim_{h = (h_1,h_2) \to 0} \frac{|f((z,w) + (h_1,h_2)) - f(z,w) - (w,z)(h_1,h_2)|}{||h||_{\infty}} = \lim_{h \to 0} \frac{|(z + h_1)(w + h_2) - zw - wh_1 - zh_2|}{||h||_{\infty}} $$

$$ = \lim_{h \to 0} \frac{|h_1h_2|}{||h||_{\infty}} = \lim_{h \to 0} \min \{ |h_1|, |h_2| \} = 0. $$

Since $(x + iy)(u + iv) = xu - yv + i(xv + yu)$, to apply Th. 2.47, we now consider complex multiplication as the $\mathbb{R}$-differentiable map

$$ g : \mathbb{R}^4 \rightarrow \mathbb{R}^2, \quad g(x,y,u,v) := (xu - yv, xv + yu), \quad (2.56) $$

$$ Dg(x,y,u,v) = \begin{pmatrix} u & -v & x & -y \\ v & u & y & x \end{pmatrix}. $$

Noting

$$ Dg(x,y,u,v) \, Dg(x,y,u,v)^t = \begin{pmatrix} u & -v & x & -y \\ v & u & y & x \end{pmatrix} \begin{pmatrix} u & v \\ -v & u \\ x & y \\ -y & x \end{pmatrix} = \begin{pmatrix} u^2 + v^2 + x^2 + y^2 \\ 0 \\ u^2 + v^2 + x^2 + y^2 \end{pmatrix}, $$

we obtain $\|Dg(x,y,u,v)\|_2 = \sqrt{u^2 + v^2 + x^2 + y^2} = \|(x,y,u,v)\|_2$ and

$$ k_{\text{rel}}(x,y,u,v) := k_{\text{rel}}(g, (x,y,u,v)) = \frac{\|Dg(x,y,u,v)\|_2 \, \|(x,y,u,v)\|_2}{\|g(x,y,u,v)\|_2} = \frac{x^2 + y^2 + u^2 + v^2}{\sqrt{x^2u^2 + y^2v^2 + x^2v^2 + y^2u^2}} = \frac{x^2 + y^2 + u^2 + v^2}{\sqrt{(x^2 + y^2)(u^2 + v^2)}}. \quad (2.57) $$
Thus, letting $z := x + iy$ and $w := u + iv$, in generalization of what occurred in (a), the relative condition becomes arbitrarily large if $\|z\|_2 \gg \|w\|_2$ or $\|z\|_2 \ll \|w\|_2$. An estimate completely analogous to the one in (a) shows $g$ to be well-conditioned in $B_\delta(x,y,u,v) = B_\delta(z,w)$ for each $\delta > 0$: We have, for each $(\tilde{z}, \tilde{w}) := (\tilde{x}, \tilde{y}, \tilde{u}, \tilde{v}) \in B_\delta(x,y,u,v)$,

$$
\|g(x,y,u,v) - g(\tilde{x}, \tilde{y}, \tilde{u}, \tilde{v})\|_2 = |zw - \tilde{z}\tilde{w}| \leq |zw - z\tilde{w}| + |z\tilde{w} - \tilde{z}\tilde{w}|
= |z|\|w - \tilde{w}\| + |\tilde{w}|\|z - \tilde{z}\|
\leq \max \{|(z, w)|_2, \|(\tilde{z}, \tilde{w})|_2\}\|(z, w) - (\tilde{z}, \tilde{w})\|_1
\leq C(\delta + \|(z, w)|_2\) \|(z, w) - (\tilde{z}, \tilde{w})\|_2
= C(\delta + \|(x, y, u, v)|_2\) \|(x, y, u, v) - (\tilde{x}, \tilde{y}, \tilde{u}, \tilde{v})\|_2
$$

with suitable $C \in \mathbb{R}^+$.}

## 3 Interpolation

### 3.1 Motivation

Given a finite number of points $(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)$, so-called data points, supporting points, or tabular points, the goal is to find a function $f$ such that $f(x_i) = y_i$ for each $i \in \{0, \ldots, n\}$. Then $f$ is called an interpolate of the data points. The data points can be measured values from a physical experiment or computed values (for example, it could be that $y_i = g(x_i)$ and $g(x)$ can, in principle, be computed, but it could be very difficult and computationally expensive ($g$ could be the solution of a differential equation) – in that case, it can also be advantageous to interpolate the data points by an approximation $f$ to $g$ such that $f$ is efficiently computable).

To have an interpolate $f$ at hand can be desirable for many reasons, including

(a) $f(x)$ can then be computed at an arbitrary point.

(b) If $f$ is differentiable, then derivatives can be computed in an arbitrary point, at least, in principle.

(c) Computation of integrals of the form $\int_a^b f(x) \, dx$, if $f$ is integrable.

(d) Determination of extreme points of $f$ (e.g. for $\mathbb{R}$-valued $f$).

(e) Determination of zeros of $f$ (e.g. for $\mathbb{R}$-valued $f$).

While a general function $f$, defined on an infinite domain, is never determined by its values on finitely many points, the situation can be different if it is known that $f$ has
additional properties, for instance, that it has to lie in a certain set of functions or that it can at least be approximated by functions from a certain function set. For example, we will see in the next section that polynomials are determined by finitely many data points.

The interpolate should be chosen such that it reflects properties expected from the exact solution (if any). For example, polynomials are not the best choice if one expects a periodic behavior or if the exact solution is known to have singularities. If periodic behavior is expected, then interpolation by trigonometric functions is usually advised (cf. Sections F.2 and F.3 of the Appendix), whereas interpolation by rational functions is often desirable if singularities are expected (cf. [FH07, Sec. 2.2]). Due to time constraints, we will only study interpolations related to polynomials in this class, but one should keep in the back of one’s mind that there are other types of interpolates that might me more suitable, depending on the circumstances.

3.2 Polynomial Interpolation

If polynomial interpolation is to mean the problem of finding a polynomial (function) that interpolates given data points, then we will see that (cf. Th. 3.4 below), in contrast to the general interpolation problem described above, the problem is purely algebraic with (in the absence of roundoff errors) an exact solution, not involving any approximation. However, for the reasons described above, one might then use the interpolating polynomial as an approximation of some more general function.

3.2.1 Polynomials and Polynomial Functions

Definition and Remark 3.1. Let $F$ be a field.

(a) We call

$$F[X] := F^{N_0}_{\text{fin}} := \{ (f : N_0 \rightarrow F) : \# f^{-1}(F \setminus \{0\}) < \infty \}$$

(3.1)

the set of polynomials over $F$ (i.e. a polynomial over $F$ is a sequence $(a_i)_{i \in N_0}$ in $F$ such that all, but finitely many, of the entries $a_i$ are 0). On $F[X]$, we have the following pointwise-defined addition and scalar multiplication:

$$\forall f,g \in F[X] \quad (f + g) : N_0 \rightarrow F, \quad (f + g)(i) := f(i) + g(i),$$

(3.2a)

$$\forall f \in F[X] \quad \forall \lambda \in F \quad (\lambda \cdot f) : N_0 \rightarrow F, \quad (\lambda \cdot f)(i) := \lambda f(i),$$

(3.2b)

where we know from Linear Algebra (cf. [Phi19a, Ex. 5.16(c)]) that, with these compositions, $F[X]$ forms a vector space over $F$ with standard basis $B := \{e_i : i \in N_0\}$, where

$$\forall i \in N_0 \quad e_i : N_0 \rightarrow F, \quad e_i(j) := \delta_{ij} := \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases}$$
In the context of polynomials, we will usually use the common notation $X^i := e_i$ and we will call these polynomials monomials. Moreover, one commonly uses the simplified notation $X := X^1 \in F[X]$ and $\lambda := \lambda X^0 \in F[X]$ for each $\lambda \in F$. Next, one defines a multiplication on $F[X]$ by letting

\[
(a_i)_{i \in \mathbb{N}_0}, (b_i)_{i \in \mathbb{N}_0} \mapsto (c_i)_{i \in \mathbb{N}_0} := (a_i)_{i \in \mathbb{N}_0} \cdot (b_i)_{i \in \mathbb{N}_0},
\]

\[
c_i := \sum_{k+l=i} a_k b_l := \sum_{(k,l) \in (\mathbb{N}_0)^2; k+l=i} a_k b_{i-k}.
\]

We then know from Linear Algebra (cf. [Phi19b, Th. 7.4(b)]) that, with the above addition and multiplication, $F[X]$ forms a commutative ring with unity, where $1 = X^0$ is the neutral element of multiplication.

(b) If $f := (a_i)_{i \in \mathbb{N}_0} \in F[X]$, then we call the $a_i \in F$ the coefficients of $f$, and we define the degree of $f$ by

\[
\deg f := \begin{cases} 
-\infty & \text{for } f \equiv 0, \\
\max \{i \in \mathbb{N}_0 : a_i \neq 0\} & \text{for } f \neq 0.
\end{cases}
\]

If $\deg f = n \in \mathbb{N}_0$ and $a_n = 1$, then the polynomial $f$ is called monic. Clearly, for each $n \in \mathbb{N}_0$, the set

\[
F[X]_n := \text{span}\{1, X, \ldots, X^n\} = \{f \in F[X] : \deg f \leq n\},
\]

of polynomials of degree at most $n$, forms an $(n + 1)$-dimensional vector subspace of $F[X]$ with basis $B_n := \{1, X, \ldots, X^n\}$.

Remark 3.2. In the situation of Def. 3.1, using the notation $X^i = e_i$, we can write addition, scalar multiplication, and multiplication in the following, perhaps, more familiar-looking forms: If $\lambda \in F$, $f = \sum_{i=0}^n f_i X^i$, $g = \sum_{i=0}^n g_i X^i$, $n \in \mathbb{N}_0$, $f_0, \ldots, f_n, g_0, \ldots, g_n \in F$, then

\[
f + g = \sum_{i=0}^n (f_i + g_i) X^i,
\]

\[
\lambda f = \sum_{i=0}^n (\lambda f_i) X^i,
\]

\[
fg = \sum_{i=0}^{2n} \left( \sum_{k+l=i} f_k g_l \right) X^i.
\]

Definition and Remark 3.3. Let $F$ be a field.

(a) For each $x \in F$, the map

\[
\epsilon_x : F[X] \rightarrow F, \quad f \mapsto \epsilon_x(f) = \epsilon_x \left( \sum_{i=0}^{\deg f} f_i X^i \right) := \sum_{i=0}^{\deg f} f_i x^i,
\]

(3.6)
is called the substitution homomorphism or evaluation homomorphism corresponding to \( x \) (indeed, \( \varepsilon_x \) is both linear and a ring homomorphism, cf. [Phi19b, Def. and Rem. 7.9]). We call \( x \in F \) a zero or a root of \( f \in F[X] \) if, and only if, \( \varepsilon_x(f) = 0 \). It is common to define

\[
\forall \ x \in F \quad f(x) := \varepsilon_x(f),
\]

(3.7)
even though this constitutes an abuse of notation, since, according to Def. and Rem. 3.1(a), \( f \in F[X] \) is a function on \( \mathbb{N}_0 \) and not a function on \( F \). However, the notation of (3.7) tends to improve readability and the meaning of \( f(x) \) is usually clear from the context.

(b) The map

\[
\phi : F[X] \longrightarrow F^F = \mathcal{F}(F, F), \quad f \mapsto \phi(f), \quad \phi(f)(x) := \varepsilon_x(f),
\]

(3.8)
constitutes a linear and unital ring homomorphism, which is injective if, and only if, \( F \) is infinite (cf. [Phi19b, Th. 7.12]). Here, \( F^F = \mathcal{F}(F, F) \) denotes the set of functions from \( F \) into \( F \), and \( \phi \) being unital means it maps 1 = \( X^0 \in F[X] \) to the constant function \( \phi(X^0) \equiv 1 \). We define

\[
\text{Pol}(F) := \phi(F[X])
\]

and call the elements of \( \text{Pol}(F) \) polynomial functions. Then \( \phi : F[X] \longrightarrow \text{Pol}(F) \) is an epimorphism (i.e. surjective) and an isomorphism if, and only if, \( F \) is infinite. If \( F \) is infinite, then we also define

\[
\forall \ P \in \text{Pol}(F) \quad \deg P := \deg \phi^{-1}(P) \in \mathbb{N}_0 \cup \{-\infty\}
\]

and

\[
\forall \ n \in \mathbb{N}_0 \quad \text{Pol}_n(F) := \phi(F[X]_n) = \{P \in \text{Pol}(F) : \deg P \leq n\}.
\]

### 3.2.2 Existence and Uniqueness

**Theorem 3.4** (Polynomial Interpolation). Let \( F \) be a field and \( n \in \mathbb{N}_0 \). Given \( n + 1 \) data points \( (x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n) \in F^2 \), such that \( x_i \neq x_j \) for \( i \neq j \), there exists a unique interpolating polynomial \( f \in F[X]_n \), satisfying

\[
\forall \ i \in \{0, 1, \ldots, n\} \quad \epsilon_{x_i}(f) = y_i.
\]

(3.9)
Moreover, one can identify \( f \) as the Lagrange interpolating polynomial, which is given by the explicit formula

\[
f := \sum_{j=0}^{n} y_j L_j, \quad \text{where} \quad \forall \ j \in \{0, 1, \ldots, n\} \quad L_j := \prod_{i=0}^{n} \frac{X - x_i}{x_j - x_i} = \prod_{i=0}^{n} \frac{X^{n-1} - x_i X^0}{x_j - x_i}
\]

(3.10)
(the \( L_j \) are called Lagrange basis polynomials).
Proof. Since \( f = f_0 X^0 + f_1 X^1 + \cdots + f_n X^n \), (3.9) is equivalent to the linear system
\[
\begin{align*}
    f_0 + f_1 x_0 + \cdots + f_n x_0^n &= y_0 \\
    f_0 + f_1 x_1 + \cdots + f_n x_1^n &= y_1 \\
    &\vdots \\
    f_0 + f_1 x_n + \cdots + f_n x_n^n &= y_n,
\end{align*}
\]
for the \( n + 1 \) unknowns \( f_0, \ldots, f_n \in F \). This linear system has a unique solution if, and only if, the determinant
\[
D := \begin{vmatrix}
1 & x_0 & x_0^2 & \cdots & x_0^n \\
1 & x_1 & x_1^2 & \cdots & x_1^n \\
& \vdots & \vdots & \ddots & \vdots \\
1 & x_n & x_n^2 & \cdots & x_n^n
\end{vmatrix}
\]
does not vanish. According to Linear Algebra, \( D \) constitutes a so-called Vandermonde determinant, satisfying (cf. [Phi19b, Th. 4.35])
\[
D = \prod_{i,j=0 \atop i>j}^n (x_i - x_j).
\]
In particular, \( D \neq 0 \), as we assumed \( x_i \neq x_j \) for \( i \neq j \), proving both existence and uniqueness of \( f \) (bearing in mind that the \( X^0, \ldots, X^n \) form a basis of \( F[X]_n \)). It remains to identify \( f \) with the Lagrange interpolating polynomial. To this end, set \( g := \sum_{j=0}^n y_j L_j \). Since, clearly, \( \deg g \leq n \), it merely remains to show \( g \) satisfies (3.9). Since the \( x_i \) are distinct, we obtain
\[
\forall k \in \{0, \ldots, n\} \quad \epsilon_{x_k}(g) = \sum_{j=0}^n y_j \epsilon_{x_k}(L_j) = \sum_{j=0}^n y_j \delta_{jk} = y_k,
\]
thereby establishing the case. \( \blacksquare \)

Example 3.5. Let \( F \) be a field. We compute the first three Lagrange basis polynomials and resulting Lagrange interpolating polynomials, given data points
\[
(x_0, y_0), (x_1, y_1), (x_2, y_2) \in F^2
\]
with \( x_0, x_1, x_2 \) being distinct: For \( n = 0 \), the product in (3.10) is empty, such that \( L_0 = 1 = X^0 \) and \( f_0 = y_0 = y_0 X^0 \). For \( n = 1 \), (3.10) yields
\[
L_0 = \frac{X - x_1}{x_0 - x_1}, \quad L_1 = \frac{X - x_0}{x_1 - x_0},
\]
and
\[
f_1 = y_0 L_0 + y_1 L_1 = y_0 \frac{X - x_1}{x_0 - x_1} + y_1 \frac{X - x_0}{x_1 - x_0} = y_0 + \frac{y_1 - y_0}{x_1 - x_0} (X - x_0),
\]
which is the straight line through \((x_0, y_0)\) and \((x_1, y_1)\). Finally, for \(n = 2\), (3.10) yields

\[
L_0 = \frac{(X - x_1)(X - x_2)}{(x_0 - x_1)(x_0 - x_2)}, \quad L_1 = \frac{(X - x_0)(X - x_2)}{(x_1 - x_0)(x_1 - x_2)},
\]

\[
L_2 = \frac{(X - x_0)(X - x_1)}{(x_2 - x_0)(x_2 - x_1)},
\]

and

\[
f_2 = y_0 L_0 + y_1 L_1 + y_2 L_2
\]

\[
= y_0 + \frac{y_1 - y_0}{x_1 - x_0} (X - x_0) + \frac{1}{x_2 - x_0} \left( \frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0} \right) (X - x_0)(X - x_1).
\]

The above formulas for \(f_0, f_1, f_2\) indicate that the interpolating polynomial can be built recursively, and we will see below (cf. (3.20a)) that this is, indeed, the case. Having a recursive formula at hand can be advantageous in many circumstances. For example, even though the complexity of building the interpolating polynomial from scratch is \(O(n^2)\) in either case, if one already has the interpolating polynomial for \(x_0, \ldots, x_n\), then, using the recursive formula (3.20a) below, one obtains the interpolating polynomial for \(x_0, \ldots, x_n, x_{n+1}\) in just \(O(n)\) steps.

Looking at the structure of \(f_2\) above, one sees that, while the first coefficient is just \(y_0\), the second one is a difference quotient (which, for \(F = \mathbb{R}\), can be seen as an approximation of some derivative \(y'(x)\)), and the third coefficient is a difference quotient of difference quotients (which, for \(F = \mathbb{R}\), can be seen as an approximation of some second derivative \(y''(x)\)). Such iterated difference quotients are called divided differences and we will see in Th. 3.17 below that the same structure does, indeed, hold for interpolating polynomials of all orders. However, we will first introduce and study a slightly more general form of polynomial interpolation, so-called Hermite interpolation, where the given points \(x_0, \ldots, x_n\) no longer need to be distinct.

### 3.2.3 Hermite Interpolation

So far, we have determined an interpolating polynomial \(f\) of degree at most \(n\) from values at \(n + 1\) distinct points \(x_0, \ldots, x_n \in F\). Alternatively, one can prescribe the values of \(f\) at less than \(n + 1\) points and, instead, additionally prescribe the values of derivatives of \(f\) (cf. Def. 3.6 below) at some of the same points, where the total number of pieces of data still needs to be \(n + 1\). For example, to determine \(f \in \mathbb{R}[X]\) of degree at most 7, one can prescribe \(f(1), f(2), f'(2), f(3), f'(3), f''(3), f'''(3), \) and \(f^{(4)}(3)\). One then says that \(x_1 = 2\) is considered with multiplicity 2 and \(x_2 = 3\) is considered with multiplicity 5. This particular kind of polynomial interpolation is known as Hermite interpolation.

While one is probably most familiar with the derivative as an analytic concept (e.g. for functions from \(\mathbb{R}\) into \(\mathbb{R}\)), for polynomials, one can define the derivative in entirely algebraic terms:
**Definition 3.6.** Let $F$ be a field. Let $n \in \mathbb{N}_0$ and $f_0, \ldots, f_n \in F$. For $f = \sum_{i=0}^n f_i X^i \in F[X]$, we define the derivative $f' \in F[X]$ of $f$ by letting

$$f' := 0 \text{ for } n = 0, \quad f' := \sum_{i=0}^{n-1} (i+1) f_{i+1} X^i \text{ for } n \geq 1.$$ 

Higher derivatives are defined recursively, in the usual way, by letting

$$f^{(0)} := f, \quad \forall k \in \mathbb{N}_0 \quad f^{(k+1)} := (f^{(k)})',$$

and we make use of the notation $f'' := f^{(2)}, f''' := f^{(3)}$.

**Proposition 3.7.** Let $F$ be a field.

(a) Forming the derivative in $F[X]$ is linear, i.e.

$$\forall f, g \in F[X] \quad \forall \lambda, \mu \in F \quad (\lambda f + \mu g)' = \lambda f' + \mu g'.$$

(b) Forming the derivative in $F[X]$ satisfies the product rule, i.e.

$$\forall f, g \in F[X] \quad (fg)' = f'g + fg'.$$

(c) For each $x \in F$ and each $n \in \mathbb{N}$, one has

$$((X - x)^n)' = n (X - x)^{n-1}.$$ 

(d) Forming higher derivatives in $F[X]$ satisfies the general Leibniz rule, i.e.

$$\forall n \in \mathbb{N}_0 \quad \forall f, g \in F[X] \quad (fg) = \sum_{k=0}^n \binom{n}{k} f^{(n-k)} g^{(k)}.$$ 

(e) For each $x \in F$, one has, setting $g^{(-1)} := 0$,

$$\forall n \in \mathbb{N}_0 \quad \forall g \in F[X] \quad ((X - x)g)^{(n)} = n g^{(n-1)} + (X - x) g^{(n)}.$$ 

**Proof.** Exercise. \(\blacksquare\)

**Theorem 3.8.** Let $F$ be a field, $n \in \mathbb{N}$, and assume char $F = 0$ or char $F \geq n$ (i.e., in $F$, $k \cdot 1 \neq 0$ for each $k \in \{1, \ldots, n-1\}$, cf. [Phi19a, Def. and Rem. 4.42]). Let $f \in F[X]$ with deg $f = n$, $x \in F$. If $m \in \{0, \ldots, n-1\}$ and, using the notation of (3.7),

$$f(x) = f'(x) = \cdots = f^{(m)}(x) = 0,$$

then $x$ is a zero of multiplicity at least $m + 1$ for $f$, i.e. (cf. [Phi19b, Cor. 7.36])

$$\exists q \in F[X] \quad \deg q = n - m - 1 \quad \land \quad f = q (X - x)^{m+1}.$$
Proof. We conduct the proof via induction on $n$: If $n = 1$, then $m = 0$, $f(x) = 0$ and, thus, $f = q(X - x)$ with $q \in F[X]$ and $\deg q = n - 1$ by [Phi19b, Prop. 7.34]. Now fix $n \in \mathbb{N}$ with $n \geq 2$ and conduct an induction over $m \in \{0, \ldots, n - 1\}$: The case $m = 0$ holds as before. Thus, let $0 < m \leq n - 1$. By the induction hypothesis for $m - 1$, there exists $g \in F[X]$ with $\deg g = n - m$ such that

$$f = g(X - x)^m. \quad (3.12)$$

Taking derivatives in (3.12) and using Prop. 3.7(b),(c) yields

$$f' = g'(X - x)^m + mg(X - x)^{m-1}. \quad (3.13)$$

As we are assuming (3.11) and $\deg f' = n - 1$, we can now apply the induction hypothesis for $n - 1$ to $f'$ to obtain $h \in F[X]$ with $\deg h = n - 1 - m$, satisfying $f' = h(X - x)^m$. Using this in (3.13) yields

$$h(X - x)^m = g'(X - x)^m + mg(X - x)^{m-1}. \quad (3.14)$$

Computing in the field of rational fractions $F(X)$ (cf. [Phi19b, Ex. 7.39(b)]), we may divide by $(X - x)^{m-1}$ to obtain

$$h(X - x) = g'(X - x) + mg.$$

As we assume $m \neq 0$ in $F$, we can solve this equation for $g$, yielding

$$g = \frac{(X - x)(h - g')}{m}. \quad (3.15)$$

Setting $q := \frac{1}{m}(h - g') \in F[X]$ and plugging (3.14) into (3.12) yields

$$f = q(X - x)^{m+1}.$$

As this implies $n = \deg f = \deg q + m + 1$ and $\deg q = n - m - 1$, both inductions and the proof are complete. ■

Example 3.9. The present example shows that, in general, one cannot expect the conclusion of Th. 3.8 to hold without the hypothesis regarding the field’s characteristic: Consider $F := \mathbb{Z}_2 = \{0, 1\}$ (i.e. $F$ is the field with two elements, $\text{char} F = 2$) and

$$f := X^3 + X^2 = X^2(X + 1), \quad f' = 3X^2 + 2X = X^2, \quad f'' = 2X = 0.$$

Then $f(0) = f'(0) = f''(0) = 0$ (i.e. we have $n = 3$ and $m = 2$ in Th. 3.8), but $x = 0$ is not a zero of multiplicity 3 of $f$ (but merely a zero of multiplicity 2).

Theorem 3.10 (Hermite Interpolation). Let $F$ be a field and $r, n \in \mathbb{N}_0$. Let $\tilde{x}_0, \ldots, \tilde{x}_r \in F$ be $r + 1$ distinct elements of $F$. Moreover, let $m_0, \ldots, m_r \in \mathbb{N}$ be such that $\sum_{k=0}^r m_k = n + 1$ and assume $y_j^{(m)} \in F$ for each $j \in \{0, \ldots, r\}$, $m \in \{0, \ldots, m_j - 1\}$. 
(a) The map
\[ A : F[X]_n \rightarrow F^{n+1}, \]
\[ A(p) := (p(x_0), \ldots, p^{(m_0-1)}(x_0), \ldots, p(x_r), \ldots, p^{(m_r-1)}(x_r)). \]
is linear.

(b) Assume \( \text{char } F = 0 \) or \( \text{char } F \geq M := \max\{m_0, \ldots, m_r\} \). Then there exists a unique polynomial \( f \in F[X]_n \) (i.e. of degree at most \( n \)) such that (using the notation of (3.7))
\[ f^{(m)}(\tilde{x}_j) = y^{(m)}_j \quad \text{for each } j \in \{0, \ldots, r\}, \ m \in \{0, \ldots, m_j - 1\}. \quad (3.15) \]

Proof. (a): If \( p, q \in F[X]_n \) and \( \lambda, \mu \in F \), then
\[ A(\lambda p + \mu q) = (\ldots, (\lambda p + \mu q)^{(l)}(\tilde{x}_k), \ldots) \overset{\text{Prop. 3.7(a)}}{=} (\ldots, (\lambda p^{(l)} + \mu q^{(l)})(\tilde{x}_k), \ldots) \]
\[ = (\ldots, \lambda p^{(l)}(\tilde{x}_k) + \mu q^{(l)}(\tilde{x}_k), \ldots) \]
\[ = \lambda(\ldots, p^{(l)}(\tilde{x}_k), \ldots) + \mu(\ldots, q^{(l)}(\tilde{x}_k), \ldots) = \lambda A(p) + \mu A(q), \]
proving the linearity of \( A \).

(b): The existence and uniqueness statement of the theorem is the same as saying that the map \( A \) of (a) is bijective. Since \( A \) is linear by (a) and since \( \dim F[X]_n = \dim F^{n+1} = n + 1 \), it suffices to show that \( A \) is injective, i.e. \( \ker A = \{0\} \). If \( A(p) = (0, \ldots, 0) \), then, by Th. 3.8, each \( \tilde{x}_k \) is a zero of \( p \) with multiplicity at least \( m_k \), such that \( p \) has at least \( n + 1 \) zeros. More precisely, Th. 3.8 implies that, if \( p \neq 0 \), then the prime factorization of \( p \) (cf. [Phi19b, Cor. 7.31]) must contain the factor \( \prod_{k=0}^{n+1}(X - \tilde{x}_k)^{m_k} \), i.e. \( \deg p \geq n + 1 \). Since \( \deg p \leq n \), we know \( p = 0 \), showing \( \ker A = \{0\} \), thereby completing the proof. \( \blacksquare \)

Example 3.11. The present example shows that, in general, one can not expect the conclusion of Th. 3.10(b) to hold without the hypothesis regarding the field's characteristic: We show that, for \( F := \mathbb{Z}_2 = \{0, 1\} \), the Hermite interpolation problem can have multiple solutions or no solutions: The 4 conditions
\[ f(0) = f'(0) = f''(0) = 0, \quad f(1) = 0 \]
are all satisfied by both \( f = 0 \in F[X]_3 \) and \( f = X^3 + X^2 \in F[X]_3 \) (cf. Ex. 3.9). On the other hand, since \( (X^3)' = 3X^2 = X^2 \) and \( (X^2)' = 2X = 0 \), one has \( f'' = 0 \) for each \( f \in F[X]_3 \), i.e. the 4 conditions
\[ f(0) = f'(0) = f''(0) = 1, \quad f(1) = 0 \]
are not satisfied by any \( f \in F[X]_3 \).

Definition 3.12. As in Th. 3.10, let \( F \) be a field and \( r, n \in \mathbb{N}_0 \); let \( \tilde{x}_0, \ldots, \tilde{x}_r \in F \) be \( r + 1 \) distinct elements of \( F \); let \( m_0, \ldots, m_r \in \mathbb{N} \) be such that \( \sum_{k=0}^{r} m_k = n + 1 \); and assume \( y^{(m)}_j \in F \) for each \( j \in \{0, \ldots, r\}, \ m \in \{0, \ldots, m_j - 1\} \). Also, as in Th. 3.10(b), assume \( \text{char } F = 0 \) or \( \text{char } F \geq M := \max\{m_0, \ldots, m_r\} \). Finally, let \( (x_0, \ldots, x_n) \in F^{n+1} \) be such that, for each \( k \in \{0, \ldots, r\} \), precisely \( m_k \) of the \( x_j \) are equal to \( \tilde{x}_k \) (i.e. \( \tilde{x}_k \) occurs precisely \( m_k \times \) times in the vector \( (x_0, \ldots, x_n) \)).
3 INTERPOLATION

(a) Let \( P[x_0, \ldots, x_n] := f \in F[X] \) denote the unique interpolating polynomial satisfying (3.15). As a caveat, we note that \( P[x_0, \ldots, x_n] \) also depends on the \( y_j^{(m)} \), which is, however, suppressed in the notation.

(b) If \( x_0, \ldots, x_n \) are all distinct (i.e. \( n = r \) and \( (x_0, \ldots, x_n) \) constitutes a permutation of \( (\tilde{x}_0, \ldots, \tilde{x}_r) \)) and \( g : F \rightarrow F \) is a function such that
\[
\forall j \in \{0, \ldots, n\} \quad g(\tilde{x}_j) = y_j^{(0)},
\]
then we define the additional notation
\[
P[g|x_0, \ldots, x_n] := P[x_0, \ldots, x_n].
\]

(c) Let \( F = \mathbb{R} \) with \( a, b \in \mathbb{R} \), \( a < b \), such that \( \tilde{x}_0, \ldots, \tilde{x}_r \in [a, b] \), and let \( g : [a, b] \rightarrow \mathbb{R} \) be an \( M \) times differentiable function such that
\[
g^{(m)}(\tilde{x}_j) = y_j^{(m)} \quad \text{for each } j \in \{0, \ldots, r\}, \quad m \in \{0, \ldots, m_j - 1\}.
\]

Then, as in (b), we define the additional notation
\[
P[g|x_0, \ldots, x_n] := P[x_0, \ldots, x_n].
\]

We remain in the situation of Def. 3.12 and denote the interpolating polynomial by \( f := P[x_0, \ldots, x_n] \). If one is only interested in the value \( f(x) \) for some \( x \in F \), rather than all the coefficients of \( f \) (or rather than evaluating \( f \) at many different elements of \( F \)), then an efficient way to obtain \( f(x) \) is by means of a so-called Neville tableau (cf. Rem. 3.14 below), making use of the recursion provided by the following Th. 3.13(a).

**Theorem 3.13.** We consider the situation of Def. 3.12.

(a) The interpolating polynomials satisfy the following recursion: For \( r = 0 \), one has
\[
P[x_0, \ldots, x_n] = P[\tilde{x}_0, \ldots, \tilde{x}_0] = \sum_{j=0}^{n} \frac{(X - \tilde{x}_0)^j}{j!} y_0^{(j)}.
\]

For \( r > 0 \), one has, for each \( k, l \in \{0, \ldots, n\} \) such that \( x_k \neq x_l \),
\[
P[x_0, \ldots, x_n] = \frac{(X - x_l)P[x_0, \ldots, \hat{x}_l, \ldots, x_n] - (X - x_k)P[x_0, \ldots, \hat{x}_k, \ldots, x_n]}{x_k - x_l}
\]
(3.16b)

(the hatted symbols \( \hat{x}_l \) and \( \hat{x}_k \) mean that the corresponding arguments are omitted).

(b) The interpolating polynomials do not depend on the order of the \( x_0, \ldots, x_n \in F \), i.e., for each permutation \( \pi \) on \( \{0, \ldots, n\} \),
\[
P[x_{\pi(0)}, \ldots, x_{\pi(n)}] = P[x_0, \ldots, x_n].
\]
Proof. (a): The case \( r = 0 \) holds, since, clearly,

\[
\forall k \in \mathbb{N}_0 \quad \left( \frac{(X - \hat{x}_0)^j}{j!} \right)^{(k)}(\hat{x}_0) = \begin{cases} 0 & \text{for } k < j, \\ 1 & \text{for } k = j, \\ 0 & \text{for } k > j, \end{cases}
\]

implying

\[
\forall m \in \{0, \ldots, n\} \quad \left( \sum_{j=0}^{n} \frac{(X - \hat{x}_0)^j}{j!} y_0^{(j)} \right)^{(m)}(\hat{x}_0) = y_0^{(m)}.
\]

We now consider the case \( r > 0 \). Letting \( p := P[x_0, \ldots, x_n] \) and letting \( q \) denote the polynomial given by the right-hand side of (3.16b), we have to show \( p = q \). Clearly, both \( p \) and \( q \) are in \( F[X]_n \). Thus, the proof is complete, if we can show that \( q \) also satisfies (3.15), i.e.

\[
q^{(m)}(\hat{x}_j) = y_j^{(m)} \quad \text{for each } j \in \{0, \ldots, r\}, m \in \{0, \ldots, m_j - 1\}. \tag{3.17}
\]

To verify (3.17), given \( k, l \in \{0, \ldots, n\} \) with \( x_k \neq x_l \), we use Prop. 3.7(e) to obtain, for each \( m \in \mathbb{N}_0 \),

\[
q^{(m)} = \frac{m P[x_0, \ldots, \hat{x}_l, \ldots, x_n]^{(m-1)} + (X - x_l)P[x_0, \ldots, \hat{x}_l, \ldots, x_n]^{(m)}}{x_k - x_l} - \frac{m P[x_0, \ldots, \hat{x}_k, \ldots, x_n]^{(m-1)} + (X - x_k)P[x_0, \ldots, \hat{x}_k, \ldots, x_n]^{(m)}}{x_k - x_l}. \tag{3.18}
\]

Now we fix \( j \in \{0, \ldots, r\} \) such that \( \hat{x}_j = x_k \) and let \( m \in \{0, \ldots, m_j - 1\} \). Then (setting \( y_{j}^{(-1)} := 0 \))

\[
q^{(m)}(\hat{x}_j) = \frac{m y_j^{(m-1)} + (x_k - x_l) y_j^{(m)} - m y_j^{(m-1)} - 0}{x_k - x_l} = y_j^{(m)}.
\]

Similarly, for \( j \in \{0, \ldots, r\} \) such that \( \hat{x}_j = x_l \): 

\[
q^{(m)}(\hat{x}_j) = \frac{m y_j^{(m-1)} + 0 - m y_j^{(m-1)} - (x_l - x_k) y_j^{(m)}}{x_k - x_l} = y_j^{(m)}.
\]

Finally, if \( j \in \{0, \ldots, r\} \) is such that \( \hat{x}_j \notin \{x_k, x_l\} \), then

\[
q^{(m)}(\hat{x}_j) = \frac{m y_j^{(m-1)} + (\hat{x}_j - x_l) y_j^{(m)} - m y_j^{(m-1)} - (\hat{x}_j - x_k) y_j^{(m)}}{x_k - x_l} = y_j^{(m)},
\]

proving that (3.17) does, indeed, hold.

(b) holds due to the fact that the interpolating polynomial \( P[x_0, \ldots, x_n] \) is determined by the conditions (3.15), which do not depend on the order of \( x_0, \ldots, x_n \) (according to Def. 3.12, this order independence of the interpolating polynomial is already built into the definition of \( P[x_0, \ldots, x_n] \)).
Remark 3.14. To use the recursion (3.16), to compute $P[x_0, \ldots, x_n](x)$ for $x \in F$, we now assume that $(x_0, \ldots, x_n)$ is ordered in such a way that identical entries are adjacent:

\[(x_0, \ldots, x_n) = (\tilde{x}_0, \ldots, \tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_1, \ldots, \tilde{x}_r, \ldots, \tilde{x}_r).\]  \hfill (3.19)

Moreover, for fixed $x \in F$, we now define

\[
\forall_{\alpha, \beta \in \{0, \ldots, n\}, \alpha \geq \beta} P_{\alpha\beta} := P[x_{\alpha-\beta}, \ldots, x_\alpha](x).
\]

Then $P_{nn} = P[x_0, \ldots, x_n](x)$ is the desired value. According to Th. 3.13(a) and (3.19),

\[
P_{\alpha\beta} = \sum_{k=0}^{m} \frac{(x - \tilde{x}_j)^k}{k!} y_j^{(k)} \quad \text{for } x_{\alpha-\beta} = x_\alpha = \tilde{x}_j \text{ and } m = \beta,
\]

\[
P_{\alpha\beta} = \frac{(x - x_{\alpha-\beta})P[x_{\alpha+1-\beta}, \ldots, x_\alpha](x) - (x - x_\alpha)P[x_{\alpha-\beta}, \ldots, x_{\alpha-1}](x)}{x_\alpha - x_{\alpha-\beta}}
\]

\[
= \frac{(x - x_{\alpha-\beta})P_{\alpha-1,\beta-1} - (x - x_\alpha)P_{\alpha-1,\beta-1}}{x_\alpha - x_{\alpha-\beta}} \quad \text{for } x_{\alpha-\beta} \neq x_\alpha.
\]

In consequence, $P_{nn}$ can be computed, in $O(n^2)$ steps, via the following Neville tableau:

\[
y^{(0)}_0 = P_{00} \quad \rightarrow \quad y^{(0)}_1 = P_{10} \rightarrow P_{11} \rightarrow \cdots \rightarrow y^{(0)}_n = P_{n0} \rightarrow P_{n1} \rightarrow \cdots \rightarrow P_{nn}
\]

Example 3.15. Consider a field $F$ with char $F \neq 2$, $n := 4$, $r := 1$, and let

\[
\tilde{x}_0 := 1, \quad \tilde{x}_1 := 2, \quad y_0 := 1, \quad y'_0 := 4, \quad y_1 := 3, \quad y'_1 := 1, \quad y''_1 := 2.
\]

Setting $x := 3$, we want to compute $P_{44} = P[1, 1, 2, 2, 2](3)$, using a Neville tableau according to Rem. 3.14. In preparation, we compute

\[
P[1, 1](3) = y_0 + (3 - 1) y^{(1)}_0 = 1 + 2 \cdot 4 = 9,
\]

\[
P[2, 2](3) = y_1 + (3 - 2) y^{(1)}_1 = 3 + 1 = 4,
\]

\[
P[2, 2, 2](3) = P[2, 2](3) + \frac{(3 - 2)^2}{2!} y^{(2)}_1 = 4 + 1 = 5.
\]
Thus, we obtain the Neville tableau

\[
\begin{align*}
P_{00} &= y_0 = 1 \\
P_{10} &= y_0 = 1 \\
P_{20} &= y_1 = 3 \rightarrow P_{21} = P[1, 2](3) = \frac{(3-1)\cdot 3 - (3-2) \cdot 1}{2-1} = 5 \rightarrow P_{22} = P[1, 1, 2](3) = \frac{2 \cdot 5 - 9}{2-1} = 1 \\
P_{30} &= y_1 = 3 \rightarrow P_{31} = P[2, 2](3) = 4 \rightarrow P_{32} = P[1, 2, 2](3) = \frac{2^2 - 5}{2-1} = 3 \\
P_{40} &= y_1 = 3 \rightarrow P_{41} = P[2, 2](3) = 4 \rightarrow P_{42} = P[2, 2, 2](3) = 5 \\
\end{align*}
\]

and, continued:

\[
\begin{align*}
\ldots & \Rightarrow P_{22} = 1 \\
\ldots & \Rightarrow P_{32} = 3 \rightarrow P_{33} = P[1, 1, 2, 2](3) = \frac{2^3 - 1}{2-1} = 5 \\
\ldots & \Rightarrow P_{42} = 5 \rightarrow P_{43} = P[1, 2, 2, 2](3) = \frac{2^5 - 3}{2-1} = 7 \rightarrow P_{44} = P[1, 1, 2, 2, 2](3) = \frac{2^7 - 5}{2-1} = 9.
\end{align*}
\]

### 3.2.4 Divided Differences and Newton’s Interpolation Formula

We now come back to the divided differences mentioned at the end of Sec. 3.2.2. We proceed to immediately define divided differences in the general context of Hermite interpolation:

**Definition 3.16.** As in Def. 3.12, let \( F \) be a field and \( r, n \in \mathbb{N}_0 \); let \( \tilde{x}_0, \ldots, \tilde{x}_r \in F \) be \( r + 1 \) distinct elements of \( F \); let \( m_0, \ldots, m_r \in \mathbb{N} \) be such that \( \sum_{k=0}^{r} m_k = n + 1 \); and assume \( y_j^{(m)} \in F \) for each \( j \in \{0, \ldots, r\} \), \( m \in \{0, \ldots, m_j - 1\} \); assume \( \text{char} \; F = 0 \) or \( \text{char} \; F \geq M := \max\{m_0, \ldots, m_r\} \); and let \((x_0, \ldots, x_n) \in F^{n+1}\) be such that, for each \( k \in \{0, \ldots, r\} \), precisely \( m_k \) of the \( x_j \) are equal to \( \tilde{x}_k \) (i.e. \( \tilde{x}_k \) occurs precisely \( m_k \) times in the vector \((x_0, \ldots, x_n)\)).

(a) We write the interpolating polynomial \( f := P[x_0, \ldots, x_n] \in F[X]_n \) in the form \( f = \sum_{i=0}^{n} a_i X^i \) with \( a_0, \ldots, a_n \in F \). The value

\[
[x_0, \ldots, x_n] := a_n \in F
\]

is called a **divided difference** of \( n \)th order (as \( P[x_0, \ldots, x_n] \), it is determined by the \( x_j \) together with the \( y_j^{(m)} \), but, once again, the dependence on the \( y_j^{(m)} \) is suppressed in the notation).

(b) If \( x_0, \ldots, x_n \) are all distinct (i.e. \( n = r \) and \( (x_0, \ldots, x_n) \) constitutes a permutation of \((\tilde{x}_0, \ldots, \tilde{x}_r)\)) and \( g : F \rightarrow F \) is a function such that

\[
\forall j \in \{0, \ldots, n\} \quad g(\tilde{x}_j) = y_j^{(0)},
\]

then we define the additional notation

\[
[g|x_0, \ldots, x_n] := [x_0, \ldots, x_n].
\]
(c) Let $F = \mathbb{R}$ with $a, b \in \mathbb{R}$, $a < b$, such that $\tilde{x}_0, \ldots, \tilde{x}_r \in [a, b]$, and let $g : [a, b] \to \mathbb{R}$ be an $M$ times differentiable function such that

$$g^{(m)}(\tilde{x}_j) = y^{(m)}_j \text{ for each } j \in \{0, \ldots, r\}, \ m \in \{0, \ldots, m_j - 1\}.$$ 

Then, as in (b), we define the additional notation

$$[g|x_0, \ldots, x_n] := [x_0, \ldots, x_n].$$

**Theorem 3.17.** In the situation of Def. 3.16, the interpolating polynomial $P[x_0, \ldots, x_n] \in F[X]^n$ can be written in the following form known as Newton’s divided difference interpolation formula:

$$P[x_0, \ldots, x_n] = \sum_{j=0}^{n} [x_0, \ldots, x_j] \omega_j,$$  \hspace{1cm} (3.20a)

where, for each $j \in \{0, 1, \ldots, n\},$

$$\omega_j := \prod_{i=0}^{j-1} (X - x_i) \in F[X]_j \subseteq F[X]_n$$  \hspace{1cm} (3.20b)

are the so-called Newton basis polynomials. In particular, (3.20a) means that the interpolating polynomials satisfy the recursive relation

$$P[x_0] = y^{(0)}_0,$$  \hspace{1cm} (3.21a)

$$P[x_0, \ldots, x_j] = P[x_0, \ldots, x_{j-1}] + [x_0, \ldots, x_j] \omega_j \text{ for } 1 \leq j \leq n.$$  \hspace{1cm} (3.21b)

**Proof.** We conduct the proof via induction on $n \in \mathbb{N}_0$: Noting $\omega_0 = 1$ (empty product), the base case ($n = 0$) is provided by the true statement

$$P[x_0] = y^{(0)}_0 = [x_0] \omega_0.$$

Now let $n > 0$. Then, according to Def. 3.16, there exist $a_0, \ldots, a_{n-1} \in F$ and $q \in F[X]_{n-1}$ such that

$$p := P[x_0, \ldots, x_n] = [x_0, \ldots, x_n] X^n + \sum_{i=0}^{n-1} a_i X^i = [x_0, \ldots, x_n] \omega_n + q.$$

Thus, $q = p - [x_0, \ldots, x_n] \omega_n$ and, to show $q = p_{n-1} := P[x_0, \ldots, x_{n-1}]$, we need to show $q$ satisfies

$$q^{(m)}(\tilde{x}_j) = y^{(m)}_j \text{ for each } j \in \{0, \ldots, r\} \setminus \{j_0\}, \ m \in \{0, \ldots, m_j - 1\},$$  \hspace{1cm} (3.22a)

$$q^{(m)}(\tilde{x}_{j_0}) = y^{(m)}_{j_0} \text{ for each } m \in \{0, \ldots, m_{j_0} - 1\} \setminus \{m_{j_0} - 1\},$$  \hspace{1cm} (3.22b)

where $j_0 \in \{0, \ldots, r\}$ is such that $x_n = \tilde{x}_{j_0}$. However, $q$ does satisfy (3.22), since (3.22) is satisfied with $q$ replaced by $p$ and, by (3.20b) in combination with Prop. 3.7(d),
\[ \omega_n^{(m)}(\tilde{x}_j) = 0 \] for each \( j \in \{0, \ldots, r\} \setminus \{j_0\} \) and each \( m \in \{0, \ldots, m_j - 1\} \), as well as for each \( m \in \{0, \ldots, m_{j_0} - 1\} \setminus \{m_{j_0} - 1\} \) for \( j = j_0 \). Thus, \( q = p_{n-1} \), showing

\[ p = [x_0, \ldots, x_n] \omega_n + q = [x_0, \ldots, x_n] \omega_n + p_{n-1} \text{ ind. hyp.} = [x_0, \ldots, x_n] \omega_n + \sum_{j=0}^{n-1} [x_0, \ldots, x_j] \omega_j, \]

thereby completing the induction. ■

We can now make use of our knowledge regarding interpolating polynomials to infer properties of the divided differences (in particular, we will see that divided differences can also be efficiently computed using Neville tableaus):

**Theorem 3.18.** In the situation of Def. 3.16, the divided differences satisfy the following:

(a) Recursion: For \( r = 0 \), one has

\[ [x_0, \ldots, x_n] = [\tilde{x}_0, \ldots, \tilde{x}_0] = \frac{y_0^{(n)}}{n!}. \] (3.23a)

For \( r > 0 \), one has, for each \( k,l \in \{0, \ldots, n\} \) such that \( x_k \neq x_l \),

\[ [x_0, \ldots, x_n] = \frac{[x_0, \ldots, \hat{x}_l, \ldots, x_n] - [x_0, \ldots, \hat{x}_k, \ldots, x_n]}{x_k - x_l} \] (3.23b)

(\text{where, as before, the hatted symbols } \hat{x}_l \text{ and } \hat{x}_k \text{ mean that the corresponding arguments are omitted}).

(b) \([x_0, \ldots, x_n]\) does not depend on the order of the \( x_0, \ldots, x_n \in F \), i.e., for each permutation \( \pi \) on \( \{0, \ldots, n\} \),

\[ [x_{\pi(0)}, \ldots, x_{\pi(n)}] = [x_0, \ldots, x_n]. \]

**Proof.** (a): Case \( r = 0 \): According to (3.16a), we have

\[ P[x_0, \ldots, x_n] = P[\tilde{x}_0, \ldots, \tilde{x}_0] = \sum_{j=0}^{n} \frac{(X - \tilde{x}_0)^j}{j!} y_0^{(j)}, \]

i.e. the coefficient of \( X^n \) is \( \frac{y_0^{(n)}}{n!} \). Case \( r > 0 \): According to (3.16b), we have, for each \( k,l \in \{0, \ldots, n\} \) such that \( x_k \neq x_l \),

\[ P[x_0, \ldots, x_n] = \frac{(X - x_l)P[x_0, \ldots, \hat{x}_l, \ldots, x_n] - (X - x_k)P[x_0, \ldots, \hat{x}_k, \ldots, x_n]}{x_k - x_l}, \]

where, comparing the coefficient of \( X^n \) on the left-hand side with the coefficient of \( X^n \) on the right-hand side immediately proves (3.23b).

(b) is immediate from Th. 3.13(b). ■
Remark 3.19. Analogous to Rem. 3.14, to use the recursion (3.23) to compute the divided difference \([x_0, \ldots, x_n]\), we now assume that \((x_0, \ldots, x_n)\) is ordered in such a way that identical entries are adjacent, i.e.

\[
(x_0, \ldots, x_n) = (\tilde{x}_0, \ldots, \tilde{x}_0, \tilde{x}_1, \ldots, \tilde{x}_1, \ldots, \tilde{x}_r, \ldots, \tilde{x}_r).
\]

Then \(P_{nn} = [x_0, \ldots, x_n]\) can be computed via the Neville tableau of Rem. 3.14, provided that we redefine

\[
P_{\alpha\beta} := [x_{\alpha-\beta}, \ldots, x_\alpha],
\]

since, with this redefinition, Th. 3.18(a) implies

\[
P_{\alpha\beta} = \frac{y_j^{(m)}}{m!} \quad \text{for } x_{\alpha-\beta} = x_\alpha = \tilde{x}_j \text{ and } m = \beta,
\]

\[
P_{\alpha\beta} = \frac{[x_{1+\alpha-\beta}, \ldots, x_\alpha] - [x_{\alpha-\beta}, \ldots, x_{\alpha-1}]}{x_\alpha - x_{\alpha-\beta}} = \frac{P_{\alpha,\beta-1} - P_{\alpha-1,\beta-1}}{x_\alpha - x_{\alpha-\beta}} \quad \text{for } x_{\alpha-\beta} \neq x_\alpha.
\]

Example 3.20. We remain in the situation of Def. 3.16.

(a) Suppose \(\text{char } F \neq 2, x_0 \neq x_1\), and we want to compute \([x_0, x_1, x_1, x_1]\) according to (3.23). First, we obtain \([x_0], [x_1]\), \([x_1, x_1]\), and \([x_1, x_1, x_1]\) from (3.23a). Then we use (3.23b) to compute, in succession,

\[
[x_0, x_1] = \frac{[x_1] - [x_0]}{x_1 - x_0}, \quad [x_0, x_1, x_1] = \frac{[x_1, x_1] - [x_0, x_1]}{x_1 - x_0},
\]

\[
[x_0, x_1, x_1, x_1] = \frac{[x_1, x_1, x_1] - [x_0, x_1, x_1]}{x_1 - x_0}.
\]

(b) Suppose \(x_0, \ldots, x_n \in F\) are all distinct and set \(y_j := y_j^{(0)}\) for each \(j \in \{0, \ldots, n\}\). Then the Neville tableau of Rem. 3.19 takes the form

\[
y_0 = [x_0]
\]

\[
y_1 = [x_1] \to [x_0, x_1]
\]

\[
y_2 = [x_2] \to [x_1, x_2] \to [x_0, x_1, x_2]
\]

\[
\vdots
\]

\[
y_{n-1} = [x_{n-1}] \to [x_{n-2}, x_{n-1}] \to \ldots \to [x_0, \ldots, x_{n-1}]
\]

\[
y_n = [x_n] \to [x_{n-1}, x_n] \to \ldots \to [x_1, \ldots, x_n] \to [x_0, \ldots, x_n]
\]

(c) As in Ex. 3.15, consider a field \(F\) with \(\text{char } F \neq 2, n := 4, r := 1\), and the given data

\[
\tilde{x}_0 = 1, \quad \tilde{x}_1 = 2,
\]

\[
y_0 = 1, \quad y_0' = 4, \quad y_1 = 3, \quad y_1' = 1, \quad y_1'' = 2.
\]
This time, we seek to obtain the complete interpolating polynomial $P[1, 1, 2, 2, 2]$, using Newton’s interpolation formula (3.20a):

$$P[1, 1, 2, 2, 2] = [1] + [1, 1](X - 1) + [1, 1, 2](X - 1)^2 + [1, 1, 2, 2](X - 1)^2(X - 2)$$

$$+ [1, 1, 2, 2, 2](X - 1)^2(X - 2)^2.$$  \hfill (3.24)

The coefficients can be computed from the Neville tableau given by Rem. 3.19:

$[1] = y_0 = 1$

$[1] = y_0 = 1 \rightarrow [1, 1] = y'_0 = 4$

$[2] = y_1 = 3 \rightarrow [2, 1] = \frac{3 - 1}{2 - 1} = 2 \rightarrow [1, 1, 2] = \frac{2 - 4}{2 - 1} = -2$

$[2] = y_1 = 3 \rightarrow [2, 2] = y''_1 = 1 \rightarrow [1, 1, 2, 2] = \frac{1 - (-2)}{2 - 1} = 1$

$[2] = y_1 = 3 \rightarrow [2, 2] = y''_1 = 1 \rightarrow [2, 2, 2] = \frac{y''_2}{2} = 1 \rightarrow [1, 1, 2, 2, 2] = \frac{1 - (-1)}{2 - 1} = 2$

and, finally,

$$\ldots [1, 1, 2, 2] = 1$$

$$\ldots [1, 2, 2, 2] = 2 \rightarrow [1, 1, 2, 2, 2] = \frac{2 - 1}{2 - 1} = 1.$$

Hence, plugging the results of the tableau into (3.24),

$$P[1, 1, 2, 2, 2] = 1 + 4(X - 1) - 2(X - 1)^2 + (X - 1)^2(X - 2) + (X - 1)^2(X - 2)^2.$$  \hfill —

Additional formulas for divided differences over a general field $F$ can be found in Appendix E. Here, we now proceed to present a result for the field $F = \mathbb{R}$, where there turns out to be a surprising relation between divided differences and integrals over simplices. We start with some preparations:

**Notation 3.21.** For $n \in \mathbb{N}$, the following notation is introduced:

$$\Sigma^n := \left\{ (s_1, \ldots, s_n) \in \mathbb{R}^n : \sum_{i=1}^n s_i \leq 1 \text{ and } 0 \leq s_i \text{ for each } i \in \{1, \ldots, n\} \right\}.$$

**Remark 3.22.** The set $\Sigma^n$ introduced in Not. 3.21 is the convex hull of the set consisting of the origin and the $n$ standard unit vectors $e_1, \ldots, e_n \in \mathbb{R}^n$, i.e. the convex hull of the $(n - 1)$-dimensional standard simplex $\Delta^{n-1}$ united with the origin (cf. [Phi19b, Def. and Rem. 1.23]):

$$\Sigma^n = \text{conv} \left( \{0\} \cup \{e_i : i \in \{1, \ldots, n\}\} \right) = \text{conv} \left( \{0\} \cup \Delta^{n-1} \right),$$

where

$$e_i := (0, \ldots, \delta_i, \ldots, 0), \quad \Delta^{n-1} := \text{conv} \{e_i : i \in \{1, \ldots, n\}\}. \quad (3.25)$$
Lemma 3.23. For each \( n \in \mathbb{N} \):

\[
\text{vol}(\Sigma^n) := \int_{\Sigma^n} 1 \, ds = \frac{1}{n!}.
\] (3.26)

Proof. Consider the change of variables \( T : \mathbb{R}^n \rightarrow \mathbb{R}^n \), \( t = T(s) \), where

\[
t_1 := s_1, \quad s_1 := t_1, \quad s_2 := t_2 - t_1,
\]

\[
t_2 := s_1 + s_2, \quad s_3 := t_3 - t_2, \quad \vdots \quad t_n := s_1 + \cdots + s_n, \quad s_n := t_n - t_{n-1}.
\] (3.27)

According to (3.27), \( T \) is clearly linear and bijective, \( \det J^{-1}_T \equiv \det T^{-1} = 1 \), and

\[
T(\Sigma^n) = \{ (t_1, \ldots, t_n) \in \mathbb{R}^n : 0 \leq t_1 \leq t_2 \leq \cdots \leq t_n \leq 1 \}.
\]

Thus, using the change of variables theorem (CVT) and the Fubini theorem (FT), we can compute \( \text{vol}(\Sigma^n) \):

\[
\text{vol}(\Sigma^n) = \int_{\Sigma^n} 1 \, ds \stackrel{(CVT)}{=} \int_{T(\Sigma^n)} \det J_{T^{-1}}(t) \, dt = \int_{T(\Sigma^n)} 1 \, dt \quad \stackrel{(FT)}{=} \quad \int_0^1 \cdots \int_0^{t_3} dt_3 \cdots \int_0^{t_2} dt_2 \cdots dt_1
\]

\[
= \int_0^1 \cdots \int_0^{t_3} \frac{t_3^{n-1}}{n!} dt_3 \cdots dt_1 = \frac{1}{n!},
\]

proving (3.26). \( \blacksquare \)

Theorem 3.24 (Hermite-Genocchi Formula). Consider the situation of Def. 3.16(c), i.e. \( r, n \in \mathbb{N}_0 \) with \( \tilde{x}_0, \ldots, \tilde{x}_r \in [a, b] \subseteq \mathbb{R} \) distinct \( a, b \in \mathbb{R}, a < b \); \( m_0, \ldots, m_r \in \mathbb{N} \) such that \( \sum_{k=0}^r m_k = n + 1 \) and \( g \in C^n[a, b] \) (this assumption is actually stronger than in Def. 3.16(c)) with

\[
g^{(m)}(\tilde{x}_j) = y^{(m)}_j \in \mathbb{R} \quad \text{for each } j \in \{0, \ldots, r\}, \quad m \in \{0, \ldots, m_j - 1\};
\]

and \( (x_0, \ldots, x_n) \in \mathbb{R}^{n+1} \) such that, for each \( k \in \{0, \ldots, r\} \), precisely \( m_k \) of the \( x_j \) are equal to \( \tilde{x}_k \). Then the divided differences satisfy the following relation, known as the Hermite-Genocchi formula:

\[
n \geq 1 \quad \Rightarrow \quad [g|x_0, \ldots, x_n] = \int_{\Sigma^n} g^{(m)} \left( x_0 + \sum_{i=1}^n s_i (x_i - x_0) \right) \, ds,
\] (3.28)

where \( \Sigma^n \) is the set according to Not. 3.21.
Proof. First, note that, for each \( s \in \Sigma^n \),
\[
a = x_0 + a - x_0 \leq x_0 + \sum_{i=1}^{n} s_i(a - x_0) \leq x_s := x_0 + \sum_{i=1}^{n} s_i(x_i - x_0) \leq x_0 + \sum_{i=1}^{n} s_i(b - x_0) \leq x_0 + b - x_0 = b,
\]
such that \( g^{(n)} \) is defined at \( x_s \). Moreover, the integral in (3.28) exists, as the continuous function \( g^{(n)} \) is bounded on the compact set \( \Sigma^n \). If \( r = 0 \) (i.e. \( x_0 = \cdots = x_n = \tilde{x}_0 \)), then
\[
\int_{\Sigma^n} g^{(n)}(x_0 + \sum_{i=1}^{n} s_i(x_i - x_0)) \, ds = \int_{\Sigma^n} g^{(n)}(x_0) \, ds = \frac{g^{(n)}(x_0)}{n!} = \frac{n!}{(3.23a)} = [g|x_0, \ldots, x_n],
\]
proving (3.28) in this case. For \( r > 0 \), we prove (3.28) via induction on \( n \in \mathbb{N} \), where, by Th. 3.18(b), we may assume \( x_0 \neq x_n \). Then, for \( n = 1 \), (3.28) is an easy consequence of the fundamental theorem of calculus (FTC) (note \( \Sigma^1 = [0, 1] \)):
\[
\int_{0}^{1} g'(x_0 + s(x_1 - x_0)) \, ds = \frac{g(x_0 + s(x_1 - x_0) - x_0)}{x_1 - x_0} \bigg|_{s=0}^{1} = \frac{g(x_0) - g(x_0)}{x_1 - x_0} = \frac{y_1 - y_0}{x_1 - x_0} = [g|x_0, x_1].
\]
Next, for the induction step, let \( n \in \mathbb{N} \) and compute
\[
\int_{\Sigma^{n+1}} g^{(n+1)}(x_0 + \sum_{i=1}^{n+1} s_i(x_i - x_0)) \, ds = \frac{1}{x_{n+1} - x_0} \int_{\Sigma^n} \left( g^{(n)} \left( x_{n+1} + \sum_{i=1}^{n} s_i(x_i - x_{n+1}) \right) - g^{(n)} \left( x_0 + \sum_{i=1}^{n} s_i(x_i - x_0) \right) \right) \, ds_{n+1} \, ds_1 \cdots ds_n
\]
\[
= \frac{[g|x_1, \ldots, x_{n+1}] - [g|x_0, \ldots, x_n]}{x_{n+1} - x_0} = [g|x_0, \ldots, x_{n+1}],
\]
thereby establishing the case. \( \blacksquare \)

**Theorem 3.25.** Under the assumptions of Th. 3.24 and letting \( p := P[g|x_0, \ldots, x_n] \in \mathbb{R}[X]_n \) be the corresponding interpolating polynomial (cf. Def. 3.12(c)), the following holds true:
(a) For each \( x \in [a, b] \):

\[
g(x) = p(x) + [g|x, x_0, \ldots, x_n] \omega_{n+1}(x),
\]

where

\[
\omega_{n+1} = \prod_{i=0}^{n} (X - x_i)
\]

(if \( x = x_0 = \cdots = x_n \), then \([g|x, x_0, \ldots, x_n]\) is only defined if \( g \) is \((n + 1)\) times differentiable – however, in this case, \( \omega_{n+1}(x) = 0 \) and \( g(x) = p(x) \) still holds, even if \([g|x, x_0, \ldots, x_n]\) is not defined).

(b) (Mean Value Theorem for Divided Differences) Given \( x \in [a, b] \), let

\[
m := \min\{x, x_0, \ldots, x_n\}, \quad M := \max\{x, x_0, \ldots, x_n\}.
\]

If \( g \in C^{n+1}[a, b] \), then there exists \( \xi = \xi(x, x_0, \ldots, x_n) \in [m, M] \) such that

\[
[g|x, x_0, \ldots, x_n] = \frac{g^{(n+1)}(\xi)}{(n+1)!}.
\]

In consequence, with \( \omega_{n+1} \) as in (a), one then also has

\[
g(x) = p(x) + \frac{g^{(n+1)}(\xi)}{(n+1)!} \omega_{n+1}(x).
\]

(c) The function

\[
\delta : [a, b]^{n+1} \longrightarrow \mathbb{R}, \quad (x_0, \ldots, x_n) \mapsto [g|x_0, \ldots, x_n],
\]

is continuous, in particular, continuous in each \( x_k, k \in \{0, \ldots, n\} \).

Proof. (a): For \( x \in [a, b] \), one obtains:

\[
p(x) + [g|x, x_0, \ldots, x_n] \omega_{n+1}(x) \overset{(3.20)}{=} \sum_{j=0}^{n} [g|x_0, \ldots, x_j] \omega_j(x) + [g|x_0, \ldots, x_n, x] \omega_{n+1}(x) \overset{(3.20)}{=} g(x).
\]

(b): If \( x = x_0 = \cdots = x_n \), then (3.30) holds due to

\[
[g|x, x_0, \ldots, x_n] \overset{(3.23a)}{=} \frac{g^{(n+1)}(x_0)}{(n+1)!} = \frac{g^{(n+1)}(x_0)}{(n+1)!}.
\]

Otherwise, we can apply Th. 3.24 with \( a, b \) replaced by \( m, M \). Using (3.28) together with the mean value theorem of integration (see, e.g., [Phi17, Th. 2.16(f)]) then yields \( \xi \in [m, M] \) such that

\[
[g|x, x_0, \ldots, x_n] = \frac{g^{(n+1)}(\xi)}{(n+1)!},
\]
proving (3.30). Plugging (3.30) into (3.29) yields (3.31).

c: For \( n = 0 \), the continuity of \( x_0 \mapsto [g|x_0] = g(x_0) \) is merely the assumed continuity of \( g \). Let \( n > 0 \). Since \( g^{(n)} \) is continuous and the continuous function \( |g^{(n)}| \) is uniformly bounded on the compact interval \([a, b] \), the continuity of \( \delta \) is due to (3.28) and the continuity of parameter-dependent integrals (see, e.g., [Phi17, Th. 2.22]). ■

### 3.2.5 Convergence and Error Estimates

We will now provide a convergence result for the approximation of a (benign) \( C^\infty \) function by interpolating polynomial functions. Recall from Def. and Rem. 3.3(b) that, given \( f \in \mathbb{R}[X] \), \( \phi(f) \in \text{Pol}(\mathbb{R}) \), \( \phi(f)(x) = f(x) \), is the corresponding polynomial function.

**Theorem 3.26.** As in Th. 3.24, let \( a, b \in \mathbb{R}, a < b \). Let \( g \in C^\infty[a, b] \) and consider a sequence \((x_k)_{k \in \mathbb{N}}\) in \([a, b] \). For each \( n \in \mathbb{N}, \) define \( p_n := \text{P}[g|x_0, \ldots, x_n] \in \mathbb{R}[X]_n \) to be the interpolating polynomial corresponding to the first \( n + 1 \) points – thus, according to Th. 3.17 and Th. 3.24,

\[
\forall \ n \in \mathbb{N} \quad p_n = \sum_{j=0}^{n} [g|x_0, \ldots, x_j] \omega_j,
\]

where

\[
[g|x_0, \ldots, x_j] = \int_{x_0}^{x_j} g^{(j)} \left( x_0 + \sum_{i=1}^{j} s_i(x_i - x_0) \right) \, ds, \quad \omega_j := \prod_{i=0}^{j-1} (X - x_i).
\]

If there exist \( K, R \in \mathbb{R}_0^+ \) such that

\[
\forall \ n \in \mathbb{N} \quad \|g^{(n)}\|_\infty \leq Kn!R^n, \quad (3.32)
\]

then one has the error estimate

\[
\forall \ n \in \mathbb{N}_0 \quad \|g - \phi(p_n)\|_\infty \leq \frac{\|g^{(n+1)}\|_\infty}{(n+1)!} \|\phi(\omega_{n+1})\|_\infty \leq KR^{n+1}(b-a)^{n+1}. \quad (3.33)
\]

If, moreover, \( R < (b-a)^{-1} \), then one also has the convergence

\[
\lim_{n \to \infty} \|g - \phi(p_n)\|_\infty = 0. \quad (3.34)
\]

**Proof.** The first estimate in (3.33) is merely (3.31), while the second estimate in (3.33) is (3.32) combined with \( \|\phi(\omega_{n+1})\|_\infty \leq (b-a)^{n+1} \). As \( R < (b-a)^{-1} \), (3.33) establishes (3.34). ■

**Example 3.27.** Suppose \( g \in C^\infty[a, b] \) is such that there exists \( C \geq 0 \) satisfying

\[
\forall \ n \in \mathbb{N} \quad \|g^{(n)}\|_\infty \leq C^n. \quad (3.35)
\]
We claim that, in that case, there are $K \in \mathbb{R}_0^+$ and $0 \leq R < (b - a)^{-1}$ satisfying (3.32) (in particular, this shows that Th. 3.26 applies to all functions of the form $g(x) = e^{kx}$, $g(x) = \sin(kx)$, $g(x) = \cos(kx)$). To verify that (3.35) implies (3.32), set $R := \frac{1}{2}(b-a)^{-1}$, choose $N \in \mathbb{N}$ sufficiently large such that $n! > (C/R)^n$ for each $n > N$, and set $K := \max\{(C/R)^m : 0 \leq m \leq N\}$. Then, for each $n \in \mathbb{N}$,

$$\|g^{(n)}\|_{\infty} \leq C^n = (C/R)^n R^n \leq K n! R^n.$$

**Remark 3.28.** When approximating $g \in C^\infty[a,b]$ by an interpolating polynomial $p_n$, the accuracy of the approximation depends on the choice of the $x_j$. If one chooses equally-spaced $x_j$, then $\omega_{n+1}(x)$ will be much larger in the vicinity of $a$ and $b$ than in the center of the interval. This behavior of $\omega_{n+1}$ can be counteracted by choosing more data points in the vicinity of $a$ and $b$. This leads to the problem of finding data points such that $\|\phi(\omega_{n+1})\|_{\infty}$ is minimized. If $a = -1$, $b = 1$, then the optimal choice for the data points is

$$x_j := \cos\left(\frac{2j + 1}{2n + 2} \pi\right) \text{ for each } j \in \{0, \ldots, n\}.$$

These $x_j$ are precisely the zeros of the so-called Chebyshev polynomials, which we will study in Sec. 3.2.6 below (cf. Ex. 3.35).

One can compute $n! [g|x_0, \ldots, x_n]$ as a numerical approximation to $g^{(n)}(x)$. We obtain the following error estimate:

**Theorem 3.29** (Numerical Differentiation). Let $a, b \in \mathbb{R}$, $a < b$, $g \in C^{n+1}[a,b]$, $n \in \mathbb{N}_0$, $x \in [a,b]$. Then, given $n + 1$ (not necessarily distinct) points $x_0, \ldots, x_n \in [a,b]$ and defining $[g|x_0, \ldots, x_n]$ as in Def. 3.16(c), the following estimate holds:

$$\left|g^{(n)}(x) - n! [g|x_0, \ldots, x_n]\right| \leq (M - m) \|g^{(n+1)}|_{[m,M]}\|_{\infty},$$

(3.36)

where $|$ denotes restriction and, as in Th. 3.25(b),

$$m := \min\{x, x_0, \ldots, x_n\}, \quad M := \max\{x, x_0, \ldots, x_n\}.$$

**Proof.** For each $\xi \in [m, M]$, we have

$$\left|g^{(n)}(x) - g^{(n)}(\xi)\right| \leq |x - \xi| \|g^{(n+1)}|_{[m,M]}\|_{\infty},$$

(3.37)

which is clear for $x = \xi$ and, otherwise, due to the mean value theorem [Phi16a, Th. 9.18]. For $n = 0$, the left-hand side of (3.36) is $|g(x) - g(x_0)|$ and (3.36) follows from (3.37). If $n \geq 1$, then we apply (3.30) with $n - 1$, yielding $n! [g|x_0, \ldots, x_n] = g^{(n)}(\xi)$ for some $\xi \in [m, M]$, such that (3.36) follows from (3.37) also in this case. ■
3.2.6 Chebyshev Polynomials

As mentioned in Rem. 3.28 above, the zeros of the Chebyshev polynomials yield optimal points for the use of polynomial interpolation. We will provide this result as well as some other properties of Chebyshev polynomials in the present section.

**Definition 3.30.** We define the Chebyshev polynomials $T_n \in \text{Pol}(\mathbb{R})$, $n \in \mathbb{N}_0$, via the following recursion:

$$T_0(x) := 1, \quad T_1(x) := x, \quad \forall \ n \geq 2 \ T_n(x) := 2xT_{n-1}(x) - T_{n-2}(x). \quad (3.38)$$

**Theorem 3.31.** The Chebyshev polynomials $T_n \in \text{Pol}(\mathbb{R})$, $n \in \mathbb{N}_0$, as defined in Def. 3.30, have the following properties:

(a) $T_n \in \text{Pol}_n(\mathbb{R})$, deg $T_n = n$, and all coefficients of $T_n$ are integers.

(b) For $n \geq 1$, the coefficient of $x^n$ in $T_n$ is $a_n(T_n) = 2^{n-1}$.

(c) $T_n$ is even for $n$ even; $T_n$ is odd for $n$ odd, i.e.

$$\forall \ x \in \mathbb{R} \ T_n(-x) = \begin{cases} T_n(x) & \text{for } n \text{ even,} \\ -T_n(x) & \text{for } n \text{ odd.} \end{cases}$$

(d) $T_n(1) = 1$, $T_n(-1) = (-1)^n$.

(e) One has

$$T_n(x) = \begin{cases} \cos(n \arccos x) & \text{for } x \in [-1, 1], \\ \cosh(n \arcosh x) & \text{for } x \in [1, \infty[, \\ (-1)^n \cosh(n \arcosh(-x)) & \text{for } x \in ]-\infty, -1], \end{cases}$$

where

$$\cosh : \mathbb{R} \rightarrow [1, \infty[, \quad \cosh(x) = \frac{e^x + e^{-x}}{2},$$

with the inverse function

$$\arcosh : [1, \infty[ \rightarrow \mathbb{R}, \quad \arcosh(x) = \ln(x + \sqrt{x^2 - 1}).$$

(f) $T_n$ has $n$ simple zeros, all in the interval $[-1, 1]$, namely, for $n \geq 1$,

$$x_k := \cos\left(\frac{2k-1}{2n} \pi\right), \quad k \in \{1, \ldots, n\}.$$  

(g) One has

$$\forall \ x \in [-1, 1] \ T_n(x) \in [-1, 1],$$

where, for $n \geq 1$,

$$T_n(x) = \begin{cases} 1 & \text{if, and only if, } x = \cos\frac{kn}{n} \text{ with } k \in \{0, \ldots, n\} \text{ even,} \\ -1 & \text{if, and only if, } x = \cos\frac{kn}{n} \text{ with } k \in \{0, \ldots, n\} \text{ odd.} \end{cases}$$
3 INTERPOLATION

Proof. Exercise. ■

Notation 3.32. Let $n \in \mathbb{N}_0$. For each $p \in \text{Pol}(\mathbb{R})$, let $a_n(p) \in \mathbb{R}$ denote the coefficient of $x^n$ in $p$.

Lemma 3.33. Let $a, b \in \mathbb{R}$ with $a < b$. Define

$$\alpha : [a, b] \rightarrow [-1, 1], \quad \alpha(x) := 2 \frac{x - a}{b - a} - 1,$$

$$\beta : [-1, 1] \rightarrow [a, b], \quad \beta(x) := -\frac{b - a}{2} x + \frac{b + a}{2}.$$

(a) $\alpha, \beta$ are bijective, strictly increasing affine functions with $\beta = \alpha^{-1}$.

(b) If $f : [a, b] \rightarrow \mathbb{R}$, then (in $[0, \infty]$):

$$\|f\|_{\infty} = \sup \{|f(x)| : x \in [a, b]\} = \sup \{|f(\beta(x))| : x \in [-1, 1]\} = \|f \circ \beta\|_{\infty}.$$

(c) If $p \in \text{Pol}_n(\mathbb{R})$, then

$$a_n(p \circ \beta) = a_n(p) \frac{(b - a)^n}{2^n}.$$

Proof. (a): Clearly, $\alpha, \beta$ are both restrictions of affine polynomial functions with $\deg \alpha = \deg \beta = 1$. They are strictly increasing (and, thus, bijective) e.g. due to $\alpha' = \frac{2}{b-a} > 0$ and $\beta' = \frac{b-a}{2} > 0$. Moreover, $\alpha(a) = -1$, $\alpha(b) = 1$, $\beta(-1) = a$, $\beta(1) = b$, which also shows $\beta = \alpha^{-1}$.

(b) holds, as the bijectivity of $\beta$ implies $|f|([a, b]) = |f \circ \beta|([-1, 1])$.

(c) is a simple consequence of the binomial theorem (cf. [Phi16a, Th. 5.16]). ■

Theorem 3.34. Let $n \in \mathbb{N}$ and $a, b \in \mathbb{R}$ with $a < b$. Then

$$\forall p \in \text{Pol}_n(\mathbb{R}) \quad (a_n(p) \neq 0 \Rightarrow \|p|_{[a,b]} \|_{\infty} \geq |a_n(p)| \frac{(b - a)^n}{2^{2n-1}})$$

and, in particular,

$$1 = \|T_n|_{[-1,1]} \|_{\infty} = \min \left\{ \|p|_{[-1,1]} \|_{\infty} : p \in \text{Pol}_n(\mathbb{R}), \deg p = n, a_n(p) = 2^{n-1} \right\}$$

(i.e. $T_n$ minimizes the $\infty$-norm on $[-1, 1]$ in the set of polynomial functions with degree $n$ and highest coefficient $2^{n-1}$).

Proof. We first consider the special case $a = -1, b = 1$, and $a_n(p) = 2^{n-1}$. Seeking a contradiction, assume $p \in \text{Pol}_n(\mathbb{R})$ with $\deg p = n, a_n(p) = 2^{n-1}$, and $\|p|_{[-1,1]} \|_{\infty} < 1$. Using Th. 3.31(g), we obtain

$$\forall k \in \{0, \ldots, n\} \quad (p - T_n) \left( \cos \frac{k\pi}{n} \right) \begin{cases} < 0 & \text{for } k \text{ even}, \\
 > 0 & \text{for } k \text{ odd}, \end{cases}$$
showing $p - T_n$ to have at least $n$ zeros due to the intermediate value theorem (cf. [Phi16a, Th. 7.57]). Since $p, T_n \in \text{Pol}_n(\mathbb{R})$ with $a_n(p) = a_n(T_n) = 2^{n-1}$, we have $p - T_n \in \text{Pol}_{n-1}(\mathbb{R})$ and the $n$ zeros yield $p - T_n = 0$ in contradiction to $p \neq T_n$. Thus, there must exist $x_0 \in [-1, 1]$ such that $|p(x_0)| \geq 1$. Now if $q \in \text{Pol}_n(\mathbb{R})$ is arbitrary with $a_n(q) \neq 0$ and $p := \frac{2^{n-1}}{a_n(q)} q$, then $p \in \text{Pol}_n(\mathbb{R})$ with $a_n(p) = 2^{n-1}$ and, due to the special case, 

$$\exists x_0 \in [-1, 1], \quad |q(x_0)| = \left| \frac{a_n(q)}{2^{n-1}} p(x_0) \right| \geq \frac{|a_n(q)|}{2^{n-1}},$$

as desired. Moreover, on the interval $[a, b]$, one then obtains 

$$\| q \|_{[a, b]} \overset{\text{Lem. 3.33(b)}}{=} \| q \circ \beta \|_\infty \overset{\text{Lem. 3.33(c)}}{\geq} |a_n(q)| \frac{(b - a)^n}{2^{2n-1}},$$

thereby completing the proof. 

**Example 3.35.** Let $n \in \mathbb{N}$ and $a, b \in \mathbb{R}$ with $a < b$. As mentioned in Rem. 3.28, minimizing the error in (3.33) leads to the problem of finding $x_0, \ldots, x_n \in [a, b]$ such that 

$$\max \left\{ \prod_{j=0}^{n} |x - x_j| : x \in [a, b] \right\}$$

is minimized. In other words, one needs to find the zeros of a monic polynomial function $p$ in $\text{Pol}_{n+1}(\mathbb{R})$ such that $p$ minimizes the $\infty$-norm on $[a, b]$. According to Th. 3.34, for $a = -1$, $b = 1$, $p = T_{n+1}/2^n$, which, according to Th. 3.31(f) has zeros 

$$x_j = \cos \left( \frac{2(j + 1) - 1}{2(n + 1)} \pi \right) = \cos \left( \frac{2j + 1}{2n + 2} \pi \right), \quad j \in \{0, \ldots, n\}.$$

In the general case, using the functions $\alpha, \beta$ of Lem. 3.33, we obtain $p = (T_{n+1} \circ \alpha) \cdot \frac{(b-a)^{n+1}}{2^{2n+1}}$ with the zeros 

$$x_j = \beta \left( \cos \left( \frac{2j + 1}{2n + 2} \pi \right) \right), \quad j \in \{0, \ldots, n\}.$$ 

**Theorem 3.36.** Let $n \in \mathbb{N}$ and $a, b \in \mathbb{R}$ with $a < b$. Assume $s \in \mathbb{R} \setminus [a, b]$ and define 

$$c_s := \frac{1}{T_n(\alpha(s))} \in \mathbb{R} \setminus \{0\}, \quad p_s := c_s (T_n \circ \alpha) \in \text{Pol}_n(\mathbb{R}),$$

where $\alpha : [a, b] \rightarrow [-1, 1]$ is as in Lem. 3.33, extended to $\alpha : \mathbb{R} \rightarrow \mathbb{R}$ (also note $\alpha(s) \notin [-1, 1]$ and $T_n(\alpha(s)) \neq 0$). Then 

$$|c_s| = \| p_s \|_{[a, b]} \|_\infty = \min \left\{ \| p \|_{[a, b]} \|_\infty : p \in \text{Pol}_n(\mathbb{R}), \deg p = n, p(s) = 1 \right\}$$

(i.e. $p_s$ minimizes the $\infty$-norm on $[a, b]$ in the set of polynomial functions with degree $n$ and value 1 at $s$). 

**Proof.** Exercise. 


3 INTERPOLATION

3.3 Spline Interpolation

3.3.1 Introduction, Definition of Splines

In the previous sections, we have used a single polynomial to interpolate given data points. If the number of data points is large, then the degree of the interpolating polynomial is likewise large. If the data points are given by a function, and the goal is for the interpolating polynomial to approximate that function, then, in general, one might need many data points to achieve a desired accuracy. We have also seen that the interpolating polynomials tend to be large in the vicinity of the outermost data points, where strong oscillations are typical. One possibility to counteract that behavior is to choose an additional number of data points in the vicinity of the boundary of the considered interval (cf. Rem. 3.28).

The advantage of the approach of the previous sections is that the interpolating polynomial is \( C^\infty \) (i.e. it has derivatives of all orders) and one can make sure (by Hermite interpolation) that, in a given point \( x_0 \), the derivatives of the interpolating polynomial \( p \) agree with the derivatives of a given function in \( x_0 \) (which, again, will typically increase the degree of \( p \)). The disadvantage is that one often needs interpolating polynomials of large degrees, which can be difficult to handle (e.g. numerical problems resulting from large numbers occurring during calculations).

Spline interpolation follows a different approach: Instead of using one single interpolating polynomial of potentially large degree, one uses a piecewise interpolation, fitting together polynomials of low degree (e.g. 1 or 3). The resulting function \( s \) will typically not be of class \( C^\infty \), but merely of class \( C^1 \) or \( C^2 \), and the derivatives of \( s \) will usually not agree with the derivatives of a given \( f \) (only the values of \( f \) and \( s \) will agree in the data points). The advantage is that one only needs to deal with polynomials of low degree. If one does not need high differentiability of the interpolating function and one is mostly interested in a good approximation with respect to \( \| \cdot \|_\infty \), then spline interpolation is usually a good option.

**Definition 3.37.** Let \( a, b \in \mathbb{R}, a < b \). Then, given \( N \in \mathbb{N} \),

\[
\Delta := (x_0, \ldots, x_N) \in \mathbb{R}^{N+1}
\]

is called a *knot vector* for \([a, b]\) if, and only if, it corresponds to a partition \( a = x_0 < x_1 < \cdots < x_N = b \) of \([a, b]\). The \( x_k \) are then also called *knots*. Let \( l \in \mathbb{N} \). A *spline* of order \( l \) for \( \Delta \) is a function \( s \in C^{l-1}[a, b] \) such that, for each \( k \in \{0, \ldots, N-1\} \), there exists a polynomial function \( p_k \in \text{Pol}_l(\mathbb{R}) \) that coincides with \( s \) on \([x_k, x_{k+1}]\). The set of all such splines is denoted by \( S_{\Delta,l} \):

\[
S_{\Delta,l} := \left\{ s \in C^{l-1}[a, b] : \forall k \in \{0, \ldots, N-1\} \exists p_k \in \text{Pol}_l(\mathbb{R}) \ p_k {\big|}_{[x_k, x_{k+1}]} = s {\big|}_{[x_k, x_{k+1}]} \right\}.
\]

Of particular importance are the elements of \( S_{\Delta,1} \) (*linear splines*) and of \( S_{\Delta,3} \) (*cubic splines*). It is also useful to define

\[
S_{\Delta,0} := \text{span} \left\{ \chi_k : k \in \{0, \ldots, N-1\} \right\},
\]

(3.41)
For each $S$ the Definition 3.39. In the situation of Def. 3.37, if $S$ \[ \text{Definition 3.39.} \]

We call the elements of $S_{\Delta,0}$ splines of order 0.

Remark and Definition 3.38. In the situation of Def. 3.37, let $k \in \{0, \ldots, N - 1\}$. If $s \in S_{\Delta,l}$, $l \in \mathbb{N}_0$, then, as $s |_{x_k,x_{k+1}}$ coincides with a polynomial function of degree at most $l$, on $[x_k,x_{k+1}]$ the derivative $s^{(l)}$ exists and is constant, i.e. there exists $\alpha_k \in \mathbb{R}$ such that $s^{(l)} |_{x_k,x_{k+1}} \equiv \alpha_k$. We extend $s^{(l)}$ to all of $[a,b]$ by defining

$$s^{(l)} := \sum_{k=0}^{N-1} \alpha_k \chi_k \in S_{\Delta,0}. \tag{3.42}$$

Definition 3.39. In the situation of Def. 3.37, if $s \in S_{\Delta,0}$, then, for each $l \in \mathbb{N}_0$, define the $l$th antiderivative $I_l(s) : [a,b] \to \mathbb{R}$ of $s$ recursively by letting

$$I_0(s) := s, \quad \forall l \in \mathbb{N}_0 \quad I_{l+1}(s)(x) := \int_a^x I_l(s)(t) \, dt.$$ 

Theorem 3.40. Let $a, b \in \mathbb{R}$, $a < b$, $N \in \mathbb{N}$, $l \in \mathbb{N}_0$, with $\Delta := (x_0, \ldots, x_N)$ being a knot vector for $[a,b]$. Then the following statements hold true:

(a) $\text{Pol}_l(\mathbb{R}) \subseteq S_{\Delta,l}$ (here, and in the following, we are sometimes a bit lax regarding the notation, implicitly identifying polynomials on $\mathbb{R}$ with their respective restriction on $[a,b]$).

(b) For each $j \in \{0, \ldots, l\}$, we have

$$s \in S_{\Delta,l} \Rightarrow s^{(j)} \in S_{\Delta,l-j}.$$ 

(c) If $s \in S_{\Delta,0}$ and $S := I_l(s)$ is the $l$th antiderivative of $s$ as defined in Def. 3.39, then $S \in S_{\Delta,l}$.

(d) $S_{\Delta,l}$ is an $(N+l)$-dimensional vector space over $\mathbb{R}$ and, if, for each $k \in \{0, \ldots, N-1\}$, $X_k := I_l(\chi_k)$ is the $l$th antiderivative of $\chi_k$, then

$$B := \{ X_k : k \in \{0, \ldots, N-1\} \} \cup \{ x^j : j \in \{0, \ldots, l-1\} \}$$

forms a basis of $S_{\Delta,l}$.

Proof. (a) is clear from the definition of $S_{\Delta,l}$.

(b): For $j \in \{0, \ldots, l-1\}$, (b) holds since $s \in C^{l-1}[a,b]$ implies $s^{(j)} \in C^{l-j-1}[a,b]$ and $p_k \in \text{Pol}_l(\mathbb{R})$ implies $p_k^{(j)} \in \text{Pol}_{l-j}(\mathbb{R})$. For $j = l$, (b) holds due to Rem. and Def. 3.38.

(c): We obtain the continuity of $I_l(s)$ from Th. D.6 of the appendix. We then apply the fundamental theorem of calculus (cf. [Phi16a, Th. 10.20(a)]) $(l-1)$ times to $I_l(s)$ to
obtain $S \in \mathcal{C}^{l-1}[a, b]$. For each $k \in \{0, \ldots, N - 1\}$, we apply the fundamental theorem of calculus $l$ times on $[x_k, x_{k+1}]$ to obtain $S|_{[x_k, x_{k+1}]} \in \mathcal{P} \mathcal{O} \mathcal{L}_l(\mathbb{R})$.

(d): $S_{\Delta,l}$ being a vector space over $\mathbb{R}$ is a consequence of both $\mathcal{C}^{l-1}[a, b]$ and $\mathcal{P} \mathcal{O} \mathcal{L}_l(\mathbb{R})$ being vector spaces over $\mathbb{R}$. Since $B \subseteq S_{\Delta,l}$ by (a),(c) and, clearly, $\#B = N + l$, it merely remains to show that $B$ is, indeed, a basis of $S_{\Delta,l}$. Let $s \in S_{\Delta,l}$. Since $s^{(l)} \in S_{\Delta,0}$, there exist $\alpha_0, \ldots, \alpha_{N-1} \in \mathbb{R}$ such that $s^{(l)}(t) = \sum_{k=0}^{N-1} \alpha_k \chi_k$. If $\tilde{s} := \sum_{k=0}^{N-1} \alpha_k x_k$, then $\tilde{s}(t) = s(t)$. We need to show $(s - \tilde{s}) \in \mathcal{P} \mathcal{O} \mathcal{L}_{l-1}(\mathbb{R})$. To this end, let $k \in \{1, \ldots, N - 1\}$. Then the fundamental theorem of calculus yields $p := (s - \tilde{s})|_{[x_{k-1}, x_k]} \in \mathcal{P} \mathcal{O} \mathcal{L}_{l-1}(\mathbb{R})$ and $q := (s - \tilde{s})|_{[x_k, x_{k+1}]} \in \mathcal{P} \mathcal{O} \mathcal{L}_{l-1}(\mathbb{R})$. However, as $s - \tilde{s}$ is $\mathcal{C}^{l-1}$, we also know

$$
\forall_{j \in \{0, \ldots, l-1\}} p^{(j)}(x_k) = q^{(j)}(x_k),
$$

implying $p = q$ due to Th. 3.10(b) about Hermite interpolation. As $k \in \{1, \ldots, N - 1\}$ was arbitrary, this shows $(s - \tilde{s}) \in \mathcal{P} \mathcal{O} \mathcal{L}_{l-1}(\mathbb{R})$ and span $B = S_{\Delta,l}$. It remains to show that $B$ is linearly independent. To this end assume

$$
\exists_{p \in \mathcal{P} \mathcal{O} \mathcal{L}_{l-1}(\mathbb{R})} \exists_{\alpha_0, \ldots, \alpha_{N-1} \in \mathbb{R}} \forall_{x \in [a, b]} p(x) + \sum_{k=0}^{N-1} \alpha_k x_k(x) = 0.
$$

Taking the derivative $l$ times then yields

$$
\forall_{x \in [a, b]} \sum_{k=0}^{N-1} \alpha_k \chi_k(x) = 0.
$$

As the $\chi_k$ are clearly linearly independent, we obtain $\alpha_0 = \cdots = \alpha_{N-1} = 0$, implying $p = 0$ as well. Thus, $B$ is linearly independent and, hence, a basis of $S_{\Delta,l}$. \hspace{1em} \Box

### 3.3.2 Linear Splines

Given a node vector as in (3.39), and values $y_0, \ldots, y_N \in \mathbb{R}$, a corresponding linear spline is a continuous, piecewise linear function $s$ satisfying $s(x_k) = y_k$. In this case, existence, uniqueness, and an error estimate are still rather simple to obtain:

**Theorem 3.41.** Let $a, b \in \mathbb{R}$, $a < b$. Then, given $N \in \mathbb{N}$, a knot vector $\Delta := (x_0, \ldots, x_N) \in \mathbb{R}^{N+1}$ for $[a, b]$, and values $(y_0, \ldots, y_N) \in \mathbb{R}^{N+1}$, there exists a unique interpolating linear spline $s \in S_{\Delta,1}$ satisfying

$$
s(x_k) = y_k \text{ for each } k \in \{0, \ldots, N\}. \tag{3.43}
$$

Moreover, $s$ is explicitly given by

$$
s(x) = y_k + \frac{y_k - y_l}{x_{k+1} - x_k} (x - x_k) \text{ for each } x \in [x_k, x_{k+1}]. \tag{3.44}
$$

If $f \in \mathcal{C}^2[a, b]$ and $y_k = f(x_k)$, then the following error estimate holds:

$$
\|f - s\|_\infty \leq \frac{1}{8} \|f''\|_\infty h^2, \tag{3.45}
$$
where
\[ h := h(\Delta) := \max \{ x_{k+1} - x_k : k \in \{0, \ldots, N - 1\} \} \]
is the mesh size of the partition \( \Delta \).

Proof. Since \( s \) defined by (3.44) clearly satisfies (3.43), is continuous on \([a,b]\), and, for each \( k \in \{0, \ldots, N\} \), the definition of (3.44) extends to a unique polynomial in \( \text{Pol}_1(\mathbb{R}) \), we already have existence. Uniqueness is equally obvious, since, for each \( k \), the prescribed values in \( x_k \) and \( x_{k+1} \) uniquely determine a polynomial in \( \text{Pol}_1(\mathbb{R}) \) and, thus, \( s \). For the error estimate, consider \( x \in [x_k, x_{k+1}] \), \( k \in \{0, \ldots, N - 1\} \). Note that, due to the inequality \( \alpha \beta \leq \left( \frac{\alpha + \beta}{2} \right)^2 \) valid for all \( \alpha, \beta \in \mathbb{R} \), one has
\[ (x - x_k)(x_{k+1} - x) \leq \left( \frac{(x - x_k) + (x_{k+1} - x)}{2} \right)^2 \leq \frac{h^2}{4}. \] (3.46)
Moreover, on \( [x_k, x_{k+1}] \), \( s \) coincides with the interpolating polynomial \( p \in \text{Pol}_1(\mathbb{R}) \) with \( p(x_k) = f(x_k) \) and \( p(x_{k+1}) = f(x_{k+1}) \). Thus, we can use (3.31) to compute
\[ |f(x) - s(x)| = |f(x) - p(x)| \leq \frac{\|f''\|_\infty}{2!} (x - x_k)(x_{k+1} - x) \leq \frac{1}{8} \|f''\|_\infty h^2, \] thereby proving (3.45).

### 3.3.3 Cubic Splines

Given a knot vector
\[ \Delta = (x_0, \ldots, x_N) \in \mathbb{R}^{N+1} \quad \text{and values} \quad (y_0, \ldots, y_N) \in \mathbb{R}^{N+1}, \] (3.47)
the task is to find an interpolating cubic spline, i.e. an element \( s \) of \( S_{\Delta,3} \) satisfying
\[ s(x_k) = y_k \quad \text{for each} \quad k \in \{0, \ldots, N\}. \] (3.48)
As we require \( s \in S_{\Delta,3} \), according to (3.40), for each \( k \in \{0, \ldots, N - 1\} \), we must find \( p_k \in \text{Pol}_3(\mathbb{R}) \) such that \( p_k |_{[x_k, x_{k+1}]} = s |_{[x_k, x_{k+1}]} \). To that end, for each \( k \in \{0, \ldots, N - 1\} \), we employ the ansatz
\[ p_k(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3 \] (3.49a)
with suitable
\[ (a_k, b_k, c_k, d_k) \in \mathbb{R}^4; \] (3.49b)
to then define \( s \) by
\[ s(x) = p_k(x) \quad \text{for each} \quad x \in [x_k, x_{k+1}]. \] (3.49c)
It is immediate that $s$ is well-defined by (3.49c) and an element of $S_{\Delta,3}$ (cf. (3.40)) satisfying (3.48) if, and only if,

\begin{align}
p_0(x_0) &= y_0, & \text{(3.50a)} \\
p_{k-1}(x_k) &= y_k = p_k(x_k) & \text{for each } k \in \{1, \ldots, N - 1\}, \quad \text{(3.50b)} \\
p_{N-1}(x_N) &= y_N, & \quad \text{(3.50c)} \\
p''_{k-1}(x_k) &= p''_k(x_k) & \text{for each } k \in \{1, \ldots, N - 1\}, \quad \text{(3.50d)} \\
p''_{N-1}(x_N) &= p''_N(x_N) & \text{for each } k \in \{1, \ldots, N - 1\}, \quad \text{(3.50e)}
\end{align}

using the convention $\{1, \ldots, N - 1\} = \emptyset$ for $N - 1 < 1$. Note that (3.50) constitutes a system of $2 + 4(N - 1) = 4N - 2$ conditions, while (3.49a) provides $4N$ variables. This already indicates that one will be able to impose two additional conditions on $s$. One typically imposes so-called boundary conditions, i.e. conditions on the values for $s$ or its derivatives at $x_0$ and $x_N$ (see (3.61) below for commonly used boundary conditions).

In order to show that one can, indeed, choose the $a_k, b_k, c_k, d_k$ such that all conditions of (3.50) are satisfied, the strategy is to transform (3.50) into an equivalent linear system. An analysis of the structure of that system will then show that it has a solution. The trick that will lead to the linear system is the introduction of

\[ s''_k := s''(x_k), \quad k \in \{0, \ldots, N\}, \quad \text{(3.51)} \]

as new variables. As we will see in the following lemma, (3.50) is satisfied if, and only if, the $s''_k$ satisfy the linear system (3.53):

**Lemma 3.42.** Given a knot vector and values according to (3.47), and introducing the abbreviations

\begin{align}
h_k &= x_{k+1} - x_k, \quad \text{for each } k \in \{0, \ldots, N - 1\}, \quad \text{(3.52a)} \\
g_k &= 6 \left( \frac{y_{k+1} - y_k}{h_k} - \frac{y_k - y_{k-1}}{h_k} \right) \quad \text{for each } k \in \{1, \ldots, N - 1\}, \quad \text{(3.52b)}
\end{align}

the $p_k \in \text{Pol}_3(\mathbb{R})$ of (3.49a) satisfy (3.50) (and that means $s$ given by (3.49c) is in $S_{\Delta,3}$ and satisfies (3.48)) if, and only if, the $N + 1$ numbers $s''_0, \ldots, s''_N$ defined by (3.51) solve the linear system consisting of the $N - 1$ equations

\begin{align}
h_k s''_{k+1} + 2(h_k + h_{k+1}) s''_{k+1} + h_{k+1} s''_{k+2} &= g_{k+1}, \quad k \in \{0, \ldots, N - 2\}, \quad \text{(3.53)}
\end{align}

and the $a_k, b_k, c_k, d_k$ are given in terms of the $x_k, y_k$ and $s''_k$ by

\begin{align}
a_k &= y_k & \text{for each } k \in \{0, \ldots, N - 1\}, \quad \text{(3.54a)} \\
b_k &= \frac{y_{k+1} - y_k}{h_k} - \frac{h_k}{6} (s''_{k+1} + 2s''_k) & \text{for each } k \in \{0, \ldots, N - 1\}, \quad \text{(3.54b)} \\
c_k &= \frac{s''_k}{2} & \text{for each } k \in \{0, \ldots, N - 1\}, \quad \text{(3.54c)} \\
d_k &= \frac{s''_{k+1} - s''_k}{6h_k} & \text{for each } k \in \{0, \ldots, N - 1\}. \quad \text{(3.54d)}
\end{align}
3 \text{ INTERPOLATION}

Proof. First, for subsequent use, from (3.49a), we obtain the first two derivatives of the $p_k$ for each $k \in \{0, \ldots, N-1\}$:

\begin{align}
  p'_k(x) &= b_k + 2c_k(x - x_k) + 3d_k(x - x_k)^2, \\
  p''_k(x) &= 2c_k + 6d_k(x - x_k).
\end{align}

(3.55a) (3.55b)

As usual, for the claimed equivalence, we prove two implications. (3.50) implies (3.53) and (3.54):

Plugging $x = x_k$ into (3.49a) yields

\[ p_k(x_k) = a_k \quad \text{for each } k \in \{0, \ldots, N-1\}. \]

(3.56)

Thus, (3.54a) is a consequence of (3.50a) and (3.50b).

Plugging $x = x_k$ into (3.55b) yields

\[ s''_k = p''_k(x_k) = 2c_k \quad \text{for each } k \in \{0, \ldots, N-1\}, \]

(3.57)

i.e. (3.54c).

Plugging (3.50c), i.e. $p''_{k-1}(x_k) = p''_k(x_k)$, into (3.55b) yields

\[ 2c_{k-1} + 6d_{k-1}(x_k - x_{k-1}) = 2c_k \quad \text{for each } k \in \{1, \ldots, N-1\}. \]

(3.58a)

When using (3.54c) and solving for $d_{k-1}$, one obtains

\[ d_{k-1} = \frac{s''_k - s''_{k-1}}{6h_{k-1}} \quad \text{for each } k \in \{1, \ldots, N-1\}, \]

(3.58b)

i.e. the relation of (3.54d) for each $k \in \{0, \ldots, N-2\}$. Moreover, (3.51) and (3.55b) imply

\[ s''_N = s''(x_N) = p''_{N-1}(x_N) = 2c_{N-1} + 6d_{N-1}h_{N-1}, \]

(3.58c)

which, after solving for $d_{N-1}$, shows that the relation of (3.54d) also holds for $k = N - 1$.

Plugging the first part of (3.50b) and (3.50c), i.e. $p_{k-1}(x_k) = y_k$ into (3.49a) yields

\[ a_{k-1} + b_{k-1}h_{k-1} + c_{k-1}h_{k-1}^2 + d_{k-1}h_{k-1}^3 = y_k \quad \text{for each } k \in \{1, \ldots, N\}. \]

(3.59a)

When using (3.54a), (3.54c), (3.54d), and solving for $b_{k-1}$, one obtains

\[ b_{k-1} = \frac{y_k - y_{k-1}}{h_{k-1}} - \frac{s''_{k-1}h_{k-1}}{2} - \frac{s''_k - s''_{k-1}}{6h_{k-1}}h_{k-1}^2 = \frac{y_k - y_{k-1}}{h_{k-1}} - \frac{h_{k-1}}{6}(s''_k + 2s''_{k-1}) \]

(3.59b)

for each $k \in \{1, \ldots, N\}$, i.e. (3.54b).

Plugging (3.50d), i.e. $p'_{k-1}(x_k) = p'_k(x_k)$, into (3.55a) yields

\[ b_{k-1} + 2c_{k-1}h_{k-1} + 3d_{k-1}h_{k-1}^2 = b_k \quad \text{for each } k \in \{1, \ldots, N-1\}. \]

(3.60a)
Finally, applying (3.54b), (3.54c), and (3.54d), one obtains

$$\frac{y_k - y_{k-1}}{h_{k-1}} - \frac{h_{k-1}}{6} (s''_k + 2s''_{k-1}) + s''_{k-1} h_{k-1} + \frac{s''_k - s''_{k-1}}{2} h_{k-1} = \frac{y_{k+1} - y_k}{h_k} - \frac{h_k}{6} (s''_{k+1} + 2s''_k),$$

(3.60b)

or, rearranged,

$$h_{k-1} s''_{k-1} + 2(h_{k-1} + h_k)s''_k + h_k s''_{k+1} = g_k, \quad k \in \{1, \ldots, N-1\},$$

(3.60c)

which is (3.53).

(3.53) and (3.54) imply (3.50):

First, note that plugging $x = x_k$ into (3.55b) again yields $s''_k = p''_k(x_k)$. Using (3.54a) in (3.49a) yields (3.50a) and the $p_k(x_k) = y_k$ part of (3.50b). Since (3.54b) is equivalent to (3.59b), and (3.59b) (when using (3.54a), (3.54c), and (3.54d)) implies (3.59a), which, in turn, is equivalent to $p_{k-1}(x_k) = y_k$ for each $k \in \{1, \ldots, N\}$, we see that (3.54) implies (3.50c) and the $p_{k-1}(x_k) = y_k$ part of (3.50b). Since (3.53) is equivalent to (3.60c) and (3.60b), and (3.60b) (when using (3.54a) – (3.54d)) implies (3.60a), which, in turn, is equivalent to $p_k''(x_k) = p''_k(x_k)$ for each $k \in \{1, \ldots, N-1\}$, we see that (3.53) and (3.54) implies (3.50d). Finally, (3.54d) is equivalent to (3.58b), which (when using (3.54c)) implies (3.58a). Since (3.58a) is equivalent to $p_{k-1}''(x_k) = p''_k(x_k)$ for each $k \in \{1, \ldots, N-1\}$, (3.54) also implies (3.50d), concluding the proof of the lemma.

As already mentioned after formulating (3.50), (3.50) yields two conditions less than there are variables. Not surprisingly, the same is true in the linear system (3.53), where there are $N-1$ equations for $N+1$ variables. As also mentioned before, this allows to impose two additional conditions. Here are some of the most commonly used additional conditions:

**Natural boundary conditions:**

$$s''(x_0) = s''(x_N) = 0.$$  \hspace{1cm} (3.61a)

**Periodic boundary conditions:**

$$s'(x_0) = s'(x_N) \quad \text{and} \quad s''(x_0) = s''(x_N).$$  \hspace{1cm} (3.61b)

**Dirichlet boundary conditions** for the first derivative:

$$s'(x_0) = y'_0 \in \mathbb{R} \quad \text{and} \quad s'(x_N) = y'_N \in \mathbb{R}. \hspace{1cm} (3.61c)$$

Due to time constraints, we will only investigate the most simple of these cases, namely natural boundary conditions, in the following.

Since (3.61a) fixes the values of $s''_0$ and $s''_N$ as 0, (3.53) now yields a linear system consisting of $N-1$ equations for the $N-1$ variables $s''_1$ through $s''_{N-1}$. We can rewrite this system in matrix form

$$A s'' = g$$  \hspace{1cm} (3.62a)
by introducing the matrix
\[
A := \begin{pmatrix}
2(h_0 + h_1) & h_1 & 2(h_1 + h_2) & h_2 & \cdots & 2(h_2 + h_3) & h_3 \\
-h_1 & 2(h_1 + h_2) & h_2 & 2(h_2 + h_3) & h_3 & \cdots & h_N \\
& \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\
& & h_{N-3} & 2(h_{N-3} + h_{N-2}) & h_{N-2} & 2(h_{N-2} + h_{N-1})
\end{pmatrix}
\] (3.62b)

and the vectors
\[
s'' := \begin{pmatrix} s''_1 \\ \vdots \\ s''_{N-1} \end{pmatrix}, \quad g := \begin{pmatrix} g_1 \\ \vdots \\ g_{N-1} \end{pmatrix}.
\] (3.62c)

The matrix \(A\) from (3.62b) belongs to an especially benign class of matrices, so-called strictly diagonally dominant matrices:

**Definition 3.43.** Let \(A \in \mathcal{M}(N, \mathbb{C})\), \(N \in \mathbb{N}\). Then \(A\) is called **strictly diagonally dominant** if, and only if,
\[
\sum_{j=1, j \neq k}^{N} |a_{kj}| < |a_{kk}| \quad \text{for each } k \in \{1, \ldots, N\}.
\] (3.63)

**Lemma 3.44.** Let \(A \in \mathcal{M}(N, \mathbb{K})\), \(N \in \mathbb{N}\). If \(A\) is strictly diagonally dominant, then, for each \(x \in \mathbb{K}^N\),
\[
\|x\|_{\infty} \leq \max \left\{ \frac{1}{|a_{kk}| - \sum_{j=1, j \neq k}^{N} |a_{kj}|} : k \in \{1, \ldots, N\} \right\} \|Ax\|_{\infty}.
\] (3.64)

In particular, \(A\) is invertible with
\[
\|A^{-1}\|_{\infty} \leq \max \left\{ \frac{1}{|a_{kk}| - \sum_{j=1, j \neq k}^{N} |a_{kj}|} : k \in \{1, \ldots, N\} \right\},
\] (3.65)

where \(\|A^{-1}\|_{\infty}\) denotes the row sum norm according to Ex. 2.18(a).

**Proof.** Exercise. ■

**Theorem 3.45.** Given a knot vector and values according to (3.47), there is a unique cubic spline \(s \in S_{\Delta,3}\) that satisfies (3.48) and the natural boundary conditions (3.61a). Moreover, \(s\) can be computed by first solving the linear system (3.62) for the \(s''_k\) (the \(h_k\) and the \(g_k\) are given by (3.52)), then determining the \(a_k, b_k, c_k,\) and \(d_k\) according to (3.54), and, finally, using (3.49) to get \(s\).
Proof. Lemma 3.42 shows that the existence and uniqueness of $s$ is equivalent to (3.62a) having a unique solution. That (3.62a) does, indeed, have a unique solution follows from Lem. 3.44, as $A$ as defined in (3.62b) is clearly strictly diagonally dominant. It is also part of the statement of Lem. 3.42 that $s$ can be computed in the described way. ■

Theorem 3.46. Let $a, b \in \mathbb{R}$, $a < b$, and $f \in C^4[a, b]$. Moreover, given $N \in \mathbb{N}$ and a knot vector $\Delta := (x_0, \ldots, x_N) \in \mathbb{R}^{N+1}$ for $[a, b]$, define

$$h_k := x_{k+1} - x_k \quad \text{for each } k \in \{0, \ldots, N-1\},$$
$$h_{\min} := \min \{h_k : k \in \{0, \ldots, N-1\}\},$$
$$h_{\max} := \max \{h_k : k \in \{0, \ldots, N-1\}\}.$$  \hspace{1cm} (3.66a)

Let $s \in S_{\Delta, 3}$ be any cubic spline function satisfying $s(x_k) = f(x_k)$ for each $k \in \{0, \ldots, N\}$. If there exists $C > 0$ such that

$$\max \left\{|s''(x_k) - f''(x_k)| : k \in \{0, \ldots, N\}\right\} \leq C \|f^{(4)}\|_{\infty} h_{\max}^2,$$  \hspace{1cm} (3.67)

then the following error estimates hold:

$$\|f - s\|_{\infty} \leq c \|f^{(4)}\|_{\infty} h_{\max}^4,$$  \hspace{1cm} (3.68a)
$$\|f' - s'\|_{\infty} \leq 2c \|f^{(4)}\|_{\infty} h_{\max}^3,$$  \hspace{1cm} (3.68b)
$$\|f'' - s''\|_{\infty} \leq 2c \|f^{(4)}\|_{\infty} h_{\max}^2,$$  \hspace{1cm} (3.68c)
$$\|f''' - s'''\|_{\infty} \leq 2c \|f^{(4)}\|_{\infty} h_{\max}^1,$$  \hspace{1cm} (3.68d)

where

$$c := \frac{h_{\max}}{h_{\min}} \left(C + \frac{1}{4}\right).$$  \hspace{1cm} (3.69)

At the $x_k$, the third derivative $s'''$ does not need to exist. In that case, (3.68d) is to be interpreted to mean that the estimate holds for all values of one-sided third derivatives of $s$ (which must exist, since $s$ agrees with a polynomial on each $[x_k, x_{k+1}]$).

Proof. Exercise. ■

Theorem 3.47. Once again, consider the situation of Th. 3.46, but, instead of (3.67), assume that the interpolating cubic spline $s$ satisfies natural boundary conditions $s''(a) = s''(b) = 0$. If $f \in C^4[a, b]$ satisfies $f''(a) = f''(b) = 0$, then it follows that $s$ satisfies (3.67) with $C = 3/4$, and, in consequence, the error estimates (3.68) with $c = h_{\max}/h_{\min}$.

Proof. We merely need to show that (3.67) holds with $C = 3/4$. Since $s$ is the interpolating cubic spline satisfying natural boundary conditions, $s''$ satisfies the linear system (3.62a). Deviding, for each $k \in \{1, \ldots, N-1\}$, the $(k-1)$th equation of that system by $3(h_{k-1} + h_k)$ yields the form

$$Bs'' = \hat{g},$$  \hspace{1cm} (3.70a)
with

$$B := \begin{pmatrix}
\frac{2}{3} & \frac{h_1}{3(h_0 + h_1)} & \frac{h_2}{3(h_1 + h_2)} & \ldots & \frac{h_{n-3}}{3(h_{n-3} + h_{n-2})} & \frac{h_{n-2}}{3(h_{n-2} + h_{n-1})} \\
\frac{h_1}{3(h_1 + h_2)} & \frac{2}{3} & \frac{h_2}{3(h_2 + h_3)} & \ldots & \frac{h_{n-3}}{3(h_{n-3} + h_{n-2})} & \frac{h_{n-2}}{3(h_{n-2} + h_{n-1})} \\
\frac{h_2}{3(h_2 + h_3)} & \frac{h_2}{3(h_2 + h_3)} & \frac{2}{3} & \ldots & \frac{h_{n-3}}{3(h_{n-3} + h_{n-2})} & \frac{h_{n-2}}{3(h_{n-2} + h_{n-1})} \\
\vdots & \ddots & \ddots & \ddots & \frac{2}{3} & \frac{h_{n-2}}{3(h_{n-2} + h_{n-1})} \\
\frac{h_{n-3}}{3(h_{n-3} + h_{n-2})} & \frac{h_{n-3}}{3(h_{n-3} + h_{n-2})} & \frac{h_{n-3}}{3(h_{n-3} + h_{n-2})} & \ldots & \frac{2}{3} & \frac{h_{n-2}}{3(h_{n-2} + h_{n-1})} \\
\frac{h_{n-2}}{3(h_{n-2} + h_{n-1})} & \frac{h_{n-2}}{3(h_{n-2} + h_{n-1})} & \frac{h_{n-2}}{3(h_{n-2} + h_{n-1})} & \ldots & \frac{h_{n-2}}{3(h_{n-2} + h_{n-1})} & \frac{2}{3}
\end{pmatrix}$$

and

$$\hat{g} := \begin{pmatrix}
\hat{g}_1 \\
\vdots \\
\hat{g}_{N-1}
\end{pmatrix} := \begin{pmatrix}
\frac{g_1}{3(h_0 + h_1)} \\
\vdots \\
\frac{g_{N-1}}{3(h_{N-2} + h_{N-1})}
\end{pmatrix}.$$  

(3.70b)

Taylor’s theorem applied to $f''$ provides, for each $k \in \{1, \ldots, N - 1\}$, $\alpha_k \in ]x_{k-1}, x_k[$ and $\beta_k \in ]x_k, x_{k+1}[$ such that the following equations (3.71) hold, where (3.71a) has been multiplied by $\frac{h_{k-1}}{3(h_{k-1} + h_k)}$ and (3.71b) has been multiplied by $\frac{h_k}{3(h_{k-1} + h_k)}$:

$$\frac{h_{k-1}}{3(h_{k-1} + h_k)} f''(x_{k-1}) + \frac{2}{3} f''(x_k) + \frac{h_k}{3(h_{k-1} + h_k)} f''(x_{k+1}) = f''(x_k) + R_k + \delta_k, \quad (3.72a)$$

where

$$R_k := \frac{1}{3} (h_k - h_{k-1}) f'''(x_k), \quad \delta_k := \frac{1}{6(h_{k-1} + h_k)} \left(h_{k-1}^3 f^{(4)}(\alpha_k) + h_k^3 f^{(4)}(\beta_k)\right). \quad (3.72b)$$

Adding (3.71a) and (3.71b) results in

$$\frac{h_{k-1}}{3(h_{k-1} + h_k)} f''(x_{k-1}) + \frac{h_k}{3(h_{k-1} + h_k)} f''(x_{k+1}) = f''(x_k) + R_k + \delta_k, \quad (3.72a)$$

Using the matrix $B$ from (3.70b), (3.72a) takes the form

$$B \begin{pmatrix}
f''(x_1) \\
\vdots \\
f''(x_{N-1})
\end{pmatrix} = \begin{pmatrix}
f''(x_1) \\
\vdots \\
f''(x_{N-1})
\end{pmatrix} + \begin{pmatrix}
R_1 \\
\vdots \\
R_{N-1}
\end{pmatrix} + \begin{pmatrix}
\delta_1 \\
\vdots \\
\delta_{N-1}
\end{pmatrix}. \quad (3.73)$$
Observing that we can write (3.76) as
\[ \frac{2 f(x_{k+1}) - f(x_k)}{h_k} = 2 f'(x_k) + h_k f''(x_k) + \frac{h_k}{3} f'''(x_k) + \frac{h_k^3}{12} f^{(4)}(\xi_k), \] (3.75a)
\[ -2 \frac{f(x_k) - f(x_{k-1})}{h_k} = -2 f'(x_k) + h_{k-1} f''(x_k) - \frac{h_{k-1}}{3} f'''(x_k) + \frac{h_{k-1}^3}{12} f^{(4)}(\eta_k). \] (3.75b)

Multiplying (3.74a) and (3.74b) by 2/h_k and 2/h_{k-1}, respectively, leads to
\[ 2 \frac{f(x_{k+1}) - f(x_k)}{h_k} = 2 f'(x_k) + h_k f''(x_k) + \frac{h_k}{3} f'''(x_k) + \frac{h_k^3}{12} f^{(4)}(\xi_k), \] (3.75a)
\[ -2 \frac{f(x_k) - f(x_{k-1})}{h_k} = -2 f'(x_k) + h_{k-1} f''(x_k) - \frac{h_{k-1}}{3} f'''(x_k) + \frac{h_{k-1}^3}{12} f^{(4)}(\eta_k). \] (3.75b)

Adding (3.75a) and (3.75b) followed by a division by h_{k-1} + h_k yields
\[ 2 \frac{f(x_{k+1}) - f(x_k)}{h_k(h_{k-1} + h_k)} - 2 \frac{f(x_k) - f(x_{k-1})}{h_k(h_{k-1} + h_k)} = f''(x_k) + R_k + \hat{\delta}_k, \] (3.76)
where R_k is as in (3.72b) and
\[ \hat{\delta}_k := \frac{1}{12(h_{k-1} + h_k)} \left( h_{k-1}^3 f^{(4)}(\eta_k) + h_k^3 f^{(4)}(\xi_k) \right). \]

Observing that
\[ \hat{g}_k = \frac{g_k}{3(h_{k-1} + h_k)} \overset{(3.70c),(3.52b),y_k = f(x_k)}{=} \frac{2 f(x_{k+1}) - f(x_k)}{h_k(h_{k-1} + h_k)} - 2 \frac{f(x_k) - f(x_{k-1})}{h_k(h_{k-1} + h_k)}, \]
we can write (3.76) as
\[ \begin{pmatrix} f''(x_1) \\ \vdots \\ f''(x_{N-1}) \end{pmatrix} = \begin{pmatrix} \hat{g}_1 \\ \vdots \\ \hat{g}_{N-1} \end{pmatrix} - \begin{pmatrix} R_1 \\ \vdots \\ R_{N-1} \end{pmatrix} - \begin{pmatrix} \hat{\delta}_1 \\ \vdots \\ \hat{\delta}_{N-1} \end{pmatrix}. \] (3.77)

Using (3.77) in the right-hand side of (3.73) and then subtracting (3.70a) from the result yields
\[ B \begin{pmatrix} f''(x_1) - s''(x_1) \\ \vdots \\ f''(x_{N-1}) - s''(x_{N-1}) \end{pmatrix} = \begin{pmatrix} \delta_1 - \hat{\delta}_1 \\ \vdots \\ \delta_{N-1} - \hat{\delta}_{N-1} \end{pmatrix}. \]

From
\[ \frac{2}{3} - \frac{h_{k-1}}{3(h_{k-1} + h_k)} = \frac{1}{3} \quad \text{for } k \in \{1, \ldots, N - 1\}, \]
we conclude that B is strictly diagonally dominant, such that Lem. 3.44 allows us to estimate
\[ \max \{ |f''(x_k) - s''(x_k)| : k \in \{0, \ldots, N\} \} \leq 3 \max \{ |\delta_k| + |\hat{\delta}_k| : k \in \{1, \ldots, N - 1\} \} \]
\[ \overset{(s)}{\leq} \frac{3}{4} h_{\text{max}}^2 \|f^{(4)}\|_{\infty}, \] (3.78)
where
\[ |\delta_k| + |\hat{\delta}_k| \leq \left( \frac{1}{6} + \frac{1}{12} \right) \frac{h^3_{k-1} + h^3_k}{h_{k-1} + h_k} \|f^{(4)}\|_\infty \leq \frac{1}{4} h_{\text{max}}^2 \|f^{(4)}\|_\infty \]
was used for the inequality at \((\ast)\). The estimate (3.78) shows that \(s\) satisfies (3.67) with \(C = 3/4\), thereby concluding the proof of the theorem.

Note that (3.68) is, in more than one way, much better than (3.45): From (3.68) we get convergence of the first three derivatives, and the convergence of \(\|f - s\|_\infty\) is also much faster in (3.68) (proportional to \(h^4\) rather than to \(h^2\) in (3.45)). The disadvantage of (3.68), however, is that we needed to assume the existence of a continuous fourth derivative of \(f\).

**Remark 3.48.** A piecewise interpolation by cubic polynomials different from spline interpolation is given by carrying out a Hermite interpolation on each interval \([x_k, x_{k+1}]\) such that the resulting function satisfies \(f(x_k) = s(x_k)\) as well as
\[ f'(x_k) = s'(x_k). \tag{3.79} \]
The result will usually not be a spline, since \(s\) will usually not have second derivatives in the \(x_k\). On the other hand the corresponding cubic spline will usually not satisfy (3.79). Thus, one will prefer one or the other form of interpolation, depending on either (3.79) (i.e. the agreement of the first derivatives of the given and interpolating function) or \(s \in C^2\) being more important in the case at hand.

We conclude the section by depicting different interpolations of the functions \(f(x) = 1/(x^2 + 1)\) and \(f(x) = \cos x\) on the interval \([-6, 6]\) in Figures 3.1 and 3.2, respectively. While \(f\) is depicted as a solid blue curve, the interpolating polynomial of degree 6 is shown as a dotted black curve, the interpolating linear spline as a solid red curve, and the interpolating cubic spline is shown as a dashed green curve. One clearly observes the large values and strong oscillations of the interpolating polynomial toward the boundaries of the interval.

\section{4 Numerical Integration}

\subsection{4.1 Introduction}

The goal is the computation of (an approximation of) the value of definite integrals
\[ \int_a^b f(x) \, dx. \tag{4.1} \]
Even if we know from the fundamental theorem of calculus that \(f\) has an antiderivative, even for simple \(f\), the antiderivative can often not be expressed in terms of so-called
elementary functions (such as polynomials, exponential functions, and trigonometric functions). An example is the important function

\[ \Phi(y) := \int_0^y \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx, \quad (4.2) \]

which can not be expressed by elementary functions.

On the other hand, as we will see, there are effective and efficient methods to compute accurate approximations of such integrals. As a general rule, one can say that numerical integration is easy, whereas symbolic integration is hard (for differentiation, one has the reverse situation – symbolic differentiation is easy, while numeric differentiation is hard).

A first approximative method for the evaluation of integrals is given by the very definition of the Riemann integral. We recall it in the form of the following Th. 4.3. First, let us formally introduce some notation that, in part, we already used when studying spline interpolation:

**Notation 4.1.** Given a real interval \([a, b]\), \(a < b\), \((x_0, \ldots, x_n) \in \mathbb{R}^{n+1}\), \(n \in \mathbb{N}\), is called a *partition* of \([a, b]\) if, and only if, \(a = x_0 < x_1 < \cdots < x_n = b\) (what was called a knot vector in the context of spline interpolation). A *tagged partition* of \([a, b]\) is a partition together with a vector \((t_1, \ldots, t_n) \in \mathbb{R}^n\) such that \(t_i \in [x_{i-1}, x_i]\) for each \(i \in \{1, \ldots, n\}\).
Given a partition $\Delta$ (with or without tags) of $[a, b]$ as above, the number

$$h(\Delta) := \max \{ h_i : i \in \{1, \ldots, n\} \}, \quad h_i := x_i - x_{i-1},$$

is called the mesh size of $\Delta$. Moreover, if $f : [a, b] \to \mathbb{R}$, then

$$\sum_{i=1}^{n} f(t_i)h_i$$

is called the Riemann sum of $f$ associated with a tagged partition of $[a, b]$.

**Notation 4.2.** Given a real interval $[a, b]$, $a < b$, we denote the set of all Riemann integrable functions $f : [a, b] \to \mathbb{R}$ by $R[a, b]$.

**Theorem 4.3.** Given a real interval $[a, b]$, $a < b$, a function $f : [a, b] \to \mathbb{R}$ is in $R[a, b]$ if, and only if, for each sequence of tagged partitions

$$\Delta_k = ((x_0^k, \ldots, x_n^k), (t_1^k, \ldots, t_{n_k}^k)), \quad k \in \mathbb{N},$$

of $[a, b]$ such that $\lim_{k \to \infty} h(\Delta_k) = 0$, the limit of associated Riemann sums

$$I(f) := \int_a^b f(x) \, dx := \lim_{k \to \infty} \sum_{i=1}^{n_k} f(t_i^k)h_i^k$$

is defined.
4 NUMERICAL INTEGRATION

exists. The limit is then unique and called the Riemann integral of \( f \). Recall that every continuous and every piecewise continuous function on \([a, b]\) is in \( R[a, b] \).

**Proof.** See, e.g., [Phi16a, Th. 10.10(d)]. ■

Thus, for each Riemann integrable function, we can use (4.6) to compute a sequence of approximations that converges to the correct value of the Riemann integral.

**Example 4.4.** To employ (4.4) to compute approximations to \( \Phi(1) \), where \( \Phi \) is the function defined in (4.2), one can use the tagged partitions \( \Delta_n \) of \([0, 1]\) given by \( x_i = t_i = i/n \) for \( i \in \{0, \ldots, n\} \), \( n \in \mathbb{N} \). Noticing \( h_i = 1/n \), one obtains

\[
I_n := \sum_{i=1}^{n} \frac{1}{n} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{i}{n} \right)^2 \right).
\]

From (4.6), we know that \( \Phi(1) = \lim_{n \to \infty} I_n \), since the integrand is continuous. For example, one can compute the following values:

<table>
<thead>
<tr>
<th>( n )</th>
<th>( I_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2426</td>
</tr>
<tr>
<td>2</td>
<td>0.2978</td>
</tr>
<tr>
<td>10</td>
<td>0.3342</td>
</tr>
<tr>
<td>100</td>
<td>0.3415</td>
</tr>
<tr>
<td>1000</td>
<td>0.3422</td>
</tr>
<tr>
<td>5000</td>
<td>0.3422</td>
</tr>
</tbody>
</table>

A problem is that, when using (4.6), a priori, we have no control on the approximation error. Thus, given a required accuracy, we do not know what mesh size we need to choose. And even though the first four digits in the table of Example 4.4 seem to have stabilized, without further investigation, we can not be certain that that is, indeed, the case. The main goal in the following is to improve on this situation by providing approximation formulas for Riemann integrals together with effective error estimates.

We will slightly generalize the problem class by considering so-called weighted integrals:

**Definition 4.5.** Given a real interval \([a, b] \), \( a < b \), each Riemann integrable function \( \rho : [a, b] \to \mathbb{R}_0^+ \), such that

\[
\int_a^b \rho(x) \, dx > 0
\]

(in particular, \( \rho \not\equiv 0 \), \( \rho \) bounded, \( 0 < \int_a^b \rho(x) \, dx < \infty \)), is called a weight function. Given a weight function \( \rho \), the weighted integral of \( f \in R[a, b] \) is

\[
I_\rho(f) := \int_a^b f(x)\rho(x) \, dx,
\]

(4.8)
that means the usual integral $I(f)$ corresponds to $\rho \equiv 1$ (note that $f\rho \in \mathbb{R}[a,b]$, cf. [Phi16a, Th. 10.18(d)]).

In the sequel, we will first study methods designed for the numerical solution of non-weighted integrals (i.e. $\rho \equiv 1$) in Sections 4.2, 4.3, and 4.5, followed by an investigation of Gaussian quadrature in Sec. 4.6, which is suitable for the numerical solution of both weighted and nonweighted integrals.

**Definition 4.6.** Given a real interval $[a,b]$, $a < b$, in a generalization of the Riemann sums (4.4), a map

$$I_n : \mathbb{R}[a,b] \rightarrow \mathbb{R}, \quad I_n(f) = (b - a) \sum_{i=0}^{n} \sigma_i f(x_i), \quad (4.9)$$

is called a quadrature formula or a quadrature rule with distinct points $x_0, \ldots, x_n \in [a,b]$ and weights $\sigma_0, \ldots, \sigma_n \in \mathbb{R}$, $n \in \mathbb{N}_0$.

**Remark 4.7.** Note that every quadrature rule constitutes a linear functional on $\mathbb{R}[a,b]$.

A decent quadrature rule should give the exact integral for polynomials of low degree. This gives rise to the next definition:

**Definition 4.8.** Given a real interval $[a,b]$, $a < b$, and a weight function $\rho : [a,b] \rightarrow \mathbb{R}_+^0$, a quadrature rule $I_n$ is defined to have the degree of accuracy $r \in \mathbb{N}_0$ if, and only if,

$$I_n(x^m) = \int_a^b x^m \rho(x) \, dx \quad \text{for each } m \in \{0, \ldots, r\}, \quad (4.10a)$$

but,

$$I_n(x^{r+1}) \neq \int_a^b x^{r+1} \rho(x) \, dx. \quad (4.10b)$$

**Remark 4.9.** (a) Due to the linearity of $I_n$, $I_n$ has degree of accuracy at least $r \in \mathbb{N}_0$ if, and only if, $I_n$ is exact for each $p \in \text{Pol}_r(\mathbb{R})$.

(b) A quadrature rule $I_n$ as given by (4.9) can never be exact for the polynomial

$$p : \mathbb{R} \rightarrow \mathbb{R}, \quad p(x) := \prod_{i=0}^{n} (x - x_i)^2, \quad (4.11)$$

since

$$I_n(p) = 0 \neq \int_a^b p(x) \rho(x) \, dx. \quad (4.12)$$

In particular, $I_n$ can have degree of accuracy at most $r = 2n + 1$. 
In view of (b), a quadrature rule $I_n$ as given by (4.9) has any degree of accuracy (i.e. at least degree of accuracy 0) if, and only if,
\[
\int_a^b \rho(x) \, dx = I_n(1) = (b - a) \sum_{i=0}^n \sigma_i,
\] (4.13)
which simplifies to
\[
\sum_{i=0}^n \sigma_i = 1 \quad \text{for } \rho \equiv 1.
\] (4.14)

### 4.2 Quadrature Rules Based on Interpolating Polynomials

In this section, we consider $\rho \equiv 1$, that means only nonweighted integrals. We can make use of interpolating polynomials to produce an abundance of useful quadrature rules.

Definition 4.10. Given a real interval $[a, b]$, $a < b$, the quadrature rule based on interpolating polynomials for the distinct points $x_0, \ldots, x_n \in [a, b]$, $n \in \mathbb{N}_0$, is defined by
\[
I_n : \mathbb{R}[a, b] \longrightarrow \mathbb{R}, \quad I_n(f) := \int_a^b p_n(x) \, dx,
\] (4.15)
where $p_n = p_n(f) \in \text{Pol}_n(\mathbb{R})$ is the interpolating polynomial corresponding to the data points $(x_0, f(x_0)), \ldots, (x_n, f(x_n))$.

Theorem 4.11. Let $[a, b]$ be a real interval, $a < b$, and let $I_n$ be the quadrature rule based on interpolating polynomials for the distinct points $x_0, \ldots, x_n \in [a, b]$, $n \in \mathbb{N}_0$, as defined in (4.15).

(a) $I_n$ is, indeed, a quadrature rule in the sense of Def. 4.6, where the weights $\sigma_i$, $i \in \{0, \ldots, n\}$, are given in terms of the Lagrange basis polynomials $L_i$ of (3.10) by
\[
\sigma_i = \frac{1}{b - a} \int_a^b L_i(x) \, dx = \frac{1}{b - a} \int_a^b \prod_{k=0}^n \frac{x - x_k}{x_i - x_k} \, dx
\]
\[
= \int_0^1 \prod_{k=0}^n \frac{t - t_k}{t_i - t_k} \, dt, \quad \text{where } t_i := \frac{x_i - a}{b - a}.
\] (4.16)

(b) $I_n$ has at least degree of accuracy $n$, i.e. it is exact for each $q \in \text{Pol}_n(\mathbb{R})$.

(c) For each $f \in C^{n+1}[a, b]$, one has the error estimate
\[
\left| \int_a^b f(x) \, dx - I_n(f) \right| \leq \frac{\|f^{(n+1)}\|\infty}{(n + 1)!} \int_a^b |\omega_{n+1}(x)| \, dx,
\] (4.17)
where $\omega_{n+1}(x) = \prod_{i=0}^n (x - x_i)$ is the Newton basis polynomial.
4 NUMERICAL INTEGRATION

Proof. (a): If \( p_n \) is as in (4.15), then, according to (3.10), for each \( x \in \mathbb{R} \),

\[
p_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x) = \sum_{i=0}^{n} f(x_i) \prod_{k=0}^{n} \frac{x-x_k}{x_i-x_k}.
\]

Comparing with (4.9) then yields (4.16), where the equality labeled (*) is due to

\[
\frac{x-x_k}{x_i-x_k} = \frac{x-a}{b-a} - \frac{x-b-a}{b-a}
\]

and the change of variables \( x \mapsto t := \frac{x-a}{b-a} \).

(b): For \( q \in \text{Pol}_n(\mathbb{R}) \), we have \( p_n(q) = q \) such that \( I_n(q) = \int_a^b q(x) \, dx \) is clear from (4.15).

(c) follows by using (4.15) in combination with (3.31).

Remark 4.12. If \( f \in C^\infty[a,b] \) and there exist \( K \in \mathbb{R}_0^+ \) and \( R < (b-a)^{-1} \) such that (3.32) holds, i.e.

\[
\|f^{(n)}\|_\infty \leq Kn!R^n \quad \text{for each } n \in \mathbb{N},
\]

then, for each sequence of distinct numbers \((x_0, x_1, \ldots)\) in \([a,b]\), the corresponding quadrature rules based on interpolating polynomials converge, i.e.

\[
\lim_{n \to \infty} \left| \int_a^b f(x) \, dx - I_n(f) \right| = 0.
\]

Indeed, \( \left| \int_a^b f(x) \, dx - I_n(f) \right| \leq \int_a^b \|f - p_n\|_\infty \, dx = (b-a) \|f - p_n\|_\infty \) and \( \lim_{n \to \infty} \|f - p_n\|_\infty = 0 \) according to (3.34).

In certain situations, (4.17) can be improved by determining the sign of the error term (see Prop. 4.14 below). This information will come in handy when proving subsequent error estimates for special cases.

Lemma 4.13. Given a real interval \([a,b]\), \( a<b \), an integrable function \( g : [a,b] \to \mathbb{R} \) that has only one sign (i.e. \( g \geq 0 \) or \( g \leq 0 \)), and \( h \in C[a,b] \), there exists \( \tau \in [a,b] \) satisfying

\[
\int_a^b h(t)g(t) \, dt = h(\tau) \int_a^b g(t) \, dt.
\]

Proof. First, suppose \( g \geq 0 \). Let \( s_m, s_M \in [a,b] \) be points, where \( h \) assumes its min (denoted by \( m_h \)) and its max (denoted by \( M_h \)), respectively. We have

\[
m_h \int_a^b g(t) \, dt \leq \int_a^b h(t)g(t) \, dt \leq M_h \int_a^b g(t) \, dt.
\]
Thus, if we define

\[ F : [a, b] \longrightarrow \mathbb{R}, \quad F(s) := h(s) \int_a^b g(t) \, dt - \int_a^b h(t)g(t) \, dt, \]

then \( F(s_m) \leq 0 \leq F(s_M) \), and the intermediate value theorem yields \( \tau \in [a, b] \) such that \( F(\tau) = 0 \), i.e. \( \tau \) satisfies (4.18). If \( g \leq 0 \), then applying the result we just proved to \(-g\) establishes the case.

**Proposition 4.14.** Let \([a, b]\) be a real interval, \( a < b \), and let \( I_n \) be the quadrature rule based on interpolating polynomials for the distinct points \( x_0, \ldots, x_n \in [a, b], n \in \mathbb{N}_0 \), as defined in (4.15). If \( \omega_{n+1}(x) = \prod_{i=0}^n (x - x_i) \) has only one sign on \([a, b]\) (i.e. \( \omega_{n+1} \geq 0 \) or \( \omega_{n+1} \leq 0 \) on \([a, b]\), see Rem. 4.15 below), then, for each \( f \in C^{n+1}[a, b] \), there exists \( \tau \in [a, b] \) such that

\[ \int_a^b f(x) \, dx - I_n(f) = \frac{f^{(n+1)}(\tau)}{(n+1)!} \int_a^b \omega_{n+1}(x) \, dx. \]  

(4.19)

In particular, if \( f^{(n+1)} \) has just one sign as well, then one can infer the sign of the error term (i.e. of the right-hand side of (4.19)).

**Proof.** From Th. 3.25, we know that, for each \( x \in [a, b] \), there exists \( \xi(x) \in [a, b] \) such that (cf. (3.31))

\[ f(x) = p_n(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \omega_{n+1}(x). \]  

(4.20)

Combining (4.20) with (3.29) yields

\[ f^{(n+1)}(\xi(x)) = [f|x, x_0, \ldots, x_n](n+1)! \]  

(4.21)

for each \( x \in [a, b] \) such that \( \omega_{n+1}(x) \neq 0 \), i.e. for each \( x \in [a, b] \) \( \{x_0, \ldots, x_n\} \). Since \( x \mapsto [f|x, x_0, \ldots, x_n] \) is continuous on \([a, b]\) by Th. 3.25(c), the function \( x \mapsto f^{(n+1)}(\xi(x)) \) is also continuous in each \( x \in [a, b] \) \( \{x_0, \ldots, x_n\} \), i.e. piecewise continuous on \([a, b]\), \( k \in \{0, \ldots, n\} \) and let \((\tau_{\nu})_{\nu \in \mathbb{N}}\) be a sequence in \([a, b] \) \( \{x_0, \ldots, x_n\} \) such that \( \lim_{\nu \to \infty} \tau_{\nu} = x_k \). From the above, we then know that (4.21) holds with \( x \) replaced by \( \tau_{\nu} \) for each \( \nu \in \mathbb{N} \). Since \((\xi(\tau_{\nu}))_{\nu \in \mathbb{N}}\) is a sequence in the compact interval \([a, b]\), there exists a subsequence \((\tau_{\nu})_{\nu \in \mathbb{N}}\) of \((\tau_{\nu})_{\nu \in \mathbb{N}}\) such that \( \xi(\tau_{\nu}) \) converges to some \( \xi_k \in [a, b] \). We now choose \( \xi(x_k) := \xi_k \) for each \( k \in \{0, \ldots, n\} \). Then the continuity of \( f^{(n+1)} \) and the continuity of \( x \mapsto [f|x, x_0, \ldots, x_n] \) imply (4.21) to also hold at each \( x = x_k \) and, in particular, \( x \mapsto f^{(n+1)}(\xi(x)) \) is continuous on \([a, b]\). Thus, integrating (4.20) and applying Lem. 4.13 yields \( \tau \in [a, b] \) satisfying

\[ \int_a^b f(x) \, dx - I_n(f) = \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\xi(x))\omega_{n+1}(x) \, dx = \frac{f^{(n+1)}(\tau)}{(n+1)!} \int_a^b \omega_{n+1}(x) \, dx, \]

proving (4.19).

**Remark 4.15.** There exist precisely 3 examples where \( \omega_{n+1} \) has only one sign on \([a, b]\), namely \( \omega_1(x) = (x-a) \geq 0 \), \( \omega_1(x) = (x-b) \leq 0 \), and \( \omega_2(x) = (x-a)(x-b) \leq 0 \). In all other cases, there exists at least one \( x_i \in [a, b] \). As \( x_i \) is a simple zero of \( \omega_{n+1} \), \( \omega_{n+1} \) changes its sign in the interior of \([a, b]\).
4.3 Newton-Cotes Formulas

We will now study quadrature rules based on interpolating polynomials with equally-spaced data points. In particular, we continue to assume $\rho \equiv 1$ (nonweighted integrals).

4.3.1 Definition, Weights, Degree of Accuracy

Definition and Remark 4.16. Given a real interval $[a, b], a < b$, a quadrature rule based on interpolating polynomials according to Def. 4.10 is called a Newton-Cotes formula, also known as a Newton-Cotes quadrature rule if, and only if, the points $x_0, \ldots, x_n \in [a, b], n \in \mathbb{N}$, are equally-spaced. Moreover, a Newton-Cotes formula is called closed or a Newton-Cotes closed quadrature rule if, and only if,

$$x_i := a + i h, \quad h := \frac{b - a}{n}, \quad i \in \{0, \ldots, n\}, \quad n \in \mathbb{N}. \quad (4.22)$$

In the present context of equally-spaced $x_i$, the discretization’s mesh size $h$ is also called step size. It is an exercise to compute the weights of the Newton-Cotes closed quadrature rules. One obtains

$$\sigma_i = \sigma_{i,n} = \frac{1}{n} \int_0^n \prod_{k=0}^n \frac{s-k}{i-k} ds. \quad (4.23)$$

Given $n$, one can compute the $\sigma_i$ explicitly and store or tabulate them for further use. The next lemma shows that one actually does not need to compute all $\sigma_i$:

Lemma 4.17. The weights $\sigma_i$ of the Newton-Cotes closed quadrature rules (see (4.23)) are symmetric, i.e.

$$\sigma_i = \sigma_{n-i} \quad \text{for each } i \in \{0, \ldots, n\}, \quad n \in \mathbb{N}. \quad (4.24)$$

Proof. Exercise. ■

If $n$ is even, then, as we will show in Th. 4.19 below, the corresponding closed Newton-Cotes formula $I_n$ is exact not only for $p \in \text{Pol}_n(\mathbb{R})$, but even for $p \in \text{Pol}_{n+1}(\mathbb{R})$. On the other hand, we will also see in Th. 4.19 that $I(x^{n+2}) \neq I_n(x^{n+2})$ as a consequence of the following preparatory lemma:

Lemma 4.18. If $[a, b], a < b$, is a real interval, $n \in \mathbb{N}$ is even, and $x_i, i \in \{0, \ldots, n\}$, are defined according to (4.22), then the antiderivative of the Newton basis polynomial $\omega := \omega_{n+1}$, i.e. the function

$$F : \mathbb{R} \rightarrow \mathbb{R}, \quad F(x) := \int_a^x \prod_{i=0}^n (y - x_i) dy, \quad (4.25)$$

satisfies

$$F(x) \begin{cases} = 0 & \text{for } x = a, \\ > 0 & \text{for } a < x < b, \\ = 0 & \text{for } x = b. \end{cases} \quad (4.26)$$
Proof. $F(a) = 0$ is clear. The remaining assertions of (4.26) are shown in several steps. Claim 1. With $h \in ]0, b - a[$ defined according to (4.22), one has

\[
\begin{align*}
\omega(x_{2j} + \tau) & > 0, \\
\omega(x_{2j+1} + \tau) & < 0, \quad \text{for each } \tau \in ]0, h[ \text{ and each } j \in \{0, \ldots, \lfloor n/2 \rfloor - 1\}. \quad (4.27)
\end{align*}
\]

Proof. As $\omega$ has degree precisely $n + 1$ and a zero at each of the $n + 1$ distinct points $x_0, \ldots, x_n$, each of these zeros must be simple. In particular, $\omega$ must change its sign at each $x_i$. Moreover, $\lim_{x \to -\infty} \omega(x) = -\infty$ due to the fact that the degree $n + 1$ of $\omega$ is odd. This implies that $\omega$ changes its sign from negative to positive at $x_0 = a$ and from positive to negative at $x_1 = a + h$, establishing the case $j = 0$. The general case (4.27) now follows by induction. \hfill \Box

Claim 2. On the left half of the interval $]a, b[,$ $|\omega|$ decays in the following sense:

\[
|\omega(x + h)| < |\omega(x)| \quad \text{for each } x \in \left[ a, \frac{a + b}{2} - h \right] \setminus \{x_0, \ldots, x_{n/2-1}\}. \quad (4.28)
\]

Proof. For each $x$ that is not a zero of $\omega$, we compute

\[
\frac{\omega(x + h)}{\omega(x)} = \frac{\prod_{i=0}^{n}(x + h - x_i)}{\prod_{i=0}^{n}(x - x_i)} = \frac{(x + h - a) \prod_{i=1}^{n}(x + h - x_i)}{(x - b) \prod_{i=0}^{n-1}(x - x_i)} = \frac{x + h - a}{x - b}.
\]

Moreover, if $a < x < \frac{a + b}{2} - h$, then

\[
|x + h - a| = x + h - a < \frac{b - a}{2} < b - x = |x - b|,
\]

proving (4.28). \hfill \Box

Claim 3. $F(x) > 0$ for each $x \in ]a, \frac{a + b}{2}].$

Proof. There exists $\tau \in ]0, h]$ and $i \in \{0, \ldots, n/2 - 1\}$ such that $x = x_i + \tau$. Thus,

\[
F(x) = \int_{x_i}^{x} \omega(y) \, dy + \sum_{k=0}^{i-1} \int_{x_k}^{x_{k+1}} \omega(y) \, dy, \quad (4.29)
\]

and it suffices to show that, for each $\tau \in ]0, h],$

\[
\begin{align*}
\int_{x_{2j}}^{x_{2j} + \tau} \omega(y) \, dy & > 0 \quad \text{if } j \in \left\{ k \in \mathbb{N}_0 : 0 \leq k \leq \frac{n/2 - 1}{2} \right\}, \quad (4.30a) \\
\int_{x_{2j+1}}^{x_{2j+1} + \tau} \omega(y) \, dy & > 0 \quad \text{if } j \in \left\{ k \in \mathbb{N}_0 : 0 \leq k \leq \frac{n/2 - 2}{2} \right\}. \quad (4.30b)
\end{align*}
\]

4 NUMERICAL INTEGRATION

77
While (4.30a) is immediate from (4.27), for (4.30b), one calculates
\[
\int_{x_{2j}}^{x_{2j+1}} \omega(y) \, dy = \int_{x_{2j}}^{x_{2j+1}} \omega(y) \, dy + \int_{x_{2j+1}}^{x_{2j+\tau}} \omega(y) \, dy > 0
\]
for each \( \tau \in [0, h] \) and each \( 0 \leq j \leq \frac{n-2}{2} \), thereby establishing the case. ▲

**Claim 4.** \( \omega \) is antisymmetric with respect to the midpoint of \([a, b] \), i.e.
\[
\omega \left( \frac{a + b}{2} + \tau \right) = -\omega \left( \frac{a + b}{2} - \tau \right) \quad \text{for each} \quad \tau \in \mathbb{R}. \tag{4.31}
\]

**Proof.** From \((a + b)/2 - x_i = -(a + b)/2 - x_{n-i} = (n/2 - i)h \) for each \( i \in \{0, \ldots, n\} \), one obtains
\[
\omega \left( \frac{a + b}{2} + \tau \right) = \prod_{i=0}^{n} \left( \frac{a + b}{2} + \tau - x_i \right) = -\prod_{i=0}^{n} \left( \frac{a + b}{2} - \tau - x_{n-i} \right) \\
= -\prod_{i=0}^{n} \left( \frac{a + b}{2} - \tau - x_i \right) = -\omega \left( \frac{a + b}{2} - \tau \right),
\]
which is (4.31). ▲

**Claim 5.** \( F \) is symmetric with respect to the midpoint of \([a, b] \), i.e.
\[
F \left( \frac{a + b}{2} + \tau \right) = F \left( \frac{a + b}{2} - \tau \right) \quad \text{for each} \quad \tau \in \mathbb{R}. \tag{4.32}
\]

**Proof.** For each \( \tau \in [0, \frac{b-a}{2}] \), we obtain
\[
F \left( \frac{a + b}{2} + \tau \right) = \int_a^{(a+b)/2-\tau} \omega(x) \, dx + \int_{(a+b)/2+\tau}^{b} \omega(x) \, dx = F \left( \frac{a + b}{2} - \tau \right),
\]
since the second integral vanishes due to (4.31). ▲

Finally, (4.26) follows from combining \( F(a) = 0 \) with Claims 3 and 5. ■

**Theorem 4.19.** If \([a, b], a < b, \) is a real interval and \( n \in \mathbb{N} \) is even, then the Newton-Cotes closed quadrature rule \( I_n : \mathbb{R}[a, b] \longrightarrow \mathbb{R} \) has degree of accuracy \( n + 1 \).

**Proof.** It is an exercise to show that \( I_n \) has at least degree of accuracy \( n + 1 \). It then remains to show that
\[
I(x^{n+2}) \neq I_n(x^{n+2}). \tag{4.33}
\]

To that end, let \( x_{n+1} \in [a, b] \setminus \{x_0, \ldots, x_n\} \), and let \( q \in \text{Pol}_{n+1}(\mathbb{R}) \) be the interpolating polynomial for \( g(x) = x^{n+2} \) and \( x_0, \ldots, x_{n+1} \), whereas \( p_n \in \text{Pol}_n(\mathbb{R}) \) is the corresponding
interpolating polynomial for \( x_0, \ldots, x_n \) (note that \( p_n \) interpolates both \( g \) and \( q \)). As we already know that \( I_n \) has at least degree of accuracy \( n + 1 \), we obtain

\[
\int_a^b p_n(x) \, dx = I_n(g) = I_n(q) = \int_a^b q(x) \, dx. \tag{4.34}
\]

By applying (3.31) with \( g(x) = x^{n+2} \) and using \( g^{(n+2)}(x) = (n+2)! \), one obtains

\[
g(x) - q(x) = \frac{g^{(n+2)}(\xi(x))}{(n+2)!} \omega_{n+2}(x) = \omega_{n+2}(x) = \prod_{i=0}^{n+1} (x-x_i) \quad \text{for each} \quad x \in [a, b]. \tag{4.35}
\]

Integrating (4.35) while taking into account (4.34) as well as using \( F \) from (4.25) yields

\[
I(g) - I_n(g) = \int_a^b \prod_{i=0}^{n+1} (x-x_i) \, dx = \int_a^b F'(x)(x-x_{n+1}) \, dx
\]

\[
= [F(x)(x-x_{n+1})]_a^b - \int_a^b F(x) \, dx \quad \tag{4.26} < 0,
\]

proving (4.33).

### 4.3.2 Rectangle Rules \((n = 0)\)

Note that, for \( n = 0 \), there is no closed Newton-Cotes formula, as (4.22) can not be satisfied with just one point \( x_0 \). Every Newton-Cotes quadrature rule with \( n = 0 \) is called a rectangle rule, since the integral \( \int_a^b f(x) \, dx \) is approximated by \( (b-a)f(x_0) \), \( x_0 \in [a, b] \), i.e. by the area of the rectangle with vertices \((a, 0), (b, 0), (a, f(x_0)), (b, f(x_0))\).

Not surprisingly, the rectangle rule with \( x_0 = (b+a)/2 \) is known as the midpoint rule.

Rectangle rules have degree of accuracy \( r \geq 0 \), since they are exact for constant functions. Rectangle rules actually all have \( r = 0 \), except for the midpoint rule with \( r = 1 \) (exercise).

In the following lemma, we provide an error estimate for the rectangle rule using \( x_0 = a \). For other choices of \( x_0 \), one obtains similar results.

**Lemma 4.20.** Given a real interval \([a, b], a < b\), and \( f \in C^1[a, b] \), there exists \( \tau \in [a, b] \) such that

\[
\int_a^b f(x) \, dx - (b-a)f(a) = \frac{(b-a)^2}{2} f'(\tau). \tag{4.36}
\]

**Proof.** As \( \omega_1(x) = (x-a) \geq 0 \) on \([a, b]\), we can apply Prop. 4.14 to get \( \tau \in [a, b] \) satisfying

\[
\int_a^b f(x) \, dx - (b-a)f(a) = f'(\tau) \int_a^b (x-a) \, dx = \frac{(b-a)^2}{2} f'(\tau),
\]

proving (4.36).
4.3.3 Trapezoidal Rule ($n = 1$)

Definition and Remark 4.21. The closed Newton-Cotes formula with $n = 1$ is called trapezoidal rule. Its explicit form is easily computed, e.g. from (4.15):

$$I_1(f) := \int_a^b \left( f(a) + \frac{f(b) - f(a)}{b - a} (x - a) \right) \, dx = \left[ f(a)x + \frac{f(b) - f(a)}{b - a} (x - a)^2 \right]_a^b = (b - a) \left( \frac{1}{2} f(a) + \frac{1}{2} f(b) \right)$$

(4.37)

for each $f \in C[a, b]$. Note that (4.37) justifies the name trapezoidal rule, as this is precisely the area of the trapezoid with vertices $(a, 0), (b, 0), (a, f(a)), (b, f(b))$.

Lemma 4.22. The trapezoidal rule has degree of accuracy $r = 1$.

Proof. Exercise. ■

Lemma 4.23. Given a real interval $[a, b], a < b,$ and $f \in C^2[a, b]$, there exists $\tau \in [a, b]$ such that

$$\int_a^b f(x) \, dx - I_1(f) = -\frac{(b - a)^3}{12} f''(\tau) = -\frac{h^3}{12} f''(\tau).$$

(4.38)

Proof. Exercise. ■

4.3.4 Simpson’s Rule ($n = 2$)

Definition and Remark 4.24. The closed Newton-Cotes formula with $n = 2$ is called Simpson’s rule. To find the explicit form, we compute the weights $\sigma_0, \sigma_1, \sigma_2$. According to (4.23), one finds

$$\sigma_0 = \frac{1}{2} \int_0^2 \left( \frac{s - 1}{-1} - \frac{s - 2}{-2} \right) \, ds = \frac{1}{4} \left[ \frac{s^3}{3} - \frac{3s^2}{2} + 2s \right]_0^2 = \frac{1}{6}.$$  

(4.39)

Next, one obtains $\sigma_2 = \sigma_0 = \frac{1}{6}$ from (4.24) and $\sigma_1 = 1 - \sigma_0 - \sigma_2 = \frac{2}{3}$ from Rem. 4.9. Plugging this into (4.9) yields

$$I_2(f) = (b - a) \left( \frac{1}{6} f(a) + \frac{2}{3} f \left( \frac{a + b}{2} \right) + \frac{1}{6} f(b) \right)$$

(4.40)

for each $f \in C[a, b]$.

Lemma 4.25. Simpson’s rule has degree of accuracy $3$.

Proof. The assertion is immediate from Th. 4.19. ■

Lemma 4.26. Given a real interval $[a, b], a < b,$ and $f \in C^4[a, b]$, there exists $\tau \in [a, b]$ such that

$$\int_a^b f(x) \, dx - I_2(f) = -\frac{(b - a)^5}{2880} f^{(4)}(\tau) = -\frac{h^5}{90} f^{(4)}(\tau).$$

(4.41)
Proof. In addition to \( x_0 = a, x_1 = (a + b)/2, x_2 = b \), we choose \( x_3 := x_1 \). Using Hermite interpolation, we know that there is a unique \( q \in \text{Pol}_3(\mathbb{R}) \) such that \( q(x_i) = f(x_i) \) for each \( i \in \{0, 1, 2\} \) and \( q'(x_3) = f'(x_3) \). As before, let \( p_2 \in \text{Pol}_2(\mathbb{R}) \) be the corresponding interpolating polynomial merely satisfying \( p_2(x_i) = f(x_i) \). As, by Lem. 4.25, \( I_2 \) has degree of accuracy 3, we know

\[
\int_a^b p_2(x) \, dx = I_2(f) = I_2(q) = \int_a^b q(x) \, dx. \tag{4.42}
\]

By applying (3.31) with \( n = 3 \), one obtains

\[
f(x) = q(x) + \frac{f^{(4)}(\xi(x))}{4!} \omega_4(x) \quad \text{for each } x \in [a,b], \tag{4.43}
\]

where

\[
\omega_4(x) = (x - a) \left( x - \frac{a + b}{2} \right)^2 (x - b). \tag{4.44}
\]

As in the proof of Prop. 4.14, it follows that the function \( x \mapsto f^{(4)}(\xi(x)) \) can be chosen continuous. Moreover, according to (4.44), \( \omega_4(x) \leq 0 \) for each \( x \in [a,b] \). Thus, integrating (4.43) and applying (4.42) as well as Lem. 4.13 yields \( \tau \in [a,b] \) satisfying

\[
\int_a^b f(x) \, dx - I_2(f) = -\frac{f^{(4)}(\tau)}{24} \frac{(b-a)^5}{120} = -\frac{h^5}{90} f^{(4)}(\tau), \tag{4.45}
\]

where it was used that

\[
\int_a^b \omega_4(x) \, dx = -\int_a^b \frac{(x-a)^2}{2} \left( 2 \left( x - \frac{a + b}{2} \right) (x - b) \right) \, dx \\
= -\frac{(b-a)^5}{24} - \int_a^b \frac{(x-a)^3}{6} \left( 2(x-b) + 4 \left( x - \frac{a + b}{2} \right) \right) \, dx \\
= -\frac{(b-a)^5}{24} - 2 \frac{(b-a)^5}{24} + \int_a^b \frac{(x-a)^4}{24} \, dx \\
= -\frac{(b-a)^5}{120}. \tag{4.46}
\]

This completes the proof of (4.41). \( \blacksquare \)

4.3.5 Higher Order Newton-Cotes Formulas

As before, let \([a,b]\) be a real interval, \( a < b \), and let \( I_n : \mathbb{R} [a,b] \rightarrow \mathbb{R} \) be the Newton-Cotes closed quadrature rule based on interpolating polynomials \( p \in \text{Pol}_n(\mathbb{R}) \). According to Th. 4.11(b), we know that \( I_n \) has degree of accuracy at least \( n \), and one might hope that, by increasing \( n \), one obtains more and more accurate quadrature rules for all
4 NUMERICAL INTEGRATION

\( f \in R[a,b] \) (or at least for all \( f \in C[a,b] \)). Unfortunately, this is not the case! As it turns out, Newton-Cotes formulas with larger \( n \) have quite a number of drawbacks. As a result, in practice, Newton-Cotes formulas with \( n > 3 \) are hardly ever used (for \( n = 3 \), one obtains Simpson’s 3/8 rule, and even that is not used all that often).

The problems with higher order Newton-Cotes formulas have to do with the fact that the formulas involve negative weights (negative weights first occur for \( n = 8 \)). As \( n \) increases, the situation becomes worse and worse and one can actually show that (see [Bra77, Sec. IV.2])

\[
\lim_{n \to \infty} \sum_{i=0}^{n} |\sigma_{i,n}| = \infty.
\] (4.47)

For polynomials and some other functions, terms with negative and positive weights cancel, but for general \( f \in C[a,b] \) (let alone general \( f \in R[a,b] \)), that is not the case, leading to instabilities and divergence. That, in the light of (4.47), convergence can not be expected is a consequence of the following Th. 4.28.

4.4 Convergence of Quadrature Rules

We first compute the operator norm of a quadrature rule in terms of its weights:

Proposition 4.27. Given a real interval \([a,b], a < b,\) and a quadrature rule

\[
I_n : C[a,b] \to \mathbb{R}, \quad I_n(f) = (b - a) \sum_{i=0}^{n} \sigma_i f(x_i),
\] (4.48)

with distinct \( x_0, \ldots, x_n \in [a,b] \) and weights \( \sigma_0, \ldots, \sigma_n \in \mathbb{R}, n \in \mathbb{N}, \) the operator norm of \( I_n \) induced by \( \| \cdot \|_{\infty} \) on \( C[a,b] \) is

\[
\|I_n\| = (b-a) \sum_{i=0}^{n} |\sigma_i|.
\] (4.49)

Proof. For each \( f \in C[0,1], \) we estimate

\[
|I_n(f)| \leq (b-a) \sum_{i=0}^{n} |\sigma_i| \|f\|_{\infty},
\]

which already shows \( \|I_n\| \leq (b-a) \sum_{i=0}^{n} |\sigma_i| \). For the remaining inequality, define \( \phi : \{0, \ldots, n\} \to \{0, \ldots, n\} \) to be a reordering such that \( x_{\phi(0)} < x_{\phi(1)} < \cdots < x_{\phi(n)} \), let \( y_i := \text{sgn}(\sigma_{\phi(i)}) \) for each \( i \in \{0, \ldots, n\} \), and

\[
f(x) := \begin{cases} 
  y_0 & \text{for } x \in [a, x_{\phi(0)}], \\
  y_i + \frac{y_{i+1} - y_i}{x_{\phi(i+1)} - x_{\phi(i)}} (x - x_{\phi(i)}) & \text{for } x \in [x_{\phi(i)}, x_{\phi(i+1)}], i \in \{0, \ldots, n-1\}, \\
  y_n & \text{for } x \in [x_{\phi(n)}, b].
\end{cases}
\]
Note that $f$ is a linear spline on $[x_{\phi(0)}, x_{\phi(n)}]$ (cf. (3.44)), possibly with constant continuous extensions on $[a, x_{\phi(0)}]$ and $[x_{\phi(n)}, b]$. In particular, $f \in C[a, b]$. Clearly, either $I_n \equiv 0$ or $\|f\|_\infty = 1$ and

$$I_n(f) = (b - a) \sum_{i=0}^{n} \sigma_i f(x_i) = (b - a) \sum_{i=0}^{n} \sigma_i \text{sgn}(\sigma_i) = (b - a) \sum_{i=0}^{n} |\sigma_i|,$$

showing the remaining inequality $\|I_n\| \geq (b - a) \sum_{i=0}^{n} |\sigma_i|$ and, thus, (4.49).

**Theorem 4.28** (Polya). Given a real interval $[a, b]$, $a < b$, and a weight function $\rho : [a, b] \to \mathbb{R}_0^+$, consider a sequence of quadrature rules

$$I_n : C[a, b] \to \mathbb{R}, \quad I_n(f) = (b - a) \sum_{i=0}^{n} \sigma_{i,n} f(x_{i,n}),$$

with distinct points $x_{0,n}, \ldots, x_{n,n} \in [a, b]$ and weights $\sigma_{0,n}, \ldots, \sigma_{n,n} \in \mathbb{R}$, $n \in \mathbb{N}$. The sequence converges in the sense that

$$\lim_{n \to \infty} I_n(f) = \int_{a}^{b} f(x)\rho(x) \, dx \quad \text{for each } f \in C[a, b]$$

if, and only if, the following two conditions are satisfied:

(i) $\lim_{n \to \infty} I_n(p) = \int_{a}^{b} p(x)\rho(x) \, dx$ holds for each polynomial $p$.

(ii) The sums of the absolute values of the weights are uniformly bounded, i.e. there exists $C \in \mathbb{R}^+$ such that

$$\sum_{i=0}^{n} |\sigma_{i,n}| \leq C \quad \text{for each } n \in \mathbb{N}.$$

**Proof.** Suppose (i) and (ii) hold true. Let $W := \int_{a}^{b} \rho(x) \, dx$. Given $f \in C[a, b]$ and $\epsilon > 0$, according to the Weierstrass Approximation Theorem G.1, there exists a polynomial $p$ such that

$$\|f - p\|_{[a, b]} \|_\infty < \epsilon := \frac{\epsilon}{C(b - a) + 1 + W}.$$  \hspace{1cm} (4.52)

Due to (i), there exists $N \in \mathbb{N}$ such that, for each $n \geq N$,

$$\left| I_n(p) - \int_{a}^{b} p(x)\rho(x) \, dx \right| < \epsilon.$$

Thus, for each $n \geq N$,

$$\left| I_n(f) - \int_{a}^{b} f(x)\rho(x) \, dx \right|
\leq \left| I_n(f) - I_n(p) \right| + \left| I_n(p) - \int_{a}^{b} p(x)\rho(x) \, dx \right| + \left| \int_{a}^{b} p(x)\rho(x) \, dx - \int_{a}^{b} f(x)\rho(x) \, dx \right|
< (b - a) \sum_{i=0}^{n} |\sigma_{i,n}| |f(x_{i,n}) - p(x_{i,n})| + \epsilon + \epsilon W \leq \epsilon \left( (b - a)C + 1 + W \right) = \epsilon,$$
proving (4.51).

Conversely, (4.51) trivially implies (i). That it also implies (ii) is much more involved. Letting $\mathcal{T} := \{I_n : n \in \mathbb{N}\}$, we have $\mathcal{T} \subseteq \mathcal{L}(C[a, b], \mathbb{R})$ and (4.51) implies

$$\sup \{ |T(f)| : T \in \mathcal{T} \} = \sup \{ |I_n(f)| : n \in \mathbb{N} \} < \infty \text{ for each } f \in C[a, b]. \quad (4.53)$$

Then the Banach-Steinhaus Th. H.4 yields

$$\sup \{ \|I_n\| : n \in \mathbb{N} \} = \sup \{ \|T\| : T \in \mathcal{T} \} < \infty, \quad (4.54)$$

which, according to Prop. 4.27, is (ii). The proof of the Banach-Steinhaus theorem is usually part of Functional Analysis. It is provided in Appendix H. As it only uses elementary metric space theory, the interested reader should be able to understand it.

While, for $\rho \equiv 1$, Condition (i) of Th. 4.28 is obviously satisfied by the Newton-Cotes formulas, since, given $q \in \text{Pol}_n(\mathbb{R})$, $I_n(q)$ is exact for each $m \geq n$, Condition (ii) of Th. 4.28 is violated due to (4.47). As the Newton-Cotes formulas are based on interpolating polynomials, and we had already needed additional conditions to guarantee the convergence of interpolating polynomials (see Th. 3.26), it should not be too surprising that Newton-Cotes formulas do not converge for all continuous functions.

So we see that improving the accuracy of Newton-Cotes formulas by increasing $n$ is, in general, not feasible. However, the accuracy can be improved using a number of different strategies, two of which will be considered in the following. In Sec. 4.5, we will study so-called composite Newton-Cotes rules that result from subdividing $[a, b]$ into small intervals, using a low-order Newton-Cotes formula on each small interval, and then summing up the results. In Sec. 4.6, we will consider Gaussian quadrature, where, in contrast to the Newton-Cotes formulas, one does not use equally-spaced points $x_i$ for the interpolation, but chooses the locations of the $x_i$ more carefully. As it turns out, one can choose the $x_i$ such that all weights remain positive even for $n \to \infty$. Then Rem. 4.9 combined with Th. 4.28 yields convergence.

### 4.5 Composite Newton-Cotes Quadrature Rules

Before studying Gaussian quadrature rules in the context of weighted integrals in Sec. 4.6, for the present section, we return to the situation of Sec. 4.3, i.e. nonweighted integrals ($\rho \equiv 1$).

#### 4.5.1 Introduction, Convergence

As discussed in the previous section, improving the accuracy of numerical integration by using high-order Newton-Cotes rules does usually not work. On the other hand, better accuracy can be achieved by subdividing $[a, b]$ into small intervals and then using a low-order Newton-Cotes rule (typically $n \leq 2$) on each of the small intervals. The composite Newton-Cotes quadrature rules are the results of this strategy.
We begin with a general definition and a general convergence result based on the convergence of Riemann sums for Riemann integrable functions.

**Definition 4.29.** Suppose, for \( f : [0,1] \rightarrow \mathbb{R}, I_n, n \in \mathbb{N}_0, \) has the form

\[
I_n(f) = \sum_{l=0}^{n} \sigma_l f(\xi_l) \tag{4.55}
\]

with \( 0 \leq \xi_0 < \xi_1 < \cdots < \xi_n \leq 1 \) and weights \( \sigma_0, \ldots, \sigma_n \in \mathbb{R} \). Given a real interval \([a,b], a < b, \) and \( N \in \mathbb{N}, \) set

\[
x_k := a + kh, \quad h := \frac{b-a}{N}, \quad \text{for each } k \in \{0, \ldots, N\},
\]

\[
x_{kl} := x_{k-1} + h \xi_l, \quad \text{for each } k \in \{1, \ldots, N\}, \ l \in \{0, \ldots, n\}. \tag{4.56a}
\]

Then

\[
I_{n,N}(f) := h \sum_{k=1}^{N} \sum_{l=0}^{n} \sigma_l f(x_{kl}), \ f : [a,b] \rightarrow \mathbb{R}, \tag{4.57}
\]

are called the composite quadrature rules based on \( I_n. \)

**Remark 4.30.** The heuristics behind (4.57) is

\[
\int_{x_k}^{x_{k-1}} f(x) \, dx = h \int_{0}^{1} f(x_{k-1} + \xi h) \, d\xi \approx h \sum_{l=0}^{n} \sigma_l f(x_{kl}). \tag{4.58}
\]

**Theorem 4.31.** Let \([a,b], a < b, \) be a real interval, \( n \in \mathbb{N}_0. \) If \( I_n \) has the form (4.55) and is exact for each constant function (i.e. it has at least degree of accuracy 0 in the language of Def. 4.8), then the composite quadrature rules (4.57) based on \( I_n \) converge for each Riemann integrable \( f : [a,b] \rightarrow \mathbb{R}, \) i.e.

\[
\forall f \in R[a,b] \lim_{N \to \infty} I_{n,N}(f) = \int_{a}^{b} f(x) \, dx. \tag{4.59}
\]

**Proof.** Let \( f \in R[a,b]. \) Introducing, for each \( l \in \{0, \ldots, n\}, \) the abbreviation

\[
S_l(N) := \sum_{k=1}^{N} \frac{b-a}{N} f(x_{kl}), \tag{4.60}
\]

(4.57) yields

\[
I_{n,N}(f) = \sum_{l=0}^{n} \sigma_l S_l(N). \tag{4.61}
\]

Note that, for each \( l \in \{0, \ldots, n\}, \) \( S_l(N) \) is a Riemann sum for the tagged partition \( \Delta_{Nl} := \{(x_N^0, \ldots, x_N^N), (x_{Nl}^0, \ldots, x_{Nl}^N)\} \) and \( \lim_{N \to \infty} h(\Delta_{Nl}) = \lim_{N \to \infty} \frac{b-a}{N} = 0. \) Thus, applying Th. 4.3,

\[
\lim_{N \to \infty} S_l(N) = \int_{a}^{b} f(x) \, dx \quad \text{for each } l \in \{0, \ldots, n\}.
\]
This, in turn, implies
\[
\lim_{N \to \infty} I_{n,N}(f) = \lim_{N \to \infty} \sum_{l=0}^{n} \sigma_l S_l(N) = \int_{a}^{b} f(x) \, dx \sum_{l=0}^{n} \sigma_l = \int_{a}^{b} f(x) \, dx
\]
as, due to hypothesis,
\[
\sum_{l=0}^{n} \sigma_l = I_n(1) = \int_{0}^{1} \, dx = 1,
\]
proving (4.59).

The convergence result of Th. 4.31 still suffers from the drawback of the original convergence result for the Riemann sums, namely that we do not obtain any control on the rate of convergence and the resulting error term. We will achieve such results in the following sections by assuming additional regularity of the integrated functions \( f \). We introduce the following notion as a means to quantify the convergence rate of composite quadrature rules:

**Definition 4.32.** Given a real interval \([a, b] \), \( a < b \), and a subset \( \mathcal{F} \) of the Riemann integrable functions on \([a, b] \), composite quadrature rules \( I_{n,N} \) according to (4.57) are said to have order of convergence \( r > 0 \) for \( f \in \mathcal{F} \) if, and only if, for each \( f \in \mathcal{F} \), there exists \( K = K(f) \in \mathbb{R}_+^\ast \) such that
\[
\left| I_{n,N}(f) - \int_{a}^{b} f(x) \, dx \right| \leq K h^r \quad \text{for each } N \in \mathbb{N} \tag{4.62}
\]
\( (h = (b - a)/N \) as before).  

### 4.5.2 Composite Rectangle Rules \((n = 0)\)

**Definition and Remark 4.33.** Using the rectangle rule \( \int_{0}^{1} f(x) \, dx \approx f(0) \) of Sec. 4.3.2 as \( I_0 \) on \([0, 1] \), the formula (4.57) yields, for \( f : [a, b] \to \mathbb{R} \), the corresponding composite rectangle rules
\[
I_{0,N}(f) = h \sum_{k=1}^{N} f(x_{k0}) \frac{x_{k0} = x_{k-1}}{h} h(f(x_0) + f(x_1) + \cdots + f(x_{N-1})). \tag{4.63}
\]

**Theorem 4.34.** For each \( f \in C^1[a, b] \), there exists \( \tau \in [a, b] \) satisfying
\[
\int_{a}^{b} f(x) \, dx - I_{0,N}(f) = \frac{(b - a)^2}{2N} f'(\tau) = \frac{b - a}{2} h f'(\tau). \tag{4.64}
\]
In particular, the composite rectangle rules have order of convergence \( r = 1 \) for \( f \in C^1[a, b] \) in terms of Def. 4.32.
Proof. Fix $f \in C^1[a,b]$. From Lem. 4.20, we obtain, for each $k \in \{1, \ldots, N\}$, a point $\tau_k \in [x_{k-1}, x_k]$ such that
\[
\int_{x_{k-1}}^{x_k} f(x) \, dx = -\frac{b-a}{N} f(x_{k-1}) = \frac{(b-a)^2}{2N^2} f'(\tau_k) = \frac{h^2}{2} f'(\tau_k). \tag{4.65}
\]
Summing (4.65) from $k = 1$ through $N$ yields
\[
\int_a^b f(x) \, dx - I_{0,N}(f) = \frac{b-a}{2} h \frac{1}{N} \sum_{k=1}^{N} f'(\tau_k).
\]
As $f'$ is continuous and
\[
\min \{ f'(x) : x \in [a,b] \} \leq \alpha := \frac{1}{N} \sum_{k=1}^{N} f'(\tau_k) \leq \max \{ f'(x) : x \in [a,b] \},
\]
the intermediate value theorem provides $\tau \in [a,b]$ satisfying $\alpha = f'(\tau)$, proving (4.64). The claimed order of convergence now follows as well, since (4.64) implies that (4.62) holds with $r = 1$ and $K := (b-a)\|f''\|/2$. \hfill \blacksquare

4.5.3 Composite Trapezoidal Rules \((n = 1)\)

**Definition and Remark 4.35.** Using the trapezoidal rule (4.37) as $I_1$ on $[0, 1]$, the formula (4.57) yields, for $f : [a, b] \to \mathbb{R}$, the corresponding **composite trapezoidal rules**
\[
I_{1,N}(f) = h \sum_{k=1}^{N} \frac{1}{2} \left( f(x_{k0}) + f(x_{k1}) \right) = h \sum_{k=1}^{N} \frac{1}{2} \left( f(x_{k-1}) + f(x_k) \right) = \frac{h}{2} \left( f(a) + 2 \left( f(x_1) + \cdots + f(x_{N-1}) \right) + f(b) \right). \tag{4.66}
\]

**Theorem 4.36.** For each $f \in C^2[a,b]$, there exists $\tau \in [a,b]$ satisfying
\[
\int_a^b f(x) \, dx - I_{1,N}(f) = -\frac{(b-a)^3}{12N^2} f''(\tau) = -\frac{b-a}{12} h^2 f''(\tau). \tag{4.67}
\]
In particular, the composite trapezoidal rules have order of convergence $r = 2$ for $f \in C^2[a,b]$ in terms of Def. 4.32.

**Proof.** Fix $f \in C^2[a,b]$. From Lem. 4.23, we obtain, for each $k \in \{1, \ldots, N\}$, a point $\tau_k \in [x_{k-1}, x_k]$ such that
\[
\int_{x_{k-1}}^{x_k} f(x) \, dx = \frac{h}{2} \left( f(x_{k-1}) + f(x_k) \right) = \frac{h^3}{12} f''(\tau_k).
\]
Summing the above equation from $k = 1$ through $N$ yields
\[
\int_a^b f(x) \, dx - I_{1,N}(f) = -\frac{b-a}{12} h^2 \frac{1}{N} \sum_{k=1}^{N} f''(\tau_k).
\]
4 NUMERICAL INTEGRATION

As \( f'' \) is continuous and

\[
\min \left\{ f''(x) : x \in [a, b] \right\} \leq \alpha := \frac{1}{N} \sum_{k=1}^{N} f''(\tau_k) \leq \max \left\{ f''(x) : x \in [a, b] \right\},
\]

the intermediate value theorem provides \( \tau \in [a, b] \) satisfying \( \alpha = f''(\tau) \), proving (4.67). The claimed order of convergence now follows as well, since (4.67) implies that (4.62) holds with \( r = 2 \) and \( K := (b - a)\|f''\|_{\infty}/12 \).

4.5.4 Composite Simpson’s Rules (\( n = 2 \))

**Definition and Remark 4.37.** Using Simpson’s rule (4.40) as \( I_2 \) on \([0, 1]\), the formula (4.57) yields, for \( f : [a, b] \rightarrow \mathbb{R} \), the corresponding composite Simpson’s rules \( I_{2,N}(f) \). It is an exercise to check that

\[
I_{2,N}(f) = \frac{h}{6} \left( f(a) + 4 \sum_{k=1}^{N} f(x_{k-\frac{1}{2}}) + 2 \sum_{k=1}^{N-1} f(x_k) + f(b) \right), \tag{4.68}
\]

where \( x_{k-\frac{1}{2}} := (x_{k-1} + x_k)/2 \) for each \( k \in \{1, \ldots, N\} \).

**Theorem 4.38.** For each \( f \in C^4[a, b] \), there exists \( \tau \in [a, b] \) satisfying

\[
\int_a^b f(x) \, dx - I_{2,N}(f) = -\frac{(b - a)^5}{2880N^4} f^{(4)}(\tau) = -\frac{b - a}{2880} h^4 f^{(4)}(\tau). \tag{4.69}
\]

In particular, the composite Simpson’s rules have order of convergence \( r = 4 \) for \( f \in C^4[a, b] \) in terms of Def. 4.32.

**Proof.** Exercise.

4.6 Gaussian Quadrature

4.6.1 Introduction

As discussed in Sec. 4.3.5, quadrature rules based on interpolating polynomials for equally-spaced points \( x_i \) are suboptimal. This fact is related to equally-spaced points being suboptimal for the construction of interpolating polynomials in general as mentioned in Rem. 3.28. In Gaussian quadrature, one invests additional effort into smarter placing of the \( x_i \), resulting in more accurate quadrature rules. For example, this will ensure that all weights remain positive as the number of \( x_i \) increases, which, in turn, will yield the convergence of Gaussian quadrature rules for all continuous functions (see Th. 4.50 below).

We will see in the following that Gaussian quadrature rules \( I_n \) are quadrature rules in the sense of Def. 4.6. However, for historical reasons, in the context of Gaussian
4 NUMERICAL INTEGRATION

It is common to use the notation

\[ I_n(f) = \sum_{i=1}^{n} \sigma_i f(\lambda_i), \quad (4.70) \]

i.e. one writes \( \lambda_i \) instead of \( x_i \) and the factor \((b-a)\) is incorporated into the weights \( \sigma_i \).

4.6.2 Orthogonal Polynomials

During Gaussian quadrature, the \( \lambda_i \) are determined as the zeros of polynomials that are orthogonal with respect to certain scalar products on the (infinite-dimensional) real vector space \( \text{Pol}(\mathbb{R}) \).

**Notation 4.39.** Given a real interval \([a, b], a < b\), and a weight function \( \rho : [a, b] \rightarrow \mathbb{R}_0^+ \) as defined in Def. 4.5, let

\[
\langle \cdot, \cdot \rangle_\rho : \text{Pol}(\mathbb{R}) \times \text{Pol}(\mathbb{R}) \rightarrow \mathbb{R}, \quad \langle p, q \rangle_\rho := \int_a^b p(x)q(x)\rho(x) \, dx, \quad (4.71a)
\]

\[
\| \cdot \|_\rho : \text{Pol}(\mathbb{R}) \rightarrow \mathbb{R}_0^+, \quad \| p \|_\rho := \sqrt{\langle p, p \rangle_\rho}. \quad (4.71b)
\]

**Lemma 4.40.** Given a real interval \([a, b], a < b\), and a weight function \( \rho : [a, b] \rightarrow \mathbb{R}_0^+ \), (4.71a) defines a scalar product on \( \text{Pol}(\mathbb{R}) \).

**Proof.** As \( \rho \in R[a, b], \rho \geq 0, \) and \( \int_a^b \rho(x) \, dx > 0, \) there must be some nontrivial interval \( J \subseteq [a, b] \) and \( \epsilon > 0 \) such that \( \rho \geq \epsilon \) on \( J \). Thus, if \( 0 \neq p \in \text{Pol}(\mathbb{R}) \), then there must be some nontrivial interval \( J_p \subseteq J \) and \( \epsilon_p > 0 \) such that \( p^2 \rho \geq \epsilon_p |J_p| > 0 \) and proving \( \langle \cdot, \cdot \rangle_\rho \) to be positive definite. Given \( p, q, r \in \text{Pol}(\mathbb{R}) \) and \( \lambda, \mu \in \mathbb{R} \), one computes

\[
\langle \lambda p + \mu q, r \rangle_\rho = \int_a^b (\lambda p(x) + \mu q(x))r(x)\rho(x) \, dx \\
= \lambda \int_a^b q(x)r(x)\rho(x) \, dx + \mu \int_a^b q(x)r(x)\rho(x) \, dx \\
= \lambda \langle p, r \rangle_\rho + \mu \langle q, r \rangle_\rho,
\]

proving \( \langle \cdot, \cdot \rangle_\rho \) to be linear in its first argument. Since, for \( p, q \in \text{Pol}(\mathbb{R}) \), one also has

\[
\langle p, q \rangle_\rho = \int_a^b p(x)q(x)\rho(x) \, dx = \int_a^b q(x)p(x)\rho(x) \, dx = \langle q, p \rangle_\rho,
\]

\( \langle \cdot, \cdot \rangle_\rho \) is symmetric as well and the proof of the lemma is complete. \( \blacksquare \)

**Remark 4.41.** Given a real interval \([a, b], a < b\), and a weight function \( \rho : [a, b] \rightarrow \mathbb{R}_0^+ \), consider \( \text{Pol}(\mathbb{R}) \) with the scalar product \( \langle \cdot, \cdot \rangle_\rho \) given by (4.71a). Then, for each \( n \in \mathbb{N}_0, q \in \text{Pol}_n(\mathbb{R}) \) (cf. [Phi19b, Def. 10.6(b)]) if, and only if,

\[
\int_a^b q(x)p(x)\rho(x) = 0 \quad \text{for each } p \in \text{Pol}_n(\mathbb{R}). \quad (4.72)
\]
At first glance, it might not be obvious that there are always \( q \in \text{Pol}(\mathbb{R}) \setminus \{0\} \) that satisfy \( (4.72) \). However, such \( q \neq 0 \) can always be found by the general procedure of Gram-Schmidt orthogonalization, known from Linear Algebra: According to [Phi19b, Th. 10.9], if \( \langle \cdot, \cdot \rangle \) is an arbitrary scalar product on \( \text{Pol}(\mathbb{R}) \), if \( (x_n)_{n \in \mathbb{N}_0} \) is a sequence in \( \text{Pol}(\mathbb{R}) \) and if \( (p_n)_{n \in \mathbb{N}_0} \) are recursively defined by

\[
p_0 := x_0, \quad \forall n \in \mathbb{N} \quad p_n := x_n - \frac{\langle x_n, p_k \rangle}{\|p_k\|^2} p_k, \quad (4.73)
\]

then the sequence \( (p_n)_{n \in \mathbb{N}_0} \) constitutes an orthogonal system in \( \text{Pol}(\mathbb{R}) \).

While Gram-Schmidt orthogonalization works in every space with a scalar product, for polynomials, the following theorem provides a more efficient recursive orthogonalization algorithm:

**Theorem 4.42.** Given a scalar product \( \langle \cdot, \cdot \rangle \) on \( \text{Pol}(\mathbb{R}) \) that satisfies

\[
\langle xp, q \rangle = \langle p, xq \rangle \quad \text{for each } p, q \in \text{Pol}(\mathbb{R}) \quad (4.74)
\]

(clearly, the scalar product from \( (4.71a) \) satisfies \( (4.74) \)), let \( \| \cdot \| \) denote the induced norm. If, for each \( n \in \mathbb{N}_0 \), \( x_n \in \text{Pol}(\mathbb{R}) \) is defined by \( x_n(x) := x^n \) and \( p_n \) is given by Gram-Schmidt orthogonalization according to \( (4.73) \), then the \( p_n \) satisfy the following recursive relation (sometimes called three-term recursion):

\[
p_0 = x_0 \equiv 1, \quad p_1 = x - \beta_0, \quad (4.75a)
\]

\[
p_{n+1} = (x - \beta_n)p_n - \gamma_n^2 p_{n-1} \quad \text{for each } n \in \mathbb{N}, \quad (4.75b)
\]

where

\[
\beta_n := \frac{\langle xp_n, p_n \rangle}{\|p_n\|^2} \quad \text{for each } n \in \mathbb{N}_0, \quad \gamma_n := \frac{\|p_n\|}{\|p_{n-1}\|} \quad \text{for each } n \in \mathbb{N}. \quad (4.75c)
\]

**Proof.** We know from [Phi19b, Th. 10.9] that the linear independence of the \( x_n \) guarantees the linear independence of the \( p_n \). We observe that \( p_0 = x_0 \equiv 1 \) is clear from \( (4.73) \) and the definition of \( x_0 \). Next, form \( (4.73) \), we compute

\[
p_1 = v_1 = x_1 - \frac{\langle x_1, p_0 \rangle}{\|p_0\|^2} p_0 = x - \frac{\langle xp_0, p_0 \rangle}{\|p_0\|^2} = x - \beta_0.
\]

It remains to consider \( n + 1 \) for each \( n \in \mathbb{N} \). Letting

\[
q_{n+1} := (x - \beta_n)p_n - \gamma_n^2 p_{n-1},
\]

the proof is complete once we have shown

\[
q_{n+1} = p_{n+1}.
\]
Clearly, 

\[ r := p_{n+1} - q_{n+1} \in \text{Pol}_n(\mathbb{R}), \]

as the coefficient of \( x^{n+1} \) in both \( p_{n+1} \) and \( q_{n+1} \) is 1. Since \( \text{Pol}_n(\mathbb{R}) \cap \text{Pol}_0(\mathbb{R}) = \{0\} \) (cf. [Phi19b, Lem. 10.7(a)]), it now suffices to show \( r \in \text{Pol}_n(\mathbb{R})^\perp \). Moreover, we know \( p_{n+1} \in \text{Pol}_n(\mathbb{R})^\perp \), since \( p_{n+1} \perp p_k \) for each \( k \in \{0, \ldots, n\} \) and \( \text{Pol}_n(\mathbb{R}) = \text{span}\{p_0, \ldots, p_n\} \). Thus, as \( \text{Pol}_n(\mathbb{R})^\perp \) is a vector space, to prove \( r \in \mathcal{P}_n^\perp \), it suffices to show

\[ q_{n+1} \in \text{Pol}_n(\mathbb{R})^\perp. \quad (4.76) \]

To this end, we start by using \( \langle p_{n-1}, p_n \rangle = 0 \) and the definition of \( \beta_n \) to compute

\[ \langle q_{n+1}, p_n \rangle = \langle (x - \beta_n)p_n - \gamma_n^2 p_{n-1}, p_n \rangle = \langle xp_n, p_n \rangle - \beta_n \langle p_n, p_n \rangle = 0. \quad (4.77a) \]

Similarly, also using (4.74) as well as the definition of \( \gamma_n \), we obtain

\[
\begin{align*}
\langle q_{n+1}, p_{n-1} \rangle &= \langle xp_n, p_{n-1} \rangle - \gamma_n^2 \|p_{n-1}\|^2 = \langle p_n, xp_{n-1} \rangle - \|p_n\|^2 \\
&= \langle p_n, xp_{n-1} - p_n \rangle = 0, \quad (4.77b)
\end{align*}
\]

where, at (\( * \)), it was used that \( xp_{n-1} - p_n \in \text{Pol}_{n-1}(\mathbb{R}) \) due to the fact that \( x^n \) is canceled. Finally, for an arbitrary \( q \in \text{Pol}_{n-2}(\mathbb{R}) \) (setting \( \text{Pol}_1(\mathbb{R}) := \{0\} \)), it holds that

\[ \langle q_{n+1}, q \rangle = \langle p_n, xq \rangle - \beta_n \langle p_n, q \rangle - \gamma_n^2 \langle p_{n-1}, q \rangle = 0, \quad (4.77c) \]

due to \( p_n \in \text{Pol}_{n-1}(\mathbb{R})^\perp \) and \( p_{n-1} \in \text{Pol}_{n-2}(\mathbb{R})^\perp \). Combining the equations (4.77) yields \( q_{n+1} \perp q \) for each \( q \in \text{span} \{\{p_n, p_{n-1}\} \cup \text{Pol}_{n-2}(\mathbb{R})\} = \text{Pol}_n(\mathbb{R}) \), establishing (4.76) and completing the proof.

**Remark 4.43.** The reason that (4.75) is more efficient than (4.73) lies in the fact that the computation of \( p_{n+1} \) according to (4.73) requires the computation of the \( n+1 \) scalar products \( \langle x_{n+1}, p_0 \rangle, \ldots, \langle x_{n+1}, p_n \rangle \), whereas the computation of \( p_{n+1} \) according to (4.75) merely requires the computation of the 2 scalar products \( \langle p_n, p_n \rangle \) and \( \langle xp_n, p_n \rangle \) occurring in \( \beta_n \) and \( \gamma_n \).

**Definition 4.44.** In the situation of Th. 4.42, we call the polynomials \( p_n, n \in \mathbb{N}_0 \), given by (4.75), the **orthogonal polynomials** with respect to \( \langle \cdot, \cdot \rangle \) (\( p_n \) is called the \( n \)th orthogonal polynomial).

**Example 4.45.** Considering \( \rho = 1 \) on \([-1, 1]\), one obtains \( \langle p, q \rangle_\rho = \int_{-1}^{1} p(x)q(x) \, dx \) and it is an exercise to check that the first four resulting orthogonal polynomials are

\[ p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = x^2 - \frac{1}{3}, \quad p_3(x) = x^3 - \frac{3}{5} x. \]

The Gaussian quadrature rules are based on polynomial interpolation with respect to the zeros of orthogonal polynomials. The following theorem provides information regarding such zeros. In general, the explicit computation of the zeros is a nontrivial task and a disadvantage of Gaussian quadrature rules (the price one has to pay for higher accuracy and better convergence properties as compared to Newton-Cotes rules).
Theorem 4.46. Given a real interval \([a, b]\), \(a < b\), and a weight function \(\rho : [a, b] \rightarrow \mathbb{R}_0^+\), let \(p_n\), \(n \in \mathbb{N}_0\), be the orthogonal polynomials with respect to \(\langle \cdot, \cdot \rangle_\rho\). For a fixed \(n \geq 1\), \(p_n\) has precisely \(n\) distinct zeros \(\lambda_1, \ldots, \lambda_n\), which all are simple and lie in the open interval \([a, b]\).

Proof. Let \(a < \lambda_1 < \lambda_2 < \cdots < \lambda_k < b\) be an enumeration of the zeros of \(p_n\) that lie in \([a, b]\) and have odd multiplicity (i.e. \(p_n\) changes sign at these \(\lambda_j\)). From \(p_n \in \text{Pol}_n(\mathbb{R})\), we know \(k \leq n\). The goal is to show \(k = n\). Seeking a contradiction, we assume \(k < n\). This implies

\[
q(x) := \prod_{i=1}^{k}(x - \lambda_i) \in \text{Pol}_{n-1}(\mathbb{R}).
\]

Thus, as \(p_n \in \text{Pol}_{n-1}(\mathbb{R})^\perp\), we get

\[
\langle p_n, q \rangle_\rho = 0. \tag{4.78}
\]

On the other hand, the polynomial \(p_nq\) has only zeros of even multiplicity on \([a, b]\), which means either \(p_nq \geq 0\) or \(p_nq \leq 0\) on the entire interval \([a, b]\), such that

\[
\langle p_n, q \rangle_\rho = \int_a^b p_n(x)q(x)\rho(x) \, dx \neq 0, \tag{4.79}
\]

in contradiction to (4.78). This shows \(k = n\), and, in particular, that all zeros of \(p_n\) are simple.

4.6.3 Gaussian Quadrature Rules

Definition 4.47. Given a real interval \([a, b]\), \(a < b\), a weight function \(\rho : [a, b] \rightarrow \mathbb{R}_0^+\), and \(n \in \mathbb{N}\), let \(\lambda_1, \ldots, \lambda_n \in [a, b]\) be the zeros of the \(n\)th orthogonal polynomial \(p_n\) with respect to \(\langle \cdot, \cdot \rangle_\rho\), and let \(L_1, \ldots, L_n\) be the corresponding Lagrange basis polynomials (cf. (3.10)), now written as polynomial functions, i.e.

\[
L_j(x) := \prod_{\substack{i=1 \\text{to } n \ \text{with} \ i \neq j}} \frac{x - \lambda_i}{\lambda_j - \lambda_i} \quad \text{for each } j \in \{1, \ldots, n\}, \tag{4.80}
\]

then the quadrature rule

\[
I_n : R[a, b] \rightarrow \mathbb{R}, \quad I_n(f) := \sum_{j=1}^{n} \sigma_j f(\lambda_j), \tag{4.81a}
\]

where

\[
\sigma_j := \langle L_j, 1 \rangle_\rho = \int_a^b L_j(x)\rho(x) \, dx \quad \text{for each } j \in \{1, \ldots, n\}, \tag{4.81b}
\]

is called the \(n\)th order Gaussian quadrature rule with respect to \(\rho\).
Theorem 4.48. Consider the situation of Def. 4.47; in particular let $I_n$ be the Gaussian quadrature rule defined according to (4.81).

(a) $I_n$ is exact for each polynomial of degree at most $2n - 1$. This can be stated in the form
\[
\langle p, 1 \rangle_\rho = \sum_{j=1}^{n} \sigma_j p(\lambda_j) \quad \text{for each } p \in \text{Pol}_{2n-1}(\mathbb{R}).
\]

Comparing with (4.9) and Rem. 4.9(b) shows that $I_n$ has degree of accuracy precisely $2n - 1$, i.e. the maximal possible degree of accuracy.

(b) All weights $\sigma_j$, $j \in \{1, \ldots, n\}$, are positive: $\sigma_j > 0$.

(c) For $\rho \equiv 1$, $I_n$ is based on interpolating polynomials for $\lambda_1, \ldots, \lambda_n$ in the sense of Def. 4.10.

Proof. (a): Recall from Linear Algebra that the remainder theorem [Phi19b, Th. 7.15] holds in the ring of polynomials $\mathbb{R}[X]$ and, thus, in $\text{Pol}(\mathbb{R}) \cong \mathbb{R}[X]$ (using that the field $\mathbb{R}$ is infinite). In particular, since $p_n$ has degree precisely $n$, given an arbitrary $p \in \text{Pol}_{2n-1}(\mathbb{R})$, there exist $q, r \in \text{Pol}_{n-1}(\mathbb{R})$ such that
\[
p = qp_n + r.
\]

Then, the relation $p_n(\lambda_j) = 0$ implies
\[
p(\lambda_j) = r(\lambda_j) \quad \text{for each } j \in \{1, \ldots, n\}.
\]

Since $r \in \text{Pol}_{n-1}(\mathbb{R})$, it must be the unique interpolating polynomial for the data
\[
(\lambda_1, r(\lambda_1)), \ldots, (\lambda_n, r(\lambda_n)).
\]

Thus,
\[
r(x) = \sum_{j=1}^{n} r(\lambda_j)L_j(x) \quad \text{(3.10)} = \sum_{j=1}^{n} p(\lambda_j)L_j(x).
\]

This allows to compute
\[
\langle p, 1 \rangle_\rho \overset{(4.82)}{=} \int_a^b (q(x)p_n(x) + r(x))\rho(x)\,dx = \langle q, p_n \rangle_\rho + \langle r, 1 \rangle_\rho
\]
\[
\overset{p_n \in \text{Pol}_{n-1}(\mathbb{R})^\perp}{=} \sum_{j=1}^{n} p(\lambda_j)\langle L_j, 1 \rangle_\rho \quad \overset{(4.81b)}{=} \sum_{j=1}^{n} \sigma_j p(\lambda_j),
\]

proving (4.82).

(b): For each $j \in \{1, \ldots, n\}$, we apply (4.82) to $L_j^2 \in \text{Pol}_{2n-2}(\mathbb{R})$ to obtain
\[
0 < \|L_j\|_\rho^2 = \langle L_j^2, 1 \rangle_\rho \overset{(4.82)}{=} \sum_{k=1}^{n} \sigma_k L_j^2(\lambda_k) \quad \overset{(3.10)}{=} \sigma_j.
\]
(c): This follows from (a): If \( f : [a, b] \rightarrow \mathbb{R} \) is given and \( p \in \text{Pol}_{n-1}(\mathbb{R}) \) is the interpolating polynomial function for the data 
\[ (\lambda_1, f(\lambda_1)), \ldots, (\lambda_n, f(\lambda_n)), \]
then 
\[ I_n(f) = \sum_{j=1}^{n} \sigma_j f(\lambda_j) = \sum_{j=1}^{n} \sigma_j p(\lambda_j) = \langle p, 1 \rangle_1 = \int_a^b p(x) \, dx, \]
which establishes the case.  

\[ \blacksquare \]

**Theorem 4.49.** The converse of Th. 4.48(a) is also true: If, for \( n \in \mathbb{N}, \lambda_1, \ldots, \lambda_n \in \mathbb{R} \) and \( \sigma_1, \ldots, \sigma_n \in \mathbb{R} \) are such that (4.82) holds, then \( \lambda_1, \ldots, \lambda_n \) must be the zeros of the \( n \)th orthogonal polynomial \( p_n \) and \( \sigma_1, \ldots, \sigma_n \) must satisfy (4.81b).

**Proof.** We first verify \( q = p_n \) for \( q : \mathbb{R} \rightarrow \mathbb{R}, \quad q(x) := \prod_{j=1}^{n} (x - \lambda_j). \) (4.87)

To that end, let \( m \in \{0, \ldots, n - 1\} \) and apply (4.82) to the polynomial \( p(x) := x^m q(x) \) (note \( p \in \text{Pol}_{n+m}(\mathbb{R}) \subseteq \text{Pol}_{2n-1}(\mathbb{R}) \)) to obtain
\[ \langle q, x^m \rangle_\rho = \langle x^m q, 1 \rangle_\rho = \sum_{j=1}^{n} \sigma_j \lambda_j^m q(\lambda_j) = 0, \]
showing \( q \in \text{Pol}_{n-1}(\mathbb{R})^\perp \) and \( q - p_n \in \text{Pol}_{n-1}(\mathbb{R})^\perp \). On the other hand, both \( q \) and \( p_n \) have degree precisely \( n \) and the coefficient of \( x^n \) is 1 in both cases. Thus, in \( q - p_n, \) \( x^n \) cancels, showing \( q - p_n \in \text{Pol}_{n-1}(\mathbb{R}) \). Since \( \text{Pol}_{n-1}(\mathbb{R}) \cap \text{Pol}_{n-1}(\mathbb{R})^\perp = \{0\} \), we have established \( q = p_n \), thereby also identifying the \( \lambda_j \) as the zeros of \( p_n \) as claimed. Finally, applying (4.82) with \( p = L_j \) yields
\[ \langle L_j, 1 \rangle_\rho = \sum_{k=1}^{n} \sigma_j L_j(\lambda_k) = \sigma_j, \]
showing that the \( \sigma_j \), indeed, satisfy (4.81b).  

\[ \blacksquare \]

**Theorem 4.50.** For each \( f \in C[a, b] \), the Gaussian quadrature rules \( I_n(f) \) of Def. 4.47 converge:
\[ \lim_{n \to \infty} I_n(f) = \int_a^b f(x) \rho(x) \, dx \quad \text{for each} \ f \in C[a, b]. \] (4.88)

**Proof.** The convergence is a consequence of Polya’s Th. 4.28: Condition (i) of Th. 4.28 is satisfied as \( I_n(p) \) is exact for each \( p \in \text{Pol}_{2n-1}(\mathbb{R}) \) and Condition (ii) of Th. 4.28 is satisfied since the positivity of the weights yields
\[ \sum_{j=1}^{n} |\sigma_{j,n}| = \sum_{j=1}^{n} \sigma_{j,n} = I_n(1) = \int_a^b \rho(x) \, dx \in \mathbb{R}^+ \] (4.89)
for each \( n \in \mathbb{N} \).  

\[ \blacksquare \]
Remark 4.51. (a) Using the linear change of variables
\[ x \mapsto u = u(x) := \frac{b - a}{2} x + \frac{a + b}{2}, \]
we can transform an integral over \([a, b]\) into an integral over \([-1, 1] \):
\[
\int_a^b f(u) \, du = \frac{b - a}{2} \int_{-1}^1 f(u(x)) \, dx = \frac{b - a}{2} \int_{-1}^1 f \left( \frac{b - a}{2} x + \frac{a + b}{2} \right) \, dx. \tag{4.90}
\]
This is useful in the context of Gaussian quadrature rules, since the computation of the zeros \(\lambda_j\) of the orthogonal polynomials is, in general, not easy. However, due to (4.90), one does not have to recompute them for each interval \([a, b]\), but one can just compute them once for \([-1, 1]\) and tabulate them. Of course, one still needs to recompute for each \(n \in \mathbb{N}\) and each new weight function \(\rho\).

(b) If one is given the \(\lambda_j\) for a large \(n\) with respect to some strange \(\rho > 0\), and one wants to exploit the resulting Gaussian quadrature rule to just compute the nonweighted integral of \(f\), one can obviously always do that by applying the rule to \(g := f/\rho\) instead of to \(f\).

As usual, to obtain an error estimate, one needs to assume more regularity of the integrated function \(f\).

Theorem 4.52. Consider the situation of Def. 4.47; in particular let \(I_n\) be the Gaussian quadrature rule defined according to (4.81). Then, for each \(f \in C^{2n}[a, b]\), there exists \(\tau \in [a, b]\) such that
\[
\int_a^b f(x) \rho(x) \, dx - I_n(f) = \int_a^b \frac{f^{(2n)}(\tau)}{(2n)!} \, dx \int_a^b p_n^2(x) \rho(x) \, dx, \tag{4.91}
\]
where \(p_n\) is the \(n\)th orthogonal polynomial, which can be further estimated by
\[
|p_n(x)| \leq (b - a)^n \quad \text{for each} \ x \in [a, b]. \tag{4.92}
\]

Proof. As before, let \(\lambda_1, \ldots, \lambda_n\) be the zeros of the \(n\)th orthogonal polynomial \(p_n\). Given \(f \in C^{2n}[a, b]\), let \(q \in \text{Pol}_{2n-1}(\mathbb{R})\) be the Hermite interpolating polynomial for the \(2n\) points
\[ \lambda_1, \lambda_1, \lambda_2, \lambda_2, \ldots, \lambda_n, \lambda_n. \]
The identity (3.29) yields
\[
f(x) = q(x) + [f | x, \lambda_1, \lambda_1, \lambda_2, \lambda_2, \ldots, \lambda_n, \lambda_n] \prod_{j=1}^n (x - \lambda_j)^2. \tag{4.93}
\]
Now we multiply (4.93) by \(\rho\), replace the generalized divided difference by using the mean value theorem for generalized divided differences (3.30), and integrate over \([a, b]\):
\[
\int_a^b f(x) \rho(x) \, dx = \int_a^b q(x) \rho(x) \, dx + \frac{1}{(2n)!} \int_a^b f^{(2n)}(\xi(x)) p_n^2(x) \rho(x) \, dx. \tag{4.94}
\]
Note that, as in the proof of Prop. 4.14, the map \( x \mapsto f^{(2n)}(\xi(x)) \) can be chosen to be continuous (in particular, integrable), and this choice is assumed here. Since \( p_n^2 \rho \geq 0 \), we can apply Lem. 4.13 to obtain \( \tau \in [a, b] \) satisfying
\[
\int_a^b f(x) \rho(x) \, dx = \int_a^b q(x) \rho(x) \, dx + \frac{f^{(2n)}(\tau)}{(2n)!} \int_a^b p_n^2(x) \rho(x) \, dx.
\] (4.95)
Since \( q \in \text{Pol}_{2n-1}(\mathbb{R}) \), \( I_n \) is exact for \( q \), i.e.
\[
\int_a^b q(x) \rho(x) \, dx = I_n(q) = \sum_{j=1}^n \sigma_j q(\lambda_j) = \sum_{j=1}^n \sigma_j f(\lambda_j) = I_n(f).
\] (4.96)
Combining (4.96) and (4.95) proves (4.91). Finally, (4.92) is clear from the representation \( p_n(x) = \prod_{j=1}^n (x - \lambda_j) \) and \( \lambda_j \in [a, b] \) for each \( j \in \{1, \ldots, n\} \).

**Remark 4.53.** One can obviously also combine the strategies of Sec. 4.5 and Sec. 4.6. The results are *composite Gaussian quadrature rules*.

## 5 Numerical Solution of Linear Systems

### 5.1 Background and Setting

**Definition 5.1.** Let \( F \) be a field. Given an \( m \times n \) matrix \( A \in \mathcal{M}(m, n, F) \), \( m, n \in \mathbb{N} \), and \( b \in \mathcal{M}(m, 1, F) \cong F^m \), the equation
\[
Ax = b
\] (5.1)
is called a *linear system* for the unknown \( x \in \mathcal{M}(n, 1, F) \cong F^n \). The matrix one obtains by adding \( b \) as the \((n + 1)\text{th}\) column to \( A \) is called the *augmented matrix* of the linear system. It is denoted by \((A|b)\). The linear system (5.1) is called *homogeneous* for \( b = 0 \) and *inhomogeneous* for \( b \neq 0 \). By
\[
\mathcal{L}(A|b) := \{ x \in F^n : Ax = b \},
\] (5.2)
we denote the set of solutions to (5.1).

We have already learned in Linear Algebra how to determine the set of solutions \( \mathcal{L}(A|b) \) for linear systems in echelon form (cf. [Phi19a, Def. 8.7]) using back substitution (cf. [Phi19a, Sec. 8.3.1]), and we have learned how to transform a general linear system into echelon form without changing the set of solutions using the Gaussian elimination algorithm (cf. [Phi19a, Sec. 8.3.3]). Moreover, we have seen how to augment the Gaussian elimination algorithm to obtain a so-called *LU decomposition* for a linear system, which is more efficient if one needs to solve a linear system \( Ax = b \), repeatedly, with fixed \( A \) and varying \( b \) (cf. [Phi19a, Sec. 8.4], in particular, [Phi19a, Sec. 8.4.4]). We recall
from [Phi19a, Th. 8.34] that, for each $A \in \mathcal{M}(m, n, F)$, there exists a permutation matrix $P \in \mathcal{M}(m, F)$ (which contains precisely one 1 in each row and one 1 in each column and only zeros, otherwise), a unipotent lower triangular matrix $L \in \mathcal{M}(m, F)$ and $\tilde{A} \in \mathcal{M}(m, n, F)$ in echelon form such that

$$PA = L\tilde{A}.$$  \hspace{1cm} (5.3)

It is an LU decomposition in the strict sense (i.e. with $\tilde{A}$ being upper triangular) for $A$ being quadratic (i.e. $m \times m$).

**Remark 5.2.** Let $F$ be a field. We recall from Linear Algebra how to make use of an LU decomposition $PA = L\tilde{A}$ to solve linear systems $Ax = b$ with fixed matrix $A \in \mathcal{M}(m, n, F)$ and varying $b \in F^n$. One has

$$x \in L(A|b) \iff PAx = L\tilde{Ax} = Pb \iff \tilde{Ax} = L^{-1}Pb \iff x \in L(\tilde{A}|L^{-1}Pb).$$  

Thus, solving $Ax = b$ is equivalent to solving $Lz = Pb$ for $z$ and then solving $\tilde{Ax} = z = L^{-1}Pb$ for $x$. Since $\tilde{A}$ is in echelon form, $\tilde{Ax} = z = L^{-1}Pb$ can be solved via back substitution; and, since $L$ is lower triangular, $Lz = Pb$ can be solved analogously via forward substitution.

Gaussian elimination and LU decomposition have the advantage that, at least in principle, they work over every field $F$ and for every matrix $A$. The algorithm is still rather simple and, to the best of my knowledge, still the most efficient one known in the absence of any further knowledge regarding $F$ and $A$. However, in the presence of roundoff errors, Gaussian elimination can become numerically unstable and so-called pivot strategies can help to improve the situation (cf. Sec. 5.2 below). Numerically, it can also be an issue that the condition numbers of $L$ and $U$ in an LU decomposition $A = LU$ can be much larger than the condition number of $A$ (cf. Rem. 5.15 below), in which case, the use of the strategy of Rem. 5.2 is numerically ill-advised. If $F = \mathbb{K}$, then the so-called QR decomposition $A = QR$ of Sec. 5.4 below is usually a numerically better option. Moreover, if $A$ has a special structure, such as being Hermitian or sparse (i.e. having many 0 entries), than it is usually much more efficient (and, thus, advisable) to use algorithms designed to exploit this special structure. In this class, we will not have time to consider algorithms designed for sparse matrices, but we will present the Cholesky decomposition for Hermitian matrices in Sec. 5.3 below.

### 5.2 Pivot Strategies for Gaussian Elimination

The Gaussian elimination algorithm can be numerically unstable even for matrices with small condition number, as demonstrated in the following example.

**Example 5.3.** Consider the linear system $Ax = b$ over $\mathbb{R}$ with

$$A := \begin{pmatrix} 0.0001 & 1 \\ 1 & 1 \end{pmatrix}, \quad b := \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \hspace{1cm} (5.4)$$
The exact solution is
\[ x = \begin{pmatrix} 10^4 \\ 9999 \\ 9999 \\ 9999 \end{pmatrix} = \begin{pmatrix} 1.00010001 \\ 0.99989999 \ldots \end{pmatrix}. \] (5.5)

We now examine what occurs if we solve the system approximately using the Gaussian elimination algorithm and a floating point arithmetic, rounding to 3 significant digits:

When applying the Gaussian elimination algorithm to \((A|b)\), we replace the second row by the sum of the second row and the first row multiplied by \(-10^4\). The exact result is
\[ \begin{pmatrix} 10^{-4} \\ 1 \\ 10^{-4} \\ -9999 \\ -9998 \end{pmatrix}. \]

However, when rounding to 3 digits, it is replaced by
\[ \begin{pmatrix} 10^{-4} \\ 1 \\ 10^{-4} \\ -10^4 \\ -10^4 \end{pmatrix}, \]
yielding the approximate solution
\[ x = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \] (5.6)

Thus, the relative error in the solution is about 10000 times the relative error of rounding. The condition number is low, namely \(\kappa_2(A) \approx 2.618\) with respect to the Euclidean norm. The error should not be amplified by a factor of 10000 if the algorithm is stable.

Now consider what occurs if we first switch the rows of \((A|b)\):
\[ \begin{pmatrix} 1 \\ 1 \\ 10^{-4} \\ 1 \\ 1 \end{pmatrix}. \]

Now Gaussian elimination and rounding yields
\[ \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} \]
with the approximate solution
\[ x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \] (5.7)

which is much better than (5.6).

---

Example 5.3 shows that it can be advantageous to switch rows even if one could proceed Gaussian elimination without switching. Similarly, from a numerical point of view, not every choice of the row number \(i\), used for switching, is equally good. It is advisable to use what is called a pivot strategy – a strategy to determine if, in the Gaussian elimination’s \(k\)th step, one should first switch the current row with another row, and, if so, which other row should be used.
A first strategy that can be used to avoid instabilities of the form that occurred in Example 5.3 is the column maximum strategy (cf. Def. 5.5 below): In the $k$th step of Gaussian elimination (i.e. when eliminating in the $k$th column), find the row number $i \geq r(k)$ (all rows with numbers less than $r(k)$ are already in echelon form and remain unchanged, cf. [Phi19a, Sec. 8.3.3]), where the pivot element (i.e. the row’s first nonzero element) has the maximal absolute value; if $i \neq r(k)$, then switch rows $i$ and $r(k)$ before proceeding. This was the strategy that resulted in the acceptable solution (5.7). However, observe what happens in the following example:

**Example 5.4.** Consider the linear system $Ax = b$ with

$$
A := \begin{pmatrix} 1 & 10000 \\ 1 & 1 \end{pmatrix}, \quad b := \begin{pmatrix} 10000 \\ 2 \end{pmatrix}.
$$

The exact solution is the same as in Example 5.3, i.e. given by (5.5) (the first row of Example 5.3 has merely been multiplied by 10000). Again, we examine what occurs if we solve the system approximately using Gaussian elimination, rounding to 3 significant digits:

As the pivot element of the first row is maximal, the column maximum strategy does not require any switching. After the first step we obtain

$$
\begin{pmatrix} 1 & 10000 & 10000 \\ 0 & -9999 & -9998 \end{pmatrix},
$$

which, after rounding to 3 digits, is replaced by

$$
\begin{pmatrix} 1 & 10000 & 10000 \\ 0 & -10000 & -10000 \end{pmatrix},
$$

yielding the unsatisfactory approximate solution

$$
\begin{pmatrix} 0 \\ 1 \end{pmatrix}.
$$

The problem here is that the pivot element of the first row is small as compared with the other elements of that row! This leads to the so-called relative column maximum strategy (cf. Def. 5.5 below), where, instead of choosing the pivot element with the largest absolute value, one chooses the pivot element such that its absolute value divided by the sum of the absolute values of all elements in that row is maximized.

In the above case, the relative column maximum strategy would require first switching rows:

$$
\begin{pmatrix} 1 & 1 & 2 \\ 1 & 10000 & 10000 \end{pmatrix}.
$$

Now Gaussian elimination and rounding yields

$$
\begin{pmatrix} 1 & 1 & 2 \\ 0 & 10000 & 10000 \end{pmatrix}.$$
once again with the acceptable approximate solution
\[ x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \]

Both pivot strategies discussed above are summarized in the following definition:

**Definition 5.5.** Let \( A = (a_{ij}) \in \mathcal{M}(m, n, \mathbb{K}), m, n \in \mathbb{N}, \) and \( b \in \mathbb{K}^m. \) We define two modified versions of the Gaussian elimination algorithm [Phi19a, Alg. 8.17] by adding one of the following actions (0) or (0’) at the beginning of the algorithm’s \( k \)th step (for each \( k \geq 1 \) such that \( r(k) < m \) and \( k \leq n \)).

As in [Phi19a, Alg. 8.17], let \( (A^{(1)}|b^{(1)}) := (A|b), r(1) := 1, \) and, for each \( k \geq 1 \) such that \( r(k) < m \) and \( k \leq n, \) transform \( (A^{(k)}|b^{(k)}) \) into \( (A^{(k+1)}|b^{(k+1)}) \) and \( r(k) \) into \( r(k+1). \) In contrast to [Phi19a, Alg. 8.17], to achieve the transformation, first perform either (0) or (0’):

(0) **Column Maximum Strategy:** Determine \( i \in \{r(k), \ldots, m\} \) such that
\[ \left| a_{ik}^{(k)} \right| = \max \left\{ \left| a_{\alpha k}^{(k)} \right| : \alpha \in \{r(k), \ldots, m\} \right\}. \]  

(0’) **Relative Column Maximum Strategy:** For each \( \alpha \in \{r(k), \ldots, m\}, \) define
\[ S_{\alpha}^{(k)} := \sum_{\beta=k}^{n} \left| a_{\alpha \beta}^{(k)} \right| + \left| b_{\alpha}^{(k)} \right|. \]  

If \( \max \left\{ S_{\alpha}^{(k)} : \alpha \in \{r(k), \ldots, m\} \right\} = 0, \) then \( (A^{(k)}|b^{(k)}) \) is already in echelon form and the algorithm is halted. Otherwise, determine \( i \in \{r(k), \ldots, m\} \) such that
\[ \frac{a_{ik}^{(k)}}{S_{i}^{(k)}} = \max \left\{ \frac{a_{i \alpha}^{(k)}}{S_{\alpha}^{(k)}} : \alpha \in \{r(k), \ldots, m\}, S_{\alpha}^{(k)} \neq 0 \right\}. \]

If \( a_{ik}^{(k)} = 0, \) then nothing needs to be done in the \( k \)th step and one merely proceeds to the next column (if any) without changing the row by setting \( r(k+1) := r(k) \) (cf. [Phi19a, Alg. 8.17(c)]). If \( a_{ik}^{(k)} \neq 0 \) and \( i = r(k), \) then no row switching is needed before elimination in the \( k \)th column (cf. [Phi19a, Alg. 8.17(a)]). If \( a_{ik}^{(k)} \neq 0 \) and \( i \neq r(k), \) then one proceeds by switching rows \( i \) and \( r(k) \) before elimination in the \( k \)th column (cf. [Phi19a, Alg. 8.17(b)]).

**Remark 5.6.** (a) Clearly, [Phi19a, Th. 8.19] remains valid for the Gaussian elimination algorithm with column maximum strategy as well as with relative column maximum strategy (i.e. the modified Gaussian elimination algorithm still transforms the augmented matrix \( (A|b) \) of the linear system \( Ax = b \) into a matrix \( (\hat{A}|\hat{b}) \) in echelon form, such that the linear system \( \hat{A}x = \hat{b} \) has precisely the same set of solutions as the original system).

(b) The relative maximum strategy is more stable in cases where the order of magnitude of matrix elements varies strongly, but it is also less efficient.
5.3 Cholesky Decomposition

Hermitian positive definite matrices $A \in \mathcal{M}(n, \mathbb{K})$ (and also Hermitian positive semidefinite matrices, see Appendix I.1) can be decomposed in the particularly simple form $A = LL^*$, where $L \in \mathcal{M}(n, \mathbb{K})$ is lower triangular:

**Theorem 5.7.** If $A \in \mathcal{M}(n, \mathbb{K})$ is Hermitian and positive definite (cf. Def. 2.21), then there exists a unique lower triangular $L \in \mathcal{M}(n, \mathbb{K})$ with positive diagonal entries (i.e. with $l_{jj} \in \mathbb{R}^+$ for each $j \in \{1, \ldots, n\}$) and

$$A = LL^*. \quad (5.10)$$

This decomposition is called the **Cholesky decomposition** of $A$.

**Proof.** The proof is conducted via induction on $n$. For $n = 1$, it is $A = (a_{11})$, $a_{11} > 0$, and $L = (l_{11})$ is uniquely given by the square root of $a_{11}$: $l_{11} = \sqrt{a_{11}} > 0$. For $n > 1$, we write $A$ in block form

$$A = \begin{pmatrix} B & v \\ v^* & a_{nn} \end{pmatrix}, \quad (5.11)$$

where $B \in \mathcal{M}(n - 1, \mathbb{K})$ is the matrix with $b_{kl} = a_{kl}$ for each $k, l \in \{1, \ldots, n - 1\}$ and $v \in \mathbb{K}^{n-1}$ is the column vector with $v_k = a_{kn}$ for each $k \in \{1, \ldots, n - 1\}$. According to Prop. C.4, $B$ is positive definite (and, clearly, $B$ is also Hermitian) and $a_{nn} > 0$. According to the induction hypothesis, there exists a unique lower triangular matrix $L' \in \mathcal{M}(n - 1, \mathbb{K})$ with positive diagonal entries and $B = L'(L')^*$. Then $L'$ is invertible and we can define

$$w := (L')^{-1} v \in \mathbb{K}^{n-1}. \quad (5.12)$$

Moreover, let

$$\alpha := a_{nn} - w^* w. \quad (5.13)$$

Then $\alpha \in \mathbb{R}$ and there exists $\beta \in \mathbb{C}$ such that $\beta^2 = \alpha$. We set

$$L := \begin{pmatrix} L' & 0 \\ w^* & \beta \end{pmatrix}. \quad (5.14)$$

Then $L$ is lower triangular,

$$L^* = \begin{pmatrix} (L')^* & w \\ 0 & \beta \end{pmatrix}, \quad (5.15)$$

and

$$LL^* = \begin{pmatrix} L'(L')^* & L'w \\ w^*(L')^* & w^*w + \beta^2 \end{pmatrix} = \begin{pmatrix} B & v \\ v^* & a_{nn} \end{pmatrix} = A.$$
It only remains to be shown one can choose $\beta \in \mathbb{R}^+$ and that this choice determines $\beta$ uniquely. Since $L'$ is a triangular matrix with positive entries on its diagonal, it is $\det L' \in \mathbb{R}^+$ and
\[
\det A = (\det L')^2 \beta^2 > 0. \tag{5.16}
\]
Thus, $\alpha = \beta^2 > 0$, and one can choose $\beta = \sqrt{\alpha} \in \mathbb{R}^+$ and $\sqrt{\alpha}$ is the only positive number with square $\alpha$.

**Remark 5.8.** From the proof of Th. 5.7, it is not hard to extract a recursive algorithm to actually compute the Cholesky decomposition of an Hermitian positive definite $A \in \mathcal{M}(n, \mathbb{K})$: One starts by setting
\[
l_{11} := \sqrt{a_{11}}, \tag{5.17a}
\]
where the proof of Th. 5.7 guarantees $a_{11} > 0$. In the recursive step, one already has constructed the entries for the first $r - 1$ rows of $L$, $1 < r \leq n$ (corresponding to the entries of $L'$ in the proof of Th. 5.7), and one needs to construct $l_{r1}, \ldots, l_{rr}$. Using (5.12), one has to solve $L'w = v$ for
\[
w = \begin{pmatrix} l_{r1} \\ \vdots \\ l_{r,r-1} \end{pmatrix}, \quad \text{where} \quad v = \begin{pmatrix} a_{1r} \\ \vdots \\ a_{r-1,r} \end{pmatrix} \quad \text{(then} \ w^* = (l_{r1} \ldots l_{r,r-1}))
\]
As $L'$ has lower triangular form, we obtain $l_{r1}, \ldots, l_{r,r-1}$ successively, using forward substitution, where
\[
l_{rj} := \frac{a_{jr} - \sum_{k=1}^{j-1} l_{jk}l_{rk}}{l_{jj}} \quad \text{for} \quad j \in \{1, \ldots, r - 1\} \tag{5.17b}
\]
and the proof of Th. 5.7 guarantees $l_{jj} > 0$. Finally, one uses (5.13) (with $n$ replaced by $r$) to obtain
\[
l_{rr} = \sqrt{\alpha} = \sqrt{a_{rr} - w^*w} = \sqrt{a_{rr} - \sum_{k=1}^{r-1} |l_{rk}|^2}, \tag{5.17c}
\]
where the proof of Th. 5.7 guarantees $\alpha > 0$.

**Example 5.9.** We compute the Cholesky decomposition for the symmetric positive definite matrix
\[
A = \begin{pmatrix} 1 & -1 & -2 \\ -1 & 2 & 3 \\ -2 & 3 & 9 \end{pmatrix}.
\]
According to the algorithm described in Rem. 5.8, we compute

\[
\begin{align*}
l_{11} &= \sqrt{a_{11}} = \sqrt{1} = 1, \\
l_{21} &= \frac{a_{12}}{l_{11}} = \frac{-1}{1} = -1, \\
l_{22} &= \sqrt{a_{22} - l_{21}^2} = \sqrt{2 - 1} = 1, \\
l_{31} &= \frac{a_{13}}{l_{11}} = \frac{-2}{1} = -2, \\
l_{32} &= \frac{a_{23} - l_{21}l_{31}}{l_{22}} = \frac{3 - (-1)(-2)}{1} = 1, \\
l_{33} &= \sqrt{a_{33} - l_{31}^2 - l_{32}^2} = \sqrt{9 - 4 - 1} = 2.
\end{align*}
\]

Thus, the Cholesky decomposition for \(A\) is

\[
LL^* = L^*L = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -2 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 2 \\ -1 & 2 & 3 \\ -2 & 3 & 9 \end{pmatrix} = A.
\]

We conclude the section by showing that a Cholesky decomposition can not exist if \(A\) is not symmetric positive semidefinite:

**Proposition 5.10.** For \(n \in \mathbb{N}\), let \(A, L \in \mathcal{M}(n, \mathbb{K})\), where \(A = LL^*\). Then \(A\) is Hermitian and positive semidefinite.

**Proof.** \(A\) is Hermitian, since

\[
A^* = (LL^*)^* = LL^* = A.
\]

\(A\) is positive semidefinite, since

\[
\forall x \in \mathbb{K}^n, \quad x^*Ax = x^*LL^*x = (L^*x)^*(L^*x) \geq 0,
\]

which completes the proof.

\[\blacksquare\]

## 5.4 QR Decomposition

### 5.4.1 Definition and Motivation

**Definition 5.11.** Let \(n \in \mathbb{N}\) and let \(A \in \mathcal{M}(n, \mathbb{K})\). Then \(A\) is called *unitary* if, and only if, \(A\) is invertible and \(A^{-1} = A^*\). Moreover, for \(\mathbb{K} = \mathbb{R}\), unitary matrices are also called *orthogonal*.

**Definition 5.12.** Let \(m, n \in \mathbb{N}\) and \(A \in \mathcal{M}(m, n, \mathbb{K})\).
(a) A decomposition

\[ A = Q \tilde{A} \]  \hspace{1cm} (5.18a)

is called a QR decomposition of \( A \) if, and only if, \( Q \in \mathcal{M}(m, \mathbb{K}) \) is unitary and \( \tilde{A} \in \mathcal{M}(m, n, \mathbb{K}) \) is in echelon form. If \( A \in \mathcal{M}(m, \mathbb{K}) \) (i.e. quadratic), then \( \tilde{A} \) is upper (i.e. right) triangular, i.e. (5.18a) is a QR decomposition in the strict sense, which is emphasized by writing \( R := \tilde{A} \):

\[ A = QR. \]  \hspace{1cm} (5.18b)

(b) Let \( r := \text{rk}(A) \) be the rank of \( A \). A decomposition

\[ A = Q \tilde{A} \]  \hspace{1cm} (5.19a)

is called an economy size QR decomposition of \( A \) if, and only if, \( Q \in \mathcal{M}(m, r, \mathbb{K}) \) is such that its columns form an orthonormal system with respect to the standard scalar product on \( \mathbb{K}^m \) and \( \tilde{A} \in \mathcal{M}(r, n, \mathbb{K}) \) is in echelon form. If \( r = n \), then \( \tilde{A} \) is upper triangular, i.e. (5.19a) is an economy size QR decomposition in the strict sense, which is emphasized by writing \( R := \tilde{A} \):

\[ A = QR. \]  \hspace{1cm} (5.19b)

Note: In the German literature, the economy size QR decomposition is usually called just QR decomposition, while our QR decomposition is referred to as the extended QR decomposition.

While, even for an invertible \( A \in \mathcal{M}(n, \mathbb{K}) \), the QR decomposition is not unique, the nonuniqueness can be precisely described:

**Proposition 5.13.** Suppose \( Q_1, Q_2, R_1, R_2 \in \mathcal{M}(n, \mathbb{K}), n \in \mathbb{N} \), are such that \( Q_1, Q_2 \) are unitary, \( R_1, R_2 \) are invertible upper triangular, and

\[ Q_1 R_1 = Q_2 R_2. \]  \hspace{1cm} (5.20)

Then there exists a diagonal matrix \( S \in \mathcal{M}(n, \mathbb{K}), S = \text{diag}(\sigma_1, \ldots, \sigma_n), |\sigma_1| = \cdots = |\sigma_n| = 1 \) (for \( \mathbb{K} = \mathbb{R} \), this means \( \sigma_1, \ldots, \sigma_n \in \{-1, 1\} \)), such that

\[ Q_2 = Q_1 S \quad \text{and} \quad R_2 = S^* R_1. \]  \hspace{1cm} (5.21)

**Proof.** As (5.20) implies \( S := Q_1^{-1} Q_2 = R_1 R_2^{-1} \), \( S \) must be both unitary and upper triangular. Thus, \( S \) must be diagonal, \( S = \text{diag}(\sigma_1, \ldots, \sigma_n) \) with \( |\sigma_k| = 1 \).

**Proposition 5.14.** Let \( n \in \mathbb{N} \) and \( A, Q \in \mathcal{M}(n, \mathbb{K}), A \) being invertible and \( Q \) being unitary. With respect to the 2-norm \( \| \cdot \|_2 \) on \( \mathbb{K}^n \), the following holds (recall Def. 2.12 of a natural matrix norm and Def. and Rem. 2.36 for the condition of a matrix):

\[ \|QA\|_2 = \|AQ\|_2 = \|A\|_2, \]

\[ \|Q\|_2 = \kappa_2(Q) = 1, \]

\[ \kappa_2(QA) = \kappa_2(AQ) = \kappa_2(A). \]
Proof. Exercise. □

Remark 5.15. Suppose the goal is to solve linear systems \( Ax = b \) with fixed \( A \in \mathcal{M}(m,n,K) \) and varying \( b \in K^n \). In Rem. 5.2, we described how to make use of a given LU decomposition \( PA = L\tilde{A} \) in this situation. Even though the strategy of Rem. 5.2 works fine in many situations, it can fail numerically due to the following stability issue: While the condition numbers, for invertible \( A \), always satisfy
\[
\kappa(PA) \leq \kappa(L)\kappa(\tilde{A}) \tag{5.22}
\]
due to (2.23) and (2.14), the right-hand side of (5.22) can be much larger than the left-hand side. In that case, it is numerically ill-advised to solve \( Lz = Pb \) and \( \tilde{A}x = z \) instead of solving \( Ax = b \). An analogous procedure can be used given a QR decomposition \( A = QA \). Solving \( Ax = QAx = b \) is equivalent to solving \( Qz = b \) for \( z \) and then solving \( Ax = z \) for \( x \). Due to \( Q^{-1} = Q^\ast \), solving for \( z = Q^\ast b \) is particularly simple. Moreover, one does not encounter the possibility of an increased condition as for the LU decomposition: Due to Prop. 5.14, the condition of using the QR decomposition is exactly the same as the condition of \( A \). Thus, if available, the QR decomposition should always be used!

5.4.2 QR Decomposition via Gram-Schmidt Orthogonalization

As just noted in Rem. 5.15, it is numerically advisable to make use of a QR decomposition if it is available. However, so far we have not addressed the question of how to obtain such a decomposition. The simplest way, but not the most numerically stable, is to use Gram-Schmidt orthogonalization. More precisely, Gram-Schmidt orthogonalization provides the economy size QR decomposition of Def. 5.12(b) as described in Th. 5.16. In Sec. 5.4.3 below, we will study the Householder method, which is more numerically stable and also provides the full QR decomposition directly.

Theorem 5.16. Let \( A \in \mathcal{M}(m,n,K) \), \( m,n \in \mathbb{N} \), with columns \( x_1, \ldots, x_n \). If \( r := \text{rk}(A) > 0 \), then define increasing functions
\[
\rho : \{1, \ldots, n\} \to \{0, \ldots, r\}, \quad \rho(k) := \dim(\text{span}\{x_1, \ldots, x_k\}), \tag{5.23a}
\mu : \{1, \ldots, r\} \to \{1, \ldots, n\}, \quad \mu(k) := \min\{j \in \{1, \ldots, n\} : \rho(j) = k\}. \tag{5.23b}
\]

(a) There exists a QR decomposition of \( A \) as well as, for \( r > 0 \), an economy size QR decomposition of \( A \). More precisely, there exist a unitary \( Q \in \mathcal{M}(m,K) \), an \( \tilde{A} \in \mathcal{M}(m,n,K) \) in echelon form, a \( Q_e \in \mathcal{M}(m,r,K) \) (for \( r > 0 \)) such that its columns form an orthonormal system with respect to the standard scalar product \( \langle \cdot, \cdot \rangle \) on \( K^m \), and (for \( r > 0 \)) an \( \tilde{A}_e \in \mathcal{M}(r,n,K) \) in echelon form such that
\[
A = Q\tilde{A}, \tag{5.24a}
A = Q_e\tilde{A}_e \quad \text{for } r > 0. \tag{5.24b}
\]
(i) For \( r > 0 \), the columns \( q_1, \ldots, q_r \in \mathbb{K}^m \) of \( Q_e \) can be computed from the columns \( x_1, \ldots, x_n \in \mathbb{K}^m \) of \( \tilde{A} \), obtaining \( v_1, \ldots, v_n \in \mathbb{K}^m \) from
\[
v_1 := x_1, \quad \forall \ k \geq 2, v_k := x_k - \sum_{i=1, v_i \neq 0}^{k-1} \frac{\langle x_k, v_i \rangle}{\|v_i\|^2} v_i, \tag{5.25}
\]
letting, for each \( k \in \{1, \ldots, r\} \), \( v_k := v_{\mu(k)} \) (i.e. the \( \tilde{v}_1, \ldots, \tilde{v}_r \) are the \( v_1, \ldots, v_n \) with each \( v_k = 0 \) omitted) and, finally, letting \( q_k := v_k / \|v_k\|_2 \) for each \( k = 1, \ldots, r \).

(ii) For \( r > 0 \), the entries \( r_{ik} \) of \( \tilde{A}_e \in \mathcal{M}(r, n, \mathbb{K}) \) can be defined by
\[
r_{ik} := \begin{cases} \|v_1\|_2 & \text{for } i = k = 1, \rho(1) = 1, \\ \langle x_i, q_i \rangle & \text{for } k \in \{2, \ldots, n\}, i < \rho(k), \\ \langle x_k, q_i \rangle & \text{for } k \in \{2, \ldots, n\}, i = \rho(k) = \rho(k-1), \\ \|\tilde{v}_{\rho(k)}\|_2 & \text{for } k \in \{2, \ldots, n\}, i = \rho(k) > \rho(k-1), \\ 0 & \text{otherwise.} \end{cases}
\]

(iii) For the columns \( q_1, \ldots, q_m \) of \( Q \), one can complete the orthonormal system \( q_1, \ldots, q_r \) of (i) by \( q_{r+1}, \ldots, q_m \in \mathbb{K}^m \) to an orthonormal basis of \( \mathbb{K}^m \).

(iv) For \( \tilde{A} \), one can use \( \tilde{A}_e \) as in (ii), completed by \( m-r \) zero rows at the bottom.

(b) For \( r > 0 \), there is a unique economy size QR decomposition of \( A \) such that all pivot elements of \( \tilde{A}_e \) are positive. Similarly, if one requires all pivot elements of \( \tilde{A} \) to be positive, then the QR decomposition of \( A \) is unique, except for the last columns \( q_{r+1}, \ldots, q_m \) of \( Q \).

**Proof.** (a): We start by showing that, for \( r > 0 \), if \( Q_e \) and \( \tilde{A}_e \) are defined by (i) and (ii), respectively, then they provide the desired economy size QR decomposition of \( A \). From [Phi19b, Th. 10.9], we know that \( v_k = 0 \) if, and only if, \( x_k \in \text{span}\{x_1, \ldots, x_{k-1}\} \). Thus, in (i), we indeed obtain \( r \) linearly independent \( \tilde{v}_k \) and \( Q_e \) is in \( \mathcal{M}(m, r, \mathbb{K}) \) with orthonormal columns. To see that \( \tilde{A}_e \) according to (ii) is in echelon form, consider its \( i \)th row, \( i \in \{1, \ldots, r\} \). If \( k < \mu(i) \), then \( \rho(k) < i \), i.e. \( r_{ik} = 0 \). Thus, \( \|\tilde{v}_{\rho(k)}\|_2 \) with \( k = \mu(i) \) is the pivot element of the \( i \)th row, and as \( \mu \) is strictly increasing, \( \tilde{A}_e \) is in echelon form. To show \( A = Q_e \tilde{A}_e \), note that, for each \( k \in \{1, \ldots, n\} \),
\[
x_k = \begin{cases} 0 & \text{for } k = 1, \rho(1) = 0, \\ \|\tilde{v}_{\rho(1)}\|_2 q_{\rho(1)} = \|v_1\|_2 q_1 & \text{for } k = 1, \rho(1) = 1, \\ \sum_{i=1}^{\rho(k)-1} \langle x_k, q_i \rangle q_i + \|\tilde{v}_{\rho(k)}\|_2 q_{\rho(k)} & \text{for } k > 1, \rho(k) > \rho(k-1) \end{cases} = \sum_{i=1}^{r} r_{ik} q_i,
\]
as a consequence of (5.25), completing the proof of the economy size QR decomposition. Moreover, given the economy size QR decomposition, (iii) and (iv) clearly provide a QR decomposition of \( A \).
(b): Recall from the proof of (a) that, for the $\tilde{A}_{\alpha}$ defined in (ii) and, thus, for the $\tilde{A}$ defined in (iv), the pivot elements are given by the $\|\tilde{v}_{\rho(k)}\|_2$ and, thus, always positive. It suffices to prove the uniqueness statement for the economy size QR decomposition, as it clearly implies the uniqueness statement for the QR decomposition. Suppose

$$A = QR = \hat{Q}\hat{R},$$

(5.26)

where $Q, \hat{Q} \in \mathcal{M}(m, r)$ with orthonormal columns, and $R, \hat{R} \in \mathcal{M}(r, n)$ are in echelon form such that all pivot elements are positive. We write (5.26) in the equivalent form

$$x_k = \sum_{i=1}^{r} r_{ik}q_i = \sum_{i=1}^{r} \tilde{r}_{ik}\tilde{q}_i \quad \text{for each } k \in \{1, \ldots, n\},$$

(5.27)

and show, by induction on $\alpha \in \{1, \ldots, r + 1\}$, that $q_\alpha = \tilde{q}_\alpha$ for $\alpha \leq r$ and $r_{ik} = \tilde{r}_{ik}$ for each $i \in \{1, \ldots, r\}$, $k \in \{1, \ldots, \min\{\mu(\alpha), n\}\}$, $\mu(r + 1) := n + 1$, as well as

$$\text{span}\{x_1, \ldots, x_{\mu(\alpha)}\} = \text{span}\{q_1, \ldots, q_{\alpha-1}, q_{\alpha}\} \quad \text{for } \alpha \leq r$$

(5.28)

(the induction goes to $r + 1$, as it might occur that $\mu(r) < n$). Consider $\alpha = 1$. For $1 \leq k < \mu(1)$, we have $x_k = 0$, and the linear independence of the $q_i$ (respectively, $\tilde{q}_i$) yields $r_{ik} = \tilde{r}_{ik} = 0$ for each $i \in \{1, \ldots, r\}$, $1 \leq k < \mu(1)$ (if any). Next,

$$0 \neq x_{\mu(1)} = r_{1,\mu(1)}q_1 = \tilde{r}_{1,\mu(1)}\tilde{q}_1,$$

(5.29)

since the echelon forms of $R$ and $\hat{R}$ imply $r_{i,\mu(1)} = \tilde{r}_{i,\mu(1)} = 0$ for each $i > 1$. Moreover,

$$\|q_1\|_2 = \|\tilde{q}_1\|_2 = 1 \quad \Rightarrow \quad |r_{1,\mu(1)}| = |\tilde{r}_{1,\mu(1)}|$$

$$\Rightarrow \quad (r_{1,\mu(1)}, \tilde{r}_{1,\mu(1)} \in \mathbb{R}^+ \quad \text{and } q_1 = \tilde{q}_1).$$

Note that (5.29) also implies \(\text{span}\{x_1, \ldots, x_{\mu(1)}\} = \text{span}\{x_{\mu(1)}\} = \text{span}\{q_1\}\). Now let $\alpha > 1$ and, by induction, assume $q_\beta = \tilde{q}_\beta$ for $1 \leq \beta < \alpha$ and $r_{ik} = \tilde{r}_{ik}$ for each $i \in \{1, \ldots, r\}$, $k \in \{1, \ldots, \mu(\alpha - 1)\}$, as well as

$$\forall \ 1 \leq \beta < \alpha \quad \text{span}\{x_1, \ldots, x_{\mu(\beta)}\} = \text{span}\{q_1, \ldots, q_\beta\}.$$ 

(5.30)

We have to show $r_{ik} = \tilde{r}_{ik}$ for each $i \in \{1, \ldots, r\}$, $k \in \{\mu(\alpha - 1) + 1, \ldots, \min\{\mu(\alpha), n\}\}$, and $q_\alpha = \tilde{q}_\alpha$ for $\alpha \leq r$, as well as (5.28). For each $k \in \{\mu(\alpha - 1) + 1, \ldots, \mu(\alpha) - 1\}$, from (5.30) with $\beta = \alpha - 1$, we know

$$\text{span}\{x_1, \ldots, x_k\} = \text{span}\{x_1, \ldots, x_{\mu(\alpha - 1)}\} = \text{span}\{q_1, \ldots, q_{\alpha-1}\},$$

(5.31)

such that (5.27) implies $r_{ik} = \tilde{r}_{ik} = 0$ for each $i > \alpha - 1$. Then, for $1 \leq i \leq \alpha - 1$, multiplying (5.27) by $q_i$ and using the orthonormality of the $q_i$ provides $r_{ik} = \tilde{r}_{ik} = \langle x_k, q_i \rangle$. For $\alpha = r + 1$, we are already done. Otherwise, a similar argument also works for $k = \mu(\alpha) \leq n$: As we had just seen, $r_{al} = \tilde{r}_{al} = 0$ for each $1 \leq l < k$. So the echelon
forms of $R$ and $\tilde{R}$ imply $r_{ik} = \tilde{r}_{ik} = 0$ for each $i > \alpha$. Then, as before, $r_{ik} = \tilde{r}_{ik} = (x_k, q_i)$ for each $i < \alpha$ by multiplying (5.27) by $q_i$, and, finally,

$$0 \neq x_k - \sum_{i=1}^{\alpha-1} r_{ik} q_i = x_{\mu(\alpha)} - \sum_{i=1}^{\alpha-1} r_{ik} q_i = r_{\alpha k} q_{\alpha} = \tilde{r}_{\alpha k} \tilde{q}_{\alpha}.$$  \hspace{1cm} (5.32)

Analogous to the case $\alpha = 1$, the positivity of $r_{\alpha k}$ and $\tilde{r}_{\alpha k}$ implies $r_{\alpha k} = \tilde{r}_{\alpha k}$ and $q_{\alpha} = \tilde{q}_{\alpha}$. To conclude the induction, we note that combining (5.32) with (5.31) verifies (5.28). \hspace{1cm} ■

**Remark 5.17.** Gram-Schmidt orthogonalization according to (4.73) becomes numerically unstable in cases where $x_k$ is “almost” in span$\{x_1, \ldots, x_{k-1}\}$, resulting in a very small $0 \neq v_k$, in which case $\|v_k\|_2^{-1}$ can become arbitrarily large. In the following section, we will study an alternative method to compute the QR decomposition that avoids this issue.

**Example 5.18.** Consider

$$A := \begin{pmatrix} 1 & 4 & 2 & 3 \\ 1 & 2 & 1 & 0 \\ 2 & 6 & 3 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix}.$$

Applying Gram-Schmidt orthogonalization (4.73) to the columns

$$x_1 := \begin{pmatrix} 1 \\ 1 \\ 2 \\ 0 \end{pmatrix}, \quad x_2 := \begin{pmatrix} 4 \\ 2 \\ 6 \\ 0 \end{pmatrix}, \quad x_3 := \begin{pmatrix} 2 \\ 1 \\ 3 \\ 1 \end{pmatrix}, \quad x_4 := \begin{pmatrix} 3 \\ 0 \\ 1 \\ 4 \end{pmatrix}$$

of $A$ yields

$$v_1 = x_1 = \begin{pmatrix} 1 \\ 1/2 \\ 0 \end{pmatrix},$$

$$v_2 = x_2 - \frac{\langle x_2, v_1 \rangle}{\|v_1\|_2^2} v_1 = x_2 - \frac{18}{6} v_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix},$$

$$v_3 = x_3 - \frac{\langle x_3, v_1 \rangle}{\|v_1\|_2^2} v_1 - \frac{\langle x_3, v_2 \rangle}{\|v_2\|_2^2} v_2 = x_3 - \frac{9}{6} v_1 - \frac{1}{2} v_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

$$v_4 = x_4 - \frac{\langle x_4, v_1 \rangle}{\|v_1\|_2^2} v_1 - \frac{\langle x_4, v_2 \rangle}{\|v_2\|_2^2} v_2 - \frac{\langle x_4, v_3 \rangle}{\|v_3\|_2^2} v_3 = x_4 - \frac{5}{6} v_1 - \frac{3}{2} v_2 - \frac{4}{1} v_3 = \begin{pmatrix} 2/3 \\ 2/3 \\ -2/3 \\ 0 \end{pmatrix}.$$
Thus,

\[ q_1 = \frac{v_1}{\|v_1\|_2} = \frac{v_1}{\sqrt{6}}, \quad q_2 = \frac{v_2}{\|v_2\|_2} = \frac{v_2}{\sqrt{2}}, \quad q_3 = \frac{v_3}{\|v_3\|_2} = v_3, \quad q_4 = \frac{v_4}{\|v_4\|_2} = \frac{v_4}{2/\sqrt{3}} \]

and

\[
Q = \begin{pmatrix}
\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{3}} \\
\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{3}} \\
\frac{2}{\sqrt{6}} & 0 & 0 & -\frac{1}{\sqrt{3}} \\
0 & 0 & 1 & 0
\end{pmatrix}.
\]

Next, we obtain

\[
\begin{align*}
q_{11} &= \|v_1\|_2 = \sqrt{6}, \quad r_{12} = \langle x_2, q_1 \rangle = 3\sqrt{6}, \quad r_{13} = \langle x_3, q_1 \rangle = 9/\sqrt{6}, \quad r_{14} = \langle x_4, q_1 \rangle = 5/\sqrt{6}, \\
r_{21} &= 0, \quad r_{22} = \|v_2\|_2 = \sqrt{2}, \quad r_{23} = \langle x_3, q_2 \rangle = 1/\sqrt{2}, \quad r_{24} = \langle x_4, q_2 \rangle = 3/\sqrt{2}, \\
r_{31} &= 0, \quad r_{32} = 0, \quad r_{33} = \|v_3\|_2 = 1, \quad r_{34} = \langle x_4, q_3 \rangle = 4, \\
r_{41} &= 0, \quad r_{42} = 0, \quad r_{43} = 0, \quad r_{44} = \|v_4\|_2 = 2/\sqrt{3},
\end{align*}
\]

that means

\[
R = \tilde{A} = \begin{pmatrix}
\frac{\sqrt{6}}{2} & 3\sqrt{6} & 9/\sqrt{6} & 5/\sqrt{6} \\
0 & \sqrt{2} & 1/\sqrt{2} & 3/\sqrt{2} \\
0 & 0 & 1 & 4 \\
0 & 0 & 0 & 2/\sqrt{3}
\end{pmatrix}.
\]

One verifies \( A = QR \).

### 5.4.3 QR Decomposition via Householder Reflections

The Householder method uses reflections to compute the QR decomposition of a matrix. It does not go through the economy size QR decomposition (as does the Gram-Schmidt method), but provides the QR decomposition directly (which can be seen as an advantage or disadvantage, depending on the circumstances). The Householder method does not suffer from the numerical instability issues discussed in Rem. 5.17. On the other hand, the Gram-Schmidt method provides \( k \) correct columns of \( Q \) in \( k \) steps, whereas, for the Householder method, in general, one might not obtain any correct columns of \( Q \) before completing the final step. We begin with some preparatory definitions and remarks:

**Definition 5.19.** A matrix \( H \in \mathcal{M}(n, \mathbb{K}) \), \( n \in \mathbb{N} \), is called a *Householder matrix* or *Householder transformation* over \( \mathbb{K} \) if, and only if, there exists \( u \in \mathbb{K}^n \) such that \( \|u\|_2 = 1 \) and

\[
H = \text{Id} - 2 uu^*, \tag{5.33}
\]

where, here and in the following, \( u \) is treated as a *column* vector. If \( H \) is a Householder matrix, then the linear map \( H : \mathbb{K}^n \to \mathbb{K}^n \) is also called a *Householder reflection* or *Householder transformation*. 

Lemma 5.20. Let $H \in \mathcal{M}(n, \mathbb{K})$, $n \in \mathbb{N}$, be a Householder matrix. Then the following holds:

(a) $H$ is Hermitian: $H^* = H$.
(b) $H$ is involutary: $H^2 = \text{Id}$.
(c) $H$ is unitary: $H^* H = \text{Id}$.

Proof. As $H$ is a Householder matrix, there exists $u \in \mathbb{K}^n$ such that $\|u\|_2 = 1$ and (5.33) is satisfied. Then

$$H^* = (\text{Id} - 2 uu^*)^* = \text{Id} - 2(u^*)^* u^* = H$$

proves (a) and

$$H^2 = (\text{Id} - 2 uu^*)(\text{Id} - 2 uu^*) = \text{Id} - 4 uu^* + 4 uu^* uu^* = \text{Id},$$

proves (b), where $u^* u = \|u\|_2^2 = 1$ was used in the last equality. Combining (a) and (b) then yields (c). □

Remark 5.21. Let $H \in \mathcal{M}(n, \mathbb{R})$, $n \in \mathbb{N}$, be a Householder matrix, with $H = \text{Id} - 2 uu^t$, $u \in \mathbb{R}^n$, $\|u\|_2 = 1$. Then the map $H : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $x \mapsto Hx$ constitutes the reflection through the hyperplane

$$V := \text{span}\{u\}^\perp = \{z \in \mathbb{R}^n : z^t u = 0\} :$$

Note that, for each $x \in \mathbb{R}^n$, $x - uu^t x \in V$:

$$(x - uu^t x)^t u = x^t u - (x^t u) u^t u \overset{u^t u = 1}{=} 0.$$ 

Also:

$$\forall z \in V \quad \langle uu^t x, z \rangle = z^t uu^t x \overset{z^t u = 0}{=} 0.$$ 

In particular, $\text{Id} - uu^t$ is the orthogonal projection onto $V$, showing that $H$ constitutes the reflection through $V$.

The key idea of the Householder method for the computation of a QR decomposition is to find a Householder reflection that transforms a given vector into the span of the first standard unit vector $e_1 \in \mathbb{K}^k$. This task then has to be performed for varying dimensions $k$. That such Householder reflections exist is the contents of the following lemma$^2$:

$^2$If one is only interested in the case $\mathbb{K} = \mathbb{R}$, then one can use the simpler version provided in Appendix Sec. 1.2.
Lemma 5.22. Let $k \in \mathbb{N}$. We consider the elements of $\mathbb{K}^k$ as column vectors. Let $e_1 \in \mathbb{K}^k$ be the first standard unit vector of $\mathbb{K}^k$. If $x \in \mathbb{K}^k \setminus \{0\}$, then

$$ u := u(x) := \frac{v(x)}{\|v(x)\|_2}, \quad v := v(x) := \begin{cases} \frac{|x_1| + x_1 \|x\|_2 e_1}{\|x\|_2} & \text{for } x_1 \neq 0, \\ \frac{x_1}{\|x\|_2} + e_1 & \text{for } x_1 = 0, \end{cases} \quad (5.34a) $$

satisfies

$$ \|u\|_2 = 1 \quad \text{and} \quad \|v\|_2 = 1 \quad \text{(5.34b)} $$

$$(\text{Id} - 2uu^*)x = \xi e_1, \quad \xi = \begin{cases} -\frac{x_1 \|x\|_2}{\|x\|_2^2} & \text{for } x_1 \neq 0, \\ -\|x\|_2^2 & \text{for } x_1 = 0. \end{cases} \quad (5.34c) $$

Proof. If $x_1 = 0$, then $v_1 = 0 + 1 = 1$, i.e. $v \neq 0$ and $u$ is well-defined. If $x_1 \neq 0$, then

$$ v_1 = \frac{x_1 (|x_1| + \|x\|_2)}{|x_1| \|x\|_2} \neq 0, $$

i.e. $v \neq 0$ and $u$ is well-defined in this case as well. Moreover, (5.34b) is immediate from the definition of $u$. To verify (5.34c), note

$$ v^* v = \begin{cases} \frac{1 + 2|x_1|^2}{\|x_1\|_2^2} + 1 = 2 + \frac{2|x_1| \|x\|_2^2}{\|x\|_2^2} & \text{for } x_1 \neq 0, \\ 1 + 0 + 1 = 2 \frac{2|x_1| \|x\|_2^2}{\|x\|_2^2} & \text{for } x_1 = 0, \end{cases} $$

$$ v^* x = \begin{cases} \|x\|_2 + |x_1| & \text{for } x_1 \neq 0, \\ \|x\|_2 + 0 = \|x\|_2 + |x_1| & \text{for } x_1 = 0. \end{cases} $$

Thus, for $x_1 = 0$, we obtain

$$ (\text{Id} - 2uu^*)x = x - \frac{2vv^* x}{v^* v} = x - v \left( \frac{\|x\|_2 + |x_1|}{1 + \frac{|x_1|}{\|x\|_2}} \right) = x - v \|x\|_2 = -\|x\|_2 e_1 = \xi e_1, $$

proving (5.34c). Similarly, for $x_1 \neq 0$, we obtain

$$ (\text{Id} - 2uu^*)x = x - \frac{|x_1| x + x_1 \|x\|_2 e_1}{|x_1|} = -\frac{x_1 \|x\|_2}{|x_1|} e_1 = \xi e_1, $$

once again proving (5.34c) and concluding the proof of the lemma.

We note that the numerator in (5.34a) is constructed in such a way that there is no subtractive cancellation of digits, independent of the signs of $\text{Re } x_1$ and $\text{Im } x_1$.

Definition 5.23. Given $A \in \mathcal{M}(m, n, \mathbb{K})$ with $m, n \in \mathbb{N}$, the Householder method is the following procedure: Let $A^{(1)} := A$, $Q^{(1)} := \text{Id}_m$ (identity matrix on $\mathbb{K}^m$), $r(1) := 1$. For $k \geq 1$, as long as $r(k) < m$ and $k \leq n$, the Householder method transforms $A^{(k)}$ into $A^{(k+1)}$, $Q^{(k)}$ into $Q^{(k+1)}$, and $r(k)$ into $r(k + 1)$ by performing precisely one of the following actions:
(a) If \( a_{ik}^{(k)} = 0 \) for each \( i \in \{r(k), \ldots, m\} \), then
\[
A^{(k+1)} := A^{(k)}, \quad Q^{(k+1)} := Q^{(k)}, \quad r(k + 1) := r(k).
\]

(b) If \( a_{r(k),k}^{(k)} \neq 0 \) but \( a_{ik}^{(k)} = 0 \) for each \( i \in \{r(k) + 1, \ldots, m\} \),
\[
A^{(k+1)} := A^{(k)}, \quad Q^{(k+1)} := Q^{(k)}, \quad r(k + 1) := r(k) + 1.
\]

(c) Otherwise, i.e. if \( x^{(k)} \notin \text{span}\{e_1\} \), where
\[
x^{(k)} := \left( a_{r(k),k}^{(k)}, \ldots, a_{m,k}^{(k)} \right)^t \in \mathbb{K}^{m-r(k)+1}
\]
and \( e_1 \) is the standard basis vector in \( \mathbb{K}^{m-r(k)+1} \), then
\[
A^{(k+1)} := H^{(k)} A^{(k)}, \quad Q^{(k+1)} := Q^{(k)} H^{(k)}, \quad r(k + 1) := r(k) + 1,
\]
where
\[
H^{(k)} := \begin{pmatrix}
\text{Id}_{r(k)-1} & 0 \\
0 & \text{Id}_{m-r(k)+1} - 2u^{(k)}(u^{(k)})^* \\
\end{pmatrix}, \quad u^{(k)} := u(x^{(k)}), \quad (5.36)
\]
with \( u(x^{(k)}) \) according to (5.34a).

**Theorem 5.24.** Given \( A \in \mathcal{M}(m,n,\mathbb{K}) \) with \( m,n \in \mathbb{N} \), the Householder method as defined in Def. 5.23 yields a QR decomposition of \( A \). More precisely, if \( r(k) = m \) or \( k = n + 1 \), then \( Q := Q^{(k)} \in \mathcal{M}(m,\mathbb{K}) \) is unitary and \( \tilde{A} := A^{(k)} \in \mathcal{M}(m,n,\mathbb{K}) \) is in echelon form, satisfying \( A = Q \tilde{A} \).

**Proof.** First note that the Householder method terminates after at most \( n \) steps. If \( 1 \leq N \leq n + 1 \) is the maximal \( k \) occurring during the Householder method, and one lets \( H^{(k)} := \text{Id}_m \) for each \( k \) such that Def. 5.23(a) or Def. 5.23(b) was used, then
\[
Q = H^{(1)} \cdots H^{(N-1)}, \quad \tilde{A} = H^{(N-1)} \cdots H^{(1)} A, \quad (5.37)
\]
Since, according to Lem. 5.20(b), \( (H^{(k)})^2 = \text{Id}_m \) for each \( k \), \( Q \tilde{A} = A \) is an immediate consequence of (5.37). Moreover, it follows from Lem. 5.20(c) that the \( H^{(k)} \) are all unitary, implying \( Q \) to be unitary as well. To prove that \( \tilde{A} \) is in echelon form, we show by induction over \( k \) that the first \( k - 1 \) columns of \( A^{(k)} \) are in echelon form as well as the first \( r(k) \) rows of \( A^{(k)} \) with \( a_{ij}^{(k)} = 0 \) for each \( (i,j) \in \{r(k),\ldots,m\} \times \{1,\ldots,k-1\} \).

For \( k = 1 \), the assertion is trivially true. By induction, we assume the assertion for \( k \) and prove it for \( k + 1 \) (for \( k = 1 \), we are not assuming anything). Observe that multiplication of \( A^{(k)} \) with \( H^{(k)} \) as defined in (5.36) does not change the first \( r(k) - 1 \) rows of \( A^{(k)} \). It does not change the first \( k - 1 \) columns of \( A^{(k)} \), either, since \( a_{ij}^{(k)} = 0 \)
for each \( (i, j) \in \{r(k), \ldots, m\} \times \{1, \ldots, k - 1\} \). Moreover, if we are in the case of Def. 5.23(c), then the choice of \( u^{(k)} \) in (5.36) and (5.34c) of Lem. 5.22 guarantee

\[
\begin{pmatrix}
  a_{r(k),k}^{(k+1)} \\
  a_{r(k)+1,k}^{(k+1)} \\
  \vdots \\
  a_{m,k}^{(k+1)}
\end{pmatrix} \in \text{span} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{5.38}
\]

In each case, after the application of (a), (b), or (c) of Def. 5.23, \( a_{ij}^{(k+1)} = 0 \) for each \( (i, j) \in \{r(k) + 1, \ldots, m\} \times \{1, \ldots, k\} \). We have to show that the first \( r(k + 1) \) rows of \( A^{(k+1)} \) are in echelon form. For Def. 5.23(a), it is \( r(k + 1) = r(k) \) and there is nothing to prove. For Def. 5.23(b),(c), we know \( a_{r(k)j}^{(k+1)} = 0 \) for each \( j \in \{1, \ldots, k\} \), while \( a_{r(k)k}^{(k+1)} \neq 0 \), showing that the first \( r(k + 1) \) rows of \( A^{(k+1)} \) are in echelon form in each case. So we know that the first \( r(k + 1) \) rows of \( A^{(k+1)} \) are in echelon form, and all elements in the first \( k \) columns of \( A^{(k+1)} \) below row \( r(k+1) \) are zero, showing that the first \( k \) columns of \( A^{(k+1)} \) are also in echelon form. As, at the end of the Householder method, \( r(k) = m \) or \( k = n + 1 \), we have shown that \( A \) is in echelon form. \( \blacksquare \)

**Example 5.25.** We apply the Householder method of Def. 5.23 to the matrix \( A \) of the previous Ex. 5.18, i.e. to

\[
A := \begin{pmatrix} 1 & 4 & 2 & 3 \\ 1 & 2 & 1 & 0 \\ 2 & 6 & 3 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix}. \tag{5.39a}
\]

We start with

\[
A^{(1)} := A, \quad Q^{(1)} := \text{Id}, \quad r(1) := 1. \tag{5.39b}
\]

Since \( x^{(1)} := (1, 1, 2, 0)^t \notin \text{span}\{(1, 0, 0, 0)^t\} \), we compute

\[
u^{(1)} := \frac{x^{(1)} + \sigma(x^{(1)}) e_1}{\|x^{(1)} + \sigma(x^{(1)}) e_1\|_2} = \frac{1}{\|x^{(1)} + \sqrt{6} e_1\|_2} \begin{pmatrix} 1 + \sqrt{6} \\ 1 \\ 2 \\ 0 \end{pmatrix}, \tag{5.39c}
\]

\[
H^{(1)} := \text{Id} - 2u^{(1)}(u^{(1)})^t = \text{Id} - \frac{1}{6 + \sqrt{6}} \begin{pmatrix} 7 + 2\sqrt{6} & 1 + \sqrt{6} & 2 + 2\sqrt{6} & 0 \\ 1 + \sqrt{6} & 1 & 2 & 0 \\ 2 + 2\sqrt{6} & 2 & 4 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{5.39d}
\]

\[
\approx \begin{pmatrix} -0.4082 & -0.4082 & -0.8165 & 0 \\ -0.4082 & 0.8816 & -0.2367 & 0 \\ -0.8165 & -0.2367 & 0.5266 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.
\]
5 NUMERICAL SOLUTION OF LINEAR SYSTEMS

\[ A^{(2)} := H^{(1)} A^{(1)} \approx \begin{pmatrix} -2.4495 & -7.3485 & -3.6742 & -2.0412 \\ 0 & -1.2899 & -0.6449 & -1.4614 \\ 0 & -0.5798 & -0.2899 & -1.9229 \\ 0 & 0 & 1 & 4 \end{pmatrix}, \] (5.39e)

\[ Q^{(2)} := Q^{(1)} H^{(1)} = H^{(1)}, \quad r(2) := r(1) + 1 = 2. \] (5.39f)

Since \( x^{(2)} := (-1.2899, -0.5798, 0)^t \notin \text{span}\{(1, 0, 0)^t\}, \) we compute

\[ u^{(2)} := \frac{x^{(2)} + \sigma(x^{(2)}) e_1}{\|x^{(2)} + \sigma(x^{(2)}) e_1\|_2} = \frac{x^{(2)} - \|x^{(2)}\|_2 e_1}{\|x^{(2)} - \|x^{(2)}\|_2 e_1\|_2} \approx \begin{pmatrix} -0.9778 \\ -0.2096 \\ 0 \end{pmatrix}, \] (5.39g)

\[ H^{(2)} := \begin{pmatrix} 1 & 0 \\ 0 & \text{Id} - 2u^{(2)} (u^{(2)})^t \end{pmatrix} \approx \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.9121 & -0.4100 & 0 \\ 0 & -0.4100 & 0.9121 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \] (5.39h)

\[ A^{(3)} = H^{(2)} A^{(2)} \approx \begin{pmatrix} -2.4495 & -7.3485 & -3.6742 & -2.0412 \\ 0 & 1.4142 & 0.7071 & 2.1213 \\ 0 & 0 & 0 & -1.1547 \\ 0 & 0 & 1 & 4 \end{pmatrix}, \] (5.39i)

\[ Q^{(3)} = Q^{(2)} H^{(2)} \approx \begin{pmatrix} -0.4082 & 0.7071 & -0.5774 & 0 \\ -0.4082 & -0.7071 & -0.5774 & 0 \\ -0.8165 & -0.0000 & 0.5774 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad r(3) = r(2) + 1 = 3. \] (5.39j)

Since \( x^{(3)} := (0, 1)^t \notin \text{span}\{(1, 0)^t\}, \) we compute

\[ u^{(3)} := \frac{x^{(3)} + \sigma(x^{(3)}) e_1}{\|x^{(3)} + \sigma(x^{(3)}) e_1\|_2} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}, \] (5.39k)

\[ H^{(3)} := \begin{pmatrix} \text{Id} & 0 \\ 0 & \text{Id} - 2u^{(3)} (u^{(3)})^t \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \] (5.39l)

\[ R = \tilde{A} = A^{(4)} = H^{(3)} A^{(3)} \approx \begin{pmatrix} -2.4495 & -7.3485 & -3.6742 & -2.0412 \\ 0 & 1.4142 & 0.7071 & 2.1213 \\ 0 & 0 & -1 & -4 \\ 0 & 0 & 0 & 1.1547 \end{pmatrix}, \] (5.39m)

\[ Q = Q^{(4)} = Q^{(3)} H^{(3)} \approx \begin{pmatrix} -0.4082 & 0.7071 & 0 & 0.5774 \\ -0.4082 & -0.7071 & 0 & 0.5774 \\ -0.8165 & 0 & 0 & -0.5774 \\ 0 & 0 & -1 & 0 \end{pmatrix}. \] (5.39n)

One verifies \( A = QR. \)
6 Iterative Methods, Solution of Nonlinear Equations

6.1 Motivation: Fixed Points and Zeros

Definition 6.1. (a) Given a function \( \varphi : A \to B \), where \( A, B \) can be arbitrary sets, \( x \in A \) is called a fixed point of \( \varphi \) if, and only if, \( \varphi(x) = x \).

(b) If \( f : A \to Y \), where \( A \) can be an arbitrary set, and \( Y \) is (a subset of) a vector space, then \( x \in A \) is called a zero (or sometimes a root) of \( f \) if, and only if, \( f(x) = 0 \).

In the following, we will study methods for finding zeros of functions in special situations. While, if formulated in the generality of Def. 6.1, the setting of fixed point and zero problems is different, in many situations, a fixed point problem can be transformed into an equivalent zero problem and vice versa (see Rem. 6.2 below). In consequence, methods that are formulated for fixed points, such as the Banach fixed point method, can often also be used to find zeros, while methods formulated for zeros, such as Newton’s method, can often also be used to find fixed points.

Remark 6.2. (a) If \( \varphi : A \to B \), where \( A, B \) are subsets of a common vector space \( X \), then \( x \in A \) is a fixed point of \( \varphi \) if, and only if, \( \varphi(x) - x = 0 \), i.e. if, and only if, \( x \) is a zero of the function \( f : A \to X \), \( f(x) := \varphi(x) - x \) (we can no longer admit arbitrary sets \( A, B \), as adding and subtracting must make sense; but note that it might happen that \( \varphi(x) - x \notin A \cup B \)).

(b) If \( f : A \to Y \), where \( A \) and \( Y \) are both subsets of some common vector space \( Z \), then \( x \in A \) is a zero of \( f \) if, and only if, \( f(x) + x = x \), i.e. if, and only if, \( x \) is a fixed point of the function \( \varphi : A \to Z \), \( \varphi(x) := f(x) + x \).

Given a function \( g \), an iterative method defines a sequence \( x_0, x_1, \ldots \), where, for \( n \in \mathbb{N}_0 \), \( x_{n+1} \) is computed from \( x_n \) (or possibly \( x_0, \ldots, x_n \)) by using \( g \). For example, “using \( g \)” can mean applying \( g \) to obtain \( x_{n+1} := g(x_n) \) or it can mean applying \( g \) and \( g' \) in a suitable way (as in the case of Newton’s method according to (6.1) and (6.5) below). The iterative method with \( x_{n+1} := g(x_n) \) is already known from the Banach fixed point theorem of Analysis 2 (cf. [Phi16b, Th. 2.29]). In the following, we will provide an introduction to the iterative method known as Newton’s method.

6.2 Newton’s Method

As mentioned above, the Banach fixed point method has already been studied in Analysis 2, [Phi16b, Sec. 2.2]. The Banach fixed point method has the advantage of being applicable in the rather general setting of a complete metric space and of not requiring
any differentiability. Moreover, convergence results and error estimates are comparatively simple to prove.

Newton’s method on the other hand is only applicable to differentiable maps between normed vector spaces and convergence results as well as error estimates are typically considerably more difficult to establish, especially in more than one dimension. However, if Newton’s method converges, than the convergence is typically considerably faster than for the Banach fixed point method and, hence, Newton’s method is preferred in such situations.

While the Banach fixed point method is designed to provide sequences that converge to a fixed point of a function \( \varphi \), Newton’s method is designed to provide sequences that converge to a zero of a function \( f \) (cf. Sec. 6.1 above).

The subject of Newton’s and related methods is a vast field and we will only be able to provide a brief introduction – see, e.g., [Deu11], [OR70, Sections 12.6,13.3,14.4] for further results regarding such methods.

### 6.2.1 Newton’s Method in One Dimension

If \( A \subseteq \mathbb{R} \) and \( f : A \rightarrow \mathbb{R} \) is differentiable, then Newton’s method is defined by the recursion

\[
x_0 \in A, \quad x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)} \quad \text{for each } n \in \mathbb{N}_0.
\]

There are clearly several issues with Newton’s method as formulated in (6.1):

(a) Is the derivative of \( f \) in \( x_n \) invertible (i.e. nonzero in the case of (6.1))?  
(b) Is \( x_{n+1} \) an element of \( A \) such that the iteration can be continued?  
(c) Does the sequence \( (x_n)_{n \in \mathbb{N}_0} \) converge?

To ensure that we can answer “yes” to all of the above questions, we need suitable hypotheses. The following theorem provides examples for sufficient conditions:

**Theorem 6.3.** Let \( I \subseteq \mathbb{R} \) be a nontrivial interval (i.e. an interval consisting of more than one point). Assume \( f : I \rightarrow \mathbb{R} \) to be differentiable with \( f'(x) \neq 0 \) for each \( x \in I \) and to have a zero \( \hat{x} \in I \). If \( f \) satisfies one of the following four conditions (a) – (d), then \( \hat{x} \) is the only zero of \( f \) in \( I \), and (6.1) well-defines a sequence \( (x_n)_{n \in \mathbb{N}_0} \) in \( I \) that converges to \( \hat{x} \), provided that \( x_0 = \hat{x} \) (in which case the sequence is constant) or \( x_0, x_1 \) are on different sides of \( \hat{x} \) with \( x_1 \in I \) or \( x_0, x_1 \) are on the same side of \( \hat{x} \) (in which case the sequence is monotone).

(a) \( f \) is strictly increasing and convex.  
(b) \( f \) is strictly decreasing and concave.  
(c) \( f \) is strictly increasing and concave.
(d) \( f \) is strictly decreasing and convex.

For (a) and (b), \( x_0, x_1 \) are on different sides of \( \hat{x} \) for \( x_0 < \hat{x} \) and \( (x_n)_{n \in \mathbb{N}_0} \) is decreasing for \( x_0 > \hat{x} \); for (c) and (d), \( x_0, x_1 \) are on different sides of \( \hat{x} \) for \( x_0 > \hat{x} \) and \( (x_n)_{n \in \mathbb{N}_0} \) is increasing for \( x_0 < \hat{x} \).

**Proof.** In each case, \( f \) is strictly monotone and, hence, injective. Thus, \( \hat{x} \) must be the only zero of \( f \). For each \( n \in \mathbb{N}_0 \), if \( f(x_n) = 0 \), then (6.1) implies \( x_{n+1} = x_n \), showing \( (x_n)_{n \in \mathbb{N}_0} \) to be constant for \( x_0 = \hat{x} \).

We now assume (a). The convexity of \( f \) yields (cf. [Phi16a, (9.37)])

\[
\forall x,y \in I \quad f(x) \geq f(y)(x - y) + f(y). \tag{6.2}
\]

We now apply (6.2) to \( x := x_{n+1} \) and \( y := x_n \), assuming \( n \in \mathbb{N}_0 \) with \( x_n, x_{n+1} \in I \), obtaining

\[
f(x_{n+1}) \geq f'(x_n)(x_{n+1} - x_n) + f(x_n) \quad (6.1) = 0. \tag{6.3}
\]

**Claim 1.** If (a) holds, then

\[ \eta : I \to \mathbb{R}, \quad \eta(x) := x - \frac{f(x)}{f'(x)}, \]

is increasing on \( [\hat{x}, \sup I] \).

**Proof.** Let \( x_1, x_2 \in I \) such that \( \hat{x} \leq x_1 < x_2 \). Then we estimate

\[
\eta(x_2) - \eta(x_1) = x_2 - x_1 + \frac{f(x_1)}{f'(x_1)} - \frac{f(x_2)}{f'(x_2)} \geq \frac{f(x_2) - f(x_1)}{f'(x_2)} + \frac{f(x_1)}{f'(x_1)} - \frac{f(x_2)}{f'(x_2)}
\]

\[
= f(x_1) \left( \frac{1}{f'(x_1)} - \frac{1}{f'(x_2)} \right) \geq 0,
\]

where we used \( f(x_1) \geq 0 \) (as \( f \) is increasing) and \( f'(x_1) \leq f'(x_2) \) (as \( f \) is convex, [Phi16a, (9.37)]) \( \blacksquare \).

If \( x_n \in I \) with \( x_n > \hat{x} \), then \( f(x_n) > 0 \) (as \( f \) is strictly increasing) and \( f'(x_n) > 0 \) (as \( f \) is strictly increasing and \( f'(x_n) \neq 0 \), implying

\[
\hat{x} = \eta(\hat{x}) \leq \eta(x_n) = x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} < x_n \quad \Rightarrow \quad x_{n+1} \in I. \tag{6.4}
\]

According to (6.4), if \( x_0 \geq \hat{x} \), then \( (x_n)_{n \in \mathbb{N}_0} \) is a decreasing sequence in \( [\hat{x}, x_0] \), converging to some \( x_* \in [\hat{x}, x_0] \) (this is the case, where \( x_0, x_1 \) are on the same side of \( \hat{x} \)). If \( x_0 < \hat{x} \) and \( x_1 \in I \), then (6.3) shows \( x_1 \geq \hat{x} \) and, then, (6.4) yields \( (x_n)_{n \in \mathbb{N}} \) is a decreasing sequence in \( [\hat{x}, x_1] \), converging to some \( x_* \in [\hat{x}, x_1] \) (this is the case, where \( x_0, x_1 \) are on different sides of \( \hat{x} \)). In each case, we then obtain

\[
x_* = \lim_{n \to \infty} x_{n+1} = \lim_{n \to \infty} \left( x_n - \frac{f(x_n)}{f'(x_n)} \right) = x_* - \frac{f(x_*)}{f'(x*)},
\]
proving \( f(x_*) = 0 \), i.e. \( x_* = \hat{x} \) (note that we did not assume the continuity of \( f' \) – however the boundedness of \( f' \) on \([\hat{x}, x_1]\), as given by its monotonicity; already suffices to imply \( f(x_*) = 0 \); but the above equation does actually hold as stated, since the monotonicity of \( f' \), together with the intermediate value theorem for derivatives (cf. [Phi16a, Th. 9.24]), does also imply the continuity of \( f' \)).

If \( f \) is strictly decreasing and concave, then \(-f\) is strictly increasing and convex. Since the zeros of \( f \) and \(-f\) are the same, and (6.1) does not change if \( f \) is replaced by \(-f\), (b) follows from (a).

If \( f \) is strictly increasing and concave, then \( g : (-I) \to \mathbb{R}, g(x) := f(-x), -I := \{-x : x \in I\}, \) is strictly decreasing and concave: For each \( x_1, x_2 \in (-I), \)
\[
    x_1 < x_2 \implies -x_2 < -x_1 \\
    \quad \implies \left( g(x_1) = f(-x_1) > f(-x_2) = g(x_2) \right) \\
    \quad \quad \quad \land \ g'(x_1) = f'(-x_1) > f'(-x_2) = g'(x_2). 
\]

As (6.1) implies
\[
    \forall_{n \in \mathbb{N}_0} -x_{n+1} = -x_n - \frac{f(x_n)}{f'(x_n)} = -x_n - \frac{g(-x_n)}{g'(-x_n)}, 
\]
(b) applies to \( g \), yielding that \(-x_n\) converges to \(-\hat{x}\) in the stated manner, i.e. \( x_n \) converges to \( \hat{x} \) in the manner stated in (c).

If \( f \) is strictly decreasing and convex, then \(-f\) is strictly increasing and concave, i.e. (d) follows from (c) in the same way that (b) follows from (a).

While Th. 6.3 shows the convergence of Newton’s method for each \( x_0 \) in the domain \( I \), it does not provide error estimates. We will prove error estimates in Th. 6.5 below in the context of the multi-dimensional Newton’s method, where the error estimates are only shown in a sufficiently small neighborhood of the zero \( x_* \). Results of this type are quite typical in the context of Newton’s method, albeit somewhat unsatisfactory, as the location of the zero is usually unknown in applications (also cf. Ex. 6.6 below).

**Example 6.4.** (a) We show that \( \cos x = x \) has a unique solution in \([0, \frac{\pi}{2}]\) and that the solution can be approximated via Newton’s method: To this end, we consider
\[
    f : [0, \pi/2] \to \mathbb{R}, \quad f(x) := \cos x - x. 
\]
Then \( f \) is differentiable with \( f'(x) = -\sin x - 1 < 0 \) on \([0, \pi/2] \) and \( f''(x) = -\cos x < 0 \) on \([0, \pi/2] \), showing \( f \) to be strictly decreasing and concave. Thus, Th. 6.3(b) applies for each \( x_0 \in [0, \pi/2] \) (see Ex. 6.6(a) below for error estimates).

(b) Theorem 6.3(a) applies to the function
\[
    f : \mathbb{R}^+ \to \mathbb{R}, \quad f(x) := x^2 - a, \quad a \in \mathbb{R}^+ 
\]
of Ex. 1.3, which is, clearly, strictly increasing and convex.
6.2.2 Newton’s Method in Several Dimensions

Newton’s method, as defined in (6.1), naturally generalizes to a differentiable function \( f : A \rightarrow \mathbb{R}^m, A \subseteq \mathbb{R}^m \), as follows:

\[
    x_0 \in A, \quad x_{n+1} := x_n - (Df(x_n))^{-1}f(x_n) \quad \text{for each } n \in \mathbb{N}_0 \tag{6.5}
\]

(indeed, the same still works if \( \mathbb{R}^m \) is replaced by some general normed vector space, but then the notion of differentiability needs to be generalized to such spaces, which is beyond the scope of the present class).

In practice, in each step of Newton’s method (6.5), one will determine \( x_{n+1} \) as the solution to the linear system

\[
    Df(x_n)x_{n+1} = Df(x_n)x_n - f(x_n) \tag{6.6}
\]

Not surprisingly, the issues formulated as (a), (b), (c) after (6.1) (i.e. invertibility of \( Df(x_n) \), \( x_{n+1} \) being in \( A \), and convergence of \( (x_n)_{n \in \mathbb{N}_0} \)), are also present in the multi-dimensional situation.

As in the 1-dimensional case, we will present a result that provides an example for sufficient conditions for convergence (cf. Th. 6.5 below). While Th. 6.3 for the 1-dimensional Newton’s method is an example of a global Newton theorem (proving convergence for each \( x_0 \) in the considered domain of \( f \)), Th. 6.5 is an example of a local Newton theorem, proving convergence under the assumption that \( x_0 \) is sufficiently close to the zero \( \hat{x} \) – however, in contrast to Th. 6.3, Th. 6.5 also includes error estimates (cf. (6.12), (6.13)).

In the literature (e.g. in the previously mentioned references [Deu11], [OR70]), one finds many variants of both local and global Newton theorems. A global Newton theorem in the a multi-dimensional setting is also provided by Th. J.1 in the appendix.

**Theorem 6.5.** Let \( m \in \mathbb{N} \) and fix a norm \( \| \cdot \| \) on \( \mathbb{R}^m \). Let \( A \subseteq \mathbb{R}^m \) be open. Moreover let \( f : A \rightarrow A \) be differentiable, let \( x_* \in A \) be a zero of \( f \), and let \( r > 0 \) be (sufficiently small) such that

\[
    B_r(x_*) = \{ x \in \mathbb{R}^m : \| x - x_* \| < r \} \subseteq A. \tag{6.7}
\]

Assume that \( Df(x_*) \) is invertible and that \( \beta > 0 \) is such that

\[
    \left\| (Df(x_*))^{-1} \right\| \leq \beta \tag{6.8}
\]

(here and in the following, we consider the induced operator norm on real \( n \times n \) matrices). Finally, assume that the map \( Df : B_r(x_*) \rightarrow \mathcal{L}(\mathbb{R}^m, \mathbb{R}^m) \) is Lipschitz continuous with Lipschitz constant \( L \in \mathbb{R}^+ \), i.e.

\[
    \| (Df)(x) - (Df)(y) \| \leq L \| x - y \| \quad \text{for each } x, y \in B_r(x_*). \tag{6.9}
\]

Then, letting

\[
    \delta := \min \left\{ r, \frac{1}{2\beta L} \right\}, \tag{6.10}
\]
Newton’s method (6.5) is well-defined for each \( x_0 \in B_\delta(x_*) \) (i.e., for each \( n \in \mathbb{N}_0 \), \( Df(x_n) \) is invertible and \( x_{n+1} \in B_\delta(x_*) \), and

\[
\lim_{n \to \infty} x_n = x_*
\]  

(6.11)

with the error estimates

\[
\|x_{n+1} - x_*\| \leq \beta L \|x_n - x_*\|^2 \leq \frac{1}{2} \|x_n - x_*\|, \\
\|x_n - x_*\| \leq \left(\frac{1}{2}\right)^{2^n-1} \|x_0 - x_*\|
\]  

(6.12)

(6.13)

for each \( n \in \mathbb{N}_0 \). Moreover, within \( B_\delta(x_*) \), the zero \( x_* \) is unique.

**Proof.** The proof is conducted via several steps.

**Claim 1.** For each \( x \in B_\delta(x_*) \), the matrix \( Df(x) \) is invertible with

\[
\left\| (Df(x))^{-1} \right\| \leq 2\beta.
\]  

(6.14)

**Proof.** Fix \( x \in B_\delta(x_*) \). We will apply Lem. 2.45 with

\( A := Df(x_*) \), \( \Delta A := Df(x) - Df(x_*) \), \( Df(x) = A + \Delta A \).  

(6.15)

To apply Lem. 2.45, we must estimate \( \Delta A \). We obtain

\[
\|\Delta A\| \leq L \|x - x_*\| < L \delta \leq \frac{1}{2\beta} \leq \frac{1}{2} \|A^{-1}\|^{-1},
\]  

(6.16)

such that we can, indeed, apply Lem. 2.45. In consequence, \( Df(x) \) is invertible with

\[
\left\| (Df(x))^{-1} \right\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\Delta A\|} \leq 2\|A^{-1}\| \leq 2\beta,
\]  

(6.17)

thereby establishing the case. ▲

**Claim 2.** For each \( x, y \in B_\epsilon(x_*) \), the following holds:

\[
\|f(x) - f(y) - Df(y)(x - y)\| \leq \frac{L}{2} \|x - y\|^2.
\]  

(6.18)

**Proof.** We apply the chain rule to the auxiliary function

\[
\phi : [0, 1] \longrightarrow \mathbb{R}^m, \quad \phi(t) := f(y + t(x - y))
\]  

(6.19)

to obtain

\[
\phi'(t) := (Df(y + t(x - y)))(x - y) \quad \text{for each } t \in [0, 1].
\]  

(6.20)

This allows us to write

\[
f(x) - f(y) - Df(y)(x - y) = \phi(1) - \phi(0) - \phi'(0) = \int_0^1 (\phi'(t) - \phi'(0)) \, dt.
\]  

(6.21)
Next, we estimate the norm of the integrand:

\[ \| \phi'(t) - \phi'(0) \| \leq \| Df(y + t(x - y)) - Df(y) \| \| x - y \| \leq L t \| x - y \|^2. \]  

(6.22)

Thus,

\[ \left\| \int_0^1 (\phi'(t) - \phi'(0)) \, dt \right\| \leq \int_0^1 \| \phi'(t) - \phi'(0) \| \, dt \leq \int_0^1 L t \| x - y \|^2 \, dt = \frac{L}{2} \| x - y \|^2. \]  

(6.23)

Combining (6.23) with (6.21) proves (6.18).

We will now show that

\[ x_n \in B_\delta(x_\ast) \]  

(6.24)

and the error estimate (6.12) holds for each \( n \in \mathbb{N}_0 \). We proceed by induction. For \( n = 0 \), (6.24) holds by hypothesis. Next, we assume that \( n \in \mathbb{N}_0 \) and (6.24) holds. Using \( f(x_\ast) = 0 \) and (6.5), we write

\[
x_{n+1} - x_\ast = x_n - x_\ast - (Df(x_n))^{-1}(f(x_n) - f(x_\ast))
= -(Df(x_n))^{-1}(f(x_n) - f(x_\ast) - Df(x_n)(x_n - x_\ast)).
\]  

(6.25)

Applying the norm to (6.25) and using (6.14) and (6.18) implies

\[
\|x_{n+1} - x_\ast\| \leq 2 \beta \frac{L}{2} \|x_n - x_\ast\|^2 = \beta L \|x_n - x_\ast\|^2 \leq \beta L \delta \|x_n - x_\ast\| \leq \frac{1}{2} \|x_n - x_\ast\|,
\]  

(6.26)

which proves \( x_{n+1} \in B_\delta(x_\ast) \) as well as (6.12).

Another induction shows that (6.12) implies, for each \( n \in \mathbb{N} \),

\[
\|x_n - x_\ast\| \leq (\beta L)^{1+2+\cdots+2^{n-1}} \|x_0 - x_\ast\|^{2^n} \leq (\beta L \delta)^{2^{n-1}} \|x_0 - x_\ast\|,
\]  

(6.27)

where the geometric sum formula \( 1 + 2 + \cdots + 2^{n-1} = \frac{2^n - 1}{2-1} = 2^n - 1 \) was used for the last inequality. Moreover, \( \beta L \delta \leq \frac{1}{2} \) together with (6.27) proves (6.13), and, in particular, (6.11).

Finally, we show that \( x_\ast \) is the unique zero in \( B_\delta(x_\ast) \): Suppose \( x_\ast, x_{\ast\ast} \in B_\delta(x_\ast) \) with \( f(x_\ast) = f(x_{\ast\ast}) = 0 \). Then

\[
\|x_{\ast\ast} - x_\ast\| = \left\| (Df(x_\ast))^{-1}(f(x_{\ast\ast}) - f(x_\ast) - Df(x_\ast)(x_{\ast\ast} - x_\ast)) \right\|.
\]  

(6.28)

Applying (6.14) and (6.18) yields

\[
\|x_{\ast\ast} - x_\ast\| \leq 2 \beta \frac{L}{2} \|x_\ast - x_{\ast\ast}\|^2 \leq \beta L \delta \|x_\ast - x_{\ast\ast}\| \leq \frac{1}{2} \|x_\ast - x_{\ast\ast}\|,
\]  

(6.29)

implying \( \|x_\ast - x_{\ast\ast}\| = 0 \) and completing the proof of the theorem.
Example 6.6. (a) We already know from Ex. 6.4(a), that
\[ f : [0, \pi/2] \longrightarrow \mathbb{R}, \quad f(x) := \cos x - x, \]
has a unique zero \( x_* \in [0, \pi/2] \) and that Newton’s method converges to \( x_* \) for each \( x_0 \in [0, \pi/2] \). Moreover,
\[
\forall x \in [0, \pi/2] \quad (-2 \leq f'(x) = -\sin x - 1 \leq -1 \quad \Rightarrow \quad \left| \frac{1}{f'(x)} \right| \leq 1,)
\]
and \( f' \) is Lipschitz continuous with Lipschitz constant \( L = 1 \). Thus, Th. 6.5 applies with \( \beta = L = 1 \). However, without further knowledge regarding the location of the zero in \([0, \pi/2]\), we do not have any lower positive bound on \( r \) to obtain \( \delta = \min \left\{ r, \frac{1}{2L} \right\} \). One possibility is to extend the domain to \( A := [-\frac{1}{2}, \frac{\pi}{2} + \frac{1}{2}] \), yielding \( \delta = r = \frac{1}{2L} = \frac{1}{2} \). Thus, in combination with the information of Ex. 6.4(a), we know that, for each \( x_0 \in [0, \pi/2] \), Newton’s method converges to the zero \( x_* \) and the error estimates (6.12), (6.13) hold if \( |x_0 - x_*| < \frac{1}{2} \) (otherwise, (6.12), (6.13) might not hold for some initial steps).

(b) Consider the map
\[ f : \mathbb{R}^2 \longrightarrow \mathbb{R}^2, \quad f \left( \begin{array}{c} x_1 \\ x_2 \end{array} \right) := \left( \begin{array}{c} x_1 \\ x_2 \end{array} \right) - \frac{1}{4} \left( \frac{\cos x_1 - \sin x_2}{\cos x_1 - 2 \sin x_2} \right). \]
Then \( f \) is differentiable with, for each \( x = (x_1, x_2) \in \mathbb{R}^2 \),
\[
Df(x) = \frac{1}{4} \begin{pmatrix} 4 + \sin x_1 & \cos x_2 \\ \sin x_1 & 4 + 2 \cos x_2 \end{pmatrix},
\]
det \( Df(x) = \frac{1}{16} \left( 16 + 4 \sin x_1 + 8 \cos x_2 + \sin x_1 \cos x_2 \right) \geq \frac{3}{16} > 0, \]
\[
Df(x)^{-1} = \frac{1}{\det Df(x)} \frac{1}{4} \begin{pmatrix} 4 + 2 \cos x_2 & -\cos x_2 \\ -\sin x_1 & 4 + \sin x_1 \end{pmatrix}.
\]
We now consider \( \mathbb{R}^2 \) with \( \cdot \) and recall from Ex. 2.18(a) that the row sum norm is the corresponding operator norm. For each \( x, y \in \mathbb{R}^2 \), we estimate, using that both \( \sin \) and \( \cos \) are Lipschitz continuous with Lipschitz constant 1,
\[
\|Df(x) - Df(y)\|_\infty = \frac{1}{4} \left\| \begin{pmatrix} |\sin x_1 - \sin y_1| & |\cos x_2 - \cos y_2| \\ |\sin x_1 - \sin y_1| & 2|\cos x_2 - \cos y_2| \end{pmatrix} \right\|_\infty \leq \frac{1}{4} (|x_1 - y_1| + 2|x_2 - y_2|) \leq \frac{3}{4} \|x - y\|_\infty,
\]
as well as
\[
\|Df(x)^{-1}\|_\infty \leq \frac{16}{3} \cdot \frac{7}{4} \cdot \|x\|_\infty.
\]
Thus, if we can establish \( f \) to have a zero in \( B_1(0) = [-1, 1] \times [-1, 1] \), then Th. 6.5 applies with \( \beta := \frac{28}{3} \) and \( L := \frac{3}{4} \) (and arbitrary \( r \in \mathbb{R}^+ \), i.e. with \( \delta = \frac{1}{2} \cdot \frac{3}{28} \cdot \frac{3}{4} = \frac{1}{14} \).
To verify that $f$ has a zero in $B_1(0)$, we note that $f(x) = 0$ is equivalent to the fixed point problem $\phi(x) = x$ with

$$\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^2, \quad \phi \left( \begin{array}{c} x_1 \\ x_2 \end{array} \right) := \frac{1}{4} \left( \begin{array}{c} \cos x_1 - \sin x_2 \\ \cos x_1 - 2 \sin x_2 \end{array} \right).$$

As $\phi$, clearly, maps $B_1(0)$ (even all of $\mathbb{R}^2$) into $B_{\frac{3}{4}}(0)$ and is a contraction, due to

$$\forall_{x, y \in \mathbb{R}^2} \| \phi(x) - \phi(y) \|_\infty \leq \frac{3}{4} \| x - y \|_\infty,$$

the Banach fixed point theorem [Phi16b, Th. 2.29] applies, showing $\phi$ to have a unique fixed point $x_*$, which must be located in $B_{\frac{3}{4}}(0)$. As we now only know Th. 6.5 to yield convergence of Newton’s method for $x_0 \in B_1(x_*)$, in practice, without more accurate knowledge of the location of $x_*$, one could cover $B_{\frac{3}{4}}(0)$ with squares of side length $\frac{1}{7}$, successively starting iterations in each of the squares, aborting an iteration if one finds it to be inconsistent with the error estimates of Th. 6.5.

7 Eigenvalues

7.1 Introductory Remarks

We recall some basic definitions and results from Linear Algebra: If $A \in \mathcal{M}(n, F)$ is an $n \times n$ matrix over the field $F$, $n \in \mathbb{N}$, then $\lambda \in F$ is called an eigenvalue of $A$ if, and only if,

$$\exists_{0 \neq v \in F^n} \quad Av = \lambda v,$$

where every $0 \neq v \in F^n$ satisfying $Av = \lambda v$ is called an eigenvector to the eigenvalue $\lambda$. The set of all eigenvalues of $A$ is called the spectrum of $A$, denoted $\sigma(A)$. According to [Phi19b, Prop. 8.2(c)], $\sigma(A)$ is precisely the set of zeros of the characteristic polynomial $\chi_A$ of $A$, where $\chi_A$ is defined by (cf. [Phi19b, Def. 8.1] and [Phi19b, Rem. 8.7(a)])

$$\chi_A := \det(X \ Id_n - A) \in F[X].$$

For $\lambda \in \sigma(A)$, its algebraic multiplicity $m_a(\lambda)$ and geometric multiplicity $m_g(\lambda)$ are (cf. [Phi19b, Def. 6.14] and [Phi19b, Th. 8.4]):

$$m_a(\lambda) := m_a(\lambda, A) := \text{multiplicity of } \lambda \text{ as a zero of } \chi_A,$$
$$m_g(\lambda) := m_g(\lambda, A) := \dim \ker(A - \lambda \text{Id}).$$

Remark 7.1. Let $F$ be a field, $n \in \mathbb{N}$, $A \in \mathcal{M}(n, F)$.

(a) Eigenvalues of $A$ and their multiplicities are invariant under similarity transformations, i.e. if $T \in \mathcal{M}(n, F)$ is invertible and $B := T^{-1}AT$, then $\sigma(B) = \sigma(A)$ and, for each $\lambda \in \sigma(A)$, $m_a(\lambda, A) = m_a(\lambda, B)$ as well as $m_g(\lambda, A) = m_g(\lambda, B)$ (e.g. due to
the fact that \( \chi_A \) and \( \ker(A - \lambda \text{Id}) \) only depend on the linear maps (in \( \mathcal{L}(F^n, F^n) \)) represented by \( A \) and \( \ker(A - \lambda \text{Id}) \), respectively, where \( A \) and \( B \) (resp. \( A - \lambda \text{Id} \) and \( T^{-1}(A - \lambda \text{Id})T \)) represent the same map, merely with respect to different bases of \( F^n \) (cf. [Phi19b, Prop. 8.2(a)]).

(b) Note that the characteristic polynomial \( \chi_A = \det(X \text{Id}_n - A) \) is always a monic polynomial of degree \( n \) (i.e. a polynomial of degree \( n \), where the coefficient of \( X^n \) is 1, cf. [Phi19b, Rem. 8.3]). In general, the eigenvalues of \( A \) will be in the algebraic closure \( \overline{F} \) of \( F \) (cf. [Phi19b, Rem. 6.12(e), Def. 7.32, Th. 7.35])). On the other hand, every monic polynomial \( f \in F[X] \) of degree \( n \geq 1 \) is the characteristic polynomial of some matrix \( A(f) \in \mathcal{M}(n, F) \), where \( A(f) \) is the so-called companion matrix of the polynomial: If \( a_1, \ldots, a_n \in F \), and

\[
f := X^n + \sum_{j=1}^{n} a_j X^{n-j},
\]

then \( \chi_{A(f)} = f \) for

\[
A(f) := \begin{pmatrix}
-a_1 & -a_2 & -a_3 & \cdots & -a_{n-1} & -a_n \\
1 & 0 & 0 & \cdots & 0 \\
0 & 1 & 0 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
0 & \cdots & 0 & 1 & 0
\end{pmatrix}
\]

(cf. [Phi19b, Rem. 6.12(d)]). In the following, we will only consider matrices over \( \mathbb{R} \) and \( \mathbb{C} \), and we know that, due to the fundamental theorem of algebra (cf. [Phi16a, Th. 8.32]), \( \mathbb{C} \) is the algebraic closure of \( \mathbb{R} \).

The computation of eigenvalues is of importance to numerous applications. A standard example from Physics is the occurrence of eigenvalues as eigenfrequencies of oscillators. More generally, one often needs to compute eigenvalues when solving (linear) differential equations (cf., e.g., [Phi16c, Th. 4.32, Rem. 4.51]). In this and many other applications, the occurrence of eigenvalues can be traced back to the fact that eigenvalues occur when transforming a matrix into suitable normal forms, e.g. into Jordan normal form (cf. [Phi19b, Th. 9.13]): To compute the Jordan normal form of a matrix, one first needs to compute its eigenvalues. For another example, recall that the computation of the spectral norm \( \|A\|_2 \) of a matrix \( A \in \mathcal{M}(m, \mathbb{K}) \) requires the computation of the absolute value of an eigenvalue of \( A \) with maximal size (cf. Th. 2.24).

**Remark 7.2.** Given that eigenvalues are precisely the roots of the characteristic polynomial, and given that, as mentioned in Rem. 7.1(b), every monic polynomial of degree at least 1 can occur as the characteristic polynomial of a matrix, it is not surprising
that computing eigenvalues is, in general, a difficult task. According to Galois theory in Algebra, for a generic polynomial of degree at least 5, it is not possible to obtain its zeros using so-called radicals (which are, roughly, zeros of polynomials of the form $X^k - \lambda, k \in \mathbb{N}, \lambda \in F$, see, e.g., [Bos13, Def. 6.1.1] for a precise definition) in finitely many steps (cf., e.g., [Bos13, Cor. 6.1.7]) and, in general, an approximation by iterative numerical methods is warranted. Having said that, let us note that the problem of computing eigenvalues is, indeed, typically easier than the general problem of computing zeros of polynomials. This is due to the fact that the difficulty of computing the zeros of a polynomial depends tremendously on the form in which the polynomial is given: It is typically hard if the polynomial is expanded into the form $\sum_{j=0}^{n} a_j X^j$, but it is easy (trivial, in fact), if the polynomial is given in a factored form $c \prod_{j=1}^{n} (X - \lambda_j)$. If the characteristic polynomial is given implicitly by a matrix, one is, in general, somewhere between the two extremes. In particular, for a large matrix, it usually makes no sense to compute the characteristic polynomial in its expanded form (this is an expensive task in itself and, in the process, one even loses the additional structure given by the matrix). It makes much more sense to use methods tailored to the computation of eigenvalues, and, if available, one should make use of additional structure (such as symmetry) a matrix might have.

### 7.2 Estimates, Localization

As discussed in Rem. 7.2, iterative numerical methods will usually need to be employed to approximate the eigenvalues of a given matrix. Iterative methods (e.g. Newton’s method of Sec. 6.2 above) will, in general, only converge if one starts sufficiently close to the exact solution. The localization results of the present section can help choosing suitable initial points.

**Definition 7.3.** For each $A = (a_{kl}) \in \mathcal{M}(n, \mathbb{C}), n \in \mathbb{N}$, the closed disks

$$G_k := G_k(A) := \left\{ z \in \mathbb{C} : |z - a_{kk}| \leq \sum_{l=1, l \neq k}^{n} |a_{kl}| \right\}, \quad k \in \{1, \ldots, n\}, \quad (7.1)$$

are called the Gershgorin disks of the matrix $A$.

**Theorem 7.4** (Gershgorin Circle Theorem). For each $A = (a_{kl}) \in \mathcal{M}(n, \mathbb{C}), n \in \mathbb{N}$, one has

$$\sigma(A) \subseteq \bigcup_{k=1}^{n} G_k(A), \quad (7.2)$$

that means the spectrum of a square complex matrix $A$ is always contained in the union of the Gershgorin disks of $A$.

**Proof.** Let $\lambda \in \sigma(A)$ and let $x \in \mathbb{C}^n \setminus \{0\}$ be a corresponding eigenvector. Moreover, choose $k \in \{1, \ldots, n\}$ such that $|x_k| = \|x\|_{\infty}$. The $k$th component of the equation
Ax = \lambda x \text{ reads}
\lambda x_k = \sum_{i=1}^{n} a_{ki} x_i

and implies

\begin{align*}
|\lambda - a_{kk}| |x_k| &\leq \sum_{i=1, i \neq k}^{n} |a_{ki}| |x_i| \\
&\leq \sum_{i=1, i \neq k}^{n} |a_{kl}| |x_k|.
\end{align*}

Dividing the estimate by |x_k| \neq 0 shows \lambda \in G_k and, thus, proves (7.2).

\textbf{Corollary 7.5.} For each \( A = (a_{kl}) \in \mathcal{M}(n, \mathbb{C}) \), \( n \in \mathbb{N} \), instead of using row sums to define the Gershgorin disks of (7.1), one can use the column sums to define

\begin{align*}
G'_k := G'_k(A) := \left\{ z \in \mathbb{C} : |z - a_{kk}| \leq \sum_{i=1, i \neq k}^{n} |a_{ik}| \right\}, \quad k \in \{1, \ldots, n\}. \tag{7.3}
\end{align*}

Then one can strengthen the statement of Th. 7.4 to

\begin{align*}
\sigma(A) \subseteq \left( \bigcup_{k=1}^{n} G_k(A) \right) \cap \left( \bigcup_{k=1}^{n} G'_k(A) \right). \tag{7.4}
\end{align*}

\textbf{Proof.} Since \( \sigma(A) = \sigma(A^t) \) and \( G_k(A^t) = G'_k(A) \), (7.4) is an immediate consequence of (7.2).

\textbf{Example 7.6.} Consider the matrix

\begin{align*}
A := \begin{pmatrix}
1 & -1 & 0 \\
2 & -1 & 0 \\
0 & 1 & 4
\end{pmatrix}.
\end{align*}

Then the Gershgorin disks according to (7.1) and (7.3) are (cf. Fig. 7.1)

\begin{align*}
G_1 = \overline{B}_1(1), & \quad G_2 = \overline{B}_2(-1), & \quad G_3 = \overline{B}_1(4), \\
G'_1 = \overline{B}_2(1), & \quad G'_2 = \overline{B}_2(-1), & \quad G'_3 = \{4\}.
\end{align*}

Here, the intersection of (7.4) is \( G_1 \cup G_2 \cup G'_3 \) and, thus, noticeably smaller than either \( G_1 \cup G_2 \cup G_3 \) or \( G'_1 \cup G'_2 \cup G'_3 \) (cf. Fig. 7.1). The spectrum is

\begin{align*}
\sigma(A) = \{-i, i, 4\}.
\end{align*}

Indeed, \( A \) is diagonalizable and, with \( T := \frac{1}{34} \begin{pmatrix}
0 & 17 + 17i & 17 - 17i \\
0 & 34i & -34i \\
4 & -2 - 8i & -2 + 8i
\end{pmatrix} \),

\begin{align*}
T^{-1}AT = \frac{1}{2} \begin{pmatrix}
2 & 3 & 17 \\
2 & -1 - i & 0 \\
2 & -1 + i & 0
\end{pmatrix} \begin{pmatrix}
1 & -1 & 0 \\
2 & -1 & 0 \\
0 & 1 & 4
\end{pmatrix} \frac{1}{34} \begin{pmatrix}
0 & 17 + 17i & 17 - 17i \\
0 & 34i & -34i \\
4 & -2 - 8i & -2 + 8i
\end{pmatrix}
\end{align*}

\begin{align*}
= \begin{pmatrix}
4 & 0 & 0 \\
0 & -i & 0 \\
0 & 0 & i
\end{pmatrix}.
\end{align*}
Figure 7.1: The Gershgorin disks of matrix $A$ of Ex. 7.6 are depicted in the complex plane together with the exact locations of $A$'s eigenvalues: The black dots are the locations of the eigenvalues, the grey disks are the rowwise disks $G_1$, $G_2$, $G_3$, whereas the columnwise disks $G'_1$, $G'_2$, $G'_3$ are not shaded (only visible for $G'_1$, since $G_2 = G'_2$ and $G'_3 = \{4\}$).

Figure 7.1 depicts the Gershgorin disks of $A$ as well as the exact location of $A$’s eigenvalues in the complex plane. Note that, while $G_3$ contains precisely one eigenvalue, $G_1$ contains none, and $G_2$ contains 2.

**Theorem 7.7.** Let $A = (a_{kl}) \in \mathcal{M}(n, \mathbb{C})$, $n \in \mathbb{N}$. Suppose $k_1, \ldots, k_n$ is an enumeration of the numbers from 1 to $n$, and let $q \in \{1, \ldots, n\}$. Moreover, suppose

$$D^{(1)} = G_{k_1} \cup \cdots \cup G_{k_q} \quad (7.5a)$$

is the union of $q$ Gershgorin disks of $A$ and

$$D^{(2)} = G_{k_{q+1}} \cup \cdots \cup G_{k_n} \quad (7.5b)$$

is the union of the remaining $n-q$ Gershgorin disks of $A$. If $D^{(1)}$ and $D^{(2)}$ are disjoint, then $D^{(1)}$ must contain precisely $q$ eigenvalues of $A$ and $D^{(2)}$ must contain precisely $n-q$ eigenvalues of $A$ (counted according to their algebraic multiplicity).

**Proof.** Let $D := \text{diag}(a_{11}, \ldots, a_{nn})$ be the diagonal matrix that has precisely the same diagonal elements as $A$. The idea of the proof is to connect $A$ with $D$ via a path (in fact, a line segment) through $\mathcal{M}(n, \mathbb{C})$. Then we will show that the claim of the theorem holds for $D$ and must remain valid along the entire path. In consequence, it must also hold for $A$. The path is defined by

$$\Phi : [0, 1] \longrightarrow \mathcal{M}(n, \mathbb{C}), \quad t \mapsto \Phi(t) = (\phi_{kl}(t)), \quad \phi_{kl}(t) := \begin{cases} a_{kl} & \text{for } k = l, \\ ta_{kl} & \text{for } k \neq l, \end{cases} \quad (7.6)$$
i.e. 
\[
\Phi(t) = \begin{pmatrix}
a_{11} & t a_{12} & \cdots & \cdots & t a_{1n} \\
t a_{21} & a_{22} & t a_{23} & \cdots & t a_{2n} \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
t a_{n-1,1} & \cdots & t a_{n-1,n-2} & a_{n-1,n-1} & t a_{n-1,n} \\
t a_{n1} & t a_{n2} & \cdots & t a_{n,n-1} & a_{nn}
\end{pmatrix}.
\]

Clearly, 
\[
\Phi(0) = D = \text{diag}(a_{11}, \ldots, a_{nn}), \quad \Phi(1) = A.
\]

We denote the corresponding Gershgorin disks as follows:

\[
\forall k \in \{1, \ldots, n\} \quad \forall t \in [0, 1] \quad G_k(t) := G_k(\Phi(t)) = \left\{ z \in \mathbb{C} : |z - a_{kk}| \leq t \sum_{l=1, l \neq k}^{n} |a_{kl}| \right\}. \quad (7.7)
\]

In consequence of (7.7), we have

\[
\forall k \in \{1, \ldots, n\} \quad \forall t \in [0, 1] \quad \{a_{kk}\} = G_k(0) \subseteq G_k(t) \subseteq G_k(1) = G_k(A) = G_k. \quad (7.8)
\]

We now consider the set

\[
S := \{ t \in [0, 1] : \Phi(t) \text{ has precisely } q \text{ eigenvalues in } D^{(1)} \}. \quad (7.9)
\]

In (7.9) as well as in the rest of the proof, we always count all eigenvalues according to their respective algebraic multiplicities. Due to (7.8), we know that precisely the \( q \) eigenvalues \( a_{k_1,k_1}, \ldots, a_{k_q,k_q} \) of \( D \) are in \( D^{(1)} \). In particular, \( 0 \in S \) and \( S \neq \emptyset \). The proof is done once we show

\[
\tau := \sup S = 1 \quad (7.10a)
\]

and

\[
\tau \in S \quad \text{(i.e. } \tau = \max S). \quad (7.10b)
\]

As \( D^{(1)} \) and \( D^{(2)} \) are compact and disjoint, they have a positive distance,

\[
\epsilon := \text{dist}(D^{(1)}, D^{(2)}) > 0. \quad (7.11)
\]

Due to the continuous dependence of the eigenvalues on the coefficients of the matrix (cf. Cor. K.3 in the Appendix), there exists \( \delta > 0 \) such that, for each \( s,t \in [0,1] \) with \( |s - t| < \delta \), there exist enumerations \( \lambda_1(s), \ldots, \lambda_n(s) \) and \( \lambda_1(t), \ldots, \lambda_n(t) \) of the eigenvalues of \( \Phi(s) \) and \( \Phi(t) \), respectively, satisfying

\[
\forall k \in \{1, \ldots, n\} \quad |\lambda_k(s) - \lambda_k(t)| < \epsilon. \quad (7.12)
\]

Combining (7.12) with (7.11) yields

\[
\forall_{s,t \in [0,1]} \quad \left( s \in S \text{ and } |s - t| < \delta \right) \implies t \in S. \quad (7.13)
\]
Clearly, (7.13) implies both (7.10a) (since, otherwise, there were \( \tau < t < 1 \) with \( |\tau - t| < \delta \)) and (7.10b) (since there are \( 0 < t < \tau \) with \( |\tau - t| < \delta \)), thereby completing the proof.

\[ \blacksquare \]

**Remark 7.8.** If \( A \in \mathcal{M}(n, \mathbb{R}), n \in \mathbb{N} \), then Th. 7.7 can sometimes help to determine which regions can contain nonreal eigenvalues: For real matrices all Gershgorin disks are centered on the real line. Moreover, if \( A \) is real, then \( \lambda \in \sigma(A) \) implies \( \bar{\lambda} \in \sigma(A) \) and they must lie in the same Gershgorin disks (as they have the same distance to the real line). Thus, if \( \lambda \) is not real, then it can not lie in any Gershgorin disks known to contain just one eigenvalue.

**Example 7.9.** For the matrix \( A \) of Example 7.6, we see that the sets \( G_1 \cup G_2 = B_1(1) \cup B_2(-1) \) and \( G_3 = B_1(4) \) are disjoint (cf. Fig. 7.1). Thus, Th. 7.7 implies that \( G_3 \) contains precisely one eigenvalue and \( G_1 \cup G_2 \) contains precisely two eigenvalues. Combining this information with Rem. 7.8 shows that \( G_3 \) must contain a single real eigenvalue, whereas \( G_1 \cup G_2 \) could contain (as, indeed, it does) two conjugated nonreal eigenvalues.

### 7.3 Power Method

Given an \( n \times n \) matrix \( A \in \mathcal{M}(n, \mathbb{C}) \) and an initial vector \( z_0 \in \mathbb{C}^n \), the power method (also known as power iteration or von Mises iteration) means applying increasingly higher powers of \( A \) to \( z_0 \) according to the definition
\[
z_k := A z_{k-1} \quad \text{for each } k \in \mathbb{N} \quad \text{(i.e. } z_k = A^k z_0 \text{ for each } k \in \mathbb{N}_0). \tag{7.14}\]
We will see below that this simple iteration can, in many situations and with suitable modifications, be useful to approximate eigenvalues and eigenvectors (cf. Th. 7.12 and Cor. 7.15 below). Before we can proof the main results, we need some preparation:

**Proposition 7.10.** The following holds true for each \( A \in \mathcal{M}(n, \mathbb{C}), n \in \mathbb{N} \):
\[
\lim_{k \to \infty} A^k = 0 \iff r(A) < 1, \tag{7.15}\]
where \( r(A) \) denotes the spectral radius of \( A \) (cf. Def. 2.23).

**Proof.** First, suppose \( \lim_{k \to \infty} A^k = 0 \). Let \( \lambda \in \sigma(A) \) and choose a corresponding eigenvector \( v \in \mathbb{C}^n \setminus \{0\} \). Then, for each norm \( \| \cdot \| \) on \( \mathbb{C}^n \) with induced matrix norm on \( \mathcal{M}(n, \mathbb{C}) \),
\[
\forall k \in \mathbb{N} \quad \| A^k v \| = \| \lambda^k v \| = |\lambda|^k \| v \| \leq \| A^k \| \| v \|, \]
implying (since \( \lim_{k \to \infty} \| A^k \| = 0 \) and \( \| v \| > 0 \))
\[
\lim_{k \to \infty} |\lambda|^k = 0, \]
which, in turn, implies \( |\lambda| < 1 \). Conversely, suppose \( r(A) < 1 \). If \( B, T \in \mathcal{M}(n, \mathbb{C}) \) and \( T \) is invertible with \( A = T B T^{-1} \), then an obvious induction shows
\[
\forall k \in \mathbb{N}_0 \quad A^k = T B^k T^{-1}. \tag{7.16}\]
According to [Phi19b, Th. 9.13], we can choose $T$ such that (7.16) holds with $B$ in Jordan normal form, i.e. with $B$ in block diagonal form

$$B = \begin{pmatrix} B_1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & B_r \\ \end{pmatrix},$$

(7.17)

\[ 1 \leq r \leq n, \]

where

$$B_j = (\lambda_j) \text{ or } B_j = \begin{pmatrix} \lambda_j & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \lambda_j & 1 \\ \end{pmatrix},$$

\[ \lambda_j \in \sigma(A), B_j \in \mathcal{M}(n_j, \mathbb{C}), n_j \in \{1, \ldots, n\} \text{ for each } j \in \{1, \ldots, r\}, \text{ and } \sum_{j=1}^{r} n_j = n. \]

Using blockwise matrix multiplication (cf. [Phi19a, Sec. 7.5]) in (7.17) yields

$$\forall k \in \mathbb{N}_0 \quad B^k = \begin{pmatrix} B_1^k & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & B_r^k \\ \end{pmatrix}.$$  

(7.18)

We now show

$$\forall j \in \{1, \ldots, r\} \quad \lim_{k \to \infty} B^k_j = 0.$$  

(7.19)

To this end, we write $B_j = \lambda_j \text{Id} + N_j$, where $N_j \in \mathcal{M}(n_j, \mathbb{C})$ is a canonical nilpotent matrix,

$$N_j = 0 \text{ (zero matrix)} \quad \text{or} \quad N_j = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & \vdots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 1 \\ 0 \\ \end{pmatrix},$$

and apply the binomial theorem to obtain

$$\forall k \in \mathbb{N} \quad B^k_j = \sum_{\alpha=0}^{k} \binom{k}{\alpha} \lambda_j^{k-\alpha} N_j^\alpha = \sum_{\alpha=0}^{\min\{k,n_j-1\}} \binom{k}{\alpha} \lambda_j^{k-\alpha} N_j^\alpha.$$  

(7.20)

Using

$$\forall k \in \mathbb{N} \quad \forall \alpha \in \{0, \ldots, k\} \quad \binom{k}{\alpha} = \frac{k!}{\alpha! (k-\alpha)!} \leq k(k-1) \cdots (k-\alpha+1) \leq k^\alpha,$$

and an arbitrary induced matrix norm on $\mathcal{M}(n, \mathbb{C})$, (7.20) implies

$$\|B^k_j\| \leq \sum_{\alpha=0}^{n_j-1} k^\alpha |\lambda_j|^{k-\alpha} \|N_j\|^\alpha \rightarrow 0 \quad \text{for } k \to \infty,$$

proving (7.19). Next, we conclude $\lim_{k \to \infty} B^k = 0$ from (7.18). Thus, as $\|A^k\| \leq \|T\| \|B^k\| \|T^{-1}\|$, the claimed $\lim_{k \to \infty} A^k = 0$ also follows and completes the proof. \blacksquare
**Definition 7.11.** Let \( A \in \mathcal{M}(n, \mathbb{C}) \), \( z \in \mathbb{C}^n \), \( n \in \mathbb{N} \). Recalling from [Phi19b, Th. 9.13(a)] that \( \mathbb{C}^n \) is the direct sum of the generalized eigenspaces of \( A \),

\[
\mathbb{C}^n = \bigoplus_{\lambda \in \sigma(A)} M(\lambda), \quad M(\lambda) := M(\lambda, A) := \ker(A - \lambda \text{Id})^{m_\lambda(\lambda)},
\]

we say that \( z \) has a nontrivial part in \( M(\mu), \mu \in \sigma(A) \), if, and only if, in the unique decomposition of \( z \) with respect to the direct sum (7.21),

\[
z = \sum_{\lambda \in \sigma(A)} z_\lambda \quad \text{with} \quad z_\mu \neq 0, \quad \forall \lambda \in \sigma(A) \setminus \{\lambda\}, \quad z_\lambda \in M(\lambda).
\] (7.22)

**Theorem 7.12.** Let \( A \in \mathcal{M}(n, \mathbb{C}) \), \( n \in \mathbb{N} \). Suppose that \( A \) has an eigenvalue that is larger in size than all other eigenvalues of \( A \), i.e.

\[
\exists \lambda \in \sigma(A) \quad \forall \mu \in \sigma(A) \setminus \{\lambda\} \quad |\lambda| > |\mu|.
\] (7.23)

Moreover, assume \( \lambda \) to be semisimple (i.e. \( A|_{M(\lambda)} \) is diagonalizable and all Jordan blocks with respect to \( \lambda \) are trivial). Choosing an arbitrary norm \( \| \cdot \| \) on \( \mathbb{C}^n \) and starting with \( z_0 \in \mathbb{C}^n \), define the sequence \((z_k)_{k \in \mathbb{N}_0}\) via the power iteration (7.14). From the \( z_k \) also define sequences \((w_k)_{k \in \mathbb{N}_0}\) and \((\alpha_k)_{k \in \mathbb{N}_0}\) by

\[
w_k := \begin{cases} 
z_k & \text{for } z_k \neq 0, \\
z_0 & \text{for } z_k = 0 \end{cases} \quad \in \mathbb{C}^n \quad \text{for each } k \in \mathbb{N}_0,
\]

\[
\alpha_k := \begin{cases} 
z_k^*A_k z_k & \text{for } z_k \neq 0, \\
z_k^*z_k & \text{for } z_k = 0 \end{cases} \quad \in \mathbb{C} \quad \text{for each } k \in \mathbb{N}_0
\]

(treating the \( z_k \) as column vectors). If \( z_0 \) has a nontrivial part in \( M(\lambda) \) in the sense of Def. 7.11, then

\[
\lim_{k \to \infty} \alpha_k = \lambda
\] (7.25a)

and there exist sequences \((\phi_k)_{k \in \mathbb{N}}\) in \( \mathbb{R} \) and \((r_k)_{k \in \mathbb{N}}\) in \( \mathbb{C}^n \) such that, for all \( k \) sufficiently large,

\[
w_k = e^{i\phi_k} \frac{v_\lambda + r_k}{\|v_\lambda + r_k\|}, \quad \|v_\lambda + r_k\| \neq 0, \quad \lim_{k \to \infty} r_k = 0,
\]

(7.25b)

where \( v_\lambda \) is an eigenvector corresponding to the eigenvalue \( \lambda \).

**Proof.** If \( \lambda = 0 \), then condition (7.23) together with \( \lambda \) being semisimple implies \( A = 0 \) and there is nothing to prove. Thus, we proceed under the assumption \( \lambda \neq 0 \). As in the proof of Prop. 7.10, we choose \( T \in \mathcal{M}(n, \mathbb{C}) \) invertible such that (7.16) holds with \( B \) in Jordan normal form. However, this time, we choose \( T \) such that the (trivial) Jordan blocks corresponding to \( \lambda \) come first, i.e.

\[
B = \begin{pmatrix} B_1 & 0 \\ B_2 & \ddots \\ \vdots & \ddots & \ddots \\ 0 & \cdots & B_r \end{pmatrix}, \quad B_1 = \lambda \text{Id} \in \mathcal{M}(n_1, \mathbb{C}),
\] (7.26)
$B_j \in \mathcal{M}(n_j, \mathbb{C})$, is a Jordan matrix to $\lambda_j \in \sigma(A) \setminus \{\lambda\}$ for each $j = 2, \ldots, r$, $r \in \{1, \ldots, n\}$, $n_1, \ldots, n_r \in \mathbb{N}$, where $\sum_{j=1}^r n_j = n$. Since $T$ is invertible, the columns $t_1, \ldots, t_n$ of $T$ form a basis of $\mathbb{C}^n$. For each $j \in \{1, \ldots, n_1\}$, we compute

$$At_j = TBT^{-1}t_j = T\lambda e_j = \lambda t_j,$$

(7.27)

showing $t_j$ to be an eigenvector corresponding to $\lambda$. Thus, $t_1, \ldots, t_{n_1}$ form a basis of $M(\lambda) = \ker(A - \lambda \text{Id})$ (here, the eigenspace is the same as the generalized eigenspace), and, if $z_0$ has a nontrivial part in $M(\lambda)$, then there exist $\zeta_1, \ldots, \zeta_n \in \mathbb{C}$ with at least one $\zeta_j \neq 0$, $j \in \{1, \ldots, n_1\}$, such that

$$z_0 = v_\lambda + \sum_{j=n_1+1}^n \zeta_j t_j, \quad v_\lambda = \sum_{j=1}^{n_1} \zeta_j t_j.$$ 

We note $v_\lambda$ to be an eigenvector corresponding to $\lambda$. Next, we compute, for each $k \in \mathbb{N}_0$,

$$A^k z_0 = T B^k T^{-1} \left( v_\lambda + \sum_{j=n_1+1}^n \zeta_j t_j \right) = \lambda^k v_\lambda + T B^k \left( \sum_{j=n_1+1}^n \zeta_j e_j \right)$$

$$= \lambda^k \left( v_\lambda + T \frac{1}{\lambda^k} B^k \left( \sum_{j=n_1+1}^n \zeta_j e_j \right) \right),$$

(7.28)

where, in the last step, we have used $\lambda \neq 0$. In consequence, letting, for each $k \in \mathbb{N}_0$,

$$r_k := T \frac{1}{\lambda^k} B^k \left( \sum_{j=n_1+1}^n \zeta_j e_j \right),$$

(7.29)

we obtain, for each $k$ such that $v_\lambda + r_k \neq 0$,

$$w_k = \frac{A^k z_0}{\|A^k z_0\|} = \frac{\lambda^k}{\|A^k z_0\|} \frac{v_\lambda + r_k}{\|v_\lambda + r_k\|}.$$ 

(7.30)

Let $\beta_k := \frac{\lambda^k}{|\lambda|^k}$. Then $|\beta_k| = 1$ and, thus, there is $\phi_k \in [0, 2\pi]$ such that $\beta_k = e^{i\phi_k}$. To prove (7.25b), it now only remains to show $\lim_{k \to \infty} r_k = 0$. However, the hypothesis (7.23) implies

$$\forall j \in \{2, \ldots, r\} \quad r \left( \frac{1}{\lambda} B_j \right) = \frac{|\lambda_j|}{|\lambda|} < 1$$

and Prop. 7.10, in turn, implies $\lim_{k \to \infty} \frac{1}{\lambda^k} B_j^k = 0$. Using this in (7.29), where $\frac{1}{\lambda^k} B^k$ is applied to a vector in span$\{e_{n_1+1}, \ldots, e_n\}$, yields

$$\lim_{k \to \infty} r_k = T \lim_{k \to \infty} \frac{1}{\lambda^k} B^k \left( \sum_{j=n_1+1}^n \zeta_j e_j \right) = 0,$$

(7.31)

finishing the proof of (7.25b). Finally, it is

$$\lim_{k \to \infty} \alpha_k = \lim_{k \to \infty} \frac{z_k^* A z_k}{z_k^* z_k} = \lim_{k \to \infty} \frac{w_k^* A w_k}{w_k^* w_k} = \lim_{k \to \infty} \frac{(v_\lambda^* + r_k^*) A (v_\lambda + r_k)}{(v_\lambda^* + r_k^*) (v_\lambda + r_k)} = \frac{\lambda v_\lambda^* v_\lambda}{v_\lambda^* v_\lambda} = \lambda,$$

establishing (7.25a) and completing the proof.
Caveat 7.13. It is emphasized that, due to the presence of the factor $e^{i\phi_k}$, (7.25b) does \emph{not} claim the convergence of the $w_k$ to an eigenvector of $\lambda$. Indeed, in general, the sequence $(w_k)_{k \in \mathbb{N}}$ cannot be expected to converge. However, as $v_\lambda$ is an eigenvector for $\lambda$, so is every $\frac{e^{i\phi_k}v_\lambda}{\|v_\lambda + r_k\|}$ (and every other scalar multiple). Thus, what (7.25b) does claim is that, for sufficiently large $k$, $w_k$ is arbitrarily close to an eigenvector of $\lambda$, namely to the eigenvector $\frac{e^{i\phi_k}v_\lambda}{\|v_\lambda\|}$ (the eigenvector that $w_k$ is close to still depends on $k$).

Remark 7.14. (a) The hypothesis of Th. 7.12, that $z_0$ have a nontrivial part in $M(\lambda)$, can be difficult to check in practice. Still, it is usually not a severe restriction: Even if, initially, $z_0$ does not satisfy the condition, roundoff errors during the computation of the $z_k = A^k z_0$ will cause the condition to be satisfied for sufficiently large $k$.

(b) The rate of convergence of the $r_k$ in Th. 7.12 (and, thus, of the $\alpha_k$ as well) is determined by the size of $|\mu|/|\lambda|$, where $\mu$ is a second largest eigenvalue (in terms of absolute value). Therefore, the convergence will be slow if $|\mu|$ and $|\lambda|$ are close. 

Even though the power method applied according to Th. 7.12 will (under the hypothesis of the theorem) only approximate the dominant eigenvalue (i.e. the largest in terms of absolute value), the following modification, known as the \emph{inverse power method}, can be used to approximate other eigenvalues:

Corollary 7.15. Let $A \in \mathcal{M}(n, \mathbb{C})$, $n \in \mathbb{N}$, and assume $\lambda \in \sigma(A)$ to be semisimple. Let $p \in \mathbb{C} \setminus \{\lambda\}$ be closer to $\lambda$ than to any other eigenvalue of $A$, i.e.

$$\forall_{\mu \in \sigma(A) \setminus \{\lambda\}} \ 0 < |\lambda - p| < |\mu - p|. \quad (7.32)$$

Set

$$G := (A - p \text{Id})^{-1}. \quad (7.33)$$

Choosing an arbitrary norm $\| \cdot \|$ on $\mathbb{C}^n$ and starting with $z_0 \in \mathbb{C}^n \setminus \{0\}$, define the sequence $(z_k)_{k \in \mathbb{N}_0}$ via the power iteration (7.14) with $A$ replaced by $G$. From the $z_k$ also define sequences $(w_k)_{k \in \mathbb{N}_0}$ and $(\alpha_k)_{k \in \mathbb{N}_0}$ by

$$w_k := \frac{z_k}{\|z_k\|} = \frac{G^k z_0}{\|G^k z_0\|} \in \mathbb{C}^n \text{ for each } k \in \mathbb{N}_0, \quad (7.34a)$$

$$\alpha_k := \begin{cases} \frac{z_k^* G z_k}{z_k^* z_k} + p & \text{for } z_k^* G z_k \neq 0, \\ 0 & \text{otherwise} \end{cases} \in \mathbb{C} \text{ for each } k \in \mathbb{N}_0 \quad (7.34b)$$

(treating the $z_k$ as column vectors). If $z_0$ has a nontrivial part in $M((\lambda - p)^{-1}, G)$ in the sense of Def. 7.11, then

$$\lim_{k \to \infty} \alpha_k = \lambda \quad (7.35a)$$

and there exist sequences $(\phi_k)_{k \in \mathbb{N}}$ in $\mathbb{R}$ and $(r_k)_{k \in \mathbb{N}}$ in $\mathbb{C}^n$ such that, for all $k$ sufficiently large,

$$w_k = e^{i\phi_k} \frac{v_\lambda + r_k}{\|v_\lambda + r_k\|}, \quad \|v_\lambda + r_k\| \neq 0, \quad \lim_{k \to \infty} r_k = 0, \quad (7.35b)$$

where $v_\lambda$ is an eigenvector corresponding to the eigenvalue $\lambda$. 

Proof. First note that (7.32) implies $p \notin \sigma(A)$, such that $G$ is well-defined. As $G$ is invertible, $z_0 \neq 0$ implies $G^k z_0 \neq 0$ for each $k \in \mathbb{N}_0$, showing the $w_k$ to be well-defined as well. The idea is to apply Th. 7.12 to $G$. Let $\mu \in \mathbb{C} \setminus \{p\}$ and $v \in \mathbb{C}^n \setminus \{0\}$. Then
\[
\begin{align*}
Av &= \mu v \\
\iff (\mu - p) v &= (A - p \text{Id}) v \\
\iff (A - p \text{Id})^{-1} v &= (\mu - p)^{-1} v, \quad (7.36)
\end{align*}
\]
showing $\mu$ to be an eigenvalue for $A$ with corresponding eigenvector $v$ if, and only if, $(\mu - p)^{-1}$ is an eigenvalue for $G = (A - p \text{Id})^{-1}$ with the same corresponding eigenvector $v$. In particular, (7.32) implies $\frac{1}{\lambda - p}$ to be the dominant eigenvalue of $G$. Next, notice that the hypothesis that $\lambda$ be semisimple for $A$ implies $M(\lambda, A)$ to have a basis of eigenvectors $v_1, \ldots, v_n$, $n_1 = m_\lambda(\lambda)$ (cf. (7.21)), which, by (7.36), then constitutes a basis of eigenvectors for $M(\frac{1}{\lambda - p}, G)$ as well, showing $\frac{1}{\lambda - p}$ to be semisimple for $G$. Thus, all hypotheses of Th. 7.12 are satisfied and we obtain $\lim_{k \to \infty} \alpha_k = (\lambda - p) + p = \lambda$, proving (7.35a). Moreover, (7.35b) also follows, where $v_\lambda$ is an eigenvector for $\frac{1}{\lambda - p}$ and $G$. However, then $v_\lambda$ is also an eigenvector for $\lambda$ and $A$ by (7.36).

Remark 7.16. In practice, one computes the $z_k$ for the inverse power method of Cor. 7.15 by solving the linear systems $(A - p \text{Id}) z_k = z_{k-1}$ for $z_k$. As the matrix $A - p \text{Id}$ does not vary, but only the right-hand side of the system varies, one should solve the systems by first computing a suitable decomposition of $A - p \text{Id}$ (e.g. a QR decomposition, cf. Rem. 5.15).

Remark 7.17. In sufficiently benign situations (say, if $A$ is diagonalizable and all eigenvalues are in $\mathbb{R}_0^+$, e.g., if $A$ is Hermitian), one can combine the inverse power method (IPM) of Cor. 7.15 with a bisection (or similar) method to obtain all eigenvalues of $A$: One starts by determining upper and lower bounds for $\sigma(A) \subseteq [m, M]$ (for example, by using Cor. 7.5). One then chooses $p := (m + M)/2$ as the midpoint and determines $\lambda = \lambda(p)$. Next, one uses the midpoints $p_1 := (m + p)/2$ and $p_2 := (p + M)/2$ and iterates the method, always choosing the midpoints of previous intervals. For each new midpoint $p$, four possibilities can occur: (i) one finds a new eigenvalue by IPM, (ii) IPM finds an eigenvalue that has already been found before, (iii) IPM fails, since $p$ itself is an eigenvalue, (iv) IPM fails since $p$ is exactly in the middle between two other eigenvalues. Cases (iii) and (iv) are unlikely, but if they do occur, they need to be recognized and treated separately. If not all eigenvalues are real, one can still use an analogous method, one just has to cover a certain compact section of the complex plane by finer and finer grids. If one were to use squares for the grids, one could use the centers of the squares to replace the midpoints of the intervals in the described procedure.

### 7.4 QR Method

The QR algorithm for the approximation of the eigenvalues of $A \in \mathcal{M}(n, \mathbb{K})$ makes use of the QR decomposition $A = QR$ of $A$ into a unitary matrix $Q$ and an upper triangular matrix $R$ (cf. Def. 5.12(a)). We already know from Th. 5.24 that $Q$ and $R$ can be obtained via the Householder method of Def. 5.23.
Algorithm 7.18 (QR Method). Given $A \in \mathcal{M}(n, \mathbb{K})$, $n \in \mathbb{N}$, we define the following algorithm for the computation of a sequence $(A^{(k)})_{k \in \mathbb{N}}$ of matrices in $\mathcal{M}(n, \mathbb{K})$: Let $A^{(1)} := A$. For each $k \in \mathbb{N}$, let

$$A^{(k+1)} := R_k Q_k,$$  \hspace{1cm} (7.37a)

where

$$A^{(k)} = Q_k R_k$$  \hspace{1cm} (7.37b)

is a QR decomposition of $A^{(k)}$.

Remark 7.19. In each step of the QR method of Alg. 7.18, one needs to obtain a QR decomposition of $A^{(k)}$. For a fully populated matrix $A^{(k)}$, the complexity of obtaining a QR decomposition is $O(n^3)$ (obtaining a QR decomposition via the Householder method of Def. 5.23 requires $O(n)$ matrix multiplications, each matrix multiplication needing $O(n^2)$ multiplications). The complexity can be improved to $O(n^2)$ by first transforming $A$ into so-called Hessenberg form (the complexity of the transformation is $O(n^3)$, cf. Sections K.2 and K.3 of the Appendix). Thus, in practice, it is always advisable to transform $A$ into Hessenberg form before applying the QR method.

Lemma 7.20. Let $A \in \mathcal{M}(n, \mathbb{K})$, $n \in \mathbb{N}$. Moreover, for each $k \in \mathbb{N}$, let matrices $A^{(k)}, R_k, Q_k \in \mathcal{M}(n, \mathbb{K})$ be defined according to Alg. 7.18. Then the following identities hold for each $k \in \mathbb{N}$:

$$A^{(k+1)} = Q_k^* A^{(k)} Q_k,$$  \hspace{1cm} (7.38a)

$$A^{(k+1)} = (Q_1 Q_2 \cdots Q_k)^* A (Q_1 Q_2 \cdots Q_k),$$ \hspace{1cm} (7.38b)

$$A^k = (Q_1 Q_2 \cdots Q_k) (R_k R_{k-1} \cdots R_1).$$ \hspace{1cm} (7.38c)

Proof. (7.38a) holds, as

$$\forall k \in \mathbb{N} \quad Q_k^* A^{(k)} Q_k = Q_k^* Q_k R_k Q_k = R_k Q_k = A^{(k+1)}.$$

A simple induction applied to (7.38a) yields (7.38b). Another induction yields (7.38c): For $k = 1$, it is merely (7.37b). For $k \in \mathbb{N}$, we compute

$$Q_1 \cdots Q_k \overbrace{Q_{k+1} R_{k+1} \cdots R_1}^{(7.38b)} (Q_1 \cdots Q_k)^* A (Q_1 \cdots Q_k) (R_k \cdots R_1)$$ \hspace{1cm} \ind. hyp. \hspace{1cm}$$AA^k = A^{k+1},$$

which completes the induction and establishes the case. \hfill \blacksquare

In the following Th. 7.22, it will be shown that, under suitable hypotheses, the matrices $A^{(k)}$ of the QR method become, asymptotically, upper triangular, with the diagonal entries converging to the eigenvalues of $A$. Theorem 7.22 is not the most general convergence result of this kind. We will remark on possible extensions afterwards and we will also indicate that some of the hypotheses are not as restrictive as they might seem at first glance.
**Notation 7.21.** Let $n \in \mathbb{N}$. By $\text{diag}(\lambda_1, \ldots, \lambda_n)$ we denote the diagonal $n \times n$ matrix with diagonal entries $\lambda_1, \ldots, \lambda_n$, i.e.

$$\text{diag}(\lambda_1, \ldots, \lambda_n) := D = (d_{kl}), \quad d_{kl} := \begin{cases} \lambda_k & \text{for } k = l, \\ 0 & \text{for } k \neq l \end{cases}$$

(we have actually used this notation before). If $A = (a_{kl})$ is an $n \times n$ matrix, then, by $\text{diag}(A)$, we denote the diagonal $n \times n$ matrix that has precisely the same diagonal entries as $A$, i.e.

$$\text{diag}(A) := D = (d_{kl}), \quad d_{kl} := \begin{cases} a_{kk} & \text{for } k = l, \\ 0 & \text{for } k \neq l. \end{cases}$$

**Theorem 7.22.** Let $A \in \mathcal{M}(n, \mathbb{K})$, $n \in \mathbb{N}$. Assume $A$ to be diagonalizable, invertible, and such that all eigenvalues $\lambda_1, \ldots, \lambda_n \in \sigma(A) \subseteq \mathbb{K}$ are distinct, satisfying

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0. \quad (7.39a)$$

Let $T \in \mathcal{M}(n, \mathbb{K})$ be invertible and such that

$$A = TDT^{-1}, \quad D := \text{diag}(\lambda_1, \ldots, \lambda_n). \quad (7.39b)$$

Assume $T^{-1}$ to have an LU decomposition (without permutations), i.e., assume

$$T^{-1} = LU, \quad (7.39c)$$

where $L, U \in \mathcal{M}(n, \mathbb{K})$, $L$ being unipotent and lower triangular, $U$ being upper triangular. If $(A^{(k)})_{k \in \mathbb{N}}$ is defined according to Alg. 7.18, then the $A^{(k)}$ are, asymptotically, upper triangular, with the diagonal entries converging to the eigenvalues of $A$, i.e.

$$\forall_{j,l \in \{1, \ldots, n\}, \ j > l} \lim_{k \to \infty} a_{jl}^{(k)} = 0 \quad (7.40a)$$

and

$$\forall_{j \in \{1, \ldots, n\}} \lim_{k \to \infty} a_{jj}^{(k)} = \lambda_j. \quad (7.40b)$$

Moreover, there exists a constant $C > 0$, such that we have the error estimates

$$\forall_{k \in \mathbb{N}} \left( \forall_{j,l \in \{1, \ldots, n\}, \ j > l} \left| a_{jl}^{(k)} \right| \leq C q^k, \quad \forall_{j \in \{1, \ldots, n\}} \left| a_{jj}^{(k)} - \lambda_j \right| \leq C q^k \right), \quad (7.40c)$$

where

$$q := \max \left\{ \frac{|\lambda_{j+1}|}{|\lambda_j|} : j \in \{1, \ldots, n-1\} \right\} < 1. \quad (7.40d)$$
Proof. For $n = 1$, there is nothing to prove (one has $A = D = A^{(k)}$ for each $k \in \mathbb{N}$). Thus, we assume $n > 1$ for the rest of the proof. We start by making a number of observations for $k \in \mathbb{N}$ being fixed. From (7.38c), we obtain a QR decomposition of $A^k$:

$$A^k = \tilde{Q}_k \tilde{R}_k, \quad \tilde{Q}_k := Q_1 \cdots Q_k, \quad \tilde{R}_k := R_k \cdots R_1. \quad (7.41)$$

On the other hand, we also have

$$A^k = TD^kLU = X_kD^kU, \quad X_k := TD^kLD^{-k}. \quad (7.42)$$

We note $X_k$ to be invertible, as it is the product of invertible matrices. Thus $X_k$ has a QR decomposition $X_k = Q_{X_k}R_{X_k}$ with invertible upper triangular matrix $R_{X_k}$. Plugging the QR decomposition of $X_k$ into (7.42), thus, yields a second QR decomposition of $A^k$, namely

$$A^k = Q_{X_k}(R_{X_k}D^kU).$$

Due to Prop. 5.13, we now obtain a unitary diagonal matrix $S_k$ such that

$$\tilde{Q}_k = Q_{X_k}S_k, \quad \tilde{R}_k = S_k^*R_{X_k}D^kU. \quad (7.43)$$

This will now allow us to compute a somewhat complicated-looking representation of $A^{(k)}$, which, nonetheless, will turn out to be useful in advancing the proof: We have, for $k \geq 2$,

$$Q_k = (Q_1 \cdots Q_{k-1})^{-1}(Q_1 \cdots Q_{k-1}Q_k) = \tilde{Q}_{k-1}Q_k = S_{k-1}^*Q_{X_{k-1}}Q_{X_k}S_k,$$

$$R_k = (R_kR_{k-1} \cdots R_1)(R_kR_{k-1} \cdots R_1)^{-1} = \tilde{R}_k\tilde{R}_{k-1}^{-1}$$

$$= S_{k-1}^*R_{X_{k-1}}D^kUU^{-1}D^{-(k-1)}R_{X_{k-1}}^{-1}S_{k-1} = S_{k-1}^*R_{X_{k-1}}DR_{X_{k-1}}^{-1}S_{k-1},$$

implying

$$A^{(k)} \overset{(7.37b)}{=} Q_kR_k = S_{k-1}^*Q_{X_{k-1}}^{-1}Q_{X_k}R_{X_k}DR_{X_{k-1}}^{-1}S_{k-1}$$

$$= S_{k-1}^*R_{X_{k-1}}R_{X_{k-1}}^{-1}Q_{X_{k-1}}^{-1}Q_{X_k}R_{X_k}DR_{X_{k-1}}^{-1}S_{k-1}$$

$$= S_{k-1}^*R_{X_{k-1}}X_{k-1}^{-1}X_kDR_{X_{k-1}}^{-1}S_{k-1}. \quad (7.44)$$

The next step is to investigate the entries of

$$X_k = TL^{(k)}, \quad L^{(k)} := D^kLD^{-k}. \quad (7.45)$$

As $L$ is lower triangular with $l_{jj} = 1$, we have

$$l^{(k)}_{jl} = \lambda_j^k l_{jl} \lambda_l^{-k} = \begin{cases} 0 & \text{for } j < l, \\ 1 & \text{for } j = l, \\ l_{jl}\left(\frac{\lambda_j}{\lambda_l}\right)^k & \text{for } j > l, \end{cases}$$

and

$$|l^{(k)}_{jl}| \begin{cases} = 0 & \text{for } j < l, \\ = 1 & \text{for } j = l, \\ \leq C_L q^k & \text{for } j > l, \end{cases}$$
where \( q \) is as in (7.40d) and
\[
C_L := \max\{|l_{ij}| : l, j \in \{1, \ldots, n\}\}.
\]
Thus, there is a lower triangular matrix \( E_k \) and a constant \( C_E \in \mathbb{R}^+ \) such that
\[
L^{(k)} = \text{Id} + E_k, \quad \|E_k\|_2 \leq C_E C_L q^k.
\]
(7.46)
Then
\[
X_k = T L^{(k)} = T + T E_k.
\]
(7.47)
Set, for \( k \geq 2 \),
\[
F_k := (E_k - E_{k-1})(\text{Id} + E_{k-1})^{-1}.
\]
(7.48)
As \( \lim_{k \to \infty} E_k = 0 \), the sequence \( \left(\| (\text{Id} + E_{k-1})^{-1}\|_2\right)_{k \in \mathbb{N}} \) is bounded, i.e. there is a constant \( C_F > 0 \) such that
\[
\forall k \geq 2, \quad \| F_k \|_2 \leq \| E_k - E_{k-1} \|_2 \| (\text{Id} + E_{k-1})^{-1}\|_2 \leq C_E C_L (q^k + q^{k-1}) C_F \leq C_1 q^k,
\]
where \( C_1 := 2 C_E C_L C_F q^{-1} \) (note \( q > 0 \)).
Now (7.48) is equivalent to
\[
\text{Id} + E_k = (\text{Id} + E_{k-1})(\text{Id} + F_k),
\]
implying
\[
X_{k-1}^{-1} X_k = (T + T E_{k-1})^{-1}(T + T E_k) = (\text{Id} + E_{k-1})^{-1}(\text{Id} + E_k) = \text{Id} + F_k.
\]
Plugging the last equality into (7.44) yields
\[
A^{(k)} = S_{k-1}^* R_{X_{k-1}}(\text{Id} + F_k) D R_{X_{k-1}}^{-1} S_{k-1} = S_{k-1}^* R_{X_{k-1}} D R_{X_{k-1}}^{-1} S_{k-1} + S_{k-1}^* R_{X_{k-1}} F_k D R_{X_{k-1}}^{-1} S_{k-1} =: A_D^{(k)} + A_F^{(k)}.
\]
(7.49)
To estimate \( \|A_F^{(k)}\|_2 \), we note
\[
\| R_{X_k} \|_2 = \| Q_{X_k}^* X_k \|_2 = \|X_k\|_2, \quad \| R_{X_k}^{-1} \|_2 = \|X_k^{-1} Q_{X_k} \|_2 = \|X_k^{-1}\|_2,
\]
since \( Q_{X_k} \) is unitary and, thus, isometric. Due to (7.46) and (7.47), \( \lim_{k \to \infty} X_k = T \), such that the sequences \( (\| R_{X_k} \|_2)_{k \in \mathbb{N}} \) and \( (\| R_{X_k}^{-1} \|_2)_{k \in \mathbb{N}} \) are bounded by some constant \( C_R > 0 \). Thus, as the \( S_k \) are also unitary,
\[
\forall k \geq 2, \quad \| A_F^{(k)} \|_2 = \| S_{k-1}^* R_{X_{k-1}} F_k D R_{X_{k-1}}^{-1} S_{k-1} \|_2 \leq C_R^2 \| D \|_2 C_1 q^k.
\]
(7.50)
Next, we observe each \( A_D^{(k)} \) to be an upper triangular matrix, as it is the product of upper triangular matrices. Moreover, each \( S_k \) is a diagonal matrix and \( \text{diag}(R_{X_k})^{-1} = \text{diag}(R_{X_k}^{-1}) \). Thus,
\[
\forall k \geq 2, \quad \text{diag}(A_D^{(k)}) = S_{k-1}^* \text{diag}(R_{X_{k-1}}) D \text{diag}(R_{X_{k-1}}^{-1}) S_{k-1} = D.
\]
(7.51)
Finally, as we know the 2-norm on \( \mathcal{M}(n, \mathbb{K}) \) to be equivalent to the max-norm on \( \mathbb{K}^{n^2} \), combining (7.49) – (7.51) proves everything that was claimed in (7.40).
Remark 7.23. (a) Hypothesis (7.39c) of Th. 7.22 that $T^{-1}$ has an LU decomposition without permutations is not as severe as it may seem. If permutations are necessary in the LU decomposition of $T^{-1}$, then Th. 7.22 and its proof remain the same, except that the eigenvalues will occur permuted in the limiting diagonal matrix $D$. Moreover, it is an exercise to show that, if the matrix $A$ in Th. 7.22 has Hessenberg form with $0 \notin \{a_{21}, \ldots, a_{n,n-1}\}$ (in view of Rem. 7.19, this is the case of most interest), then hypothesis (7.39c) can always be satisfied: As an intermediate step, one can show that, in this case,

$$\forall j \in \{1, \ldots, n-1\} \quad \text{span}\{e_1, \ldots, e_j\} \cap \text{span}\{t_{j+1}, \ldots, t_n\} = \{0\},$$

where $e_1, \ldots, e_{n-1}$ are the standard unit vectors, and $t_1, \ldots, t_n$ are the eigenvectors of $A$ that form the columns of $T$.

(b) If $A$ has several eigenvalues of equal modulus (e.g. multiple eigenvalues or nonreal eigenvalues of a real matrix), then the sequence of the QR method can no longer be expected to be asymptotically upper triangular, only block upper triangular. For example, if $|\lambda_1| = |\lambda_2| > |\lambda_3| = |\lambda_4| = |\lambda_5|$, then the asymptotic form will look like

$$\begin{pmatrix}
* & * & * & * & * \\
* & * & * & * & * \\
* & * & * & * & * \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix},$$

where the upper diagonal $2 \times 2$ block still has eigenvalues $\lambda_1, \lambda_2$ and the lower diagonal $3 \times 3$ block still has eigenvalues $\lambda_3, \lambda_4, \lambda_5$ (see [Cia89, Sec. 6.3] and references therein).

Remark 7.24. (a) The efficiency of the QR method can be improved significantly by the introduction of so-called spectral shifts. One replaces (7.37) by

$$A^{(k+1)} := R_kQ_k + \mu_k \text{Id}, \quad (7.52a)$$

where

$$A^{(k)} - \mu_k \text{Id} = Q_k R_k \quad (7.52b)$$

is a QR decomposition of $A^{(k)} - \mu_k \text{Id}$ and $\mu_k \in \mathbb{C}$ is the shift parameter of the $k$th step. As it turns out, using the lowest diagonal entry of the previous iteration as shift parameter, $\mu_k := a_{nn}^{(k)}$, is a simple choice that yields good results (see [SK11, Sec. 5.5.2]). According to [HB09, Sec. 27.3], an even better choice, albeit less simple, is to choose $\mu_k$ as the eigenvalue of

$$\begin{pmatrix}
a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} & a_{n,n-1}^{(k)} \\
a_{n-1,n}^{(k)} & a_{n,n-1}^{(k)} & 0 \\
a_{n,n-1}^{(k)} & 0 & 0
\end{pmatrix}$$

that is closest to $a_{nn}^{(k)}$. If one needs to find nonreal eigenvalues of a real matrix, the related strategy of so-called double shifts is warranted (see [SK11, Sec. 5.5.3]).
(b) The above-described QR method with shifts is particularly successful if applied to Hessenberg matrices in combination with a strategy known as deflation. After each step, each element of the first subdiagonal is tested if it is sufficiently close to 0 to be treated as “already converged”, i.e. to be replaced by 0. Then, if $a_{j(j-1)}^{(k)}$, $2 \leq j \leq n$, has been replaced by 0, the QR method is applied separately to the upper diagonal $(j-1) \times (j-1)$ block and the lower diagonal $(n-j+1) \times (n-j+1)$ block. If $j = 2$ or $j = n$ (this is what usually happens if the shift strategy as described in (a) is applied), then $a_{11}^{(k)} = \lambda_1$ or $a_{nn}^{(k)} = \lambda_n$ is already correct and one has to continue with an $(n-1) \times (n-1)$ matrix.

8 Minimization

8.1 Motivation, Setting

Minimization (more generally, optimization) is a vast field, and we can only cover a few selected aspects in this class. We will consider minimization problems of the form

$$\min f(a), \quad f : A \rightarrow \mathbb{R}. \quad (8.1)$$

The function $f$ to be minimized is often called the objective function or the objective functional of the problem, even though this seems to be particularly common in the context of constrained minimization (see below).

In Sec. 6.1, we discussed the tasks of finding zeros and fixed points of functions, and we found that, in most situations, both problems are equivalent. Minimization provides a third perspective onto this problem class: If $F : A \rightarrow Y$, where $(Y, \| \cdot \|)$ is a normed vector space, and $\alpha > 0$, then

$$F(x) = 0 \iff \|F(x)\|^\alpha = 0 = \min \{\|F(a)\|^\alpha : a \in A\}. \quad (8.2)$$

More generally, given $y \in Y$, we have the equivalence

$$F(x) = y \iff \|F(x) - y\|^\alpha = 0 = \min \{\|F(a) - y\|^\alpha : a \in A\}. \quad (8.2)$$

However, minimization is more general: If we seek the (global) min of the function $f : A \rightarrow \mathbb{R}$, $f(a) := \|F(a) - y\|^\alpha$, then it will yield a solution to $F(x) = y$, provided such a solution exists. On the other hand, $f$ can have a minimum even if $F(x) = y$ does not have a solution. If $Y = \mathbb{R}^n$, $n \in \mathbb{N}$, with the 2-norm, and $\alpha = 2$, then the minimization of $f$ is known as a least squares problem. The least squares problem is called linear if $A$ is a normed vector space over $\mathbb{R}$ and $F : A \rightarrow Y$ is linear, and nonlinear otherwise.

One distinguishes between constrained and unconstrained minimization, where constrained minimization can be seen as restricting the objective function $f : A \rightarrow \mathbb{R}$ to some subset of $A$, determined by the constraints. Constraints can be given explicitly or implicitly. If $A = \mathbb{R}$, then an explicit constraint might be to seek nonnegative
solutions. Implicit constraints can be given in the form of equation constraints – for example, if \( A = \mathbb{R}^n \), then one might want to minimize \( f \) on the set of all solutions to \( Ma = b \), where \( M \) is some real \( m \times n \) matrix. In this class, we will restrict ourselves to \textit{unconstrained} minimization. In particular, we will not cover the simplex method, a method for so-called linear programming, which consists of minimizing a linear \( f \) with affine constraints in finite dimensions (see, e.g., [HH92, Sec. 9] or [Str08, Sec. 5]). For constrained nonlinear optimization see, e.g., [QSS07, Sec. 7.3]). The optimal control of differential equations is an example of constrained minimization in infinite dimensions (see, e.g., [Phi09]).

We will only consider methods that are suitable for unconstrained nonlinear minimization, noting that even the \textit{linear} least squares problem is a nonlinear minimization problem. Both methods involving derivatives (such as gradient methods) and derivative-free methods are of interest. Derivative-based methods typically show faster convergence rates, but, on the other hand, they tend to be less robust, need more function evaluations (i.e. each step of the method is more costly), and have a smaller scope of applicability (e.g., the function to be minimized must be differentiable). Below, we will study derivative-based methods in Sections 8.3 and 8.4, whereas derivative-free methods can be found in Appendix L.1 and L.2.

8.2 Local Versus Global and the Role of Convexity

\textbf{Definition 8.1.} Let \((X, \| \cdot \|)\) be a normed vector space, \( A \subseteq X \), and \( f : A \longrightarrow \mathbb{R} \).

(a) Given \( x \in A \), \( f \) has a (\textit{strict}) \textit{global min} at \( x \) if, and only if, \( f(x) \leq f(y) \) \((f(x) < f(y))\) for each \( y \in A \setminus \{x\} \).

(b) Given \( x \in X \), \( f \) has a (\textit{strict}) \textit{local min} at \( x \) if, and only if, there exists \( r > 0 \) such that \( f(x) \leq f(y) \) \((f(x) < f(y))\) for each \( y \in \{y \in A : \|y - x\| < r\} \setminus \{x\} \).

Basically all deterministic minimization methods (in particular, all methods considered in this class) have the drawback that they will find \textit{local} minima rather then global minima. This problem vanishes if every local min is a global min, which is, indeed, the case for convex functions. For this reason, we will briefly study convex functions in this section, before treating concrete minimization methods thereafter.

\textbf{Definition 8.2.} A subset \( C \) of a real vector space \( X \) is called \textit{convex} if, and only if, for each \((x, y) \in C^2\) and each \( 0 \leq \alpha \leq 1 \), one has \( \alpha x + (1 - \alpha) y \in C \). If \( A \subseteq X \), then \( \text{conv} \ A \) is the intersection of all convex sets containing \( A \) and is called the \textit{convex hull} of \( A \) (clearly, the intersection of convex sets is always convex and, thus, the convex hull of a set is always convex, cf. [Phi19b, Def. and Rem. 1.23]).

\textbf{Definition 8.3.} Let \( C \) be a convex subset of a real vector space \( X \), \( f : C \longrightarrow \mathbb{R} \).
(a) $f$ is called \textit{convex} if, and only if,
\[ f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y). \quad (8.3) \]

(b) $f$ is called \textit{strictly convex} if, and only if, the inequality in (8.3) is strict whenever $x \neq y$ and $0 < \alpha < 1$.

(c) $f$ is called \textit{strongly} or \textit{uniformly\footnote{In the literature, strong convexity is often defined as a special case of a more general notion of uniform convexity.} convex} if, and only if, there exists a norm $\| \cdot \|$ on $X$ and $\mu \in \mathbb{R}^+$ such that
\[ f(\alpha x + (1-\alpha) y) + \mu \alpha (1-\alpha) \|x-y\|^2 \leq \alpha f(x) + (1-\alpha) f(y). \quad (8.4) \]

Remark 8.4. It is immediate from Def. 8.3 that uniform convexity implies strict convexity, which, in turn, implies convexity. In general, a convex function need not be uniformly convex, and a strictly convex function need not be uniformly convex (cf. Ex. 8.5(b)), and a strictly convex function need not be uniformly convex (cf. Ex. 8.5(d)).

Example 8.5. (a) We know from Calculus (cf. \cite[Prop. 9.32(b)]{Phi16a}) that a continuous function $f : [a, b] \rightarrow \mathbb{R}$, that is twice differentiable on $]a, b[$, is strictly convex if $f'' > 0$ on $]a, b[$. Thus, $f : \mathbb{R}^+_0 \rightarrow \mathbb{R}$, $f(x) = x^p$, $p > 1$, is strictly convex (note $f''(x) = p(p-1)x^{p-2} > 0$ for $x > 0$).

(b) Let $C$ be a convex subset of a normed vector space $(X, \| \cdot \|)$ over $\mathbb{R}$. Then $\| \cdot \| : C \rightarrow \mathbb{R}$ is convex by the triangle inequality, but not strictly convex if $C$ contains a segment $S$ of a one-dimensional subspace: If $0 \neq x_0 \in X$ and $0 \leq a < b$ ($a, b \in \mathbb{R}$) such that $S = \{tx_0 : a \leq t \leq b\} \subseteq C$, then
\[ \forall 0 \leq \alpha \leq 1 \quad \| \alpha ax_0 + (1-\alpha) bx_0 \| = (\alpha a + (1-\alpha)b) \|x_0\| = \alpha \|ax_0\| + (1-\alpha) \|bx_0\|. \]

If $\langle \cdot, \cdot \rangle$ is an inner product on $X$ and $\|x\| = \sqrt{\langle x, x \rangle}$, then
\[ f : C \rightarrow \mathbb{R}, \quad f(x) := \|x\|^2, \]
is uniformly (in particular, strictly) convex: Indeed, if $x, y \in C$ and $0 \leq \alpha \leq 1$, then, using
\[ (1-\alpha) - \alpha(1-\alpha) = 1 - 2\alpha + \alpha^2 = (1-\alpha)^2, \quad (8.5) \]
we obtain
\[ f(\alpha x + (1-\alpha)y) = \alpha^2\|x\|^2 + 2\alpha(1-\alpha)\langle x, y \rangle + (1-\alpha)^2\|y\|^2 \]
\[ \overset{(8.5)}{=} \alpha\|x\|^2 + (1-\alpha)\|y\|^2 - \alpha(1-\alpha)\left(\|x\|^2 - 2\langle x, y \rangle + \|y\|^2\right) \]
\[ = \alpha f(x) + (1-\alpha)f(y) - \alpha(1-\alpha)\left(\|x\|^2 - 2\langle x, y \rangle + \|y\|^2\right) \]
\[ = \alpha f(x) + (1-\alpha)f(y) - \alpha(1-\alpha)\langle x-y, x-y \rangle \]
\[ = \alpha f(x) + (1-\alpha)f(y) - \alpha(1-\alpha)\|x-y\|^2, \]
showing (8.4) to hold with $\mu = 1$. 
(c) The objective function of the linear least squares problem is convex and, if the linear function $F$ is injective, even strictly convex (exercise).

(d) Consider $f : \mathbb{R}^+ \rightarrow \mathbb{R}$, $f(x) := x^4$. Then $f$ is strictly convex by (a). We claim that $f$ is not uniformly convex: Indeed, if $f$ were uniformly convex, then, by (8.8) below, it had to hold that, with $\mu > 0$,

$$\forall x, y \in \mathbb{R}^+, \quad x^4 - y^4 = f(x) - f(y) \geq f'(y)(x - y) + \mu (x - y)^2 = 4y^3(x - y) + \mu (x - y)^2.$$ 

However, if we let $x := \frac{2}{n}$ and $y := \frac{1}{n}$, then the above inequality is equivalent to

$$\frac{16}{n^4} - \frac{1}{n^4} \geq \frac{4}{n^3} \cdot \frac{1}{n} + \frac{\mu}{n^2} \iff \frac{11}{n^4} \geq \frac{\mu}{n^2} \iff \frac{11}{\mu} \geq n^2,$$

showing it will always be violated for sufficiently large $n$.

**Theorem 8.6.** Let $n \in \mathbb{N}$ and let $C \subseteq \mathbb{R}^n$ be convex and open. Moreover, let $f : C \rightarrow \mathbb{R}$ be continuously differentiable. Then the following holds:

(a) $f$ is convex if, and only if,

$$\forall x, y \in C \quad f(x) - f(y) \geq \nabla f(y) (x - y). \quad (8.6)$$

(b) $f$ is strictly convex if, and only if,

$$\forall x, y \in C \quad (x \neq y \Rightarrow f(x) - f(y) > \nabla f(y) (x - y)). \quad (8.7)$$

(c) $f$ is uniformly convex if, and only if,

$$\forall x, y \in C \quad f(x) - f(y) \geq \nabla f(y) (x - y) + \mu \|x - y\|^2 \quad (8.8)$$

for some $\mu \in \mathbb{R}^+$ and some norm $\| \cdot \|$ on $\mathbb{R}^n$.

In each case, the stated condition is still sufficient for the stated form of convexity if the partials of $f$ exist, but are not necessarily continuous.

**Proof.** Suppose, (8.8) with $\mu \in \mathbb{R}_0^+$ and some norm $\| \cdot \|$ on $\mathbb{R}^n$. Fix $x, y \in C$, $\alpha \in [0, 1]$, and define $z := \alpha x + (1 - \alpha)y \in C$. Then

$$f(x) - f(z) \geq \nabla f(z) (x - z) + \mu \|x - z\|^2, \quad (8.9a)$$

$$f(y) - f(z) \geq \nabla f(z) (y - z) + \mu \|y - z\|^2. \quad (8.9b)$$

We multiply (8.9a) by $\alpha$, (8.9b) by $(1 - \alpha)$, and add the resulting inequalities, also using

$$x - z = (1 - \alpha)(x - y), \quad y - z = \alpha(y - x),$$
to obtain:
\[
\alpha f(x) + (1 - \alpha) f(y) - f(\alpha x + (1 - \alpha) y) \\
= \alpha f(x) + (1 - \alpha) f(y) - \alpha f(z) - (1 - \alpha) f(z) \\
\geq \alpha \nabla f(z) \left( (1 - \alpha) (x - y) \right) - (1 - \alpha) \nabla f(z) (\alpha x - y) \\
+ \mu \alpha (1 - \alpha)^2 \|x - y\|^2 + \mu \alpha^2 (1 - \alpha) \|x - y\|^2 \\
= \mu \alpha (1 - \alpha) \|x - y\|^2.
\] (8.10)

Thus, if (8.8) holds with \( \mu > 0 \), then (8.10) holds with \( \mu > 0 \), showing \( f \) to be uniformly convex; if (8.6) holds, then (8.8) and (8.10) hold with \( \mu = 0 \), showing \( f \) to be convex; if (8.7) holds and \( \alpha \in [0, 1], x \neq y \), then (8.10) holds with \( \mu = 0 \) and strict inequality, showing \( f \) to be strictly convex.

Now assume \( f \) to be uniformly convex. Then there exist \( \mu \in \mathbb{R}^+ \) and some norm \( \| \cdot \| \) on \( \mathbb{R}^n \) such that, for each \( x, y \in C \) and each \( \alpha \in [0, 1], \)
\[
f(y + \alpha (x - y)) = f(\alpha x + (1 - \alpha) y) \\
\leq \alpha f(x) + (1 - \alpha) f(y) - \mu \alpha (1 - \alpha) \|x - y\|^2
\] (8.11a)
or, rearranged for \( \alpha > 0, \)
\[
\frac{f(y + \alpha (x - y)) - f(y)}{\alpha} \leq f(x) - f(y) - \mu (1 - \alpha) \|x - y\|^2.
\] (8.11b)

Thus, according to the mean value theorem (cf. [Phi16b, Th. 4.35]), there exists \( s(\alpha) \in ]0, \alpha[ \) such that
\[
\nabla f(y + s(\alpha)(x - y))(x - y) = \frac{f(y + \alpha (x - y)) - f(y)}{\alpha} \leq f(x) - f(y) - \mu (1 - \alpha) \|x - y\|^2.
\] (8.11c)

Taking the limit for \( \alpha \to 0 \) and using the assumed continuity of \( \nabla f \) then yields
\[
\nabla f(y)(x - y) \leq f(x) - f(y) - \mu \|x - y\|^2,
\] (8.11d)
proving (8.8). Setting \( \mu := 0 \) in (8.11) shows (8.6) to hold for convex \( f \). It remains to prove the validity of (8.7) for \( f \) being strictly convex. The argument of (8.11) does not suffice in this case, as we lose the strict inequality, when taking the limit in (8.11c). Thus, we assume \( f \) to be strictly convex, \( x \neq y \), and proceed as follows: Letting
\[
z := \frac{1}{2} (x + y) \in C,
\]
the strict convexity of \( f \) yields
\[
f(z) = f\left( \frac{1}{2}x + \frac{1}{2} y\right) < \frac{1}{2} f(x) + \frac{1}{2} f(y)
\]
and, thus, we obtain
\[
\nabla f(y)(x - y) = 2 \nabla f(y)(z - y) \leq 2 \left( f(z) - f(y) \right) \leq f(x) - f(y),
\]
completing the proof of (8.7).
The set of mins of $C$

Assume the convexity of $C$. So, if $x \in f(b)$, then this is the unique local min of $f$, then it follows that $f(x) = f(b)$.

Thus, due to the convexity of $C$, $x_0 + \alpha (x - x_0) = (1 - \alpha) x_0 + \alpha x$.

Then, due to the convexity of $C$, $x_0 + \alpha (x - x_0) \in C$ for each $\alpha \in [0, 1]$. Moreover, for sufficiently small $\alpha$, namely for each $\alpha \in R := [0, \min\{1, r/\|x_0 - x\|\}]$, one has $x_0 + \alpha (x - x_0) \in C_r$. As $x_0$ is a local min and $f$ is convex, for each $\alpha \in R$, one obtains:

$$f(x_0) \leq f((1 - \alpha) x_0 + \alpha x) \leq (1 - \alpha) f(x_0) + \alpha f(x).$$

(8.13)

After subtracting $f(x_0)$ and dividing by $\alpha > 0$, (8.13) yields $f(x_0) \leq f(x)$, showing that $x_0$ is actually a global min as claimed.

(b): Let $x_0 \in C$ be a min of $f$. From (a), we already know that $x_0$ must be a global min. So, if $x \in C$ is any min of $f$, then it follows that $f(x) = f(x_0)$. If $\alpha \in [0, 1]$, then the convexity of $f$ implies that

$$f((1 - \alpha) x_0 + \alpha x) \leq (1 - \alpha) f(x_0) + \alpha f(x) = f(x_0).$$

(8.14)

As $x_0$ is a global min, (8.14) implies that $(1 - \alpha) x_0 + \alpha x$ is also a global min for each $\alpha \in [0, 1]$, showing that the set of mins of $f$ is convex as claimed.

Theorem 8.9. Let $(X, \|\cdot\|)$ be a normed vector space over $\mathbb{R}$, $C \subseteq X$, and $f : C \rightarrow \mathbb{R}$. Assume $C$ is a convex set, and $f$ is a strictly convex function. If $x \in C$ is a local min of $f$, then this is the unique local min of $f$, and, moreover, it is strict.
Proof. According to Th. 8.8, every local min of $f$ is also a global min of $f$. Seeking a contradiction, assume there is $y \in C$, $y \neq x$, such that $y$ is also a min of $f$. As $x$ and $y$ are both global mins, $f(x) = f(y)$ is implied. Define $z := \frac{1}{2}(x + y)$. Then $z \in C$ due to the convexity of $C$. Moreover, due to the strict convexity of $f$,

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

in contradiction to $x$ being a global min. Thus, $x$ must be the unique min of $f$, also implying that the min must be strict. 

\[\blacksquare\]

### 8.3 Gradient-Based Descent Methods

We are concerned with differentiable objective functions $f$ defined on (subsets of) $\mathbb{R}^n$.

**Definition 8.10.** Let $G \subseteq \mathbb{R}^n$ be open, $n \in \mathbb{N}$, and $f : G \rightarrow \mathbb{R}$.

(a) Given $x^1 \in G$, a **stepsize direction iteration** defines a sequence $(x^k)_{k \in \mathbb{N}}$ in $G$ by setting

$$\forall \ k \in \mathbb{N} \quad x^{k+1} := x^k + t_k u^k,$$

where $t_k \in \mathbb{R}^+$ is the **stepsize** and $u^k \in \mathbb{R}^n$ is the direction of the $k$th step. In general, $t_k$ and $u^k$ will depend on $G$, $f$, $x^1, \ldots, x^k, t_1, \ldots, t_{k-1}, u^1, \ldots, u^{k-1}$. Different choices of $t_k$ and $u^k$ correspond to different methods – some examples are provided in (b) – (d) below.

(b) A stepsize direction iteration as in (a) is called a **descent method** for $f$ if, and only if, for each $k \in \mathbb{N}$, either $u^k = 0$ or

$$\phi_k : I_k \rightarrow \mathbb{R}, \quad \phi_k(t) := f(x^k + t u^k), \quad I_k := \{ t \in \mathbb{R}_+^* : x^k + t u^k \in G \},$$

is strictly decreasing in some neighborhood of 0.

(c) If $f$ is differentiable, then a stepsize direction iteration as in (a) is called **gradient method** for $f$ or **method of steepest descent** if, and only if,

$$\forall \ k \in \mathbb{N} \quad u^k = -\nabla f(x^k),$$

i.e. if the directions are always chosen as the antigradient of $f$ in $x^k$.

(d) If $f$ is differentiable, then the **conjugate gradient method** is a stepsize direction iteration, where

$$\forall \ k \in \{2, \ldots, n\} \quad u^k = -\nabla f(x^k) + \beta_k u^{k-1},$$

and the $\beta_k \in \mathbb{R}^+$ are chosen such that $(u^k)_{k \in \{1, \ldots, n\}}$ forms an orthogonal system with respect to a suitable inner product.
Remark 8.11. In the situation of Def. 8.10 above, if \( f \) is differentiable, then its directional derivative at \( x^k \in G \) in the direction \( u^k \in \mathbb{R}^n \) is

\[
\phi_k'(0) = \lim_{t \downarrow 0} \frac{f(x^k + tu^k) - f(x^k)}{t} = \nabla f(x^k) \cdot u^k.
\]

(8.20)

If \( \|u^k\|_2 = 1 \), then the Cauchy-Schwarz inequality implies

\[
|\phi_k'(0)| = |\nabla f(x^k) \cdot u^k| \leq \|\nabla f(x^k)\|_2,
\]

(8.21)

justifying the name method of steepest descent for the gradient method of Def. 8.10(c). The next lemma shows it is, indeed, a descent method as defined in Def. 8.10(b), provided that \( f \) is continuously differentiable.

Notation 8.12. If \( X \) is a real vector space and \( x, y \in X \), then

\[
[x, y] := \{(1 - t)x + ty : t \in [0, 1]\} = \text{conv}\{x, y\}
\]

(8.22)

denotes the line segment between \( x \) and \( y \).

Lemma 8.13. Let \( G \subseteq \mathbb{R}^n \) be open, \( n \in \mathbb{N} \), and \( f : G \to \mathbb{R} \) continuously differentiable. If

\[
\forall \ k \in \mathbb{N} \quad \nabla f(x^k) \cdot u^k < 0,
\]

(8.23)

then each stepsize direction iteration as in Def. 8.10(a) is a descent method according to Def. 8.10(b). In particular, the gradient method of Def. 8.10(c) is a descent method.

Proof. According to (8.20) and (8.23), we have

\[
\phi_k'(0) < 0
\]

(8.24)

for \( \phi_k \) of (8.17). As \( \phi_k' \) is continuous in 0, (8.24) implies \( \phi_k \) to be strictly decreasing in a neighborhood of 0.

We continue to remain in the setting of Def. 8.10. Suppose we are in step \( k \) of a descent method, i.e. we know \( \phi_k \) of (8.17) to be strictly decreasing in a neighborhood of 0. Then the question remains of how to choose the next stepsize \( t_k \)? One might want to choose \( t_k \) such that \( \phi_k \) has a global (or at least a local) min at \( t_k \). A local min can, for instance, be obtained by applying the golden section search of Alg. L.5 in the Appendix. However, if function evaluations are expensive, then one might want to resort to some simpler method, even if it does not provide a local min of \( \phi_k \). Another potential problem with Alg. L.5 is that, in general, one has little control over the size of \( t_k \), which makes convergence proofs difficult. The following Armijo rule has the advantage of being simple and, at the same time, relating the size of \( t_k \) to the size of \( \nabla f(x^k) \):

Algorithm 8.14 (Armijo Rule). Let \( G \subseteq \mathbb{R}^n \) be open, \( n \in \mathbb{N} \), and \( f : G \to \mathbb{R} \) differentiable. The Armijo rule has two parameters, namely \( \sigma, \beta \in [0, 1] \). If \( x^k \in G \) and \( u^k \in \mathbb{R}^n \) is either 0 or such that \( \nabla f(x^k) \cdot u^k < 0 \), then set

\[
t_k := \begin{cases} 
1 & \text{for } u^k = 0, \\
\max \{ \beta^l : l \in \mathbb{N}_0, \ x^k + \beta^l u^k \in G, \text{ and } \beta^l \text{ satisfies (8.26)} \} & \text{for } u^k \neq 0,
\end{cases}
\]

(8.25)
Theorem 8.17. Let

\[
f(x^k + \beta^l u^k) \leq f(x^k) + \sigma \beta^l \nabla f(x^k) \cdot u^k.
\]  

(8.26)

In practise, since \((\beta^l)_{l \in \mathbb{N}_0}\) is strictly decreasing, one merely has to test the validity of \(x^k + \beta^l u^k \in G\) and (8.26) for \(l = 0, 1, 2, \ldots\) stopping once (8.26) holds for the first time.

Lemma 8.15. Under the assumptions of Alg. 8.14, if \(u^k \neq 0\), then the max in (8.25) exists. In particular, the Armijo rule is well-defined.

Proof. Let \(u^k \neq 0\). Since \((\beta^l)_{l \in \mathbb{N}_0}\) is decreasing, we only have to show that an \(l \in \mathbb{N}_0\) exists such that \(x^k + \beta^l u^k \in G\) and (8.26) holds. Since \(\lim_{l \to \infty} \beta^l = 0\) and \(G\) is open, \(x^k + \beta^l u^k \in G\) holds for each sufficiently large \(l\). Seeking a contradiction, we now assume

\[
\exists N \in \mathbb{N} \quad \forall l \geq N \quad \left( x^k + \beta^l u^k \in G \quad \text{and} \quad f(x^k + \beta^l u^k) > f(x^k) + \sigma \beta^l \nabla f(x^k) \cdot u^k \right).
\]

Thus, as \(f\) is differentiable,

\[
\nabla f(x^k) \cdot u^k \overset{(8.20)}{=} \lim_{l \to \infty} \frac{f(x^k + \beta^l u^k) - f(x^k)}{\beta^l} \geq \sigma \nabla f(x^k) \cdot u^k.
\]

Then the assumption \(\nabla f(x^k) \cdot u^k < 0\) implies \(1 \leq \sigma\), in contradiction to \(0 < \sigma < 1\). □

In preparation for a convergence theorem for the gradient method (Th. 8.17 below), we prove the following lemma:

Lemma 8.16. Let \(G \subseteq \mathbb{R}^n\) be open, \(n \in \mathbb{N}\), and \(f : G \to \mathbb{R}\) continuously differentiable. Suppose \(x \in G\), \(u \in \mathbb{R}^n\), and consider sequences \((x^k)_{k \in \mathbb{N}}\) in \(G\), \((u^k)_{k \in \mathbb{N}}\) in \(\mathbb{R}^n\), \((t_k)_{k \in \mathbb{N}}\) in \(\mathbb{R}^+\), satisfying

\[
\lim_{k \to \infty} x^k = x, \quad \lim_{k \to \infty} u^k = u, \quad \lim_{k \to \infty} t_k = 0.
\]  

(8.27)

Then

\[
\lim_{k \to \infty} \frac{f(x^k + t_k u^k) - f(x^k)}{t_k} = \nabla f(x) \cdot u.
\]  

(8.28)

Proof. As \(G\) is open, (8.27) guarantees

\[
\exists \epsilon > 0 \quad \exists N \in \mathbb{N} \quad \forall k > N \quad y^k := x^k + t_k u^k \in B_\epsilon(x^k) \subseteq G.
\]

Thus, we can apply the mean value theorem to obtain

\[
\forall k > N \quad \exists \xi^k \in [x^k, y^k] \quad f(x^k + t_k u^k) - f(x^k) = t_k \nabla f(\xi^k) \cdot u^k.
\]  

(8.29)

Since \(\lim_{k \to \infty} \xi^k = x\) and \(\nabla f\) is continuous, (8.29) implies (8.28). □

Theorem 8.17. Let \(f : \mathbb{R}^n \to \mathbb{R}\) be continuously differentiable, \(n \in \mathbb{N}\). If the sequence \((x^k)_{k \in \mathbb{N}}\) is given by the gradient method of Def. 8.10(c) together with the Armijo rule according to Alg. 8.14, then every cluster point \(x\) of the sequence is a stationary point of \(f\).
Proof. If there is \( k_0 \in \mathbb{N} \) such that \( u^{k_0} = -\nabla f(x^{k_0}) = 0 \), then \( \lim_{k \to \infty} x^k = x^{k_0} \) is clear from (8.16) and we are already done. Thus, we may now assume \( u^k \neq 0 \) for each \( k \in \mathbb{N} \). Let \( x \in \mathbb{R}^n \) be a cluster point of \( (x^k)_{k \in \mathbb{N}} \) and \( (x^{kj})_{j \in \mathbb{N}} \) a subsequence such that \( x = \lim_{j \to \infty} x^{kj} \). The continuity of \( f \) implies \( f(x) = \lim_{j \to \infty} f(x^{kj}) \). But then, as \( (f(x^k))_{k \in \mathbb{N}} \) is decreasing by the Armijo rule, one also obtains

\[
f(x) = \lim_{j \to \infty} f(x^{kj}) = \lim_{k \to \infty} f(x^k).
\]

Thus, \( \lim_{k \to \infty} (f(x^{k+1}) - f(x^k)) = 0 \), and, using the Armijo rule again, we have

\[
0 < \sigma t_k \| \nabla f(x^k) \|^2_2 \leq f(x^k) - f(x^{k+1}) \to 0,
\]

where \( \sigma \in [0,1[ \) is the parameter from the Armijo rule. Seeking a contradiction, we now assume \( \nabla f(x) \neq 0 \). Then the continuity of \( \nabla f \) yields \( \lim_{j \to \infty} \nabla f(x^{kj}) = \nabla f(x) \neq 0 \), which, together with (8.30) implies \( \lim_{j \to \infty} t_{kj} = 0 \), where \( \beta \in [0,1[ \) is the parameter from the Armijo rule and \( l_j \in \mathbb{N}_0 \) is chosen according to (8.25). Thus, for each sufficiently large \( j, m_j := l_j - 1 \in \mathbb{N}_0 \) and

\[
0 < \sigma t_k \| \nabla f(x^k) \|^2_2 \leq f(x^k) - f(x^{k+1}) \to 0,
\]

where \( \sigma \in [0,1[ \) is the parameter from the Armijo rule. Seeking a contradiction, we now assume \( \nabla f(x) \neq 0 \). Then the continuity of \( \nabla f \) yields \( \lim_{j \to \infty} \nabla f(x^{kj}) = \nabla f(x) \neq 0 \), which, together with (8.30) implies \( \lim_{j \to \infty} t_{kj} = 0 \), where \( \beta \in [0,1[ \) is the parameter from the Armijo rule and \( l_j \in \mathbb{N}_0 \) is chosen according to (8.25). Thus, for each sufficiently large \( j, m_j := l_j - 1 \in \mathbb{N}_0 \) and

\[
\frac{f(x^{kj} + \beta^{m_j} u^{kj}) - f(x^{kj})}{\beta^{m_j}} > \sigma \nabla f(x^{kj}) \cdot u^{kj}
\]

(8.31)

which, rearranged, reads

\[
\frac{f(x^{kj} + \beta^{m_j} u^{kj}) - f(x^{kj})}{\beta^{m_j}} > \sigma \nabla f(x^{kj}) \cdot u^{kj}
\]

(here we have used that \( x^{kj} + \beta^{m_j} u^{kj} \in G = \mathbb{R}^n \)). Taking the limit for \( j \to \infty \) and observing \( \lim_{j \to \infty} \beta^{m_j} = 0 \) in connection with Lem. 8.16 then yields

\[
-\| \nabla f(x) \|^2_2 \geq -\sigma \| \nabla f(x) \|^2_2,
\]

i.e. \( \sigma \geq 1 \). This contradiction to \( \sigma < 1 \) completes the proof of \( \nabla f(x) = 0 \).

As a caveat, it is underlined that Th. 8.17 does not claim the sequence \( (x^k)_{k \in \mathbb{N}} \) to have cluster points. It merely states that, if cluster points exist, then they must be stationary points. For convex differentiable functions, stationary points are minima:

**Proposition 8.18.** If \( C \subseteq \mathbb{R}^n \) is open and convex, \( n \in \mathbb{N} \), and \( f : C \to \mathbb{R} \) is differentiable and (strictly) convex, then \( f \) has a (strict) global min at each stationary point (of which there is at most one in the strictly convex case, cf. Th. 8.9).

**Proof.** Let \( x \in C \) be a stationary point of \( f \), i.e. \( \nabla f(x) = 0 \). Let \( y \in C, y \neq x \). Since \( C \) is open and convex, there exists an open interval \( I \subseteq \mathbb{R} \) such that \( 0,1 \in I \) and \( \{x + t(y - x) : t \in I\} \subseteq C \). Consider the auxiliary function

\[
\phi : I \to \mathbb{R}, \quad \phi(t) := f(x + t(y - x)).
\]

(8.31)
We claim that $\phi$ is (strictly) convex if $f$ is (strictly) convex: Indeed, let $s, t \in I$, $s \neq t$, and $\alpha \in ]0, 1[$. Then
\[
\phi(\alpha s + (1-\alpha)t) = f(x + (\alpha s + (1-\alpha)t)(y - x)) \\
= f(\alpha(x + s(y - x)) + (1-\alpha)(x + t(y - x))) \\
\leq \alpha f(x + s(y - x)) + (1-\alpha)f(x + t(y - x)) \\
= \alpha\phi(s) + (1-\alpha)\phi(t) \tag{8.32}
\]
with strict inequality if $f$ is strictly convex. By the chain rule, $\phi$ is differentiable with
\[
\forall t \in I \quad \phi'(t) = \nabla f(x + t(y - x)) \cdot (y - x).
\]
Thus, we know from Calculus (cf. [Phi16a, Prop. 9.31]) that $\phi'$ is (strictly) increasing and, as $\phi'(0) = 0$, $\phi$ must be (strictly) increasing on $I \cap \mathbb{R}_0^+$. Thus, $f(y) \geq f(x)$ ($f(y) > f(x)$), showing $x$ to be a (strict) global min. \hfill \blacksquare

**Corollary 8.19.** Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be (strictly) convex and continuously differentiable, $n \in \mathbb{N}$. If the sequence $(x^k)_{k \in \mathbb{N}}$ is given by the gradient method of Def. 8.10(c) together with the Armijo rule according to Alg. 8.14, then every cluster point $x$ of the sequence is a (strict) global min of $f$ (in particular, there is at most one cluster point in the strictly convex case).

**Proof.** Everything is immediate when combining Th. 8.17 with Prop. 8.18. \hfill \blacksquare

A disadvantage of the (hypotheses) of the previous results is that they do not guarantee the convergence of the gradient method (or even the existence of local minima). An important class of functions, where we can, indeed, guarantee both (together with an error estimate that shows a linear convergence rate) is the class of convex quadratic functions, which we will study next.

**Definition 8.20.** Let $Q \in \mathcal{M}(n, \mathbb{R})$ be symmetric and positive definite, $a \in \mathbb{R}^n$, $n \in \mathbb{N}$, $\gamma \in \mathbb{R}$. Then the function
\[
f : \mathbb{R}^n \rightarrow \mathbb{R}, \quad f(x) := \frac{1}{2} x^t Q x + a^t x + \gamma, \tag{8.33}
\]
where $a, x$ are considered column vectors, is called the *quadratic function* given by $Q$, $a$, and $\gamma$.

**Proposition 8.21.** Let $Q \in \mathcal{M}(n, \mathbb{R})$ be symmetric and positive definite, $a \in \mathbb{R}^n$, $n \in \mathbb{N}$, $\gamma \in \mathbb{R}$. Then the function $f$ of (8.33) has the following properties:

(a) $f$ is uniformly convex.

(b) $f$ is differentiable with
\[
\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad \nabla f(x) = x^t Q + a^t. \tag{8.34}
\]
(c) $f$ has a unique and strict global min at $x^* := -Q^{-1}a$.

(d) Given $x, u \in \mathbb{R}^n$, $u \neq 0$, the function

$$
\phi : \mathbb{R} \rightarrow \mathbb{R}, \quad \phi(t) := f(x + tu),
$$

has a unique and strict global min at $t^* := -\frac{(x^Q + a^u)}{u^Q u}$. 

Proof. (a): Since $\langle x, y \rangle := x^T Q y$ defines an inner product on $\mathbb{R}^n$, the uniform convexity of $f$ follows from Ex. 8.5(b): Indeed, if we let $\forall x \in \mathbb{R}^n \parallel x \parallel := \sqrt{x^T Q x}$, then Ex. 8.5(b) implies, for each $x, y \in \mathbb{R}^n$ and each $\alpha \in [0, 1]$: 

$$
\begin{align*}
&f(\alpha x + (1 - \alpha) y) + \frac{1}{2} \alpha (1 - \alpha) \parallel x - y \parallel^2 \\
&= \frac{1}{2}(\alpha x + (1 - \alpha) y)^T Q(\alpha x + (1 - \alpha) y) + a^T(\alpha x + (1 - \alpha) y) + \gamma \\
&\quad + \frac{1}{2} \alpha (1 - \alpha) \parallel x - y \parallel^2 \\
&\leq \frac{1}{2} \alpha x^T Q x + \frac{1}{2} (1 - \alpha) y^T Q y + \alpha a^T x + (1 - \alpha) a^T y + \alpha \gamma + (1 - \alpha) \gamma \\
&= \alpha f(x) + (1 - \alpha) f(y).
\end{align*}
$$

(b): Exercise.

c) follows from (a), (b), and Prop. 8.18, as $f$ has a stationary point at $x^* = -Q^{-1}a$.

(d): The computation in (8.32) with $y - x$ replaced by $u$ shows $\phi$ to be strictly convex. Moreover,

$$
\phi' : \mathbb{R} \rightarrow \mathbb{R}, \quad \phi'(t) = \nabla f(x + tu)u = (x^T + tu^T)Q u + a^T u.
$$

Then the assertion follows from Prop. 8.18, as $\phi$ has a stationary point at $t^* := -\frac{(x^Q + a^u)}{u^Q u}$. 

Theorem 8.22 (Kantorovich Inequality). Let $Q \in \mathcal{M}(n, \mathbb{R})$ be symmetric and positive definite, $n \in \mathbb{N}$. If $\alpha := \min \sigma(Q) \in \mathbb{R}^+$ and $\beta := \max \sigma(Q) \in \mathbb{R}^+$ are the smallest and largest eigenvalues of $Q$, respectively, then

$$
\forall x \in \mathbb{R}^n \setminus \{0\} \quad F(x) := \frac{(x^T x)^2}{(x^T Q x)(x^T Q^{-1} x)} \geq \frac{4\alpha \beta}{(\alpha + \beta)^2} = 4 \left( \sqrt{\frac{\alpha}{\beta}} + \sqrt{\frac{\beta}{\alpha}} \right)^{-2}.
$$

Proof. Since $Q$ is symmetric positive definite, we can write the eigenvalues in the form $0 < \alpha = \lambda_1 \leq \cdots \leq \lambda_n = \beta$ with corresponding orthonormal eigenvectors $v_1, \ldots, v_n$. Fix $x \in \mathbb{R}^n \setminus \{0\}$. As the $v_1, \ldots, v_n$ form a basis,

$$
\exists \mu_1, \ldots, \mu_n \in \mathbb{R} \quad \left( x = \sum_{i=1}^n \mu_i v_i, \quad \mu := \sum_{i=1}^n \mu_i^2 > 0 \right).
$$
As the \( v_1, \ldots, v_n \) are orthonormal eigenvectors for the \( \lambda_1, \ldots, \lambda_n \), we obtain

\[
F(x) = \frac{\mu^2}{(\sum_{i=1}^n \lambda_i \mu_i^2)(\sum_{i=1}^n \lambda_i^{-1} \mu_i^2)},
\]

which, letting \( \gamma_i := \mu_i^2 / \mu \), becomes

\[
F(x) = \frac{1}{(\sum_{i=1}^n \gamma_i \lambda_i)(\sum_{i=1}^n \gamma_i \lambda_i^{-1})}.
\]

Noting \( 0 \leq \gamma_i \leq 1 \) and \( \sum_{i=1}^n \gamma_i = 1 \), we see that

\[
\lambda := \sum_{i=1}^n \gamma_i \lambda_i
\]
is a convex combination of the \( \lambda_i \), i.e. \( \alpha \leq \lambda \leq \beta \). We would now like to show

\[
\sum_{i=1}^n \gamma_i \lambda_i^{-1} \leq \frac{\alpha + \beta - \lambda}{\alpha \beta}.
\]

(8.37)

To this end, observe that \( l : \mathbb{R} \to \mathbb{R}, l(\lambda) := \frac{\alpha + \beta - \lambda}{\alpha \beta} \), represents the line through the points \( P_\alpha := (\alpha, \alpha^{-1}) \) and \( P_\beta := (\beta, \beta^{-1}) \), and that the set

\[
C := \{ (\lambda, \mu) \in \mathbb{R}^2 : \alpha \leq \lambda \leq \beta, \mu \leq l(\lambda) \},
\]

below that line, is a convex subset of \( \mathbb{R}^2 \). The function \( \varphi : \mathbb{R}^+ \to \mathbb{R}^+, \varphi(\lambda) := \lambda^{-1} \), is strictly convex (e.g., due to \( \varphi''(\lambda) = 2\lambda^{-3} > 0 \)), implying \( P_i := (\lambda_i, \lambda_i^{-1}) \in C \) for each \( i \in \{1, \ldots, n\} \) (cf. [Phi16a, Prop. 9.29]). In consequence, \( P := \sum_{i=1}^n \gamma_i P_i = (\bar{\lambda}, \sum_{i=1}^n \gamma_i \lambda_i^{-1}) \in C \), i.e. \( \sum_{i=1}^n \gamma_i \lambda_i^{-1} \leq l(\bar{\lambda}) \), proving (8.37). Thus,

\[
F(x) = \frac{1}{\overline{\lambda}(\sum_{i=1}^n \gamma_i \lambda_i^{-1})} \geq \frac{\alpha \beta}{\overline{\lambda}(\alpha + \beta - \overline{\lambda})}.
\]

(8.38)

To estimate \( F(x) \) further, we determine the global min of the differentiable function

\[
\phi : [\alpha, \beta] \to \mathbb{R}, \quad \phi(\lambda) = \frac{\alpha \beta}{\overline{\lambda}(\alpha + \beta - \lambda)}.
\]

The derivative is

\[
\phi' : [\alpha, \beta] \to \mathbb{R}, \quad \phi'(\lambda) = -\frac{\alpha \beta(\alpha + \beta - 2\lambda)}{\lambda^2(\alpha + \beta - \lambda)^2}.
\]

Let \( \lambda_0 := \frac{\alpha + \beta}{2} \). Clearly, \( \phi'(\lambda) < 0 \) for \( \lambda < \lambda_0 \), \( \phi'(\lambda_0) = 0 \), and \( \phi'(\lambda) > 0 \) for \( \lambda > \lambda_0 \), showing that \( \phi \) has its global min at \( \lambda_0 \). Using this fact in (8.38) yields

\[
F(x) \geq \frac{\alpha \beta}{\lambda_0(\alpha + \beta - \lambda_0)} = \frac{4\alpha \beta}{(\alpha + \beta)^2},
\]

(8.39)

completing the proof of (8.36).
Theorem 8.23. Let \( Q \in \mathcal{M}(n, \mathbb{R}) \) be symmetric and positive definite, \( a \in \mathbb{R}^n \), \( n \in \mathbb{N} \), \( \gamma \in \mathbb{R} \), and let \( f \) be the corresponding quadratic function according to (8.33). If the sequence \((x^k)_{k \in \mathbb{N}}\) is given by the gradient method of Def. 8.10(c), where (cf. Prop. 8.21(d))
\[
\forall k \in \mathbb{N} \quad (u^k \neq 0 \implies t_k := -\frac{(x^k)^tQ + a^t}{(u^k)^tQu^k} > 0),
\]
then \( \lim_{k \to \infty} x^k = x^* \), where \( x^* = -Q^{-1}a \) is the global min of \( f \) according to Prop. 8.21(c). Moreover, we have the error estimate
\[
\forall k \in \mathbb{N} \quad f(x^{k+1}) - f(x^*) \leq \left(\frac{\beta - \alpha}{\alpha + \beta}\right)^2 \left(f(x^k) - f(x^*)\right) \leq \left(\frac{\beta - \alpha}{\alpha + \beta}\right)^2 \left(f(x^1) - f(x^*)\right),
\]
where \( \alpha := \min \sigma(Q) \in \mathbb{R}^+ \) and \( \beta := \max \sigma(Q) \in \mathbb{R}^+ \).

Proof. Introducing the abbreviations
\[
\forall k \in \mathbb{N} \quad g^k := (\nabla f(x^k))^t = Qx^k + a,
\]
we see that, indeed, for \( u^k \neq 0 \),
\[
t_k = \frac{(g^k)^t g^k}{(g^k)^t Qu^k} > 0.
\]
If there is \( k_0 \in \mathbb{N} \) such that \( u^{k_0} = -\nabla f(x^{k_0}) = 0 \), then, clearly, \( \lim_{k \to \infty} x^k = x^{k_0} = x^* \) and we are already done. Thus, we may now assume \( u^k \neq 0 \) for each \( k \in \mathbb{N} \). It suffices to show the first estimate in (8.41), as this implies the second estimate as well as convergence to \( x^* \) (where the latter claim is due to the uniform convexity of \( f \) according to Prop. 8.21(a) in combination with Cor. 8.7). Using the formulas for \( g^k \) and \( t_k \) from above, we have
\[
\forall k \in \mathbb{N} \quad x^{k+1} = x^k - t_k g^k = x^k - \frac{(g^k)^t g^k}{(g^k)^t Qu^k} g^k,
\]
impling
\[
\forall k \in \mathbb{N} \quad f(x^{k+1}) = \frac{1}{2} (x^k - t_k g^k)^t Q(x^k - t_k g^k) + a^t(x^k - t_k g^k) + \gamma
\]
\[
= f(x^k) - \frac{1}{2} t_k (x^k)^t Qg^k - \frac{1}{2} t_k (g^k)^t Qx^k + \frac{1}{2} t_k^2 (g^k)^t Qg^k - t_k a^t g^k
\]
\[
= f(x^k) - t_k (Qx^k + a)^t g^k + \frac{1}{2} t_k^2 (g^k)^t Qg^k
\]
\[
= f(x^k) - \frac{1}{2} \frac{(g^k)^t g^k}{(g^k)^t Qu^k} \left(\frac{(g^k)^t g^k}{(g^k)^t Qu^k}\right) (Qx^k + a)^t g^k,
\]
where it was used that \( (g^k)^t Qx^k = ((g^k)^t Qx^k)^t = (x^k)^t Qg^k \in \mathbb{R} \). Next, we claim that
\[
\forall k \in \mathbb{N} \quad f(x^{k+1}) - f(x^*) = \left(1 - \frac{(g^k)^t g^k}{(g^k)^t Qu^k} \right) (f(x^k) - f(x^*)).
\]
which, in view of (8.42), is equivalent to
\[
\frac{1}{2} ((g^k)^t g^k)^2 = \frac{((g^k)^t g^k)^2}{((g^k)^t Q g^k)((g^k)^t Q^{-1} g^k)} (f(x^k) - f(x^*))
\]
and
\[
2(f(x^k) - f(x^*)) = (g^k)^t Q^{-1} g^k. \tag{8.44}
\]
Indeed,
\[
2(f(x^k) - f(x^*)) = (x^k)^t Q x^k + 2a^t x^k + 2\gamma - (a^t Q^{-1} Q^{-1} a - 2a^t Q^{-1} a + 2\gamma)
\]
\[
= (x^k)^t g^k + a^t x^k + a^t Q^{-1} a
\]
\[
= ((x^k)^t Q + a^t)Q^{-1}(Q x^k + a) = (g^k)^t Q^{-1} g^k,
\]
which is (8.44). Applying the Kantorovich inequality (8.36) with \(x = g^k\) to (8.43) now yields
\[
\forall \, k \in \mathbb{N} \quad f(x^{k+1}) - f(x^*) \leq \left( 1 - \frac{4\alpha\beta}{(\alpha + \beta)^2} \right) (f(x^k) - f(x^*))
\]
\[
= \left( \frac{\beta - \alpha}{\alpha + \beta} \right)^2 (f(x^k) - f(x^*)),
\]
thereby establishing the case. \(\blacksquare\)

### 8.4 Conjugate Gradient Method

The conjugate gradient (CG) method is an iterative method designed for the solution of linear systems
\[
Q x = b, \quad \tag{8.45}
\]
where \(Q \in \mathcal{M}(n, \mathbb{R})\) is symmetric and positive definite, \(b \in \mathbb{R}^n\). While (in exact arithmetic), it will find the exact solution in at most \(n\) steps, it is most suitable for problems where \(n\) is large and/or \(Q\) is sparse, the idea being to stop the iteration after \(k < n\) steps, obtaining a sufficiently accurate approximation to the exact solution. According to Prop. 8.21(c), (8.45) is equivalent to finding the global min of the quadratic function given by \(Q\) and \(-b\) (and \(0 \in \mathbb{R}\)), namely of
\[
\langle \cdot, \cdot \rangle_Q : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}, \quad \langle x, y \rangle_Q := x^t Q x - b^t x. \tag{8.47}
\]

**Remark 8.24.** If \(Q \in \mathcal{M}(n, \mathbb{R})\) is symmetric and positive definite, then the map
\[
\langle \cdot, \cdot \rangle_Q : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}, \quad \langle x, y \rangle_Q := x^t Q y,
\]

clearly forms an inner product on \(\mathbb{R}^n\). We call \(x, y \in \mathbb{R}^n\) \(Q\)-orthogonal if, and only if, they are orthogonal with respect to the inner product of (8.47), i.e. if, and only if, \(\langle x, y \rangle_Q = 0\).
Proposition 8.25. Let $Q \in \mathcal{M}(n, \mathbb{R})$ be symmetric and positive definite, $b \in \mathbb{R}^n$, $n \in \mathbb{N}$.
If $x^1, \ldots, x^{n+1} \in \mathbb{R}^n$ are given by a stepsize direction iteration according to Def. 8.10(a), if the directions $u^1, \ldots, u^n \in \mathbb{R}^n \setminus \{0\}$ are $Q$-orthogonal, i.e.
\[ \forall k \in \{2, \ldots, n\} \quad \forall j < k \quad \langle u^k, u^j \rangle_Q = 0, \tag{8.48} \]
and if $t_1, \ldots, t_n \in \mathbb{R}$ are according to (8.40) with a replaced by $-b$, then $x^{n+1} = Q^{-1}b$, i.e. the iteration converges to the exact solution to (8.45) in at most $n$ steps (here, in contrast to Def. 8.10(a), we do not require $t_k \in \mathbb{R}^+$ – however, if, indeed, $t_k < 0$, then one could replace $u^k$ by $-u^k$ and $t_k$ by $-t_k > 0$ to remain in the setting of Def. 8.10(a)).

Proof. Let $f$ be according to (8.46). As in the proof of Th. 8.22, we set
\[ \forall k \in \{1, \ldots, n+1\} \quad g^k := (\nabla f(x^k))^t = Qx^k - b. \tag{8.49} \]
It then suffices to show
\[ \forall k \in \{1, \ldots, n\} \quad \forall j \leq k \quad (g^{k+1})^t u^j = 0 : \tag{8.50} \]
Since $u^1, \ldots, u^n$ are $Q$-orthogonal, they must be linearly independent. Moreover, by (8.49), $g^{n+1}$ is orthogonal to each $u^1, \ldots, u^n$. In consequence, if $g^{n+1} \neq 0$, then we have $n + 1$ linearly independent vectors in $\mathbb{R}^n$, which is impossible. Thus, $g^{n+1} = 0$, i.e. $Qx^{n+1} = b$, proving the proposition. So it merely remains to verify (8.49). We start with the case $j = k$: Indeed,
\[ (g^{k+1})^t u^k = (Qx^{k+1} - b)^t u^k \tag{8.40}, (8.50) \]
\[ = (Qx^k - ((x^k)^t Q - b^t)u^k)^t u^k - b^t u^k = 0. \tag{8.51} \]
Moreover,
\[ \forall j < i \leq k \leq n \quad (g^{i+1} - g^i)^t u^j = (Qx^{i+1} - Qx^i)^t u^j = t_i (u^i)^t Q u^j = 0, \tag{8.51} \]
implying
\[ \forall k \in \{1, \ldots, n\} \quad \forall j \leq k \quad (g^{k+1})^t u^j = (g^{j+1})^t u^j + \sum_{i=j+1}^{k} (g^{i+1} - g^i)^t u^j \tag{8.50}, (8.51) \]
which establishes (8.49) and completes the proof.

Proposition 8.25 suggests to choose $Q$-orthogonal search directions. We actually already know methods for orthogonalization from Numerical Analysis I, namely the Gram-Schmidt method and the three-term recursion. The following CG method is an efficient way to carry out the steps of the stepsize direction iteration simultaneous with a $Q$-orthogonalization:
Algorithm 8.26 (CG Method). Let \( Q \in \mathcal{M}(n, \mathbb{R}) \) be symmetric and positive definite, \( b \in \mathbb{R}^n, n \in \mathbb{N} \). The conjugate gradient (CG) method then defines a finite sequence \((x^1, \ldots, x^{n+1})\) in \( \mathbb{R}^n \) via the following recursion: Initialization: Choose \( x^1 \in \mathbb{R}^n \) (barring further information on \( Q \) and \( b \), choosing \( x^1 = 0 \) is fine) and set
\[
g^1 := Qx^1 - b, \quad u^1 := -g^1. \tag{8.52}
\]
Iteration: Let \( k \in \{1, \ldots, n\} \). If \( u^k = 0 \), then set
\[
t_k := 1, \quad x^{k+1} := x^k, \quad g^{k+1} := g^k, \quad u^{k+1} := u^k \text{ (or, in practise, stop the iteration).}
\]
If \( u^k \neq 0 \), then set
\[
t_k := \frac{\|g^k\|^2_2}{\langle u^k, u^k \rangle_Q}, \tag{8.53a}
\]
\[
x^{k+1} := x^k + t_k u^k, \tag{8.53b}
\]
\[
g^{k+1} := g^k + t_k Qu^k, \tag{8.53c}
\]
\[
\beta_k := \frac{\|g^{k+1}\|^2_2}{\|g^k\|^2_2}, \tag{8.53d}
\]
\[
u^{k+1} := -g^{k+1} + \beta_k u^k. \tag{8.53e}
\]


(a) We will see in (8.54d) below that \( u^k \neq 0 \) implies \( g^k \neq 0 \), such that Alg. 8.26 is well-defined.

(b) The most expensive operation in (8.53) is the matrix multiplication \( Qu^k \), which is \( O(n^2) \). As the result is needed in both (8.53a) and (8.53c), it should be temporarily stored in some vector \( z^k := Qu^k \). If the matrix \( Q \) is sparse, it might be possible to reduce the cost of carrying out (8.53) significantly – for example, if \( Q \) is tridiagonal, then the cost reduces from \( O(n^2) \) to \( O(n) \).

Theorem 8.28. Let \( Q \in \mathcal{M}(n, \mathbb{R}) \) be symmetric and positive definite, \( b \in \mathbb{R}^n, n \in \mathbb{N} \). If \( x^1, \ldots, x^{n+1} \in \mathbb{R}^n \) as well as the related vectors \( u^k, g^k \in \mathbb{R}^n \) and numbers \( t_k, \beta_k \in \mathbb{R}_0^+ \) are given by Alg. 8.26, then the following orthogonality relations hold:
\[
\forall_{k, j \in \{1, \ldots, n+1\}, j < k}, \quad \langle u^k, u^j \rangle_Q = 0, \tag{8.54a}
\]
\[
\forall_{k, j \in \{1, \ldots, n+1\}, j < k}, \quad (g^k)^t g^j = 0, \tag{8.54b}
\]
\[
\forall_{k, j \in \{1, \ldots, n+1\}, j < k}, \quad (g^k)^t u^j = 0, \tag{8.54c}
\]
\[
\forall_{k \in \{1, \ldots, n+1\}}, \quad (g^k)^t u^k = -\|g^k\|^2_2. \tag{8.54d}
\]
Moreover,
\[
x^{n+1} = Q^{-1}b, \tag{8.55}
\]
i.e. the CG method converges to the exact solution to (8.45) in at most \( n \) steps.
Proof. Let
\[ m := \min \left( \{ k \in \{1, \ldots, n+1\} : u^k = 0 \} \cup \{n+1\} \right). \] (8.56)

A first induction shows
\[ \forall k \in \{1, \ldots, m\} \quad g^k = Q x^k - b : \] (8.57)

Indeed, \( g^1 = Q x^1 - b \) holds by initialization and, for \( 1 \leq k < m \),
\[ g^{k+1} = g^k + t_k Q u^k \text{ ind. hyp.} \]
which completes the induction proof of (8.54).

It now suffices to show that (8.54) holds with \( n + 1 \) replaced by \( m \): If \( m < n + 1 \), then \( u^m = 0 \), i.e. \( g^m = 0 \) by (8.54d), implying \( x^m = Q^{-1} b \) by (8.57). Then (8.54) and (8.55) also follow as stated. Next, note
\[ \forall k \in \{1, \ldots, m-1\} \quad t_k = \frac{\|g^k\|^2}{\langle u^k, u^k \rangle_Q} = -\frac{(g^k)^t u^k}{\langle u^k, u^k \rangle_Q}, \] (8.58)
which is consistent with (8.40). If \( m = n + 1 \), then Prop. 8.25 applies according to (8.54a) and (8.58), implying the validity of (8.55). It, thus, remains to verify (8.54) holds with \( n + 1 \) replaced by \( m \). We proceed via induction on \( k = 1, \ldots, m \): For \( k = 1 \), (8.54a)–(8.54c) are trivially true and (8.54d) holds as \( (g^1)^t u^1 = -\|g^1\|^2 \) is true by initialization. Thus, let \( 1 \leq k < m \). As in the proof of Prop. 8.25, we obtain (8.49) from (8.54a), that means the induction hypothesis for (8.54a) implies (8.54c) for \( k + 1 \). Next, we verify (8.54d) for \( k + 1 \):
\[ (g^{k+1})^t u^{k+1} = (g^{k+1})^t (-g^{k+1} + \beta_k u^k) = -\|g^{k+1}\|^2 + (g^k + t_k Q u^k)^t \beta_k u^k \]
\[ \text{ind. hyp.} \quad -\|g^{k+1}\|^2 - \beta_k \|g^k\|^2 + \beta_k \|g^k\|^2 = -\|g^{k+1}\|^2. \]

It remains to show (8.54a) and (8.54b) for \( k + 1 \). We start with (8.54b):
\[ (g^{k+1})^t g^1 = -(g^{k+1})^t u^1 \quad (8.54c) \quad 0 \]
and, for \( j \in \{2, \ldots, k\} \),
\[ (g^{k+1})^t g^j = (g^{k+1})^t (-u^j + \beta_{j-1} u^{j-1}) \quad (8.54c) \quad 0. \]

Finally, to obtain (8.54a) for \( k + 1 \), we note, for each \( j \in \{1, \ldots, k\} \),
\[ \langle u^{k+1}, u^j \rangle_Q = (-g^{k+1} + \beta_k u^k)^t Q u^j = -\frac{1}{t_j} (g^{k+1})^t (g^{j+1} - g^j) + \beta_k (u^k)^t Q u^j. \] (8.59)

Thus, for \( j < k \), (8.59) together with (8.54b) (for \( k + 1 \)) and (8.54a) (for \( k \), i.e. the induction hypothesis) yields \( \langle u^{k+1}, u^j \rangle_Q = 0 \). For \( j = k \), we compute
\[ \langle u^{k+1}, u^k \rangle_Q \quad (8.59),(8.54b) \quad \frac{\|g^{k+1}\|^2}{t_k} + \beta_k \frac{\|g^k\|^2}{t_k} = 0, \]
which completes the induction proof of (8.54).
Theorem 8.29. Let $Q \in \mathcal{M}(n, \mathbb{R})$ be symmetric and positive definite, $b \in \mathbb{R}^n$, $n \in \mathbb{N}$. If $x^1, \ldots, x^{n+1} \in \mathbb{R}^n$ are given by Alg. 8.26 and $x^* := Q^{-1}b$, then the following error estimate holds:

$$\forall k \in \{1, \ldots, n\} \quad \|x^{k+1} - x^*\|_Q \leq 2 \left( \frac{\sqrt{\kappa_2(Q)} - 1}{\sqrt{\kappa_2(Q)} + 1} \right)^k \|x^1 - x^*\|_Q,$$

(8.60)

where $\| \cdot \|_Q$ denotes the norm induced by $\langle \cdot, \cdot \rangle_Q$ and $\kappa_2(Q) = \|Q\|_2\|Q^{-1}\|_2 = \frac{\max \sigma(Q)}{\min \sigma(Q)}$ is the spectral condition of $Q$.

Proof. See, e.g., [DH08, Th. 8.17]. ■

Remark 8.30. (a) According to Th. 8.28, Alg. 8.26 finds the exact solution after at most $n$ steps. However, it might find the exact solution even earlier: According to [Sco11, Cor. 9.10], if $Q$ has precisely $m$ distinct eigenvalues, then Alg. 8.26 finds the exact solution after at most $m$ steps; and it is stated in [GK99, p. 225] that the same also holds if one starts at $x^1 = 0$ and $b$ is the linear combination of $m$ eigenvectors of $Q$.

(b) If one needs to solve $Ax = b$ with invertible $A \in \mathcal{M}(n, \mathbb{R})$, where $A$ is not symmetric positive definite, one can still make use of the CG method, namely by applying it to $A^tAx = A^tb$ (noting $Q := A^tA$ to be symmetric positive definite). However, a disadvantage then is that the condition of $Q$ can be much larger than the condition of $A$.

A Representations of Real Numbers, Rounding Errors

A.1 $b$-Adic Expansions of Real Numbers

We are mostly used to representing real numbers in the decimal system. For example, we write

$$x = \frac{395}{3} = 131.\overline{6} = 1 \cdot 10^2 + 3 \cdot 10^1 + 1 \cdot 10^0 + \sum_{n=1}^{\infty} 6 \cdot 10^{-n}. \quad (A.1a)$$

The decimal system represents real numbers as, in general, infinite series of decimal fractions. Digital computers represent numbers in the dual system, using base 2 instead of 10. For example, the number from (A.1a) has the dual representation

$$x = 10000011.10 = 2^7 + 2^1 + 2^0 + \sum_{n=0}^{\infty} 2^{-(2n+1)}. \quad (A.1b)$$

Representations with base 16 (hexadecimal) and 8 (octal) are also of importance when working with digital computers. More generally, each natural number $b \geq 2$ can be used as a base.
In most cases, it is understood that we work only with decimal representations such that there is no confusion about the meaning of symbol strings like 101.01. However, in general, 101.01 could also be meant with respect to any other base, and, the number represented by the same string of symbols does obviously depend on the base used. Thus, when working with different representations, one needs some notation to keep track of the base.

**Notation A.1.** Given a natural number \( b \geq 2 \) and finite sequences
\[
(d_{N_1}, d_{N_1-1}, \ldots, d_0) \in \{0, \ldots, b - 1\}^{N_1+1},
\]
\[
(e_1, e_2, \ldots, e_{N_2}) \in \{0, \ldots, b - 1\}^{N_2},
\]
\[
(p_1, p_2, \ldots, p_{N_3}) \in \{0, \ldots, b - 1\}^{N_3},
\]
\( N_1, N_2, N_3 \in \mathbb{N}_0 \) (where \( N_2 = 0 \) or \( N_3 = 0 \) is supposed to mean that the corresponding sequence is empty), the respective string
\[
(d_{N_1}d_{N_1-1}\ldots d_0)_{b}
\]
for \( N_2 = N_3 = 0 \),
\[
(d_{N_1}d_{N_1-1}\ldots d_0.e_1\ldots e_{N_2}p_1\ldots p_{N_3})_{b}
\]
for \( N_2 + N_3 > 0 \)
represents the number
\[
\sum_{\nu=0}^{N_1} d_{\nu} b^{\nu} + \sum_{\nu=1}^{N_2} e_{\nu} b^{-\nu} + \sum_{\alpha=0}^{\infty} \sum_{\nu=1}^{N_3} p_{\nu} b^{-N_2-\alpha N_3-\nu}.
\]

**Example A.2.** For the number from (A.1), we get
\[
x = (131.6)_{10} = (10000011.11)_{2} = (83.A)_{16}
\]
(for the hexadecimal system, it is customary to use the symbols 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F).

**Definition A.3.** Given a natural number \( b \geq 2 \), an integer \( N \in \mathbb{Z} \), and a sequence \((d_{N}, d_{N-1}, d_{N-2}, \ldots)\) of nonnegative integers such that \( d_\nu \in \{0, \ldots, b - 1\} \) for each \( \nu \in \{N, N-1, N-2, \ldots\} \), the expression
\[
\sum_{\nu=0}^{\infty} d_{\nu} b^{N-\nu}
\]
is called a \( b \)-adic series. The number \( b \) is called the base or the radix, and the numbers \( d_\nu \) are called digits.

**Definition A.4.** If \( b \geq 2 \) is a natural number and \( x \in \mathbb{R}_0^+ \) is the value of the \( b \)-adic series given by (A.6), than one calls the \( b \)-adic series a \( b \)-adic expansion of \( x \).

**Lemma A.5.** Given a natural number \( b \geq 2 \), consider the \( b \)-adic series given by (A.6). Then
\[
\sum_{\nu=0}^{\infty} d_{N-\nu} b^{N-\nu} \leq b^{N+1},
\]
and, in particular, the \( b \)-adic series converges to some \( x \in \mathbb{R}_0^+ \). Moreover, equality in (A.7) holds if, and only if, \( d_\nu = b - 1 \) for every \( \nu \in \{N, N-1, N-2, \ldots\} \).
Proof. One estimates, using the formula for the value of a geometric series:

$$\sum_{\nu=0}^{\infty} d_{N-\nu} b^{N-\nu} \leq \sum_{\nu=0}^{\infty} (b-1) b^{N-\nu} = (b-1) b^N \sum_{\nu=0}^{\infty} b^{-\nu} = (b-1) b^N \frac{1}{1 - \frac{1}{b}} = b^{N+1}. \quad (A.8)$$

Note that (A.8) also shows that equality is achieved if all $d_n$ are equal to $b-1$. Conversely, if there is $n \in \{N, N-1, N-2, \ldots\}$ such that $d_n < b-1$, then there is $\tilde{n} \in \mathbb{N}$ such that $d_{N-\tilde{n}} < b-1$ and one estimates

$$\sum_{\nu=0}^{\infty} d_{N-\nu} b^{N-\nu} < \sum_{\nu=0}^{\tilde{n}-1} d_{N-\nu} b^{N-\nu} + (b-1) b^{N-\tilde{n}} + \sum_{\nu=\tilde{n}+1}^{\infty} d_{N-\nu} b^{N-\nu} \leq b^{N+1}, \quad (A.9)$$

showing that the inequality in (A.7) is strict. 

\[\blacksquare\]

### A.2 Floating-Point Numbers Arithmetic, Rounding

Numerical problems are often related to the computation of (approximations of) real numbers. The $b$-adic representations of real numbers considered above, in general, need infinitely many digits. However, computer memory can only store a finite amount of data. This implies the need to introduce representations that use only strings of finite length and only a finite number of symbols. Clearly, given a supply of $s \in \mathbb{N}$ symbols and strings of length $l \in \mathbb{N}$, one can represent a maximum of $s^l < \infty$ numbers. The representation of this form most commonly used to approximate real numbers is the so-called floating-point representation:

**Definition A.6.** Let $b \in \mathbb{N}$, $b \geq 2$, $l \in \mathbb{N}$, and $N_-, N_+ \in \mathbb{Z}$ with $N_- \leq N_+$. Then, for each

$$\begin{align*}
(x_1, \ldots, x_l) & \in \{0, 1, \ldots, b-1\}^l, \\
N & \in \mathbb{Z} \text{ satisfying } N_- \leq N \leq N_+,
\end{align*} \quad (A.10a)$$

the strings

$$0.x_1x_2\ldots x_l \cdot b^N \quad \text{and} \quad -0.x_1x_2\ldots x_l \cdot b^N \quad (A.11)$$

are called *floating-point representations* of the (rational) numbers

$$x := b^N \sum_{\nu=1}^{l} x_{\nu} b^{-\nu} \quad \text{and} \quad -x, \quad (A.12)$$

respectively, provided that $x_1 \neq 0$ if $x \neq 0$. For floating-point representations, $b$ is called the *radix* (or *base*), $0.x_1x_2\ldots x_l$ (i.e. $|x|/b^N$) is called the *significand* (or *mantissa* or *coefficient*), $x_1, \ldots, x_l$ are called the *significant digits*, and $N$ is called the *exponent*. The number $l$ is sometimes called the *precision*. Let $\mathcal{f}_l(b, N_-, N_+) \subseteq \mathbb{Q}$ denote the set of all rational numbers that have a floating point representation of precision $l$ with respect to base $b$ and exponent between $N_-$ and $N_+$. 

Remark A.7. To get from the floating-point representation (A.11) to the form of (A.3), one has to shift the radix point of the significand by a number of places equal to the value of the exponent — to the right if the exponent is positive or to the left if the exponent is negative.

Remark A.8. If one restricts Def. A.6 to the case \( N_\nu = N_\nu + \), then one obtains what is known as fixed-point representations. However, for many applications, the required numbers vary over sufficiently many orders of magnitude to render fixed-point representations impractical.

Remark A.9. Given the assumptions of Def. A.6, for the numbers in (A.12), it always holds that

\[
b^{N-1} \leq b^{N-1} \leq \sum_{\nu=1}^{l} x_\nu b^{N-\nu} = |x| \leq (b-1) \sum_{\nu=1}^{l} b^{N-\nu} =: \text{max}(l, b, N_+) \text{ Lem. A.5} < b^{N},
\]

(A.13)

provided that \( x \neq 0 \). In other words:

\[
\text{fl}_l(b, N_-, N_+) \subseteq \mathbb{Q} \cap \left( [-\text{max}(l, b, N_+), -b^{N-1}] \cup \{0\} \cup [b^{N-1}, \text{max}(l, b, N_+)] \right).
\]

(A.14)

Obviously, \( \text{fl}_l(b, N_-, N_+) \) is not closed under arithmetic operations. A result of absolute value bigger than \( \text{max}(l, b, N_+) \) is called an overflow, whereas a nonzero result of absolute value less than \( b^{N-1} \) is called an underflow. In practice, the result of an underflow is usually replaced by 0.

Definition A.10. Let \( b \in \mathbb{N}, b \geq 2, l \in \mathbb{N}, N \in \mathbb{Z}, \) and \( \sigma \in \{-1, 1\} \). Then, given

\[
x = \sigma b^N \sum_{\nu=1}^{\infty} x_\nu b^{-\nu},
\]

(A.15)

where \( x_\nu \in \{0, 1, \ldots, b-1\} \) for each \( \nu \in \mathbb{N} \) and \( x_1 \neq 0 \) for \( x \neq 0 \), define

\[
\text{rd}_l(x) := \begin{cases} \sigma b^N \sum_{\nu=1}^{l} x_\nu b^{-\nu} & \text{for } x_{l+1} < b/2, \\ \sigma b^N (b^{-l} + \sum_{\nu=1}^{l} x_\nu b^{-\nu}) & \text{for } x_{l+1} \geq b/2. \end{cases}
\]

(A.16)

The number \( \text{rd}_l(x) \) is called \( x \) rounded to \( l \) digits.

Remark A.11. We note that the notation \( \text{rd}_l(x) \) of Def. A.10 is actually not entirely correct, since \( \text{rd}_l \) is actually a function of the sequence \( (\sigma, N, x_1, x_2, \ldots) \) rather than of \( x \): It can actually occur that \( \text{rd}_l \) takes on different values for different representations of the same number \( x \) (for example, using decimal notation, consider \( x = 0.349 = 0.350 \), yielding \( \text{rd}_l(0.349) = 0.3 \) and \( \text{rd}_l(0.350) = 0.4 \)). However, from basic results on \( b \)-adic expansions of real numbers (see [Phi16a, Th. 7.99]), we know that \( x \) can have at most two different \( b \)-adic representations and, for the same \( x \), \( \text{rd}_l(x) \) can vary at most \( b^N b^{-l} \). Since always writing \( \text{rd}_l(\sigma, N, x_1, x_2, \ldots) \) does not help with readability, and since writing \( \text{rd}_l(x) \) hardly ever causes confusion as to what is meant in a concrete case, the abuse of notation introduced in Def. A.10 is quite commonly employed.
Lemma A.12. Let $b \in \mathbb{N}$, $b \geq 2$, $l \in \mathbb{N}$. If $x \in \mathbb{R}$ is given by (A.15), then $\text{rd}_l(x) = \sigma b^{N'} \sum_{\nu=1}^l x'_\nu b^{-\nu}$, where $N' \in \{N, N+1\}$ and $x'_\nu \in \{0, 1, \ldots, b-1\}$ for each $\nu \in \mathbb{N}$ and $x'_1 \neq 0$ for $x \neq 0$. In particular, $\text{rd}_l$ maps $\bigcup_{k=1}^\infty \text{fl}_k(b, N_-, N_+)$ into $\text{fl}_l(b, N_-, N_+ + 1)$.

Proof. For $x_{l+1} < b/2$, there is nothing to prove. Thus, assume $x_{l+1} \geq b/2$. Case (i): There exists $\nu \in \{1, \ldots, l\}$ such that $x_\nu < b - 1$. Then, letting $\nu_0 := \max\{\nu \in \{1, \ldots, l\} : x_\nu < b - 1\}$, one finds $N' = N$ and

$$x'_\nu = \begin{cases} x_\nu & \text{for } 1 \leq \nu < \nu_0, \\ x_{\nu_0} + 1 & \text{for } \nu = \nu_0, \\ 0 & \text{for } \nu_0 < \nu \leq l. \end{cases}$$

(A.17a)

Case (ii): $x_\nu = b - 1$ holds for each $\nu \in \{1, \ldots, l\}$. Then one obtains $N' = N + 1$,

$$x'_\nu := \begin{cases} 1 & \text{for } \nu = 1, \\ 0 & \text{for } 1 < \nu \leq l, \end{cases}$$

(A.17b)

thereby concluding the proof. $\blacksquare$

When composing floating point numbers of a fixed precision $l \in \mathbb{N}$ by means of arithmetic operations such as ‘+’, ‘−’, ‘·’, and ‘:’, the exact result is usually not representable as a floating point number with the same precision $l$: For example, $0.434 \cdot 10^4 + 0.705 \cdot 10^{-1} = 0.43400705 \cdot 10^4$, which is not representable exactly with just 3 significant digits. Thus, when working with floating point numbers of a fixed precision, rounding will generally be necessary.

The following Notation A.13 makes sense for general real numbers $x$, $y$, but is intended to be used with numbers from some $\text{fl}_l(b, N_-, N_+)$, i.e. numbers given in floating point representation (in particular, with a finite precision).

Notation A.13. Let $b \in \mathbb{N}$, $b \geq 2$. Assume $x, y \in \mathbb{R}$ are given in a form analogous to (A.15). We then define, for each $l \in \mathbb{N}$,

$$x \diamond_l y := \text{rd}_l(x \diamond y),$$

(A.18)

where $\diamond$ can stand for any of the operations ‘+’, ‘−’, ‘·’, and ‘:’.

Remark A.14. Consider $x, y \in \text{fl}_l(b, N_-, N_+)$. If $x \diamond y \in \bigcup_{k=1}^\infty \text{fl}_k(b, N_-, N_+)$, then, according to Lem. A.12, $x \diamond_l y$ as defined in (A.18) is either in $\text{fl}_l(b, N_-, N_+)$ or in $\text{fl}_l(b, N_-, N_+ + 1)$ with the digits given by (A.17b) (which constitutes an overflow).

Remark A.15. The definition of (A.18) should only be taken as an example of how floating-point operations can be realized. On concrete computer systems, the rounding operations implemented can be different from the one considered here. However, one would still expect a corresponding version of Rem. A.14 to hold.
Caveat A.16. Unfortunately, the associative laws of addition and multiplication as well as the law of distributivity are lost for arithmetic operations of floating-point numbers. More precisely, even if \( x, y, z \in \mathbb{R} \) and one assumes that no overflow or underflow occurs, then there are examples, where \( (x + y) + z \neq x + (y + z) \), \( (x \cdot y) \cdot z \neq x \cdot (y \cdot z) \), and \( x \cdot (y + z) \neq x \cdot y + x \cdot z \).

For \( l = 1, b = 10, N_+ = 0, N_+ = 1, x = 0.1 \cdot 10^1, y = 0.4 \cdot 10^0, z = 0.1 \cdot 10^0 \), one has

\[
(x + y) + z = \text{rd}_1(0.1 \cdot 10^1 + 0.4 \cdot 10^0) + 0.1 \cdot 10^0 = \text{rd}_1(0.14 \cdot 10^1) + 0.1 \cdot 10^0 \\
= \text{rd}_1(0.1 \cdot 10^1 + 0.1 \cdot 10^0) = 0.1 \cdot 10^1,
\]

but

\[
x + (y + z) = 0.1 \cdot 10^1 + \text{rd}_1(0.4 \cdot 10^0 + 0.1 \cdot 10^0) = \text{rd}_1(0.1 \cdot 10^1 + 0.5 \cdot 10^0) \\
= \text{rd}_1(0.15 \cdot 10^1) = 0.2 \cdot 10^1.
\]

(A.19)

For \( l = 1, b = 10, N_+ = 0, N_+ = 1, x = 0.8 \cdot 10^1, y = 0.3 \cdot 10^0, z = 0.5 \cdot 10^0 \), one has

\[
(x \cdot y) \cdot z = \text{rd}_1(0.8 \cdot 10^1 \cdot 0.3 \cdot 10^0) \cdot 0.5 \cdot 10^0 = \text{rd}_1(0.24 \cdot 10^1) \cdot 0.5 \cdot 10^0 \\
= \text{rd}_1(0.2 \cdot 10^1 \cdot 0.5 \cdot 10^0) = 0.1 \cdot 10^1,
\]

but

\[
x \cdot (y \cdot z) = 0.8 \cdot 10^1 \cdot \text{rd}_1(0.3 \cdot 10^0 \cdot 0.5 \cdot 10^0) = 0.8 \cdot 10^1 \cdot \text{rd}_1(0.15 \cdot 10^0) \\
= \text{rd}_1(0.8 \cdot 10^1 \cdot 0.2 \cdot 10^0) = 0.2 \cdot 10^1.
\]

(A.20)

For \( l = 1, b = 10, N_+ = 0, N_+ = 0, x = 0.5 \cdot 10^0, y = 0.3 \cdot 10^0 \), one has

\[
x \cdot (y + z) = 0.5 \cdot 10^0 \cdot \text{rd}_1(0.3 \cdot 10^0 + 0.3 \cdot 10^0) \\
= 0.5 \cdot 10^0 \cdot \text{rd}_1(0.6 \cdot 10^0) = 0.3 \cdot 10^0,
\]

but

\[
x \cdot y + x \cdot z = \text{rd}_1(0.5 \cdot 10^0 \cdot 0.3 \cdot 10^0) + \text{rd}_1(0.5 \cdot 10^0 \cdot 0.3 \cdot 10^0) \\
= 0.2 \cdot 10^0 + 0.2 \cdot 10^0 = 0.4 \cdot 10^0.
\]

(A.21)

Lemma A.17. Let \( b \in \mathbb{N}, b \geq 2 \). Suppose

\[
x = b^N \sum_{\nu=1}^{\infty} x_{\nu} b^{-\nu}, \quad y = b^M \sum_{\nu=1}^{\infty} y_{\nu} b^{-\nu},
\]

where \( N, M \in \mathbb{Z}; x_{\nu}, y_{\nu} \in \{0, 1, \ldots, b - 1\} \) for each \( \nu \in \mathbb{N} \); and \( x_1, y_1 \neq 0 \).

(a) If \( N > M \), then \( x \geq y \).

(b) If \( N = M \) and there is \( n \in \mathbb{N} \) such that \( x_n > y_n \) and \( x_\nu = y_\nu \) for each \( \nu \in \{1, \ldots, n - 1\} \), then \( x \geq y \).
Proof. (a): One estimates

\[ x - y = b^N \sum_{\nu=1}^{\infty} x_\nu b^{-\nu} - b^M \sum_{\nu=1}^{\infty} y_\nu b^{-\nu} \geq b^{N-1} - b^{N-1} \sum_{\nu=1}^{\infty} (b-1)b^{-\nu} = b^{N-1} - b^{N-1} \left( \frac{1}{1 - b^{-1}} - 1 \right) = 0. \]  

(A.26a)

(b): One estimates

\[ x - y = b^N \sum_{\nu=1}^{\infty} x_\nu b^{-\nu} - b^M \sum_{\nu=1}^{\infty} y_\nu b^{-\nu} = b^N \sum_{\nu=n}^{\infty} x_\nu b^{-\nu} - b^N \sum_{\nu=n}^{\infty} y_\nu b^{-\nu} \geq b^{N-n} - b^{N-n} \sum_{\nu=1}^{\infty} (b-1)b^{-\nu} \quad \text{as in (A.26a)} = 0, \]  

(A.26b)

concluding the proof of the lemma. ■

Lemma A.18. Let \( b \in \mathbb{N}, \ b \geq 2, \ l \in \mathbb{N} \). Then, for each \( x, y \in \mathbb{R} \):

(a) \( \text{rd}_l(x) = \text{sgn}(x) \text{rd}_l(|x|) \).

(b) \( 0 \leq x < y \) implies \( 0 \leq \text{rd}_l(x) \leq \text{rd}_l(y) \).

(c) \( 0 \geq x > y \) implies \( 0 \geq \text{rd}_l(x) \geq \text{rd}_l(y) \).

Proof. (a): If \( x \) is given by (A.15), one obtains from (A.16):

\[ \text{rd}_l(x) = \sigma \text{rd}_l \left( b^N \sum_{\nu=1}^{\infty} x_\nu b^{-\nu} \right) = \text{sgn}(x) \text{rd}_l(|x|). \]

(b): Suppose \( x \) and \( y \) are given as in (A.25). Then Lem. A.17(a) implies \( N \leq M \).

Case \( x_{l+1} < b/2 \): In this case, according to (A.16),

\[ \text{rd}_l(x) = b^N \sum_{\nu=1}^{l} x_\nu b^{-\nu}. \]  

(A.27)

For \( N < M \), we estimate

\[ \text{rd}_l(y) \geq b^M \sum_{\nu=1}^{l} y_\nu b^{-\nu} \quad \text{Lem. A.17(a)} \geq b^N \sum_{\nu=1}^{l} x_\nu b^{-\nu} = \text{rd}_l(x), \]  

(A.28)

which establishes the case. We claim that (A.28) also holds for \( N = M \), now due to Lem. A.17(b): This is clear if \( x_\nu = y_\nu \) for each \( \nu \in \{1, \ldots, l\} \). Otherwise, define

\[ n := \min \{ \nu \in \{1, \ldots, l\} : x_\nu \neq y_\nu \}. \]  

(A.29)
Then \( x < y \) and Lem. A.17(b) yield \( x_n < y_n \). Moreover, another application of Lem. A.17(b) then implies (A.28) for this case.

Case \( x_{i+1} \geq b/2 \): In this case, according to Lem. A.12,

\[
\text{rd}_l(x) = b^{N'} \sum_{\nu=1}^{l} x'_\nu b^{-\nu},
\]

where either \( N' = N \) and the \( x'_\nu \) are given by (A.17a), or \( N' = N + 1 \) and the \( x'_\nu \) are given by (A.17b). For \( N < M \), we obtain

\[
\text{rd}_l(y) \geq b^{M} \sum_{\nu=1}^{l} y'_\nu b^{-\nu} \overset{(\ast)}{\geq} b^{N'} \sum_{\nu=1}^{l} x'_\nu b^{-\nu} = \text{rd}_l(x),
\]

where, for \( N' < M \), (\ast) holds by Lem. A.17(a), and, for \( N' = N + 1 = M \), (\ast) holds by (A.17b). It remains to consider \( N = M \). If \( x_\nu = y_\nu \) for each \( \nu \in \{1, \ldots, l\} \), then \( x < y \) and Lem. A.17(b) yield \( b/2 \leq x_{i+1} \leq y_{i+1} \), which, in turn, yields \( \text{rd}_l(y) = \text{rd}_l(x) \). Otherwise, once more define \( n \) according to (A.29). As before, \( x < y \) and Lem. A.17(b) yield \( x_n < y_n \). From Lem. A.12, we obtain \( N' = N \) and that the \( x'_\nu \) are given by (A.17a) with \( \nu_0 \geq n \). Then the values from (A.17a) show that (A.31) holds true once again.

(c) follows by combining (a) and (b).

**Definition and Remark A.19.** Let \( b \in \mathbb{N}, b \geq 2, l \in \mathbb{N}, \) and \( N_-, N_+ \in \mathbb{Z} \) with \( N_- \leq N_+ \). If there exists a smallest positive number \( \epsilon \in \text{fl}_l(b,N_-,N_+) \) such that

\[
1 + \epsilon \neq 1,
\]

then it is called the *relative machine precision*. We will show that, for \( N_+ < -l + 1 \), one has \( 1 + y = 1 \) for every \( 0 < y \in \text{fl}_l(b,N_-,N_+) \), such that there is no positive number in \( \text{fl}_l(b,N_-,N_+) \) satisfying (A.32), whereas, for \( N_+ \geq -l + 1 \):

\[
\epsilon = \begin{cases} 
 b^{N_- - 1} & \text{for } -l + 1 < N_- , \\
 [b/2]b^{-l} & \text{for } -l + 1 \geq N_- 
\end{cases}
\]

(for \( x \in \mathbb{R}, \lfloor x \rfloor := \min \{k \in \mathbb{Z} : x \leq k\} \) is called ceiling of \( x \) or \( x \) rounded up):

First, one notices that \( 0 < y < [b/2]b^{-l} \) implies

\[
1 + y = b^l \sum_{i=1}^{\infty} y_i b^{-i},
\]

where

\[
y_i \begin{cases} 
 = 1 & \text{for } i = 1 , \\
 < [b/2] & \text{for } i = l + 1 , \\
 = 0 & \text{for } 1 < i < l + 1 , \\
 \in \{0, \ldots, b - 1\} & \text{for } l + 1 < i.
\end{cases}
\]
As \( y_l+1 < b/2 \), the definition of rounding in (A.16) yields:

\[
1 +_l y = \text{rd}_l (1 + y) = b^l \sum_{i=1}^{l} y_i b^{-i} = b^l b^{-1} = 1,
\]

(A.35)

From Rem. A.9, we know

\[
\text{fl}_l (b, N_-, N_+) \subset \mathbb{Q} \cap (-b^{N_-}, -b^{N_- - 1}) \cup \{0\} \cup [b^{N_- - 1}, b^{N_+}].
\]

(A.36)

In particular, every element in \( \text{fl}_l (b, N_-, N_+) \) is less than \( b^{N_+} \).

If \( N_+ < -l + 1 \), then \( b^{N_+} \leq b^{-l} < \lceil b/2 \rceil b^{-l} \). Thus, using what we have shown above, \( 1 +_l y = 1 \) for each \( 0 < y \in \text{fl}_l (b, N_-, N_+) \).

Now let \( N_+ \geq -l + 1 \).

One notices next that

\[
1 + [b/2] b^{-l} = 0 \cdot x_1 \ldots x_{l+1} \cdot b^l = b^l \sum_{i=1}^{l+1} x_i b^{-i},
\]

(A.37)

where

\[
x_i := \begin{cases} 
1 & \text{for } i = 1, \\
\left\lceil b/2 \right\rceil & \text{for } i = l + 1, \\
0 & \text{for } i \neq 1, l + 1.
\end{cases}
\]

(A.38)

Using (A.37) and the definition of rounding in (A.16), we have

\[
1 +_l [b/2] b^{-l} = \text{rd}_l (1 + [b/2] b^{-l}) = 1 + b^{-l+1} > 1.
\]

(A.39)

Now let \(-l + 1 < N_- \). Due to (A.14), \( b^{N_- - 1} \) is the smallest positive number contained in \( \text{fl}_l (b, N_-, N_+) \). In addition, due to \( b^{N_- - 1} \geq b^{-l+1} > \lceil b/2 \rceil b^{-l} \), we have

\[
1 + l b^{N_- - 1} = \text{rd}_l (1 + b^{N_- - 1}) \geq \text{rd}_l (1 + [b/2] b^{-l}) \overset{(\text{A.39})}{=} 1 + b^{-l+1} > 1,
\]

(A.40)

implying \( \epsilon = b^{N_- - 1} \).

For \(-l + 1 \geq N_- \), we have \([b/2] b^{-l} \in \text{fl}_l (b, N_-, N_+) \) due to \([b/2] b^{-l} = 0 \cdot \lceil b/2 \rceil \cdot b^{-l+1} \) (where \( N_+ \geq -l + 1 \) was used as well). As we have already shown that \( 1 +_l y = 1 \) for each \( 0 < y < [b/2] b^{-l} \), \( \epsilon = [b/2] b^{-l} \) is now a consequence of (A.39).

### A.3 Rounding Errors

**Definition A.20.** Let \( v \in \mathbb{R} \) be the exact value and let \( a \in \mathbb{R} \) be an approximation for \( v \). The numbers

\[
e_a := |v - a|, \quad e_r := \frac{e_a}{|v|}
\]

(A.41)

are called the absolute error and the relative error, respectively, where the relative error is only defined for \( v \neq 0 \). It can also be useful to consider variants of the absolute and relative error, respectively, where one does not take the absolute value. Thus, in the literature, one finds the definitions with and without the absolute value.
Proposition A.21. Let \( b \in \mathbb{N} \) be even, \( b \geq 2 \), \( l \in \mathbb{N} \), \( N \in \mathbb{Z} \), and \( \sigma \in \{-1, 1\} \). Suppose \( x \in \mathbb{R} \) is given by (A.15), i.e.

\[
x = \sigma b^N \sum_{\nu=1}^{\infty} x_\nu b^{-\nu}, \tag{A.42}
\]

where \( x_\nu \in \{0, 1, \ldots, b-1\} \) for each \( \nu \in \mathbb{N} \) and \( x_1 \neq 0 \) for \( x \neq 0 \),

(a) The absolute error of rounding to \( l \) digits satisfies

\[
e_a(x) = |\text{rd}_l(x) - x| \leq \frac{b^{N-l}}{2}.
\]

(b) The relative error of rounding to \( l \) digits satisfies, for each \( x \neq 0 \),

\[
e_r(x) = \frac{|\text{rd}_l(x) - x|}{|x|} \leq \frac{b^{l+1}}{2}.
\]

(c) For each \( x \neq 0 \), one also has the estimate

\[
\frac{|\text{rd}_l(x) - x|}{|\text{rd}_l(x)|} \leq \frac{b^{l+1}}{2}.
\]

Proof. (a): First, consider the case \( x_{l+1} < b/2 \): One computes

\[
e_a(x) = |\text{rd}_l(x) - x| = -\sigma \left( \text{rd}_l(x) - x \right) = b^N \sum_{\nu=l+1}^{\infty} x_\nu b^{-\nu}
\]

\[= b^{N-l-1} x_{l+1} + b^N \sum_{\nu=l+2}^{\infty} x_\nu b^{-\nu}
\]

\[\leq b^{N-l-1} \left( \frac{b}{2} - 1 \right) + b^{N-l-1} = \frac{b^{N-l}}{2}. \tag{A.43a}
\]

It remains to consider the case \( x_{l+1} \geq b/2 \). In that case, one obtains

\[
\sigma \left( \text{rd}_l(x) - x \right) = b^{N-l} - b^N x_{l+1} b^{-l-1} - b^N \sum_{\nu=l+2}^{\infty} x_\nu b^{-\nu}
\]

\[= b^{N-l-1}(b - x_{l+1}) - b^N \sum_{\nu=l+2}^{\infty} x_\nu b^{-\nu} \leq \frac{b^{N-l}}{2}. \tag{A.43b}
\]

Due to \( 1 \leq b - x_{l+1} \), one has \( b^{N-l-1} \leq b^{N-l-1}(b - x_{l+1}) \). Therefore, \( b^N \sum_{\nu=l+2}^{\infty} x_\nu b^{-\nu} \leq b^{N-l-1} \) together with (A.43b) implies \( \sigma \left( \text{rd}_l(x) - x \right) \geq 0 \), i.e. \( e_a(x) = \sigma \left( \text{rd}_l(x) - x \right) \).

(b): Since \( x \neq 0 \), one has \( x_1 \geq 1 \). Thus, \( |x| \geq b^{N-1} \). Then (a) yields

\[
e_r = \frac{e_a}{|x|} \leq \frac{b^{N-l} b^{-N+1}}{2} = \frac{b^{l+1}}{2}. \tag{A.44}
\]
A REPRESENTATIONS OF REAL NUMBERS, Rounding Errors

(c): Again, \( x_1 \geq 1 \) as \( x \neq 0 \). Thus, (A.16) implies \( |\text{rd}_1(x)| \geq b^{N-1} \). This time, (a) yields

\[
\varepsilon_a \frac{|\text{rd}_1(x)|}{|\text{rd}_1(x)|} \leq b^{N-1} \frac{b^{-N+1}}{2} = \frac{b^{-l+1}}{2},
\]

establishing the case and completing the proof of the proposition.

\[\square\]

Corollary A.22. In the situation of Prop. A.21, let, for \( x \neq 0 \):

\[
\epsilon_l(x) := \frac{\text{rd}_l(x) - x}{x}; \quad \eta_l(x) := \frac{\text{rd}_l(x) - x}{\text{rd}_l(x)}.
\]

Then

\[
\max \{ |\epsilon_l(x)|, |\eta_l(x)| \} \leq \frac{b^{-l+1}}{2} =: \tau_l.
\]

The number \( \tau_l \) is called the relative computing precision of floating-point arithmetic with \( l \) significant digits.

\[\square\]

Remark A.23. From the definitions of \( \epsilon_l(x) \) and \( \eta_l(x) \) in (A.46), one immediately obtains the relations

\[
\text{rd}_l(x) = x \left( 1 + \epsilon_l(x) \right) = \frac{x}{1 - \eta_l(x)} \quad \text{for} \quad x \neq 0.
\]

In particular, according to (A.18), one has for floating-point operations (for \( x \odot y \neq 0 \)):

\[
x \odot_l y = \text{rd}_l(x \odot y) = (x \odot y) \left( 1 + \epsilon_l(x \odot y) \right) = \frac{x \odot y}{1 - \eta_l(x \odot y)}.
\]

One can use the formulas of Rem. A.23 to perform what is sometimes called a forward analysis of the rounding error. This technique is illustrated in the next example.

Example A.24. Let \( x := 0.9995 \cdot 10^0 \) and \( y := -0.9984 \cdot 10^0 \). Computing the sum with precision 3 yields

\[
\text{rd}_3(x) +_3 \text{rd}_3(y) = 0.100 \cdot 10^1 +_3 (-0.998 \cdot 10^0) = \text{rd}_3(0.2 \cdot 10^{-2}) = 0.2 \cdot 10^{-2}.
\]

Letting \( \epsilon := \epsilon_3(\text{rd}_3(x) + \text{rd}_3(y)) \), applying the formulas of Rem. A.23 provides:

\[
\begin{align*}
\text{rd}_3(x) +_3 \text{rd}_3(y) & \overset{(A.49)}{=} \left( \text{rd}_3(x) + \text{rd}_3(y) \right) (1 + \epsilon) \\
& \overset{(A.48)}{=} \left( x(1 + \epsilon_3(x)) + y(1 + \epsilon_3(y)) \right) (1 + \epsilon) \\
& = (x + y) + \varepsilon_a,
\end{align*}
\]

where

\[
\varepsilon_a = x(\epsilon + \epsilon_3(x)(1 + \epsilon)) + y(\epsilon + \epsilon_3(y)(1 + \epsilon)).
\]
Using (A.46) and plugging in the numbers yields:

\[
\epsilon = \frac{rd_3(x) + rd_3(y) - (rd_3(x) + rd_3(y))}{rd_3(x) + rd_3(y)} = \frac{0.002 - 0.002}{0.002} = 0,
\]

\[
e_3(x) = \frac{1 - 0.9995}{0.9995} = \frac{1}{1999} = 0.00050\ldots,
\]

\[
e_3(y) = \frac{-0.998 + 0.9984}{-0.9984} = \frac{1}{2496} = -0.00040\ldots,
\]

\[
e_a = x\epsilon_3(x) + y\epsilon_3(y) = 0.0005 + 0.0004 = 0.0009.
\]

The corresponding relative error is

\[
e_r = \frac{e_a}{|x| + |y|} = \frac{0.0009}{0.0011} = \frac{9}{11} = 0.81.
\]

Thus, \(e_r\) is much larger than both \(e_r(x)\) and \(e_r(y)\). This is an example of subtractive cancellation of digits, which can occur when subtracting numbers that are almost identical. If possible, such situations should be avoided in practice (cf. Examples A.25 and A.26 below).

**Example A.25.** Let us generalize the situation of Example A.24 to general \(x, y \in \mathbb{R}\setminus\{0\}\) and a general precision \(l \in \mathbb{N}\). The formulas in (A.50) and (A.51) remain valid if one replaces 3 by \(l\). Moreover, with \(b = 10\) and \(l \geq 2\), one obtains from (A.47):

\[
|\epsilon| \leq 0.5 \cdot 10^{-l+1} \leq 0.5 \cdot 10^{-1} = 0.05.
\]

Thus,

\[
|e_a| \leq |x| (|\epsilon + 1.05|\epsilon_l(x)|) + |y| (|\epsilon + 1.05|\epsilon_l(y)|),
\]

and

\[
e_r = \frac{|e_a|}{|x| + |y|} \leq \frac{|x|}{|x| + |y|} (|\epsilon| + 1.05|\epsilon_l(x)|) + \frac{|y|}{|x| + |y|} (|\epsilon| + 1.05|\epsilon_l(y)|).
\]

One can now distinguish three (not completely disjoint) cases:

(a) If \(|x + y| < \max\{|x|, |y|\}\) (in particular, if \(\text{sgn}(x) = -\text{sgn}(y)\)), then \(e_r\) is typically larger (potentially much larger, as in Example A.24) than \(|\epsilon_l(x)|\) and \(|\epsilon_l(y)|\). Subtractive cancellation falls into this case. If possible, this should be avoided (cf. Example A.26 below).

(b) If \(\text{sgn}(x) = \text{sgn}(y)\), then \(|x + y| = |x| + |y|\), implying

\[
e_r \leq |\epsilon| + 1.05 \max\{|\epsilon_l(x)|, |\epsilon_l(y)|\}
\]

i.e. the relative error is at most of the same order of magnitude as the number \(|\epsilon| + \max\{|\epsilon_l(x)|, |\epsilon_l(y)|\}\).

(c) If \(|y| \ll |x|\) (resp. \(|x| \ll |y|\)), then the bound for \(e_r\) is predominantly determined by \(|\epsilon_l(x)|\) (resp. \(|\epsilon_l(y)|\)) (error *damping*).
As the following Example A.26 illustrates, subtractive cancellation can often be avoided by rearranging an expression into a mathematically equivalent formula that is numerically more stable.

**Example A.26.** Given \(a, b, c \in \mathbb{R}\) satisfying \(a \neq 0\) and \(4ac \leq b^2\), the quadratic equation

\[
ax^2 + bx + c = 0
\]

has the solutions

\[
x_1 = \frac{1}{2a} \left( -b - \text{sgn}(b) \sqrt{b^2 - 4ac} \right), \quad x_2 = \frac{1}{2a} \left( -b + \text{sgn}(b) \sqrt{b^2 - 4ac} \right).
\]

If \(|4ac| \ll b^2\), then, for the computation of \(x_2\), one is in the situation of Example A.25(a), i.e. the formula is numerically unstable. However, due to \(x_1 x_2 = c/a\), one can use the equivalent formula

\[
x_2 = \frac{2c}{-b - \text{sgn}(b) \sqrt{b^2 - 4ac}},
\]

where subtractive cancellation can not occur.

## B Operator Norms and Natural Matrix Norms

**Theorem B.1.** For a linear map \(A : X \to Y\) between two normed vector spaces \((X, \| \cdot \|)\) and \((Y, \| \cdot \|)\) over \(K\), the following statements are equivalent:

(a) \(A\) is bounded.

(b) \(\|A\| < \infty\).

(c) \(A\) is Lipschitz continuous.

(d) \(A\) is continuous.

(e) There is \(x_0 \in X\) such that \(A\) is continuous at \(x_0\).

**Proof.** Since every Lipschitz continuous map is continuous and since every continuous map is continuous at every point, “(c) \(\Rightarrow\) (d) \(\Rightarrow\) (e)” is clear.

“(e) \(\Rightarrow\) (c)”: Let \(x_0 \in X\) be such that \(A\) is continuous at \(x_0\). Thus, for each \(\epsilon > 0\), there is \(\delta > 0\) such that \(\|x - x_0\| < \delta\) implies \(\|Ax - Ax_0\| < \epsilon\). As \(A\) is linear, for each \(x \in X\) with \(\|x\| < \delta\), one has \(\|Ax\| = \|A(x + x_0) - Ax_0\| < \epsilon\), due to \(\|x + x_0 - x_0\| = \|x\| < \delta\). Moreover, one has \(\|(\delta x)/2\| \leq \delta/2 < \delta\) for each \(x \in X\) with \(\|x\| \leq 1\). Letting \(L := 2\epsilon/\delta\), this means that \(\|Ax\| = \|A((\delta x)/2)/(\delta/2)\| < 2\epsilon/\delta = L\) for each \(x \in X\) with \(\|x\| \leq 1\). Thus, for each \(x, y \in X\) with \(x \neq y\), one has

\[
\|Ax - Ay\| = \|A(x - y)\| = \|x - y\| \left\| A \left( \frac{x - y}{\|x - y\|} \right) \right\| < L \|x - y\|. \tag{B.1}
\]
Together with the fact that \( \|Ax - Ay\| \leq L\|x - y\| \) is trivially true for \( x = y \), this shows that \( A \) is Lipschitz continuous.

“(c) \( \Rightarrow \) (b)”: As \( A \) is Lipschitz continuous, there exists \( L \in \mathbb{R}^+ \) such that \( \|Ax - Ay\| \leq L\|x - y\| \) for each \( x, y \in X \). Considering the special case \( y = 0 \) and \( \|x\| = 1 \) yields \( \|Ax\| \leq L\|x\| = L \), implying \( \|A\| \leq L < \infty \).

“(b) \( \Rightarrow \) (c)”: Let \( \|A\| < \infty \). We will show
\[
\|Ax - Ay\| \leq \|A\| \|x - y\| \quad \text{for each } x, y \in X.
\] (B.2)

For \( x = y \), there is nothing to prove. Thus, let \( x \neq y \). One computes
\[
\frac{\|Ax - Ay\|}{\|x - y\|} = \left\| A \left( \frac{x - y}{\|x - y\|} \right) \right\| \leq \|A\|
\] (B.3)
as \( \left\| \frac{x - y}{\|x - y\|} \right\| = 1 \), thereby establishing (B.2).

“(b) \( \Rightarrow \) (a)”: Let \( \|A\| < \infty \) and let \( M \subseteq X \) be bounded. Then there is \( r > 0 \) such that \( M \subseteq B_r(0) \). Moreover, for each \( 0 \neq x \in M \):
\[
\frac{\|Ax\|}{\|x\|} = \left\| A \left( \frac{x}{\|x\|} \right) \right\| \leq \|A\|
\] (B.4)
as \( \left\| \frac{x}{\|x\|} \right\| = 1 \). Thus \( \|Ax\| \leq \|A\|\|x\| \leq r\|A\| \), showing that \( A(M) \subseteq \overline{B}_r\|A\|(0) \). Thus, \( A(M) \) is bounded, thereby establishing the case.

“(a) \( \Rightarrow \) (b)”: Since \( A \) is bounded, it maps the bounded set \( \overline{B}_1(0) \subseteq X \) into some bounded subset of \( Y \). Thus, there is \( r > 0 \) such that \( A(\overline{B}_1(0)) \subseteq B_r(0) \subseteq Y \). In particular, \( \|Ax\| < r \) for each \( x \in X \) satisfying \( \|x\| = 1 \), showing \( \|A\| \leq r < \infty \). \( \blacksquare \)

**Theorem B.2.** Let \( X \) and \( Y \) be normed vector spaces over \( \mathbb{K} \).

(a) The operator norm does, indeed, constitute a norm on the set of bounded linear maps \( \mathcal{L}(X,Y) \).

(b) If \( A \in \mathcal{L}(X,Y) \), then \( \|A\| \) is the smallest Lipschitz constant for \( A \), i.e. \( \|A\| \) is a Lipschitz constant for \( A \) and \( \|Ax - Ay\| \leq L\|x - y\| \) for each \( x, y \in X \) implies \( \|A\| \leq L \).

**Proof.** (a): If \( A = 0 \), then, in particular, \( Ax = 0 \) for each \( x \in X \) with \( \|x\| = 1 \), implying \( \|A\| = 0 \). Conversely, \( \|A\| = 0 \) implies \( Ax = 0 \) for each \( x \in X \) with \( \|x\| = 1 \). But then \( Ax = \|x\| A(x/\|x\|) = 0 \) for every \( 0 \neq x \in X \), i.e. \( A = 0 \). Thus, the operator norm is positive definite. If \( A \in \mathcal{L}(X,Y) \), \( \lambda \in \mathbb{R} \), and \( x \in X \), then
\[
\|(\lambda A)x\| = \|\lambda(Ax)\| = |\lambda| \|Ax\|,
\] (B.5)
yielding
\[
\|\lambda A\| = \sup \left\{ \|(\lambda A)x\| : x \in X, \|x\| = 1 \right\} = \sup \left\{ |\lambda| \|Ax\| : x \in X, \|x\| = 1 \right\} = |\lambda| \|A\|,
\] (B.6)
showing that the operator norm is homogeneous of degree 1. Finally, if \( A, B \in \mathcal{L}(X,Y) \) and \( x \in X \), then
\[
\| (A + B)x \| = \| Ax + Bx \| \leq \| Ax \| + \| Bx \|, \tag{B.7}
\]
yielding
\[
\| A + B \| = \sup \{ \| (A + B)x \| : x \in X, \| x \| = 1 \}
\leq \sup \{ \| Ax \| + \| Bx \| : x \in X, \| x \| = 1 \}
\leq \sup \{ \| Ax \| : x \in X, \| x \| = 1 \} + \sup \{ \| Bx \| : x \in X, \| x \| = 1 \}
= \| A \| + \| B \|, \tag{B.8}
\]
showing that the operator norm also satisfies the triangle inequality, thereby completing the verification that it is, indeed, a norm.

(b): To see that \( \| A \| \) is a Lipschitz constant for \( A \), we have to show
\[
\| Ax - Ay \| \leq \| A \| \| x - y \| \text{ for each } x,y \in X. \tag{B.9}
\]
For \( x = y \), there is nothing to prove. Thus, let \( x \neq y \). One computes
\[
\frac{\| Ax - Ay \|}{\| x - y \|} = \left\| A \left( \frac{x - y}{\| x - y \|} \right) \right\| \leq \| A \|
\]
as \( \left\| \frac{x - y}{\| x - y \|} \right\| = 1 \), thereby establishing (B.9). Now let \( L \in \mathbb{R}^+ \) be such that \( \| Ax - Ay \| \leq L \| x - y \| \) for each \( x,y \in X \). Specializing to \( y = 0 \) and \( \| x \| = 1 \) implies \( \| Ax \| \leq L \| x \| = L \), showing \( \| A \| \leq L \).

C Linear Algebra

C.1 Positive Semidefinite Matrices

Theorem C.1. Let \( n \in \mathbb{N} \). If \( A \in \mathcal{M}(n, \mathbb{K}) \) is Hermitian, then all eigenvalues \( \lambda_k \) of \( A \) are real. Moreover, there is an ordered orthonormal basis \( (v_1, \ldots, v_n) \) of \( \mathbb{K}^n \) such that \( v_k \in \mathbb{K}^n \) is an eigenvector for \( \lambda_k \in \mathbb{R} \) (e.g. with respect to the standard scalar product of \( \mathbb{K}^n \)), i.e. such that, in particular, for each \( k \in \{1, \ldots, n\} \),
\[
Av_k = \lambda_k v_k, \tag{C.1a}
\]
\[
v_k^* v_l = v_k \cdot v_l = \begin{cases} 0 & \text{for } k \neq l, \\ 1 & \text{for } k = l \end{cases} \tag{C.1b}
\]
for each \( k,l \in \{1, \ldots, n\} \).

Proof. See [Phi19b, Th. 10.30].

Theorem C.2. Let \( A \in \mathcal{M}(n, \mathbb{C}) \) be Hermitian positive semidefinite (resp. Hermitian positive definite), \( n \in \mathbb{N} \) (cf. Def. 2.21).
(a) All eigenvalues of $A$ are nonnegative (resp. positive) real numbers.

(b) $\det A \geq 0$ (resp. $\det A > 0$).

**Proof.** (a): Let $\lambda \in \mathbb{C}$ be an eigenvalue of $A$. According to Th. C.1, $\lambda \in \mathbb{R}$ and there is an eigenvector $x \in \mathbb{C}^n \setminus \{0\}$ for $\lambda$. Thus,

$$0 \leq x^* Ax = x^* \lambda x = \lambda \|x\|^2_2,$$

where the inequality is strict in the case where $A$ is positive definite. As $\|x\|^2_2 \in \mathbb{R}^+_0$, $\lambda \in \mathbb{R}^+_0$ and even $\lambda \in \mathbb{R}^+$ if $A$ is positive definite.

(b) follows from (a), as $\det A$ is the product of the eigenvalues of $A$. ■

**Notation C.3.** If $A = (a_{ij})$ is an $n \times n$ matrix, $n \in \mathbb{N}$, then, for $1 \leq k \leq l \leq n$, let

$$A^{kl} := \begin{pmatrix} a_{kk} & \cdots & a_{kl} \\ \vdots & \ddots & \vdots \\ a_{lk} & \cdots & a_{ll} \end{pmatrix}$$

denote the $(1 + l - k) \times (1 + l - k)$ principal submatrix of $A$, i.e.

$$\forall_{k,l \in \{1, \ldots, n\}, \ 1 \leq k \leq l \leq n} \forall_{i,j \in \{1, \ldots, 1+l-k\}} a_{ij}^{kl} := a_{i+k-1,j+k-1}. \quad (C.3)$$

**Proposition C.4.** Let $A = (a_{\alpha\beta}) \in \mathcal{M}(n, \mathbb{C})$, $n \in \mathbb{N}$. Then $A$ is positive (semi-)definite if, and only if, every principal submatrix $A^{kl}$ of $A$, $1 \leq k \leq l \leq n$, is positive (semi-)definite.

**Proof.** If all principal submatrices of $A$ are positive (semi-)definite, then, as $A = A^{nn}$, $A$ is positive (semi-)definite. Now assume $A$ to be positive (semi-)definite and fix $k, l \in \{1, \ldots, n\}$ with $1 \leq k \leq l \leq n$. Let $x = (x_k, \ldots, x_l) \in \mathbb{C}^{1+l-k} \setminus \{0\}$ and extend $x$ to $\mathbb{C}^n$ by 0, calling the extended vector $y$:

$$y = (y_1, \ldots, y_n) \in \mathbb{C}^n \setminus \{0\}, \quad y_\alpha = \begin{cases} x_\alpha & \text{for } k \leq \alpha \leq l, \\ 0 & \text{otherwise.} \end{cases} \quad (C.5)$$

We now consider $x$ and $y$ as column vectors and compute

$$x^t A^{kl} x = \sum_{\alpha, \beta = k}^l a_{\alpha\beta} x_\alpha x_\beta = \sum_{\alpha, \beta = 1}^n a_{\alpha\beta} y_\alpha y_\beta \geq 0, \quad (C.6)$$

showing $A^{kl}$ to be positive (semi-)definite. ■
D Integration

D.1 \( \mathbb{K}^m \)-Valued Integration

In the proof of the well-conditionedness of a continuously differentiable function \( f : U \rightarrow \mathbb{R}^m, U \subseteq \mathbb{R}^n \), in Th. 2.47, we make use of \( \mathbb{R}^m \)-valued integrals. In particular, in the estimate (2.42), we use that \( \| f_0^1 f \| \leq f_0^1 \| f \| \) for each norm on \( \mathbb{R}^m \). While this is true for vector-valued integrals in general, the goal here is only to provide a proof for our special situation (except that we also allow \( \mathbb{C}^m \)-valued integrals in the following treatment). For a general treatment of vector-valued integrals, see, for example, [Alt06, Sec. A1] or [Yos80, Sec. V.5].

Definition D.1. Let \( A \subseteq \mathbb{R}^n \) be (Lebesgue) measurable, \( m,n \in \mathbb{N} \).

(a) A function \( f : A \rightarrow \mathbb{K}^m \) is called (Lebesgue) measurable (respectively, (Lebesgue) integrable) if, and only if, each coordinate function \( f_j = \pi_j \circ f : A \rightarrow \mathbb{K} \), \( j = 1, \ldots, m \), is (Lebesgue) measurable (respectively, (Lebesgue) integrable), which, for \( \mathbb{K} = \mathbb{C} \), means if, and only if, each \( \Re f_j \) and each \( \Im f_j, j = 1, \ldots, m \), is (Lebesgue) measurable (respectively, (Lebesgue) integrable).

(b) If \( f : A \rightarrow \mathbb{K}^m \) is integrable, then

\[
\int_A f := \left( \int_A f_1, \ldots, \int_A f_m \right) \in \mathbb{K}^m
\]  

is the (\( \mathbb{K}^m \)-valued) (Lebesgue) integral of \( f \) over \( A \).

Remark D.2. The linearity of the \( \mathbb{K} \)-valued integral implies the linearity of the \( \mathbb{K}^m \)-valued integral.

Theorem D.3. Let \( A \subseteq \mathbb{R}^n \) be measurable, \( m,n \in \mathbb{N} \). Then \( f : A \rightarrow \mathbb{K}^m \) is measurable in the sense of Def. D.1(a) if, and only if, \( f^{-1}(O) \) is measurable for each open subset \( O \) of \( \mathbb{K}^m \).

Proof. Assume \( f^{-1}(O) \) is measurable for each open subset \( O \) of \( \mathbb{K}^m \). Let \( j \in \{1, \ldots, m\} \). If \( O_j \subseteq \mathbb{K} \) is open in \( \mathbb{K} \), then \( O := \pi_j^{-1}(O) = \{ x \in \mathbb{K}^m : x_j \in O_j \} \) is open in \( \mathbb{K}^m \). Thus, \( f_j^{-1}(O_j) = f^{-1}(O) \) is measurable, showing that each \( f_j \) is measurable, i.e. \( f \) is measurable. Now assume \( f \) is measurable, i.e. each \( f_j \) is measurable. Since every open \( O \subseteq \mathbb{K}^m \) is a countable union of open sets of the form \( O = O_1 \times \cdots \times O_m \) with each \( O_j \) being an open subset of \( \mathbb{K} \), it suffices to show that the preimages of such open sets are measurable. So let \( O \) be as above. Then \( f^{-1}(O) = \bigcap_{j=1}^m f_j^{-1}(O_j) \), showing that \( f^{-1}(O) \) is measurable.

Corollary D.4. Let \( A \subseteq \mathbb{R}^n \) be measurable, \( m,n \in \mathbb{N} \). If \( f : A \rightarrow \mathbb{K}^m \) is measurable, then \( \| f \| : A \rightarrow \mathbb{R} \) is measurable.

Proof. If \( O \subseteq \mathbb{R} \) is open, then \( \| \cdot \|^{-1}(O) \) is an open subset of \( \mathbb{K}^m \) by the continuity of the norm. In consequence, \( \| f \|^{-1}(O) = f^{-1}(\| \cdot \|^{-1}(O)) \) is measurable.
Theorem D.5. Let $A \subseteq \mathbb{R}^n$ be measurable, $m, n \in \mathbb{N}$. For each norm $\| \cdot \|$ on $\mathbb{K}^m$ and each integrable $f : A \to \mathbb{K}^m$, the following holds:

$$\left\| \int_A f \right\| \leq \int_A \| f \|. \quad \text{(D.2)}$$

Proof. First assume that $B \subseteq A$ is measurable, $y \in \mathbb{K}^m$, and $f = y \chi_B$, where $\chi_B$ is the characteristic function of $B$ (i.e. the $f_j$ are $y_j$ on $B$ and 0 on $A \setminus B$). Then

$$\left\| \int_A f \right\| = \left\| (y_1 \lambda_n(B), \ldots, y_m \lambda_n(B)) \right\| = \lambda_n(B) \| y \| = \int_A \| f \|, \quad \text{ (D.3)}$$

where $\lambda_n$ denotes $n$-dimensional Lebesgue measure on $\mathbb{R}^n$. Next, consider the case that $f$ is a so-called simple function, that means $f$ takes only finitely many values $y_1, \ldots, y_N \in \mathbb{K}^m$, $N \in \mathbb{N}$, and each preimage $B_j := f^{-1}\{y_j\} \subseteq \mathbb{R}^n$ is measurable. Then $f = \sum_{j=1}^N y_j \chi_{B_j}$, where, without loss of generality, we may assume that the $B_j$ are pairwise disjoint. We obtain

$$\left\| \int_A f \right\| \leq \sum_{j=1}^N \left\| \int_A y_j \chi_{B_j} \right\| = \sum_{j=1}^N \int_A \| y_j \chi_{B_j} \| = \int_A \sum_{j=1}^N \| y_j \chi_{B_j} \| \quad \text{(D.4)}$$

where, at (*), it was used that, as the $B_j$ are disjoint, the integrands of the two integrals are equal at each $x \in A$.

Now, if $f$ is integrable, then each $\text{Re} \ f_j$ and each $\text{Im} \ f_j$, $j \in \{1, \ldots, m\}$, is integrable (i.e., for each $j \in \{1, \ldots, m\}$, $\text{Re} \ f_j, \text{Im} \ f_j \in L^1(A)$ and there exist sequences of simple functions $\phi_{j,k} : A \to \mathbb{R}$ and $\psi_{j,k} : A \to \mathbb{R}$ such that $\lim_{k \to \infty} \| \phi_{j,k} - \text{Re} \ f_j \|_{L^1(A)} = \lim_{k \to \infty} \| \psi_{j,k} - \text{Im} \ f_j \|_{L^1(A)} = 0$. In particular, for each $j \in \{1, \ldots, m\}$,

$$0 \leq \lim_{k \to \infty} \left| \int_A \phi_{j,k} - \int_A \text{Re} \ f_j \right| \leq \lim_{k \to \infty} \| \phi_{j,k} - \text{Re} \ f_j \|_{L^1(A)} = 0, \quad \text{ (D.5a)}$$

$$0 \leq \lim_{k \to \infty} \left| \int_A \psi_{j,k} - \int_A \text{Im} \ f_j \right| \leq \lim_{k \to \infty} \| \psi_{j,k} - \text{Im} \ f_j \|_{L^1(A)} = 0, \quad \text{ (D.5b)}$$

and also, for each $j \in \{1, \ldots, m\}$,

$$0 \leq \lim_{k \to \infty} \| \phi_{j,k} + i\psi_{j,k} - f_j \|_{L^1(A)}$$

$$\leq \lim_{k \to \infty} \| \phi_{j,k} - \text{Re} \ f_j \|_{L^1(A)} + \lim_{k \to \infty} \| \psi_{j,k} - \text{Im} \ f_j \|_{L^1(A)} = 0. \quad \text{ (D.6)}$$

Thus, letting, for each $k \in \mathbb{N}$, $\phi_k := (\phi_{1,k}, \ldots, \phi_{m,k})$ and $\psi_k := (\psi_{1,k}, \ldots, \psi_{m,k})$, we
obtain

\[ \left\| \int_A f \right\| = \left\| \left( \int_A f_1, \ldots, \int_A f_m \right) \right\| 
= \left\| \left( \lim_{k \to \infty} \int_A \phi_{1,k} + i \lim_{k \to \infty} \int_A \psi_{1,k}, \ldots, \lim_{k \to \infty} \int_A \phi_{m,k} + i \lim_{k \to \infty} \int_A \psi_{m,k} \right) \right\| 
= \lim_{k \to \infty} \left\| \int_A (\phi_k + i \psi_k) \right\| 
= \lim_{k \to \infty} \int_A \left\| \phi_k + i \psi_k \right\| \overset{(*)}{=} \int_A \left\| f \right\|, \tag{D.7} \]

where the equality at \((*)\) holds due to \(\lim_{k \to \infty} \left\| \phi_k + i \psi_k \right\| - \left\| f \right\|_{L^1(A)} = 0\), which, in turn, is verified by

\[ 0 \leq \int_A \left\| \phi_k + i \psi_k \right\| - \left\| f \right\| \leq \int_A \left\| \phi_k + i \psi_k - f \right\| \leq C \int_A \left\| \phi_k + i \psi_k - f \right\|_1 \]

\[ = C \int_A \sum_{j=1}^m |\phi_{j,k} + i \psi_{j,k} - f_j| \to 0 \quad \text{for} \quad k \to \infty, \tag{D.8} \]

with \(C \in \mathbb{R}^+\) since the norms \(\| \cdot \|\) and \(\| \cdot \|_1\) are equivalent on \(\mathbb{K}^n\). ■

## D.2 Continuity

**Theorem D.6.** Let \(a, b \in \mathbb{R}, a < b\). If \(f : [a, b] \to \mathbb{R}\) is (Lebesgue-) integrable, then

\[ F : [a, b] \to \mathbb{R}, \quad F(t) := \int_a^t f, \tag{D.9} \]

is continuous.

**Proof.** We show \(F\) to be continuous at each \(t \in [a, b]\): Let \((t_k)_{k \in \mathbb{N}}\) be a sequence in \([a, b]\) such that \(\lim_{k \to \infty} t_k = t\). Let \(f_k := f \chi_{[a, t_k]}\). Then \(f_k \to f \chi_{[a, t]}\) (pointwise everywhere, except, possibly, at \(t\)) and the dominated convergence theorem (DCT) yields (note \(|f_k| \leq |f|\))

\[ \lim_{k \to \infty} F(t_k) = \lim_{k \to \infty} \int_a^{t_k} f \overset{\text{DCT}}{=} \int_a^t f = F(t), \]

proving the continuity of \(F\) at \(t\). ■

## E Divided Differences

The present section provides some additional results regarding the divided differences of Def. 3.16:
Theorem E.1 (Product Rule). Let $F$ be a field and consider the situation of Def. 3.16. Moreover, for each $j \in \{0, \ldots, r\}$, $m \in \{0, \ldots, m_j - 1\}$, let $y_{f,j}^{(m)}, y_{g,j}^{(m)}, y_{h,j}^{(m)} \in F$ be such that

$$y_{h,j}^{(m)} = \sum_{k=0}^{m} \binom{m}{k} y_{f,j}^{(m-k)} y_{g,j}^{(k)}. \quad (E.1)$$

Let $P_f, P_g, P_h \in F[X]_n$ be the interpolating polynomials corresponding to the case, where the $y_j^{(m)}$ of Def. 3.16 are replaced by $y_{f,j}^{(m)}, y_{g,j}^{(m)}, y_{h,j}^{(m)}$, respectively, with $[x_0, \ldots, x_n]_f, [x_0, \ldots, x_n]_g, [x_0, \ldots, x_n]_h$ denoting the corresponding divided differences. Then one has

$$[x_0, \ldots, x_n]_h = \sum_{i=0}^{n} [x_0, \ldots, x_i]_f [x_i, \ldots, x_n]_g \quad (E.2)$$

(technically, in particular, the product rule holds if $y_{f,j}^{(m)} = f^{(m)}(x_j)$, $y_{g,j}^{(m)} = g^{(m)}(x_j)$, $y_{h,j}^{(m)} = h^{(m)}(x_j)$, and $h = fg$, where $f, g \in F[X]$ or $F = \mathbb{R}$ with $f, g : \mathbb{R} \to \mathbb{R}$ differentiable sufficiently often).

Proof. According to (3.20) and (b), we have

$$P_f = \sum_{i=0}^{n} [x_0, \ldots, x_i]_f \omega_{f,i}, \quad P_g = \sum_{j=0}^{n} [x_j, \ldots, x_n]_g \omega_{g,j},$$

where, for each $i, j \in \{0, 1, \ldots, n\}$,

$$\omega_{f,i} := \prod_{k=0}^{i-1} (X - x_k), \quad \omega_{g,j} := \prod_{k=n}^{j+1} (X - x_k) = \prod_{k=j+1}^{n} (X - x_k).$$

Consider the polynomials

$$P_f P_g = \sum_{i,j=0}^{n} [x_0, \ldots, x_i]_f [x_j, \ldots, x_n]_g \omega_{f,i} \omega_{g,j} \in F[X]_{2n}$$

and

$$Q := \sum_{\substack{i,j=0, \atop i \leq j}}^{n} [x_0, \ldots, x_i]_f [x_j, \ldots, x_n]_g \omega_{f,i} \omega_{g,j} \in F[X]_n,$$

where the degree estimates for $P_f P_g$ and $Q$ are due to

$$\forall_{i,j \in \{0, \ldots, n\}} \deg(\omega_{f,i} \omega_{g,j}) = i + n - j \leq \begin{cases} 2n, \quad & \text{for } i < j, \\ n, \quad & \text{for } i \leq j. \end{cases}$$
We will show $Q = P_h$, but, first, we note, for each $j \in \{0, \ldots, r\}$, $m \in \{0, \ldots, m_j - 1\}$,

$$(P_f P_g)^{(m)}(\tilde{x}_j) = \sum_{k=0}^{m} \binom{m}{k} P_f^{(m-k)}(\tilde{x}_j) P_g^{(k)}(\tilde{x}_j) = \sum_{k=0}^{m} \binom{m}{k} y_{f,j}^{(m-k)} y_{g,j}^{(k)} = y_{h,j}^{(m)} \quad \text{(E.3)}$$

If $i > j$, then $\omega_{f,i} \omega_{g,j}$ has the divisor (i.e. the factor) $\prod_{k=0}^{n}(X - x_k)$, i.e.

$$\forall \alpha \in \{0, \ldots, r\} \quad \forall m \in \{0, \ldots, m_\alpha - 1\} \quad (\omega_{f,i} \omega_{g,j})(\tilde{x}_\alpha) = 0,$$

implying

$$\forall \alpha \in \{0, \ldots, r\} \quad \forall m \in \{0, \ldots, m_\alpha - 1\} \quad Q(\tilde{x}_\alpha) = (P_f P_g)^{(m)}(\tilde{x}_\alpha) = y_{h,\alpha}^{(m)},$$

and, thus, $Q = P_h$. In $Q$, the summands with $\deg(\omega_{f,i} \omega_{g,j}) = n$ are precisely those with $i = j$, thereby proving (E.2).

\begin{proof}

The following result is proved under the additional assumption that $x_0, \ldots, x_n \in F$ are all distinct:

\textbf{Proposition E.2.} Let $F$ be a field and consider the situation of Def. 3.16 with $r = n \in \mathbb{N}_0$, i.e. with $\tilde{x}_0, \ldots, \tilde{x}_n \in F$ all being distinct. Moreover, let $(x_0, \ldots, x_n) := (\tilde{x}_0, \ldots, \tilde{x}_n)$. Then, according to Def. 3.12, $P[x_0, \ldots, x_n]$ denotes the unique polynomial $f \in P[X]^n$ that, given $y_0, \ldots, y_n \in F$, satisfies

$$\forall j \in \{0, \ldots, n\} \quad f(x_j) = y_j^0 := y_j,$$

and, according to Def. 3.16, $[x_0, \ldots, x_n]$ denotes its coefficient of $X^n$. Then

$$[x_0, \ldots, x_n] = \sum_{i=0}^{n} y_i \prod_{\substack{m=0 \atop m \neq i}}^{n} \frac{1}{x_i - x_m} \quad \text{(E.5)}$$

\begin{proof}

The proof is carried out by induction on $n \in \mathbb{N}_0$: For $n = 0$, one has

$$\sum_{i=0}^{0} y_i \prod_{\substack{m=0 \atop m \neq i}}^{0} \frac{1}{x_i - x_m} = y_0 = [x_0],$$

as desired. Thus, let $n > 0$.

\textbf{Claim 1.} It suffices to show that the following identity holds true:

$$\sum_{i=0}^{n} y_i \prod_{\substack{m=1 \atop m \neq i}}^{n} \frac{1}{x_i - x_m} = \frac{1}{x_n - x_0} \left( \sum_{i=1}^{n} y_i \prod_{\substack{m=1 \atop m \neq i}}^{n} \frac{1}{x_i - x_m} - \sum_{i=0}^{n-1} y_i \prod_{\substack{m=1 \atop m \neq i}}^{n-1} \frac{1}{x_i - x_m} \right). \quad \text{(E.6)}$$

\end{proof}
Proof. One computes as follows (where (E.5) is used, it holds by induction):

\[
\begin{align*}
\sum_{i=0}^{n} y_i \prod_{m=0, m\neq i}^{n} \frac{1}{x_i - x_m} &= \frac{1}{x_n - x_0} \left( \sum_{i=1}^{n} y_i \prod_{m=1, m\neq i}^{n} \frac{1}{x_i - x_m} - \sum_{i=0}^{n-1} y_i \prod_{m=0, m\neq i}^{n-1} \frac{1}{x_i - x_m} \right) \\
&= \frac{1}{x_n - x_0} \left( [x_1, \ldots, x_n] - [x_0, \ldots, x_{n-1}] \right) \\
&= [x_0, \ldots, x_n],
\end{align*}
\]

(E.6)

showing that (E.6) does, indeed, suffice. ▲

So it remains to verify (E.6). To this end, we show that, for each \(i \in \{0, \ldots, n\}\), the coefficients of \(y_i\) on both sides of (E.6) are identical.

Case \(i = n\): Since \(y_n\) occurs only in first sum on the right-hand side of (E.6), one has to show that

\[
\prod_{m=0, m\neq n}^{n} \frac{1}{x_i - x_m} = \frac{1}{x_n - x_0} \prod_{m=1, m\neq n}^{n} \frac{1}{x_i - x_m}.
\]

(E.7)

However, (E.7) is obviously true.

Case \(i = 0\): Since \(y_0\) occurs only in second sum on the right-hand side of (E.6), one has to show that

\[
\prod_{m=0, m\neq 0}^{n} \frac{1}{x_0 - x_m} = -\frac{1}{x_n - x_0} \prod_{m=0, m\neq 0}^{n-1} \frac{1}{x_0 - x_m}.
\]

(E.8)

Once again, (E.8) is obviously true.

Case \(0 < i < n\): In this case, one has to show

\[
\prod_{m=0, m\neq i}^{n} \frac{1}{x_i - x_m} = \frac{1}{x_n - x_0} \left( \prod_{m=1, m\neq i}^{n} \frac{1}{x_i - x_m} - \prod_{m=0, m\neq i}^{n-1} \frac{1}{x_i - x_m} \right).
\]

(E.9)

Multiplying (E.9) by

\[
\prod_{m=0, m\neq i, 0, n}^{n} (x_i - x_m)
\]

shows that it is equivalent to

\[
\frac{1}{x_i - x_0} \frac{1}{x_i - x_n} = \frac{1}{x_n - x_0} \left( \frac{1}{x_i - x_n} - \frac{1}{x_i - x_0} \right),
\]

which is also clearly valid. ■
F Discrete Fourier Transform and Trigonometric Interpolation

F.1 Discrete Fourier Transform (DFT)

Definition F.1. Let \( n \in \mathbb{N} = \{1, 2, 3, \ldots\} \). The discrete Fourier transform (DFT) is the map
\[
\mathcal{F} : \mathbb{C}^n \rightarrow \mathbb{C}^n, \quad (f_0, \ldots, f_{n-1}) \mapsto (d_0, \ldots, d_{n-1}),
\]
where
\[
\forall k \in \{0, \ldots, n-1\} \quad d_k := \frac{1}{n} \sum_{j=0}^{n-1} f_j e^{-\frac{ijk\pi}{n}},
\]
\( i \in \mathbb{C} \) denoting the imaginary unit.

Important applications of DFT include the approximation and interpolation of periodic functions (Sec. F.2 and Sec. F.3), practical applications include digital data transmission (cf. Rem. F.9). First, we need to study some properties of DFT and its inverse map.

Remark F.2. Introducing the vectors
\[
f := \begin{pmatrix} f_0 \\
0 \\ \vdots \\
f_{n-1} \end{pmatrix} \in \mathbb{C}^n,
\]
\[
d := \begin{pmatrix} d_0 \\
0 \\ \vdots \\
d_{n-1} \end{pmatrix} \in \mathbb{C}^n,
\]
we can write (F.1) in matrix form as
\[
\mathcal{F} : \mathbb{C}^n \rightarrow \mathbb{C}^n, \quad d := \mathcal{F}(f) = \frac{1}{n} A f,
\]
where
\[
A = \begin{pmatrix}
1 & 1 & 1 & \ldots & 1 \\
1 & \omega & \omega^2 & \ldots & \omega^{n-1} \\
1 & \omega^2 & \omega^4 & \ldots & \omega^{2(n-1)} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & \omega^{n-1} & \omega^{2(n-1)} & \ldots & \omega^{(n-1)^2} \\
\end{pmatrix},
\]
that means \( A = (a_{kj}) \in \mathcal{M}(n, \mathbb{C}) \) is a symmetric complex matrix with
\[
\forall k,j \in \{0, \ldots, n-1\} \quad a_{kj} = \omega^{kj} = e^{\frac{ijk\pi}{n}}, \quad \omega := e^{\frac{2\pi}{n}}.
\]
In (F.2), \( A = (\bar{a}_{kj}) \in \mathcal{M}(n, \mathbb{C}) \) denotes the complex conjugate matrix of \( A \).

Proposition F.3. Let \( A \in \mathcal{M}(n, \mathbb{C}), \ n \in \mathbb{N} \), be the matrix defined in (F.3a), and set \( B := \frac{1}{\sqrt{n}} \overline{A} \).
(a) The columns of $B$ form an orthonormal basis with respect to the standard inner product on $\mathbb{C}^n$.

(b) The matrix $B$ is unitary, i.e. invertible with $B^{-1} = B^*$.

(c) DFT, as defined in (F.1), is invertible and its inverse map, called inverse DFT, is given by

$$F^{-1} : \mathbb{C}^n \rightarrow \mathbb{C}^n, \quad (d_0, \ldots, d_{n-1}) \mapsto (f_0, \ldots, f_{n-1}), \quad (F.4a)$$

where

$$\forall \quad f_k := \sum_{j=0}^{n-1} d_j e^{\frac{ijk2\pi}{n}}. \quad (F.4b)$$

(d) If $d, f \in \mathbb{C}^n$ are such that $d = F(f)$, then

$$\sum_{j=0}^{n-1} |d_j|^2 = \frac{1}{n} \sum_{j=0}^{n-1} |f_j|^2, \quad (F.5a)$$

i.e.

$$\|F(f)\|_2 = \frac{1}{\sqrt{n}} \|f\|_2. \quad (F.5b)$$

Proof. (a): For each $k = 0, \ldots, n-1$, we let $b_k$ denote the $k$th column of $B$. Then, noting $|\omega| = 1$,

$$\langle b_k, b_k \rangle = \sum_{j=0}^{n-1} b_{jk} \bar{b}_{jk} = \frac{1}{n} \sum_{j=0}^{n-1} |\omega|^{2jk} = \frac{n}{n} = 1.$$

Moreover, for each $k, l \in \{0, \ldots, n-1\}$ and $k \neq l$, we note $\omega^{l-k} \neq 1$, but $(\omega^{l-k})^n = 1$, and we use the geometric sum to obtain

$$\langle b_k, b_l \rangle = \sum_{j=0}^{n-1} b_{jk} \bar{b}_{jl} = \frac{1}{n} \sum_{j=0}^{n-1} \omega^{-jk} \omega^{jl} = \frac{1}{n} \sum_{j=0}^{n-1} (\omega^{l-k})^j = \frac{1}{n} \cdot \frac{1 - (\omega^{l-k})^n}{1 - \omega^{l-k}} = 0,$$

proving (a).

(b) is immediate due to (a) and the equivalences (i) and (iii) of [Phi19b, Pro. 10.26].

(c) follows from (b), as

$$\forall \quad d \in \mathbb{C}^n \quad f := F^{-1}(d) = \left( \frac{1}{\sqrt{n}} B \right)^{-1} d = \sqrt{n} B^* d \quad (F.6)$$

and

$$\forall \quad k, j \in \{0, \ldots, n-1\} \quad \sqrt{n} b_{kj}^* = \frac{\sqrt{n}}{\sqrt{n}} a_{kj} = e^{\frac{ijk2\pi}{n}}.$$ 

(d): According to [Phi19b, Pro. 10.26(viii)], $B$ is isometric with respect to $\| \cdot \|_2$. Thus,

$$\|F(f)\|_2 = \left\| \frac{1}{\sqrt{n}} Bf \right\|_2 = \frac{1}{\sqrt{n}} \|f\|_2,$$

proving (F.5b). Then (F.5a) also follows, as it is merely (F.5b) squared.
As it turns out, the inverse DFT of Prop. F.3(c) has the useful property that there are several simple ways to express it in terms of (forward) DFT:

**Proposition F.4.** Let \( n \in \mathbb{N} \) and let \( \mathcal{F} \) denote DFT as defined in (F.1).

(a) One has
\[
\forall \ d = (d_0, \ldots, d_{n-1}) \in \mathbb{C}^n \quad \mathcal{F}^{-1}(d) = n \mathcal{F}(d_0, d_{n-1}, d_{n-2}, \ldots, d_1).
\]

(b) Applying complex conjugation, one has
\[
\forall \ d \in \mathbb{C}^n \quad \mathcal{F}^{-1}(d) = n \overline{\mathcal{F}(d)}.
\]

(c) Let \( \sigma \) be the map that swaps the real and imaginary parts of a complex number, i.e.
\[
\sigma : \mathbb{C} \rightarrow \mathbb{C}, \quad \sigma(a + bi) := b + ai = i(a + bi).
\]

Using componentwise application, we can also consider \( \sigma \) to be defined on \( \mathbb{C}^n \):
\[
\sigma : \mathbb{C}^n \rightarrow \mathbb{C}^n, \quad \sigma(z_0, \ldots, z_{n-1}) := (\sigma(z_0), \ldots, \sigma(z_{n-1})) = i\mathbb{Z}.
\]

Then one has
\[
\forall \ d \in \mathbb{C}^n \quad \mathcal{F}^{-1}(d) = n \sigma(\mathcal{F}(\sigma d)).
\]

**Proof.** (a): We let \( f = (f_0, \ldots, f_{n-1}) := n \mathcal{F}(d_0, d_{n-1}, d_{n-2}, \ldots, d_1) \) and compute for each \( k = 0, \ldots, n-1 \):
\[
f_k \overset{(F.1b)}{=} d_0 + \sum_{j=1}^{n-1} d_{n-j} e^{-\frac{i j k 2\pi}{n}} = d_0 + \sum_{j=1}^{n-1} d_j e^{-\frac{i j (n-j) 2\pi}{n}} = d_0 + \sum_{j=1}^{n-1} d_j e^{-\frac{i j k 2\pi}{n}} = \sum_{j=0}^{n-1} d_j e^{\frac{i j k 2\pi}{n}},
\]
which, according to (F.4), proves (a).

(b): Letting \( f := n \mathcal{F}(\overline{d}) \), we obtain for each \( k = 0, \ldots, n-1 \):
\[
f_k \overset{(F.1b)}{=} \sum_{j=0}^{n-1} \overline{d}_j e^{-\frac{i j k 2\pi}{n}} = \sum_{j=0}^{n-1} d_j e^{\frac{i j k 2\pi}{n}}.
\]

Taking the complex conjugate of \( f_k \) and comparing with (F.4) proves (b).

(c): Using the \( \mathbb{C} \)-linearity of DFT and (b), we compute
\[
n \sigma(\mathcal{F}(\sigma d)) = n \sigma(\mathcal{F}(i \overline{d})) = n i \overline{\mathcal{F}(d)} = n \overline{\mathcal{F}(d)} \overset{(b)}{=} \mathcal{F}^{-1}(d),
\]
which proves (c).
F.2 Approximation of Periodic Functions

As a first application of the discrete Fourier transform, we consider the approximation of periodic functions; how this is used in the field of data transmission is indicated in Rem. F.9 below. The approximation is based on the Fourier series representation of (sufficiently regular) periodic functions, which we will proceed to, briefly, consider as a preparation. We will make use of the following well-known relations between the exponential function and the trigonometric functions sine and cosine (see [Phi16a, (8.46)]):

\[
\forall z \in \mathbb{C} \quad e^{iz} = \cos z + i \sin z \quad \text{(Euler formula)}, \tag{F.7a}
\]

\[
\forall z \in \mathbb{C} \quad \cos z = \frac{e^{iz} + e^{-iz}}{2}, \tag{F.7b}
\]

\[
\forall z \in \mathbb{C} \quad \sin z = \frac{e^{iz} - e^{-iz}}{2i}. \tag{F.7c}
\]

**Definition F.5.** Let \( L \in \mathbb{R}^+ \) and let \( f : [0, L] \to \mathbb{R} \) be integrable.

(a) For each \( j \in \mathbb{Z} \), we define numbers

\[
a_j := a_j(f) := \frac{1}{L} \int_0^L f(t) \cos \left(-\frac{j2\pi t}{L}\right) \, dt \in \mathbb{R}, \tag{F.8a}
\]

\[
b_j := b_j(f) := \frac{1}{L} \int_0^L f(t) \sin \left(-\frac{j2\pi t}{L}\right) \, dt \in \mathbb{R}, \tag{F.8b}
\]

\[
c_j := c_j(f) := \frac{1}{L} \int_0^L f(t) e^{-i\frac{j2\pi t}{L}} \, dt \in \mathbb{C}. \tag{F.8c}
\]

We call the \( a_j, b_j \) the **real Fourier coefficients** of \( f \) and the \( c_j \) the **complex Fourier coefficients** of \( f \). As a consequence of (F.7a), one has

\[
\forall j \in \mathbb{Z} \quad c_j = a_j + ib_j. \tag{F.9}
\]

(b) For each \( n \in \mathbb{N} \), we define the functions

\[
S_n := S_n(f) : [0, L] \to \mathbb{R}, \quad S_n(x) := \sum_{j=-n}^{n} c_j e^{\frac{ij2\pi x}{L}} \tag{F.10}
\]

and consider the series

\[
(S_n)_{n \in \mathbb{N}}, \quad \text{i.e.} \quad \forall x \in [0, L] \quad (S_n(x))_{n \in \mathbb{N}} =: \sum_{j=-\infty}^{\infty} c_j e^{\frac{ij2\pi x}{L}}. \tag{F.11}
\]

We call the \( S_n \) the **Fourier partial sums** of \( f \) and the series of (F.11) the **Fourier series** of \( f \).
Remark F.6. (a) We verify that the Fourier partial sums $S_n$ of (F.10) are, indeed, real-valued. The key observation is an identity that we will prove in a form so that we can use it both here and, later, in the proof of Th. F.13(a). To this end, we fix $n \in \mathbb{N}$ and assume, for each $j \in \{-n, \ldots, n\}$, we have numbers $a_j, b_j \in \mathbb{R}$ and $c_j \in \mathbb{C}$ (not necessarily given by (F.8)), satisfying

$$\forall \ j \in \{-n, \ldots, n\} \quad c_j = a_j + ib_j \quad \text{and} \quad \forall \ j \in \{-n+1, \ldots, n-1\} \setminus \{0\} \quad (a_j = a_{-j}, \quad b_j = -b_{-j}). \quad (F.12)$$

Then, for each $x \in \mathbb{R}$, we compute

$$\sum_{j=-n}^{n} c_j e^{ij2\pi x/L} = \sum_{j=-n}^{n} (a_j + i b_j) e^{ij2\pi x/L} = (a_{-n} + i b_{-n}) e^{-in2\pi x/L} + a_0 + ib_0 + (a_n + i b_n) e^{in2\pi x/L} + \sum_{j=1}^{n-1} a_j \left( e^{ij2\pi x/L} + e^{-ij2\pi x/L} \right) + i \sum_{j=1}^{n-1} b_j \left( e^{ij2\pi x/L} - e^{-ij2\pi x/L} \right)$$

$$= (a_{-n} + i b_{-n}) e^{-in2\pi x/L} + a_0 + ib_0 + (a_n + i b_n) e^{in2\pi x/L} + 2 \sum_{j=1}^{n-1} a_j \cos \left( \frac{j2\pi x}{L} \right) - 2 \sum_{j=1}^{n-1} b_j \sin \left( \frac{j2\pi x}{L} \right). \quad (F.13)$$

We now come back to the case, where $a_j, b_j, c_j$ are defined by (F.8). Then (F.12) holds due to (F.9) together with the symmetry of cosine (i.e. $\cos(z) = \cos(-z)$) and the antisymmetry of sine (i.e. $\sin(z) = -\sin(-z)$), where we have $a_j = a_{-j}$ and $b_j = -b_{-j}$ even for $j = n$. Thus, for each $x \in [0, L]$, we use (F.13) to obtain

$$S_n(x) = \sum_{j=-n}^{n} c_j e^{ij2\pi x/L} = a_0 + 2 \sum_{j=1}^{n} a_j \cos \left( \frac{j2\pi x}{L} \right) - 2 \sum_{j=1}^{n} b_j \sin \left( \frac{j2\pi x}{L} \right),$$

showing $S_n(x)$ to be real-valued.

(b) In general, the question, whether (and in which sense) the Fourier series of $f$, as defined in (F.11), converges to $f$ does not have an easy answer. We will provide some sufficient conditions in the following Th. F.7. For convergence results under weaker assumptions, see, e.g., [Rud87, Sec. 4.26, Sec. 5.11] or [Heu08, Th. 136.2, Th. 137.1, Th. 137.2].

Theorem F.7. Let $L \in \mathbb{R}^+$ and let $f : [0, L] \rightarrow \mathbb{R}$ be integrable and periodic in the sense that $f(0) = f(L)$. Moreover, assume that $f$ is continuously differentiable (i.e. $f \in C^1([0, L])$) with $f'(0) = f'(L)$. Then the following convergence results for the Fourier series $(S_n)_{n \in \mathbb{N}}$ of $f$, with $S_n$ according to (F.10), hold:
(a) The Fourier series of \( f \) converges uniformly to \( f \). In particular, one has the pointwise convergence
\[
\forall \ x \in [0, L] \quad f(x) = \sum_{j=-\infty}^{\infty} c_j e^{ij2\pi x / L} = \lim_{n \to \infty} S_n(x).
\] (F.14)

(b) The Fourier series of \( f \) converges to \( f \) in \( L^p([0, L]) \) for each \( p \in [1, \infty) \).

Proof. (a): For the main work of the proof, we refer to [Wer11, Th. IV.2.9], where (a) is proved for the case \( L = 2\pi \). Here, we only verify that the case \( L = 2\pi \) then also implies the general case of \( L \in \mathbb{R}^+ \): Assume (a) holds for \( L = 2\pi \), let \( L \in \mathbb{R}^+ \) be arbitrary, and \( f \in C^1([0, L]) \) with \( f(0) = f(L) \) as well as \( f'(0) = f'(L) \). Define
\[
g : [0, 2\pi] \to \mathbb{R}, \quad g(t) := f \left( \frac{t L}{2\pi} \right).
\]
Then \( g \) is continuously differentiable with
\[
g' : [0, 2\pi] \to \mathbb{R}, \quad g'(t) := \frac{L}{2\pi} f' \left( \frac{t L}{2\pi} \right).
\]
Also \( g(0) = f(0) = f(L) = g(2\pi) \) and \( g'(0) = \frac{L}{2\pi} f'(0) = \frac{L}{2\pi} f'(L) = g'(2\pi) \), i.e. we know (a) to hold for \( g \). Next, we note \( f(t) = g \left( \frac{t L}{2\pi} \right) \) for each \( t \in [0, L] \) and compute, for each \( j \in \mathbb{Z} \),
\[
c_j(f) = \frac{1}{L} \int_0^L f(t) e^{-ij2\pi t / L} \ dt = \frac{1}{L} \int_0^L g \left( \frac{2\pi}{L} t \right) e^{-ij2\pi t / L} \ dt
\]
\[
= \frac{1}{2\pi} \int_0^{2\pi} g(s) e^{-ij s} \ ds = c_j(g),
\]
where we used the change of variables \( s := \frac{t L}{2\pi} \). Thus,
\[
\lim_{n \to \infty} \left\| S_n(f) - f \right\|_\infty = \lim_{n \to \infty} \sup \left\{ \left| f(x) - \sum_{j=-n}^{n} c_j(f) e^{ij2\pi x / L} \right| : x \in [0, L] \right\}
\]
\[
= \lim_{n \to \infty} \sup \left\{ \left| g \left( \frac{x 2\pi}{L} \right) - \sum_{j=-n}^{n} c_j(f) e^{ij2\pi x / L} \right| : x \in [0, L] \right\}
\]
\[
= \lim_{n \to \infty} \sup \left\{ \left| g(t) - \sum_{j=-n}^{n} c_j(g) e^{ijt} \right| : t \in [0, 2\pi] \right\}
\]
\[
= \lim_{n \to \infty} \left\| S_n(g) - g \right\|_\infty \overset{(a)}{=} 0,
\]
which completes the proof of (a).
(b): According to (a), we have convergence in 
\[ L^\infty[0,L] \] (note that the continuity of \( f \) 
and the compactness of \([0,L]\) imply \( f \in L^\infty[0,L] \)). As a consequence of the Hölder inequality (cf. [Phi17, Th. 2.42]), one then also obtains, for each \( p \in [1,\infty[ \),
\[
\lim_{n \to \infty} ||S_n(f) - f||_p = L^{\frac{1}{p}} \lim_{n \to \infty} ||S_n(f) - f||_\infty = 0,
\]
thereby establishing the case. 

For a periodic function \( f \), applying the inverse discrete Fourier transform can yield a useful approximation of the values of \( f \) at equidistant points:

**Example F.8.** Let \( L \in \mathbb{R}^+ \) and let \( f : [0,L] \rightarrow \mathbb{R} \) be integrable and such that \( f(0) = f(L) \). Given \( n \in \mathbb{N} \), we now define the equidistant points
\[
\forall \ k \in \{0,...,2n+1\} \quad x_k := kh, \quad h := \frac{L}{2n + 1}.
\] (F.15)

Plugging \( x := x_k \) into (F.10) then yields the following approximations \( f_k \) of \( f(x_k) \),
\[
\forall \ k \in \{0,...,2n\} \quad f_k := S_n(f)(x_k) = \sum_{j=-n}^{n} c_j e^{\frac{i j k 2\pi}{2n+1}} = \sum_{j=0}^{2n} c_{j-n} e^{\frac{i j k 2\pi}{2n+1}}
\] (F.16)
where we have omitted \( f(x_{2n+1}) \) in (F.16), since we know \( f(x_{2n+1}) = f(L) = f(0) = f(x_0) \), due to the assumed periodicity of \( f \). Comparing (F.16) with (F.4), we see that
\[
\left( f_0, e^{\frac{i 2\pi}{2n+1}} f_1, \ldots, e^{\frac{i 2n \cdot 2\pi}{2n+1}} f_{2n} \right) = \mathcal{F}^{-1}(c_{-n}, \ldots, c_n).
\] (F.17)

If Th. F.7(a) holds (e.g. under the hypotheses of Th. F.7), then, letting
\[
F_n := (f(x_0), \ldots, f(x_{2n})), \quad G_n := (f_0, \ldots, f_{2n}),
\]
we conclude
\[
\lim_{n \to \infty} \|F_n - G_n\|_\infty = \lim_{n \to \infty} \sup \left\{ \left| f(x_k) - S_n(f)(x_k) \right| : k \in \{0, \ldots, 2n\} \right\} \overset{\text{Th. F.7(a)}}{=} 0. \] (F.18)

**Remark F.9.** A physical application of the above-described approximation of periodic functions in connection with DFT is digital data transmission. Given an analog signal, represented by a function \( f \), the signal is compressed into finitely many \( c_j \) according to (F.16) and (F.8) (physically, this can be accomplished using high-pass and low-pass filters). The digital signal consisting of the \( c_j \) is then transmitted and, at the remote location, one wants to reconstruct the original \( f \), at least approximately. According to (F.17) and (F.18), an approximation of \( f \) can be obtained by applying inverse DFT to the \( c_j \) (combined with a suitable interpolation method).
Another useful application of DFT is the interpolation of periodic functions by means of trigonometric functions.

**Definition F.10.** Let $L \in \mathbb{R}^+$, $n \in \mathbb{N}$, and $d := (d_0, \ldots, d_{n-1}) \in \mathbb{C}^n$. Then the function

$$ p := p(d) : \mathbb{R} \rightarrow \mathbb{C}, \quad p(x) := \sum_{k=0}^{n-1} d_k e^{\frac{ik2\pi x}{L}}, \quad (F.19) $$

is called a *trigonometric polynomial* (which is justified due to Euler’s formula). As it turns out, the following modification will be more useful for our purposes here: We let

$$ r := r(d) : \mathbb{R} \rightarrow \mathbb{C}, \quad r(x) := e^{-\frac{inx}{L}} p(x) = e^{-\frac{inx}{L}} \sum_{k=0}^{n-1} d_k e^{\frac{ik2\pi x}{L}} = e^{-\frac{inx}{L}} \sum_{k=0}^{n-1} d_k e^{\frac{i(k-n/2)2\pi x}{L}}. \quad (F.20) $$

**Theorem F.11.** In the situation of Def. F.10, let $x_0, \ldots, x_{n-1}$ be defined by $x_j := jL/n$ for each $j \in \{0, \ldots, n-1\}$. Moreover, let $z := (z_0, \ldots, z_{n-1}) \in \mathbb{C}^n$.

(a) The function $r : \mathbb{R} \rightarrow \mathbb{C}$ of Def. F.10 satisfies

$$ \forall \quad r(x_j) = z_j \quad (F.21) $$

if, and only if,

$$ \mathcal{F}((-1)^0 z_0, \ldots, (-1)^{n-1} z_{n-1}) = d. \quad (F.22) $$

(b) Let $m \in \mathbb{N}$. If $f : [0, L] \rightarrow \mathbb{C}$ is an $m$ times continuously differentiable function such that $f(0) = f(L)$, and the function $r : \mathbb{R} \rightarrow \mathbb{C}$ of Def. F.10 satisfies (F.21) with $z_j := f(x_j)$, then there exists $c_m \in \mathbb{R}^+$ such that one has the following error estimate with respect to the $L^2$-norm:

$$ \|r - f\|_2 \leq \frac{c_m (\|f\|_2 + \|f^{(m)}\|_2)}{n^m} \quad (F.23) $$

- recall that, for a square-integrable function $g : [0, L] \rightarrow \mathbb{C}$, the $L^2$-norm is defined by $\|g\|_2 := \left(\int_0^L |g|^2\right)^{1/2}$.

**Proof.** (a): According to (F.20) and using $x_j = jL/n$, (F.21) is equivalent to

$$ \forall \quad z_j = e^{-\frac{inx_j}{L}} \sum_{k=0}^{n-1} d_k e^{\frac{ik2\pi}{n}} = e^{-\frac{in\pi j}{n}} \sum_{k=0}^{n-1} d_k e^{\frac{ik2\pi}{n}} = (-1)^j \sum_{k=0}^{n-1} d_k e^{\frac{ik2\pi}{n}}. $$
which, according to (F.4), is equivalent to

\[ \mathcal{F}^{-1}(d) = ((-1)^0 z_0, \ldots, (-1)^{n-1} z_{n-1}) . \]

Applying \( \mathcal{F} \) to the previous equation completes the proof of (a).

(b) is stated as [Pla06, Th. 3.10], where it is referred to [SV01] for the proof. It actually turns out to be a special case of the more general result [HB09, Th. 52.6]. ■

Example F.12. Consider

\[ f : [0, 1] \to \mathbb{R}, \quad f(x) := \begin{cases} x & \text{for } 0 \leq x \leq \frac{1}{2}, \\ 1 - x & \text{for } \frac{1}{2} \leq x \leq 1. \end{cases} \]

It is an exercise to use different values of \( n \) (e.g. \( n = 4 \) and \( n = 16 \)) to compute and plot the corresponding \( r \) satisfying the conditions of Th. F.11(a) with \( z_j = f(x_j), x_j = j/n, \) for instance by using MATLAB.

When approximating real-valued periodic functions \( f : [0, L] \to \mathbb{R}, \) it makes sense to use real-valued trigonometric functions. Indeed, we can choose real-valued trigonometric functions such that we can make use of the above results:

Theorem F.13. Let \( L \in \mathbb{R}^+ \), let \( n \in \mathbb{N} \) be even, and

\[ (a, b) := (a_0, \ldots, a_{\frac{n}{2}}, b_1, \ldots, b_{\frac{n}{2}-1}) \in \mathbb{R}^n. \]

Define the trigonometric function

\[ t := t(a, b) : \mathbb{R} \to \mathbb{R}, \]

\[ t(x) := a_0 + 2 \sum_{k=1}^{\frac{n}{2}-1} a_k \cos \left( \frac{k2\pi x}{L} \right) + b_k \sin \left( \frac{k2\pi x}{L} \right) + a_{\frac{n}{2}} \cos \left( \frac{n\pi x}{L} \right). \quad (F.24) \]

Furthermore, let \( d = (d_0, \ldots, d_{n-1}) \in \mathbb{C}^n \), let the function \( r = r(d) \) be defined according to (F.20), and assume the relations

\[ d_0 = a_0, \quad d_{\frac{n}{2}} = a_{\frac{n}{2}}, \]

\[ \forall k \in \{1, \ldots, \frac{n}{2}-1\} \quad d_{\frac{n}{2}-k} = a_k + ib_k, \quad d_{\frac{n}{2}+k} = a_k - ib_k. \quad (F.25) \]

(a) The above conditions imply \( t = \text{Re} r. \)

(b) If \( x_0, \ldots, x_{n-1} \in [0, L] \) are defined by \( x_j := jL/n \) for each \( j \in \{0, \ldots, n-1\} \), \( z := (z_0, \ldots, z_{n-1}) \in \mathbb{R}^n \), and \( r \) satisfies the conditions of Th. F.11(a), then

\[ \forall j \in \{0, \ldots, n-1\} \quad t(x_j) = z_j \quad (F.26) \]

and, moreover,

\[ \forall k \in \{1, \ldots, \frac{n}{2}-1\} \quad a_k = \frac{1}{n} \sum_{j=0}^{n-1} z_j \cos \left( \frac{jk2\pi}{n} \right), \quad b_k = \frac{1}{n} \sum_{j=0}^{n-1} z_j \sin \left( \frac{jk2\pi}{n} \right). \quad (F.27) \]
Let \( f : [0, L] \to \mathbb{R} \). If \( f \) and \( r \) satisfy the conditions of Th. F.11(b), then there exists \( c_m \in \mathbb{R}^+ \) such that the error estimate (F.23) holds with \( r \) replaced by \( t \).

Proof. (a): To apply (F.13), we let
\[
c_{n/2} := 0, \quad c_k := \begin{cases} d_k & \text{if } k \in \{-n/2, \ldots, n/2-1\}; \\
0 & \text{otherwise},
\end{cases}
\]
Then, for each \( k \in \{1, \ldots, n/2 - 1\} \),
\[
\text{Re}(c_k) = \text{Re}(d_{n/2 + k}) \overset{(F.25)}{=} \text{Re}(d_{n/2 - k}) = \text{Re}(c_{-k}) \quad (F.28a)
\]
as well as
\[
\text{Im}(c_k) = \text{Im}(d_{n/2 + k}) \overset{(F.25)}{=} -\text{Im}(d_{n/2 - k}) = -\text{Im}(c_{-k}), \quad (F.28b)
\]
i.e. (F.12) is satisfied and, hence, (F.13) applies. We also note
\[
c_{-n/2} = d_0 \overset{(F.25)}{=} a_{n/2}, \quad c_0 = d_{n/2} \overset{(F.25)}{=} a_0, \quad c_{n/2} = 0. \quad (F.28c)
\]
We use the above in (F.13) (with \( n \) replaced by \( n/2 \) and \( j \) replaced by \( k \)) to obtain, for each \( x \in \mathbb{R} \),
\[
r(x) = \sum_{k=0}^{n-1} d_k e^{\frac{(k-n/2)2\pi x}{L}} = \sum_{k=-n/2}^{n/2-1} d_{k+n/2} e^{\frac{ik2\pi x}{L}} = \sum_{k=-n/2}^{n/2} c_k e^{\frac{ik2\pi x}{L}} \overset{(F.13)}{=} c_{-n/2} e^{-\frac{in\pi x}{L}} + c_n e^{\frac{in\pi x}{L}} + 2 \sum_{k=1}^{n/2-1} \text{Re}(c_k) \cos \left( \frac{k2\pi x}{L} \right) - 2 \sum_{k=1}^{n/2-1} \text{Im}(c_k) \sin \left( \frac{k2\pi x}{L} \right)
\]
\[
= a_0 + 2 \sum_{k=1}^{n/2-1} \left( a_k \cos \left( \frac{k2\pi x}{L} \right) + b_k \sin \left( \frac{k2\pi x}{L} \right) \right)
\]
\[
+ a_{n/2} \left( \cos \left( \frac{n\pi x}{L} \right) - i \sin \left( \frac{n\pi x}{L} \right) \right), \quad (F.28)
\]
proving \( t = \text{Re} r \).

(b): (F.26) is an immediate consequence of (a). Furthermore, as a consequence of (F.22) and (F.1b), we have, for each \( k \in \{0, \ldots, n\} \),
\[
d_k = \frac{1}{n} \sum_{j=0}^{n-1} z_j (-1)^j e^{-\frac{ijk2\pi}{n}}. \quad (F.29)
\]
Thus, for each $k \in \{1, \ldots, \frac{n}{2} - 1\}$,

$$a_k = \frac{1}{2} (d_{\frac{n}{2} - k} + d_{\frac{n}{2} + k})$$

(F.25)

$$= \frac{1}{2n} \sum_{j=0}^{n-1} z_j (-1)^j \left( e^{\frac{-ij2\pi}{n}} + e^{\frac{ijk2\pi}{n}} \right)$$

(F.29)

$$= \frac{1}{n} \sum_{j=0}^{n-1} z_j \cos \left( \frac{jk2\pi}{n} \right),$$

(F.7b)

thereby establishing the case. Analogously, for each $k \in \{1, \ldots, \frac{n}{2} - 1\}$,

$$b_k = \frac{1}{2i} (d_{\frac{n}{2} - k} - d_{\frac{n}{2} + k})$$

(F.25)

$$= \frac{1}{2in} \sum_{j=0}^{n-1} z_j (-1)^j \left( e^{\frac{ijk2\pi}{n}} - e^{\frac{-ijk2\pi}{n}} \right)$$

(F.29)

$$= \frac{1}{n} \sum_{j=0}^{n-1} z_j \sin \left( \frac{jk2\pi}{n} \right),$$

(F.7c)

thereby completing the proof of (F.27).

(c): According to (a), we have $t = \text{Re} r$. Since $f$ and $f^{(m)}$ are real-valued, we obtain from Th. F.11(b):

$$\|t - f\|_2^2 = \int_0^L |\text{Re} r - f|^2 \leq \int_0^L (|\text{Re} r - f|^2 + |\text{Im} r|^2) = \int_0^L |r - f|^2$$

(F.23)

$$\leq \left( c_m \left( \|f\|_2 + \|f^{(m)}\|_2 \right) \right)^2,$$

which establishes the case.

Remark F.14. To make use of Th. F.13 to interpolate a real-valued function $f$ on $[0, L]$, one would proceed as follows:

(i) Choose $n \in \mathbb{N}$ even and compute the equidistant $x_j := jL/n \in [0, L]$.

(ii) Compute the corresponding values $z_j := f(x_j)$.

(iii) Compute $d := \mathcal{F}((-1)^0 z_0, \ldots, (-1)^{n-1} z_{n-1})$ according to (F.22).

(iv) Compute the $a_k$ and $b_k$ from (F.25), i.e. from $a_k = \frac{1}{2} (d_{\frac{n}{2} - k} + d_{\frac{n}{2} + k})$ and $b_k = \frac{1}{2i} (d_{\frac{n}{2} - k} - d_{\frac{n}{2} + k})$.

(v) Obtain the real-valued interpolating function $t$ from (F.24).
F.4 Fast Fourier Transform (FFT)

If one computes DFT using matrix multiplication as described in Rem. F.2, then one needs $O(n^2)$ complex multiplications to apply $\mathcal{F}$ to a vector with $n$ components. However, if one is able to choose $n$ to be a power of 2, then one can apply a method called fast Fourier transform (FFT) to reduce the complexity to just $O(n \log_2 n)$ complex multiplications (cf. Rem. F.23 below). We will study the FFT method in the present section.

Remark F.15. (a) FFT can also be used to apply inverse DFT by making use of Prop. F.4.

(b) There are variants of FFT that make use of the prime factorization of $n$ and which can be applied if $n$ is not a power of 2 (see [SK11, p. 161] and the references given there). However, these variants fail to be fast if $n$ is (a large) prime (or, more generally, if $n$ has few large prime factors). A different approach that is also fast for large prime numbers $n$ is to extend the vectors to be transformed to larger vectors of size $m$, where $m$ is a power of 2. One can simply do the extension by so-called zero padding, i.e. by adding components with value zero to the original vector. However, as $\omega$ in (F.3) depends on $n$, the details of this approach are somewhat tricky. In the literature, this is known as Bluestein’s algorithm and it involves convolutions and the so-called chirp z-transform (see [Blu68, RSR69]). We will not study the details of these approaches in this class, i.e. FFT as presented below applies only to the case where $n$ is a power of 2.

---

FFT is based on the following Th. F.17, which allows to compute the DFT of a vector of length $2n$ from the DFT of two vectors of length $n$. In preparation for Th. F.17, we introduce some notation.

**Notation F.16.** If we want to emphasize the dimension on which a DFT operator operates, we write the dimension as a superscript of the operator, denoting DFT from $\mathbb{C}^n$ to $\mathbb{C}^n$, $n \in \mathbb{N}$, by $\mathcal{F}^{(n)} : \mathbb{C}^n \rightarrow \mathbb{C}^n$. For each $k \in \{0, \ldots, n-1\}$, we let $\mathcal{F}_k^{(n)} := \pi_k^{(n)} \circ \mathcal{F}^{(n)} : \mathbb{C}^n \rightarrow \mathbb{C}$, where $\pi_k^{(n)} : \mathbb{C}^n \rightarrow \mathbb{C}$ is the projection $(z_0, \ldots, z_{n-1}) \mapsto z_k$, denote the $k$th coordinate function of $\mathcal{F}^{(n)}$.

**Theorem F.17.** Let $n \in \mathbb{N}$. Then, for each $k \in \{0, \ldots, n-1\}$, it holds that

\[
\forall \quad (f_0, \ldots, f_{2n-1}) \in \mathbb{C}^{2n} \quad \mathcal{F}_k^{(n)}(f_0, f_1, \ldots, f_{n-1}) + e^{-\frac{ik\pi}{n}} \mathcal{F}_k^{(n)}(f_n, f_{n+1}, \ldots, f_{2n-1}) = \mathcal{F}_k^{(2n)}(f_0, f_n, f_1, f_{n+1}, \ldots, f_{n-1}, f_{2n-1}) \tag{F.30a}
\]

and

\[
\forall \quad (f_0, \ldots, f_{2n-1}) \in \mathbb{C}^{2n} \quad \mathcal{F}_k^{(n)}(f_0, f_1, \ldots, f_{n-1}) - e^{-\frac{ik\pi}{n}} \mathcal{F}_k^{(n)}(f_n, f_{n+1}, \ldots, f_{2n-1}) = \mathcal{F}_{n+k}^{(2n)}(f_0, f_n, f_1, f_{n+1}, \ldots, f_{n-1}, f_{2n-1}) \tag{F.30b}
\]
Proof. Given \((f_0, \ldots, f_{2n-1}) \in \mathbb{C}^{2n}\), we compute

\[
\mathcal{F}_k^{(2n)}(f_0, f_n, f_1, f_{n+1}, \ldots, f_{n-1}, f_{2n-1})
\]

\[
= \frac{1}{2n} \left( \sum_{j=0}^{n-1} f_j e^{-i(2j+1)\pi 2n} + \sum_{j=0}^{n-1} f_{n+j} e^{-i(2j+1)\pi 2n/n} \right)
\]

\[
= \frac{1}{2n} \left( \sum_{j=0}^{n-1} f_j e^{-i(2j+1)\pi 2n/n} + \sum_{j=0}^{n-1} f_{n+j} e^{-i(2j+1)\pi 2n/n} \right)
\]

\[
= \mathcal{F}_k^{(n)}(f_0, f_1, \ldots, f_{n-1}) + e^{-i\pi n/n} \mathcal{F}_k^{(n)}(f_n, f_{n+1}, \ldots, f_{2n-1}),
\]

proving \((F.30a)\), and

\[
\mathcal{F}_{k+n}^{(2n)}(f_0, f_n, f_1, f_{n+1}, \ldots, f_{n-1}, f_{2n-1})
\]

\[
= \frac{1}{2n} \left( \sum_{j=0}^{n-1} f_j e^{-i(2j+1)(n+k)\pi 2n} + \sum_{j=0}^{n-1} f_{n+j} e^{-i(2j+1)(n+k)\pi 2n/n} \right)
\]

\[
= \frac{1}{2n} \left( \sum_{j=0}^{n-1} f_j e^{-i(2j+1)(n+k)\pi 2n/n} + \sum_{j=0}^{n-1} f_{n+j} e^{-i(2j+1)(n+k)\pi 2n/n} \right)
\]

\[
= \mathcal{F}_k^{(n)}(f_0, f_1, \ldots, f_{n-1}) - e^{-i\pi n/n} \mathcal{F}_k^{(n)}(f_n, f_{n+1}, \ldots, f_{2n-1}),
\]

proving \((F.30b)\).

\[\blacksquare\]

Example F.18. (a) If \(n = 2^q, q \in \mathbb{N}\), then one can apply Th. F.17 \(q\) times to reduce an application of \(\mathcal{F}^{(n)}\) to \(n\) applications of \(\mathcal{F}^{(1)}\). Moreover, we observe that \(\mathcal{F}^{(1)}\) is merely the identity on \(\mathbb{C}\): According to \((F.1)\):

\[
\forall \ f_0 \in \mathbb{C} \quad \mathcal{F}^{(1)}(f_0) = \frac{1}{2} f_0 e^0 = f_0.
\]

(b) For \(n = 8 = 2^3\), one can build the result of an application of \(\mathcal{F}^{(n)}\) from trivial
one-dimensional transforms in 3 steps as illustrated in the scheme in (F.31) below.

\[
\begin{array}{cccccccc}
  f_0 & f_4 & f_2 & f_6 & f_1 & f_5 & f_3 & f_7 \\
  \mathcal{F}^{(1)}(f_0) & \mathcal{F}^{(1)}(f_4) & \mathcal{F}^{(1)}(f_2) & \mathcal{F}^{(1)}(f_6) & \mathcal{F}^{(1)}(f_1) & \mathcal{F}^{(1)}(f_5) & \mathcal{F}^{(1)}(f_3) & \mathcal{F}^{(1)}(f_7) \\
  \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
  \mathcal{F}^{(2)}(f_0, f_4) & \mathcal{F}^{(2)}(f_2, f_6) & \mathcal{F}^{(2)}(f_1, f_5) & \mathcal{F}^{(2)}(f_3, f_7) \\
  \downarrow & \downarrow & \downarrow & \downarrow \\
  \mathcal{F}^{(4)}(f_0, f_2, f_4, f_6) & \mathcal{F}^{(4)}(f_1, f_3, f_5, f_7) \\
  \downarrow & \downarrow \\
  \mathcal{F}^{(8)}(f_0, f_1, f_2, f_3, f_4, f_5, f_6, f_7) & & & & & & & & & & & \text{(F.31)}
\end{array}
\]

The scheme in (F.31) also illustrates that getting all the indices right needs some careful accounting. It turns out that this can be done quite elegantly and efficiently by a trick known as bit reversal. Before studying bit reversal and its application to FFT systematically, let us see how it works in the present example: In (F.31), we want to compute \( \mathcal{F}^{(8)}(f_0, \ldots, f_7) \) and need to figure out the order of indices we need in the first row of (F.31). Using the bit reversal algorithm we can do that directly, without having to compute (and store) the intermediate lines of the scheme: We start by writing the numbers 0, \ldots, 7 in their binary representation instead of in their usual decimal representation:

\[
\begin{align*}
(0)_{10} &= (000)_2, & (1)_{10} &= (001)_2, & (2)_{10} &= (010)_2, & (3)_{10} &= (011)_2, \\
(4)_{10} &= (100)_2, & (5)_{10} &= (101)_2, & (6)_{10} &= (110)_2, & (7)_{10} &= (111)_2. & \text{(F.32)}
\end{align*}
\]

Then we reverse the bits of each number in (F.32) (and we also provide the resulting decimal representation):

\[
\begin{align*}
(0)_{10} &= (000)_2, & (4)_{10} &= (100)_2, & (2)_{10} &= (010)_2, & (6)_{10} &= (110)_2, \\
(1)_{10} &= (001)_2, & (5)_{10} &= (101)_2, & (3)_{10} &= (011)_2, & (7)_{10} &= (111)_2. & \text{(F.33)}
\end{align*}
\]

We note that, as if by magic, we have obtained the correct order needed in the first row of (F.31). In the following, we will prove that this procedure works in general.

**Definition F.19.** For each \( q \in \mathbb{N}_0 \), we denote \( \mathcal{M}_q := \mathbb{Z}_{2^q} = \{0, \ldots, 2^q - 1\} \). As every \( n \in \mathcal{M}_q \) has a (unique) binary representation \( n = \sum_{j=0}^{q-1} b_j 2^j \), \( b_j \in \{0, 1\} \), we can define bit reversal (in \( q \) dimensions) as the map \( \rho_q : \mathcal{M}_q \rightarrow \mathcal{M}_q, \rho_q \left( \sum_{j=0}^{q-1} b_j 2^j \right) := \sum_{j=0}^{q-1} b_{q-1-j} 2^j \). \( \text{(F.34)} \)

**Lemma F.20.** Let \( q \in \mathbb{N}_0 \).

(a) The bit reversal map \( \rho_q : \mathcal{M}_q \rightarrow \mathcal{M}_q \) is bijective with \( \rho_q^{-1} = \rho_q \).
(b) One has
\[ \forall_{n \in \mathcal{M}_q} \rho_q(n) = \rho_{q+1}(2n). \]

(c) One has
\[ \forall_{n \in \mathcal{M}_q} 2^q + \rho_q(n) = \rho_{q+1}(2n + 1). \]

Proof. (a): Since, for each \( j \in \{0, \ldots, q - 1\} \), one has \( q - 1 - (q - 1 - j) = j \), this is immediate from (F.34).

(b): If \( n \in \mathcal{M}_q \), then \( 0 \leq n \leq 2^q - 1 \), i.e. \( 0 \leq 2n \leq 2^{q+1} - 2 < 2^{q+1} - 1 \) and \( 2n \in \mathcal{M}_{q+1} \).

Moreover, for \( n = \sum_{j=0}^{q-1} b_j 2^j \in \mathcal{M}_q \), \( b_j \in \{0, 1\} \), we compute
\[
\rho_{q+1}(2n) = \rho_{q+1} \left( 2 \sum_{j=0}^{q-1} b_j 2^j \right) = \rho_{q+1} \left( 0 \cdot 2^0 + \sum_{j=1}^{q-1} b_{j-1} 2^j + b_{q-1} 2^q \right)
\]
\[ = \sum_{j=0}^{q-1} b_{q-1-j} 2^j = \rho_q(n), \]
proving (b).

(c): If \( n \in \mathcal{M}_q \), then \( 0 \leq n \leq 2^q - 1 \), i.e. \( 0 \leq 2n + 1 \leq 2^{q+1} - 1 \) and \( 2n + 1 \in \mathcal{M}_{q+1} \).

Moreover, for \( n = \sum_{j=0}^{q-1} b_j 2^j \in \mathcal{M}_q \), \( b_j \in \{0, 1\} \), we compute
\[
\rho_{q+1}(2n + 1) = \rho_{q+1} \left( 1 + 2 \sum_{j=0}^{q-1} b_j 2^j \right) = \rho_{q+1} \left( 1 \cdot 2^0 + \sum_{j=1}^{q-1} b_{j-1} 2^j + b_{q-1} 2^q \right)
\]
\[ = \sum_{j=0}^{q-1} b_{q-1-j} 2^j + 1 \cdot 2^q = 2^q + \rho_q(n), \]
proving (c). \( \square \)

Definition F.21. We define the fast Fourier transform (FFT) algorithm for the computation of the \( n \)-dimensional discrete Fourier transform \( F^{(n)} \) in the case \( n = 2^q \), \( q \in \mathbb{N}_0 \), by the following recursion: We start with the \( n = 2^q \) complex numbers \( z_0, \ldots, z_{n-1} \in \mathbb{C} \) (we will see below (cf. Th. F.22) that, to compute \( d := F^{(n)}(f) \), \( f \in \mathbb{C}^n \), one has to choose \( z_0 := f_{\rho_q(0)}, z_1 := f_{\rho_q(1)}, \ldots, z_{n-1} := f_{\rho_q(n-1)} \):)
\[ d^{(0,0)} := z_0, \ldots, d^{(0,2^q-1)} := z_{2^q-1} = z_{n-1}. \]  \hspace{1cm} (F.35a)
Then, for each \( 1 \leq r \leq q \), one defines recursively the following \( 2^{q-r} \) vectors \( d^{(r,0)}, \ldots, d^{(r,2^{q-r}-1)} \in \mathbb{C}^{2^r} \) by
\[
d^{(r,j)} := d^{(r-1,2j)} + e^{-\frac{2\pi i}{2^r} j} d^{(r-1,2j+1)} - \frac{2}{2^r} d^{(r-1,2j+1)}, \hspace{1cm} (F.35b)
\]
\[
d^{(r,j)} := d^{(r,2j)} - e^{-\frac{2\pi i}{2^r} j} d^{(r,2j+1)}, \hspace{1cm} (F.35b)
\]
(note that the recursion ends after \( q \) steps with the single vector \( d^{(q,0)} \in \mathbb{C}^n \)).
Theorem F.22. Let $n = 2^q$, $q \in \mathbb{N}_0$, and $z = (z_0, \ldots, z_{n-1}) \in \mathbb{C}^n$. If, for $r \in \{0, \ldots, q\}$, $j \in \{0, \ldots, 2^{q-r} - 1\}$, the $d^{(r,j)}$ are defined by the FFT algorithm of Def. F.21, then

$$\forall_{r \in \{0, \ldots, q\}} \forall_{j \in \{0, \ldots, 2^{q-r} - 1\}} d^{(r,j)} = \mathcal{F}^{(2^r)}(z_{j2^r + \rho_r(0)}, z_{j2^r + \rho_r(1)}, \ldots, z_{j2^r + \rho_r(2^{q-r} - 1)}).
$$

(F.36)

In particular,

$$d^{(q,0)} = \mathcal{F}^{(n)}(z_{\rho_q(0)}, z_{\rho_q(1)}, \ldots, z_{\rho_q(2^q - 1)}),$$

(F.37)

i.e., to compute $\mathcal{F}^{(n)}(f_0, \ldots, f_{n-1})$ for a given $f \in \mathbb{C}^n$, one has to set

$$z := (f_{\rho_q(0)}, \ldots, f_{\rho_q(n-1)})$$

(F.38)

to obtain

$$d^{(q,0)} = \mathcal{F}^{(n)}(f_0, \ldots, f_{n-1}).$$

(F.39)

Proof. The proof of (F.36) is conducted via induction with respect to $r \in \{0, \ldots, q\}$. For $r = 0$, (F.36) reads

$$\forall_{j \in \{0, \ldots, 2^{q-r} - 1\}} d^{(0,j)} = \mathcal{F}^{(1)}(z_j),$$

(F.40)

which is correct due to (F.35a) and Example F.18(a). If $r \in \{1, \ldots, q\}$ and $j \in \{0, \ldots, 2^{q-r} - 1\}$, then, for each $k \in \{0, \ldots, 2^{r-1} - 1\}$, we calculate

$$d^{(r,j)} = \frac{1}{2} \left( d^{(r-1,2j)} + e^{-\frac{i\pi\alpha}{2^{r-1}}} d^{(r-1,2j+1)} \right)$$

(F.35b)

ind.hyp.

$$= \frac{1}{2} \mathcal{F}_k^{(2^{r-1})} (z_{2j2^{r-1} + \rho_{r-1}(0)}, z_{2j2^{r-1} + \rho_{r-1}(1)}, \ldots, z_{2j2^{r-1} + \rho_{r-1}(2^{q-r} - 1)} + \frac{1}{2} e^{-\frac{i\pi\alpha}{2^{r-1}}} \mathcal{F}_k^{(2^{r-1})} (z_{2(j+1)2^{r-1} + \rho_{r-1}(0)}, \ldots, z_{(2j+1)2^{r-1} + \rho_{r-1}(2^{q-r} - 1)})$$

$$= \frac{1}{2} \mathcal{F}_k^{(2^{r-1})} (z_{2j2^r + \rho_r(0)}, \ldots, z_{j2^r + \rho_r(2^{q-r} - 1)}) + \frac{1}{2} e^{-\frac{i\pi\alpha}{2^{r-1}}} \mathcal{F}_k^{(2^{r-1})} (z_{j2^r + 2^r + \rho_r(0)}, \ldots, z_{j2^r + 2^r + \rho_r(2^{q-r} - 1)})

(F.41)

where the equality at (*) holds due to (F.30a) and Lem. F.20(b),(c), since, for each $\alpha \in \{0, \ldots, 2^{r-1} - 1\}$,

$$j2^r + \rho_{r-1}(\alpha) = j2^r + \rho_r(2\alpha)$$

and

$$j2^r + 2^r + \rho_{r-1}(\alpha) = j2^r + \rho_r(2\alpha + 1).$$

(F.42)

Similarly,

$$d^{(r,j)}_{2^{r-1}+k} = \frac{1}{2} \left( d^{(r-1,2j)} - e^{-\frac{i\pi\alpha}{2^{r-1}}} d^{(r-1,2j+1)} \right)$$

(F.35b)

ind.hyp.

$$= \frac{1}{2} \mathcal{F}_k^{(2^{r-1})} (z_{j2^r + \rho_r(0)}, \ldots, z_{j2^r + \rho_r(2^{q-r} - 1)})$$

$$- \frac{1}{2} e^{-\frac{i\pi\alpha}{2^{r-1}}} \mathcal{F}_k^{(2^{r-1})} (z_{j2^r + 2^r + \rho_r(0)}, \ldots, z_{j2^r + 2^r + \rho_r(2^{q-r} - 1)})

(F.30a),(F.42)

$$= \mathcal{F}_k^{(2^r)} (z_{j2^r + \rho_r(0)}, z_{j2^r + \rho_r(1)}, \ldots, z_{j2^r + \rho_r(2^{q-r} - 1)}),$$

(F.43)
completing the induction proof of (F.36).

Cleary, (F.36) turns into (F.37) for $r = q, j = 0$; and (F.39) follows from (F.37) and (F.38) as $\rho_q \circ \rho_q = \text{Id}$.

\[\square\]

Remark F.23. As mentioned at the beginning of the section, FFT reduces the number of complex multiplications needed to compute an $n$-dimensional DFT from $O(n^2)$ (when done by matrix multiplication) to $O(n \log_2 n)$ (an actually quite significant improvement – even for a moderate $n$ like $n = 8 = 2^3$, one has $n^2 = 64$ versus $n \log_2 n = 24$, i.e. one gains almost a factor 3; the larger $n$ gets the more essential it becomes to use FFT).

Let us check the $O(n \log_2 n)$ claim for FFT as computed by (F.35) for $n = 2^q$, $q \in \mathbb{N}$, (where we need to estimate the number of multiplications involved in the executions of (F.35b)):

(i) In a preparatory step, one computes the $q - 1$ factors $e^{-\frac{2 \pi i}{2^q}}$, $\ldots$, $e^{-\frac{2 \pi i}{2^{q-1}}}$ by starting with $e^{-\frac{2 \pi i}{2^q}}$ and using

\[\forall_{r \in \{0, \ldots, q-2\}} e^{-\frac{2 \pi i}{2^r}} = \left(e^{-\frac{2 \pi i}{2^{r+1}}}ight)^2.\]

Thus, the number of multiplications used in this step is

\[N_1 := q - 2 < q.\]

It is used that $e^{-\frac{2 \pi i}{2^0}} = -1$ and $e^{-\frac{2 \pi i}{2^{q-1}}}$ is assumed to be given (it has to be precomputed once, which adds a constant number of multiplications, depending on the desired accuracy).

Then (F.35b) has to be executed $q$ times, namely once for each $r = 1, \ldots, q$. We compute the number of multiplications needed to obtain the $d^{(r,j)}$:

(ii) Starting from the precomputed $\theta_r := e^{-\frac{2 \pi i}{2^r}}$, one uses successive multiplications with $\theta_r$ to obtain the factors $\theta_r^k = e^{-\frac{2 \pi i}{2^{r-1}}} k, k = 1, \ldots, 2^{r-1} - 1$. Thus, the number of multiplications for this step is

\[N_{2,r} := 2^{r-1} - 2 < 2^{r-1}.\]

(iii) The computation of each $d^{(r,j)}$ needs two multiplications (one by the $e$-factor and one by $\frac{1}{2}$). As the range for $k$ is $\{0, \ldots, 2^{r-1} - 1\}$ and the range for $j$ is $\{0, \ldots, 2^{q-r} - 1\}$, the total number of multiplications for this step is

\[N_{3,r} := 2 \cdot 2^{r} \cdot 2^{q-r} = 2^{q+1}.\]

Finally, summing all multiplications from steps (i) – (iii), one obtains

\[N = N_1 + \sum_{r=1}^{q} \left(N_{2,r} + N_{3,r}\right) < q + \sum_{r=1}^{q} \left(2^{r-1} + 2^{q+1}\right) = q + 2^q - 1 + q \cdot 2^{q+1}\]

\[< \log_2 n + n + 2n \log_2 n,\]

which is $O(n \log_2 n)$ as claimed.
G The Weierstrass Approximation Theorem

An important result in the context of the approximation of continuous functions by polynomials, which we will need for subsequent use, is the following Weierstrass approximation Th. G.1. Even though related, this topic is not strictly part of the theory of interpolation.

**Theorem G.1** (Weierstrass Approximation Theorem). Let $a, b \in \mathbb{R}$ with $a < b$. For each continuous function $f \in C[a, b]$ and each $\epsilon > 0$, there exists a polynomial $p : \mathbb{R} \rightarrow \mathbb{R}$ such that $\| f - p \|_{[a, b]} < \epsilon$, where $p|_{[a, b]}$ denotes the restriction of $p$ to $[a, b]$.

Theorem G.1 will be a corollary of the fact that the Bernstein polynomials corresponding to $f \in C[0, 1]$ (see Def. G.2) converge uniformly to $f$ on $[0, 1]$ (see Th. G.3 below).

**Definition G.2.** Given $f : [0, 1] \rightarrow \mathbb{R}$, define the Bernstein polynomials $B_n f$ corresponding to $f$ by

$$B_n f : \mathbb{R} \rightarrow \mathbb{R}, \quad (B_n f)(x) := \sum_{\nu=0}^{n} f \left( \frac{\nu}{n} \right) \binom{n}{\nu} x^\nu (1 - x)^{n-\nu} \text{ for each } n \in \mathbb{N}. \quad \text{(G.1)}$$

**Theorem G.3.** For each $f \in C[0, 1]$, the sequence of Bernstein polynomials $(B_n f)_{n \in \mathbb{N}}$ corresponding to $f$ according to Def. G.2 converges uniformly to $f$ on $[0, 1]$, i.e.

$$\lim_{n \rightarrow \infty} \| f - (B_n f) |_{[0,1]} \|_{\infty} = 0. \quad \text{(G.2)}$$

**Proof.** We begin by noting

$$(B_n f)(0) = f(0) \quad \text{and} \quad (B_n f)(1) = f(1) \quad \text{for each } n \in \mathbb{N}. \quad \text{(G.3)}$$

For each $n \in \mathbb{N}$ and $\nu \in \{0, \ldots, n\}$, we introduce the abbreviation

$$q_{n\nu}(x) := \binom{n}{\nu} x^\nu (1 - x)^{n-\nu}. \quad \text{(G.4)}$$

Then

$$1 = \left( x + (1 - x) \right)^n = \sum_{\nu=0}^{n} q_{n\nu}(x) \quad \text{for each } n \in \mathbb{N} \quad \text{(G.5)}$$

implies

$$f(x) - (B_n f)(x) = \sum_{\nu=0}^{n} \left( f(x) - f \left( \frac{\nu}{n} \right) \right) q_{n\nu}(x) \quad \text{for each } x \in [0, 1], \ n \in \mathbb{N},$$

and

$$|f(x) - (B_n f)(x)| \leq \sum_{\nu=0}^{n} \left| f(x) - f \left( \frac{\nu}{n} \right) \right| q_{n\nu}(x) \quad \text{for each } x \in [0, 1], \ n \in \mathbb{N}. \quad \text{(G.6)}$$
As \( f \) is continuous on the compact interval \([0, 1]\), it is uniformly continuous, i.e. for each \( \epsilon > 0 \), there exists \( \delta > 0 \) such that, for each \( x \in [0, 1] \), \( n \in \mathbb{N} \), \( \nu \in \{0, \ldots, n\} \):

\[
|x - \frac{\nu}{n}| < \delta \quad \Rightarrow \quad |f(x) - f\left(\frac{\nu}{n}\right)| < \frac{\epsilon}{2}.
\]

(G.7)

For the moment, we fix \( x \in [0, 1] \) and \( n \in \mathbb{N} \) and define

\[
N_1 := \left\{ \nu \in \{0, \ldots, n\} : \left|x - \frac{\nu}{n}\right| < \delta \right\},
\]

\[
N_2 := \left\{ \nu \in \{0, \ldots, n\} : \left|x - \frac{\nu}{n}\right| \geq \delta \right\}.
\]

Then

\[
\sum_{\nu \in N_1} \left| f(x) - f\left(\frac{\nu}{n}\right) \right| q_\nu(x) \leq \frac{\epsilon}{2} \sum_{\nu \in N_1} q_\nu(x) \leq \frac{\epsilon}{2} \sum_{\nu = 0}^{n} q_\nu(x) = \frac{\epsilon}{2},
\]

(G.8)

and with \( M := \|f\|_{\infty} \),

\[
\sum_{\nu \in N_2} \left| f(x) - f\left(\frac{\nu}{n}\right) \right| q_\nu(x) \leq \sum_{\nu \in N_2} \left| f(x) - f\left(\frac{\nu}{n}\right) \right| q_\nu(x) \frac{(x - \frac{\nu}{n})^2}{\delta^2} \leq \frac{2M^2}{\delta^2} \sum_{\nu = 0}^{n} q_\nu(x) \left(x - \frac{\nu}{n}\right)^2.
\]

(G.9)

To compute the sum on the right-hand side of (G.9), observe

\[
\left(x - \frac{\nu}{n}\right)^2 = x^2 - 2x \frac{\nu}{n} + \left(\frac{\nu}{n}\right)^2
\]

(G.10)

and

\[
\sum_{\nu = 0}^{n} \binom{n}{\nu} x^\nu (1 - x)^{n - \nu} \frac{\nu}{n} = x \sum_{\nu = 1}^{n} \binom{n - 1}{\nu - 1} x^{\nu - 1} (1 - x)^{(n - 1) - (\nu - 1)} = x
\]

(G.11)

as well as

\[
\sum_{\nu = 0}^{n} \binom{n}{\nu} x^\nu (1 - x)^{n - \nu} \frac{\nu^2}{n} = \frac{x^2}{n} \sum_{\nu = 1}^{n} (\nu - 1) \binom{n - 1}{\nu - 1} x^{\nu - 1} (1 - x)^{(n - 1) - (\nu - 1)} + \frac{x}{n}
\]

\[
= \frac{x^2}{n} (n - 1) \sum_{\nu = 2}^{n} (\nu - 2) x^{\nu - 2} (1 - x)^{(n - 2) - (\nu - 2)} + \frac{x}{n}
\]

\[
= x^2 \left(1 - \frac{1}{n}\right) + \frac{x}{n} = x^2 + \frac{x}{n} (1 - x).
\]

(G.12)

Thus, we obtain

\[
\sum_{\nu = 0}^{n} q_\nu(x) \left(x - \frac{\nu}{n}\right)^2 \leq \frac{1}{4n} \quad \text{for each } x \in [0, 1], \ n \in \mathbb{N},
\]

\[
\sum_{\nu = 0}^{n} q_\nu(x) \left(x - \frac{\nu}{n}\right)^2 \leq \frac{1}{4n} \quad \text{for each } x \in [0, 1], \ n \in \mathbb{N},
\]
and together with (G.9):
\[ \sum_{\nu \in \mathbb{N}_2} \left| f(x) - f\left(\frac{\nu}{n}\right) \right| q_{n\nu}(x) \leq \frac{2M}{\delta^2} \frac{1}{4n} < \frac{\epsilon}{2} \quad \text{for each } x \in [0,1], \ n > \frac{M}{\delta^2 \epsilon}. \] (G.13)

Combining (G.6), (G.8), and (G.13) yields
\[ \left| f(x) - (B_n f)(x) \right| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \quad \text{for each } x \in [0,1], \ n > \frac{M}{\delta^2 \epsilon}, \]
proving the claimed uniform convergence.

Proof of Th. G.1. Define
\[ \phi : [a,b] \rightarrow [0,1], \ \phi(x) := \frac{x - a}{b - a}, \]
\[ \phi^{-1} : [0,1] \rightarrow [a,b], \ \phi^{-1}(x) := (b - a)x + a. \]
Given \( \epsilon > 0 \), and letting
\[ g : [0,1] \rightarrow \mathbb{R}, \ g(x) := f(\phi^{-1}(x)). \]
Th. G.3 provides a polynomial \( q : \mathbb{R} \rightarrow \mathbb{R} \) such that \( \| g - q \|_{[0,1]} \|_{\infty} < \epsilon. \) Defining
\[ p : \mathbb{R} \rightarrow \mathbb{R}, \ p(x) := q(\phi(x)) = q\left( \frac{x - a}{b - a} \right) \]
(having extended \( \phi \) in the obvious way) yields a polynomial \( p \) such that
\[ |f(x) - p(x)| = |g(\phi(x)) - q(\phi(x))| < \epsilon \quad \text{for each } x \in [a,b] \]
as needed.

\section{H Baire Category and the Uniform Boundedness Principle}

\textbf{Theorem H.1} (Baire Category Theorem). Let \( \emptyset \neq X \) be a nonempty complete metric space, and, for each \( k \in \mathbb{N} \), let \( A_k \) be a closed subset of \( X \). If
\[ X = \bigcup_{k=1}^{\infty} A_k, \] (H.1)
then there exists \( k_0 \in \mathbb{N} \) such that \( A_{k_0} \) has nonempty interior: \( \text{int} \ A_{k_0} \neq \emptyset. \)
Proof. Seeking a contradiction, we assume (H.1) holds with closed sets $A_k$ such that $\text{int } A_k = \emptyset$ for each $k \in \mathbb{N}$. Then, for each nonempty open set $O \subseteq X$ and each $k \in \mathbb{N}$, the set $O \setminus A_k = O \cap (X \setminus A_k)$ is open and nonempty (as $O \setminus A_k = \emptyset$ implied $O \subseteq A_k$, such that the points in $O$ were interior points of $A_k$). Since $O \setminus A_k$ is open and nonempty, we can choose $x \in O \setminus A_k$ and $0 < \epsilon < \frac{1}{k}$ such that

$$\overline{B}_\epsilon(x) \subseteq O \setminus A_k.$$  \hspace{1cm} (H.2)

As one can choose $B_\epsilon(x)$ as a new $O$, starting with arbitrary $x_0 \in X$ and $\epsilon_0 > 0$, one can inductively construct sequences of points $x_k \in X$, numbers $\epsilon_k > 0$, and corresponding balls such that

$$\overline{B}_{\epsilon_k}(x_k) \subseteq B_{\epsilon_{k-1}}(x_{k-1}) \setminus A_k, \quad \epsilon_k \leq \frac{1}{k} \quad \text{for each } k \in \mathbb{N}. \hspace{1cm} (H.3)$$

Then $l \geq k$ implies $x_l \in B_{\epsilon_k}(x_k)$ for each $k, l \in \mathbb{N}$, such that $(x_k)_{k \in \mathbb{N}}$ constitutes a Cauchy sequence due to $\lim_{k \to \infty} \epsilon_k = 0$. The assumed completeness of $X$ provides a limit $x = \lim_{k \to \infty} x_k \in X$. The nested form (H.3) of the balls $B_{\epsilon_k}(x_k)$ implies $x \in \overline{B}_{\epsilon_k}(x_k)$ for each $k \in \mathbb{N}$ on the one hand and $x \notin A_k$ for each $k \in \mathbb{N}$ on the other hand. However, this last conclusion is in contradiction to (H.1). \hfill \blacksquare

Remark H.2. The purpose of this remark is to explain the name Baire Category Theorem in Th. H.1. To that end, we need to introduce some terminology that will not be used again outside the present remark.

Let $X$ be a metric space.

Recall that a subset $D$ of $X$ is called dense if, and only if, $\overline{D} = X$, or, equivalently, if, and only if, for each $x \in X$ and each $\epsilon > 0$, one has $D \cap B_\epsilon(x) \neq \emptyset$. A subset $E$ of $X$ is called nowhere dense if, and only if, $X \setminus \overline{E}$ is dense.

A subset $M$ of $X$ is said to be of first category or meager if, and only if, $M = \bigcup_{k=1}^{\infty} E_k$ is a countable union of nowhere dense sets $E_k$, $k \in \mathbb{N}$. A subset $S$ of $X$ is said to be of second category or nonmeager or fat if, and only if, $S$ is not of first category. Caveat: These notions of category are completely different from the notion of category occurring in the more algebraic discipline called category theory.

Finally, the reason Th. H.1 carries the name Baire Category Theorem lies in the fact that it implies that every nonempty complete metric space $X$ is of second category (considered as a subset of itself): If $X$ were of first category, then $X = \bigcup_{k=1}^{\infty} E_k$ with nowhere dense sets $E_k$. As the $E_k$ are nowhere dense, the complements $X \setminus \overline{E_k}$ are dense, i.e. $\overline{E_k}$ has empty interior, and $X = \bigcup_{k=1}^{\infty} \overline{E_k}$ were a countable union of closed sets with empty interior in contradiction to Th. H.1.

Theorem H.3 (Uniform Boundedness Principle). Let $X$ be a nonempty complete metric space and let $Y$ be a normed vector space. Consider a set of continuous functions $\mathcal{F} \subseteq C(X,Y)$. If

$$M_x := \sup \{ \| f(x) \| : f \in \mathcal{F} \} < \infty \quad \text{for each } x \in X, \hspace{1cm} (H.4)$$
then there exists $x_0 \in X$ and $\epsilon_0 > 0$ such that
\[
\sup \{ M_x : x \in \overline{B}_{\epsilon_0}(x_0) \} < \infty. \tag{H.5}
\]
In other words, if a collection of continuous functions from $X$ into $Y$ is bounded pointwise in $X$, then it is uniformly bounded on an entire ball.

**Proof.** If $\mathcal{F} = \emptyset$, then $M_x = \sup \emptyset := \inf \mathbb{R}_0^+ = 0$ for each $x \in X$ and there is nothing to prove. Thus, let $\mathcal{F} \neq \emptyset$. For $f \in \mathcal{F}$, $k \in \mathbb{N}$, as both $f$ and the norm are continuous, the set
\[
A_{k,f} := \{ x \in X : \| f(x) \| \leq k \} \tag{H.6}
\]
is an continuous inverse image of the closed set $[0, k]$ and, hence, closed. Since arbitrary intersections of closed sets are closed, so is
\[
A_k := \bigcap_{f \in \mathcal{F}} A_{k,f} = \{ x \in X : \| f(x) \| \leq k \text{ for each } f \in \mathcal{F} \}. \tag{H.7}
\]
Now (H.4) implies
\[
X = \bigcup_{k=1}^{\infty} A_k \tag{H.8}
\]
and the Baire Category Th. H.1 provides $k_0 \in \mathbb{N}$ such that $A_{k_0}$ has nonempty interior. Thus, there exist $x_0 \in A_{k_0}$ and $\epsilon_0 > 0$ such that $\overline{B}_{\epsilon_0}(x_0) \subseteq A_{k_0}$, and, in consequence,
\[
\sup \{ M_x : x \in \overline{B}_{\epsilon_0}(x_0) \} \leq k_0, \tag{H.9}
\]
proving (H.5).

We now specialize the uniform boundedness principle to bounded linear maps:

**Theorem H.4** (Banach-Steinhaus Theorem). *Let $X$ be a Banach space (i.e. a complete normed vector space) and let $Y$ be a normed vector space. Consider a set of bounded linear functions $\mathcal{T} \subseteq \mathcal{L}(X, Y)$. If
\[
\sup \{ \| T(x) \| : T \in \mathcal{T} \} < \infty \text{ for each } x \in X, \tag{H.10}
\]
then $\mathcal{T}$ is a bounded subset of $\mathcal{L}(X, Y)$, i.e.
\[
\sup \{ \| T \| : T \in \mathcal{T} \} < \infty. \tag{H.11}
\]

**Proof.** To apply Th. H.3, define, for each $T \in \mathcal{T}$:
\[
f_T : X \rightarrow \mathbb{R}, \quad f_T(x) := \| Tx \|. \tag{H.12}
\]
Then $\mathcal{F} := \{ f_T : T \in \mathcal{T} \} \subseteq C(X, \mathbb{R})$ and (H.10) implies that $\mathcal{F}$ satisfies (H.4). Thus, if $M_x$ is as in (H.4), we obtain $x_0 \in X$ and $\epsilon_0 > 0$ such that
\[
\sup \{ M_x : x \in \overline{B}_{\epsilon_0}(x_0) \} < \infty, \tag{H.13}
\]
i.e. there exists \( C \in \mathbb{R}^+ \) satisfying
\[
\|Tx\| \leq C \text{ for each } T \in T \text{ and each } x \in X \text{ with } \|x - x_0\| \leq \epsilon_0. \tag{H.14}
\]
In consequence, for each \( x \in X \setminus \{0\} \),
\[
\|Tx\| = \frac{\|x\|}{\epsilon_0} \cdot \left\| T \left( x_0 + \epsilon_0 \frac{x}{\|x\|} \right) - Tx_0 \right\| \leq \frac{\|x\|}{\epsilon_0} \cdot 2C, \tag{H.15}
\]
showing \( \|T\| \leq \frac{2C}{\epsilon_0} \) for each \( T \in T \), thereby establishing the case. \( \blacksquare \)

I Matrix Decomposition

I.1 Cholesky Decomposition of Positive Semidefinite Matrices

In this section, we show that the results of Sec. 5.3 (except uniqueness of the decomposition) extend to Hermitian positive semidefinite matrices.

**Theorem I.1.** Let \( \Sigma \in \mathcal{M}(n, \mathbb{K}) \) be Hermitian positive semidefinite, \( n \in \mathbb{N} \). Then there exists a lower triangular \( A \in \mathcal{M}(n, \mathbb{K}) \) with nonnegative diagonal entries (i.e. with \( a_{jj} \in \mathbb{R}_0^+ \) for each \( j \in \{1, \ldots, n\} \)) and
\[
\Sigma = AA^*. \tag{I.1}
\]
In extension of the term for the positive definite case, we still call this decomposition a Cholesky decomposition of \( \Sigma \). It is recalled from Th. 5.7 that, if \( \Sigma \) is positive definite, then the diagonal entries of \( A \) can be chosen to be all positive and this determines \( A \) uniquely.

**Proof.** The positive definite case was proved in Th. 5.7. The general (positive semidefinite) case can be deduced from the positive definite case as follows: If \( \Sigma \in \mathcal{M}(n, \mathbb{K}) \) is Hermitian positive semidefinite, then each \( \Sigma_k := \Sigma + \frac{1}{k} \text{Id} \in \mathcal{M}(n, \mathbb{K}), k \in \mathbb{N}, \) is positive definite:
\[
x^* \Sigma_k x = x^* \Sigma x + \frac{x^* x}{k} > 0 \quad \text{for each } 0 \neq x \in \mathbb{K}^n. \tag{I.2}
\]
Thus, for each \( k \in \mathbb{N} \), there exists a lower triangular matrix \( A_k \mathcal{M}(n, \mathbb{K}) \) with positive diagonal entries such that \( \Sigma_k = A_k A_k^* \).

On the other hand, \( \Sigma_k \) converges to \( \Sigma \) with respect to each norm on \( \mathbb{K}^n \) (since all norms on \( \mathbb{K}^n \) are equivalent). In particular, \( \lim_{k \to \infty} \|\Sigma_k - \Sigma\|_2 = 0 \). Thus,
\[
\|A_k\|_2^2 = r(\Sigma_k) = \|\Sigma_k\|_2, \tag{I.3}
\]
such that \( \lim_{k \to \infty} \|\Sigma_k - \Sigma\|_2 = 0 \) implies that the set \( K := \{A_k : k \in \mathbb{N}\} \) is bounded with respect to \( \|\cdot\|_2 \). Thus, the closure of \( K \) in \( \mathbb{K}^n \) is compact, which implies \( (A_k)_{k \in \mathbb{N}} \) has a convergent subsequence \( (A_{k_l})_{l \in \mathbb{N}} \), converging to some matrix \( A \in \mathbb{K}^n \). As this
convergence is with respect to the norm topology on $\mathbb{K}^{n^2}$, each entry of the $A_k$ must converge (in $\mathbb{K}$) to the respective entry of $A$. In particular, $A$ is lower triangular with all nonnegative diagonal entries. It only remains to show $AA^* = \Sigma$. However,

$$\Sigma = \lim_{l \to \infty} \Sigma_k = \lim_{l \to \infty} A_k A_k^* = AA^*, \quad (I.4)$$

which establishes the case. ■

Example I.2. The decomposition

$$\begin{pmatrix} 0 & 0 \\ \sin x & \cos x \end{pmatrix} \begin{pmatrix} 0 & \sin x \\ 0 & \cos x \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{for each } x \in \mathbb{R} \quad (I.5)$$

shows a Hermitian positive semidefinite matrix (which is not positive definite) can have uncountably many different Cholesky decompositions.

Theorem I.3. Let $\Sigma \in \mathcal{M}(n, \mathbb{K})$ be Hermitian positive semidefinite, $n \in \mathbb{N}$. Define the index set

$$I := \{(i, j) \in \{1, \ldots, n\}^2 : j \leq i\}. \quad (I.6)$$

Then a matrix $A = (A_{ij})$ providing a Cholesky decomposition $\Sigma = AA^*$ of $\Sigma = (\sigma_{ij})$ is obtained via the following algorithm, defined recursively over $I$, using the order $(1, 1) < (2, 1) < \cdots < (n, 1) < (2, 2) < \cdots < (n, 2) < \cdots < (n, n)$ (which corresponds to traversing the lower half of $\Sigma$ by columns from left to right):

$$A_{11} := \sqrt{\sigma_{11}}. \quad (I.7a)$$

For $(i, j) \in I \setminus \{(1, 1)\}$:

$$A_{ij} := \begin{cases} \left(\sigma_{ij} - \sum_{k=1}^{j-1} A_{ik} \overline{A}_{jk}\right)/A_{jj} & \text{for } i > j \text{ and } A_{jj} \neq 0, \\ 0 & \text{for } i > j \text{ and } A_{jj} = 0, \\ \sqrt{\sigma_{ii} - \sum_{k=1}^{i-1} |A_{ik}|^2} & \text{for } i = j. \end{cases} \quad (I.7b)$$

Proof. A lower triangular $n \times n$ matrix $A$ provides a Cholesky decomposition of $\Sigma$ if, and only if,

$$AA^* = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{21} & A_{22} & \cdots & A_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{pmatrix} \begin{pmatrix} \overline{A}_{11} & \overline{A}_{21} & \cdots & \overline{A}_{n1} \\ \overline{A}_{21} & \overline{A}_{22} & \cdots & \overline{A}_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ \overline{A}_{n1} & \overline{A}_{n2} & \cdots & \overline{A}_{nn} \end{pmatrix} = \Sigma, \quad (I.8)$$

i.e. if, and only if, the $n(n+1)/2$ lower half entries of $A$ constitute a solution to the following (nonlinear) system of $n(n+1)/2$ equations:

$$\sum_{k=1}^{j} A_{ik} \overline{A}_{jk} = \sigma_{ij}, \quad (i, j) \in I. \quad (I.9a)$$
Using the order on $I$ introduced in the statement of the theorem, (I.9a) takes the form

\[
|A_{11}|^2 = \sigma_{11},
\]
\[
A_{21}A_{11} = \sigma_{21},
\]
\[
\vdots
\]
\[
A_{n1}A_{11} = \sigma_{n1},
\]
\[
|A_{21}|^2 + |A_{22}|^2 = \sigma_{22},
\]
\[
A_{31}A_{21} + A_{32}A_{22} = \sigma_{32},
\]
\[
\vdots
\]
\[
|A_{n1}|^2 + \cdots + |A_{nn}|^2 = \sigma_{nn}.
\]

(I.9b)

From Th. I.1, we know (I.9) must have at least one solution with $A_{jj} \in \mathbb{R}_+^+$ for each $j \in \{1, \ldots, n\}$. In particular, $\sigma_{jj} \in \mathbb{R}_+^+$ for each $j \in \{1, \ldots, n\}$ (this is also immediate from $\Sigma$ being positive semidefinite, since $\sigma_{jj} = e_j^* \Sigma e_j$, where $e_j$ denotes the $j$th standard unit vector of $\mathbb{K}^n$). We need to show that (I.7) yields a solution to (I.9). The proof is carried out by induction on $n$. For $n = 1$, we have $\sigma_{11} \in \mathbb{R}_+^+$ and $A_{11} = \sqrt{\sigma_{11}} \in \mathbb{R}_+^+$, i.e. there is nothing to prove. Now let $n > 1$. If $\sigma_{11} > 0$, then

\[
A_{11} = \sqrt{\sigma_{11}}, \quad A_{21} = \sigma_{21}/A_{11}, \quad \ldots, \quad A_{n1} = \sigma_{n1}/A_{11}
\]

(I.10a)

is the unique solution to the first $n$ equations of (I.9b) satisfying $A_{11} \in \mathbb{R}_+^+$, and this solution is provided by (I.7). If $\sigma_{11} = 0$, then $\sigma_{21} = \cdots = \sigma_{n1} = 0$: Otherwise, let $s := (\sigma_{21} \ldots \sigma_{n1})^* \in \mathbb{K}^{n-1} \setminus \{0\}$, $\alpha \in \mathbb{R}$, and note

\[
(\alpha, s^*) \begin{pmatrix} 0 & s^* \\ s & \sum_{n-1} \end{pmatrix} \begin{pmatrix} \alpha \\ s \end{pmatrix} = (\alpha, s^*) \begin{pmatrix} \sum_{n-1} & s^*s \\ \alpha s + \sum_{n-1}s \end{pmatrix} = 2\alpha\|s\|^2 + s^*\sum_{n-1}s < 0
\]

for $\alpha < -s^*\sum_{n-1}s/(2\|s\|^2_2)$, in contradiction to $\Sigma$ being positive semidefinite. Thus,

\[
A_{11} = A_{21} = \cdots = A_{n1} = 0
\]

(I.10b)

is a particular solution to the first $n$ equations of (I.9b), and this solution is provided by (I.7). We will now denote the solution to (I.9) given by Th. I.1 by $B_{11}, \ldots, B_{nn}$ to distinguish it from the $A_{ij}$ constructed via (I.7).

In each case, $A_{11}, \ldots, A_{n1}$ are given by (I.10), and, for $(i, j) \in I$ with $i, j \geq 2$, we define

\[
\tau_{ij} := \sigma_{ij} - A_{i1}A_{j1} \in \mathbb{K} \quad \text{for each } (i, j) \in J := \{(i, j) \in I : i, j \geq 2\}.
\]

(I.11)

To be able to proceed by induction, we show that the Hermitian $(n-1) \times (n-1)$ matrix

\[
T := \begin{pmatrix} \tau_{22} & \cdots & \tau_{n2} \\ \vdots & \ddots & \vdots \\ \tau_{n2} & \cdots & \tau_{nn} \end{pmatrix}
\]

(I.12)
is positive semidefinite. If $\sigma_{11} = 0$, then (I.10b) implies $\tau_{ij} = \sigma_{ij}$ for each $(i, j) \in J$ and $T$ is positive semidefinite, as $\Sigma$ being positive semidefinite implies

$$
(\pi_2 \ldots \pi_n)T \begin{pmatrix} \pi_2 \\ \vdots \\ \pi_n \end{pmatrix} = \begin{pmatrix} 0 & \pi_2 & \ldots & \pi_n \end{pmatrix} \Sigma \begin{pmatrix} 0 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \geq 0 \quad (I.13)
$$

for each $(x_2, \ldots, x_n) \in \mathbb{K}^{n-1}$. If $\sigma_{11} > 0$, then (I.10a) holds as well as $B_{11} = A_{11}, \ldots, B_{n1} = A_{n1}$. Thus, (I.11) implies

$$
\tau_{ij} := \sigma_{ij} - A_{i1}A_{j1} = \sigma_{ij} - B_{i1}B_{j1} \quad \text{for each } (i, j) \in J. \quad (I.14)
$$

Then (I.9) with $A$ replaced by $B$ together with (I.14) implies

$$
\sum_{k=2}^{j} B_{ik}B_{jk} = \tau_{ij} \quad \text{for each } (i, j) \in J \quad (I.15)
$$

or, written in matrix form,

$$
\begin{pmatrix} B_{22} & \vdots & B_{n2} \\ \vdots & \ddots & \vdots \\ B_{n2} & \ldots & B_{nn} \end{pmatrix} \begin{pmatrix} B_{22} & \ldots & B_{n2} \\ \vdots & \ddots & \vdots \\ B_{n2} & \ldots & B_{nn} \end{pmatrix} = \begin{pmatrix} \tau_{22} & \ldots & \tau_{n2} \\ \vdots & \ddots & \vdots \\ \tau_{n2} & \ldots & \tau_{nn} \end{pmatrix} = T, \quad (I.16)
$$

which, once again, establishes $T$ to be positive semidefinite (since, for each $x \in \mathbb{K}^{n-1}$, $x^*Tx = x^*BB^*x = (B^*x)^*(B^*x) \geq 0$).

By induction, we now know the algorithm of (I.7) yields a (possibly different from (I.16) for $\sigma_{11} > 0$) decomposition of $T$:

$$
CC^* = \begin{pmatrix} C_{22} & \vdots & C_{n2} \\ \vdots & \ddots & \vdots \\ C_{n2} & \ldots & C_{nn} \end{pmatrix} \begin{pmatrix} C_{22} & \ldots & C_{n2} \\ \vdots & \ddots & \vdots \\ C_{n2} & \ldots & C_{nn} \end{pmatrix} = \begin{pmatrix} \tau_{22} & \ldots & \tau_{n2} \\ \vdots & \ddots & \vdots \\ \tau_{n2} & \ldots & \tau_{nn} \end{pmatrix} = T \quad (I.17)
$$

or

$$
\sum_{k=2}^{j} C_{ik}C_{jk} = \tau_{ij} = \sigma_{ij} - A_{i1}A_{j1} \quad \text{for each } (i, j) \in J, \quad (I.18)
$$

where

$$
C_{22} := \sqrt{\tau_{22}} = \begin{cases} \sqrt{\sigma_{22}} & \text{for } \sigma_{11} = 0, \\ B_{22} & \text{for } \sigma_{11} > 0, \end{cases} \quad (I.19a)
$$

and, for $(i, j) \in J \setminus \{(2, 2)\}$,

$$
C_{ij} := \begin{cases} \left( \tau_{ij} - \sum_{k=2}^{j-1} C_{ik}C_{jk} \right)/C_{jj} & \text{for } i > j \text{ and } C_{jj} \neq 0, \\ 0 & \text{for } i > j \text{ and } C_{jj} = 0, \end{cases} \quad \sqrt{\tau_{ii} - \sum_{k=2}^{i-1} |C_{ik}|^2} \text{ for } i = j. \quad (I.19b)
$$
Substituting $\tau_{ij} = \sigma_{ij} - A_{ii}A_{jj}$ from (I.11) into (I.19) and comparing with (I.7), an induction over $J$ with respect to the order introduced in the statement of the theorem shows $A_{ij} = C_{ij}$ for each $(i, j) \in J$. In particular, since all $C_{ij}$ are well-defined by induction, all $A_{ij}$ are well-defined by (I.7) (i.e. all occurring square roots exist as real numbers). It also follows that $\{A_{ij} : (i, j) \in I\}$ is a solution to (I.9): The first $n$ equations are satisfied according to (I.10); the remaining $(n - 1)n/2$ equations are satisfied according to (I.18) combined with $C_{ij} = A_{ij}$. This concludes the proof that (I.7) furnishes a solution to (I.9).

**I.2 Multiplication with Householder Matrix over $\mathbb{R}$**

If $\mathbb{K} = \mathbb{R}$, then one can use the following, simpler, version of Lem. 5.22:

**Lemma I.4.** Let $k \in \mathbb{N}$. We consider the elements of $\mathbb{R}^k$ as column vectors. Let $e_1 \in \mathbb{R}^k$ be the first standard unit vector of $\mathbb{R}^k$. If $x \in \mathbb{R}^k$ and $x \notin \text{span}\{e_1\}$, then

\[
\begin{align*}
\mathbf{u} &:= \frac{x + \sigma e_1}{\|x + \sigma e_1\|_2} \quad \text{for} \quad \sigma = \pm \|x\|_2, \\
\end{align*}
\]

satisfies

\[
\begin{align*}
\|\mathbf{u}\|_2 &= 1 \quad \text{and} \\
(\mathbf{I} - 2\mathbf{u}^t\mathbf{u})x &= -\sigma e_1. \\
\end{align*}
\]

**Proof.** First, note that $x \notin \text{span}\{e_1\}$ implies $\|x + \sigma e_1\|_2 \neq 0$ such that $\mathbf{u}$ is well-defined. Moreover, (I.21) is immediate from the definition of $\mathbf{u}$. To verify (I.22), note

\[
\begin{align*}
\|x + \sigma e_1\|_2^2 &= \|x\|_2^2 + 2\sigma e_1^t x + \sigma^2 = 2(x + \sigma e_1)^t x \\
\end{align*}
\]

due to the definition of $\sigma$. Together with the definition of $\mathbf{u}$, this yields

\[
\begin{align*}
2\mathbf{u}^t x &= \frac{2(x + \sigma e_1)^t x}{\|x + \sigma e_1\|_2} = \|x + \sigma e_1\|_2 \\
\end{align*}
\]

and

\[
\begin{align*}
2\mathbf{u}^t x &= x + \sigma e_1, \\
\end{align*}
\]

proving (I.22).

**Remark I.5.** In (I.20), one has the freedom to choose the sign of $\sigma$. It is numerically advisable to choose

\[
\sigma = \sigma(x) := \begin{cases} 
\|x\|_2 & \text{for } x_1 \geq 0, \\
-\|x\|_2 & \text{for } x_1 < 0,
\end{cases}
\]

to avoid subtractive cancellation of digits.
J Newton’s Method

A drawback of Theorems 6.3 and 6.5 is the assumption of the existence and (for Th. 6.5) location of the zero \( x^\star \), which (especially for Th. 6.5) can be very difficult to verify in cases where one would like to apply Newton’s method. In the following variant Th. J.1, the hypotheses are such the existence of a zero can be proved as well as the convergence of Newton’s method to that zero.

**Theorem J.1.** Let \( m \in \mathbb{N} \) and fix a norm \( \| \cdot \| \) on \( \mathbb{R}^m \). Let \( A \subseteq \mathbb{R}^m \) be open and convex. Moreover let \( f : A \to A \) be differentiable, and assume \( Df \) to satisfy the Lipschitz condition

\[
\forall \ x, y \in A \quad \| (Df)(x) - (Df)(y) \| \leq L \| x - y \|, \quad L \in \mathbb{R}^+_0.
\]  

(J.1)

Assume \( Df(x) \) to be invertible for each \( x \in A \), satisfying

\[
\forall \ x \in A \quad \| (Df(x))^{-1} \| \leq \beta \in \mathbb{R}^+_0,
\]  

(J.2)

where, once again, we consider the induced operator norm on the real \( n \times n \) matrices. If \( x_0 \in A \) and \( r \in \mathbb{R}^+ \) are such that \( x_1 \in \overline{B}_r(x_0) \subseteq A \) (\( x_1 \) according to (6.5)) and

\[
\alpha := \| (Df(x_0))^{-1}(f(x_0)) \|
\]  

(J.3)

satisfies

\[
h := \frac{\alpha \beta L}{2} < 1
\]  

(J.4a)

and

\[
r \geq \frac{\alpha}{1 - h},
\]  

(J.4b)

then Newton’s method (6.5) is well-defined for \( x_0 \) (i.e., for each \( n \in \mathbb{N}_0 \), \( Df(x_n) \) is invertible and \( x_{n+1} \in A \)) and converges to a zero of \( f \): More precisely, one has

\[
\forall \ n \in \mathbb{N}_0 \quad x_n \in \overline{B}_r(x_0),
\]  

(J.5)

\[
\lim_{n \to \infty} x_n = x^\star \in A,
\]  

(J.6)

\[
f(x^\star) = 0.
\]  

(J.7)

One also has the error estimate

\[
\forall \ n \in \mathbb{N} \quad \| x_n - x^\star \| \leq \alpha \frac{h^{2^n - 1}}{1 - h^{2^n}}.
\]  

(J.8)

**Proof.** Analogous to (6.18) in the proof of Th. 6.5, we obtain

\[
\forall \ x, y \in A \quad \| f(x) - f(y) - Df(y)(x - y) \| \leq \frac{L}{2} \| x - y \|^2
\]  

(J.9)
from the Lipschitz condition (J.1) and the assumed convexity of $A$. We now show via induction on $n \in \mathbb{N}_0$ that

$$\forall n \in \mathbb{N}_0 \left( x_{n+1} \in \overline{B}_r(x_0) \land \|x_{n+1} - x_n\| \leq \alpha h^{2n-1} \right).$$  \hfill (J.10)

For $n = 0$, $x_1 \in \overline{B}_r(x_0)$ by hypothesis and $\|x_1 - x_0\| \leq \alpha$ by (6.5) and (J.3). Now let $n \geq 1$. From (6.5), we obtain

$$\forall n \in \mathbb{N}, \quad Df(x_{n-1})(x_n) = Df(x_{n-1})(x_{n-1}) - f(x_{n-1})$$

and, as we may apply $f$ to $x_n$ by the induction hypothesis,

$$f(x_n) = f(x_{n-1}) - Df(x_{n-1})(x_n - x_{n-1}).$$

Together with (J.9), this yields

$$\|f(x_n)\| \leq \frac{L}{2}\|x_n - x_{n-1}\|^2.$$

Thus,

$$\|x_{n+1} - x_n\| \leq \left\| \left( Df(x_n) \right)^{-1} \right\| \|f(x_n)\| \leq \beta \|f(x_n)\| \leq \frac{\beta L}{2} \|x_n - x_{n-1}\|^2 \leq \frac{\beta L}{2} \alpha^2 h^{2n-2} = \frac{\alpha}{1-h} h^{2n-1},$$

completing the induction step for the second part of (J.10). Next, we use the second part of (J.10) to estimate

$$\|x_{n+1} - x_0\| \leq \sum_{k=0}^{n} \|x_{k+1} - x_k\| \leq \sum_{k=0}^{n} h^{2k-1} \leq \sum_{k=0}^{\infty} h^k = \frac{\alpha}{1-h} \leq r,$$

showing $x_{n+1} \in \overline{B}_r(x_0)$ and completing the induction.

We use

$$h^{2n+k-1} = h^{2n-1} h^{2k-2} = h^{2n-1} (h^{2n})^{2k-1} \leq h^{2n-1} (h^{2n})^k$$

to estimate, for each $m, n \in \mathbb{N},$

$$\|x_{n+m} - x_n\| \leq \sum_{k=0}^{m-1} \|x_{n+k+1} - x_{n+k}\| \leq \alpha \sum_{k=0}^{m-1} h^{2n+k-1} \leq \alpha h^{2n-1} \sum_{k=0}^{\infty} (h^{2n})^k = \alpha \frac{h^{2n-1}}{1-h^{2n}} \to 0 \quad \text{for } n \to \infty,$$

showing $(x_n)_{n \in \mathbb{N}}$ to be a Cauchy sequence and implying the existence of

$$x_* = \lim_{n \to \infty} x_n \in \overline{B}_r(x_0).$$

For fixed $n \in \mathbb{N}$ and $m \to \infty$, (J.15) proves (J.8). Finally,

$$\|f(x_*)\| = \lim_{n \to \infty} \|f(x_n)\| \leq \frac{L}{2} \lim_{n \to \infty} \|x_n - x_{n-1}\|^2 \leq \frac{L \alpha^2}{2} \lim_{n \to \infty} (h^{2n-1})^2 = 0$$

shows $f(x_*) = 0$, completing the proof.
K Eigenvalues

K.1 Continuous Dependence on Matrix Coefficients

The eigenvalues of a complex matrix depend, in a certain sense, continuously on the coefficients of the matrix, which can be seen as a consequence of the fact that the zeros of a complex polynomial depend, in the same sense, continuously on its coefficients. The precise results will be given in Th. K.2 and Cor. K.3 below. The proof is based on Rouché’s theorem of Complex Analysis, which we now state (not in its most general form) for the convenience of the reader (Rouché’s theorem is usually proved as a simple consequence of the residue theorem).

Theorem K.1 (Rouché). Let $D := B_r(z_0) = \{z \in \mathbb{C} : |z - z_0| < r\}$ be a disk in $\mathbb{C}$ ($z_0 \in \mathbb{C}, r \in \mathbb{R}^+$) with boundary $\partial D$. If $G \subseteq \mathbb{C}$ is an open set containing the closed disk $\overline{D} = D \cup \partial D$ and if $f, g : G \to \mathbb{C}$ are holomorphic functions satisfying

$$\forall z \in \partial D \quad |f(z) - g(z)| < |f(z)|,$$

(K.1)

then $N_f = N_g$, where $N_f, N_g$ denote the number of zeros that $f$ and $g$ have in $D$, respectively, and where each zero is counted with its multiplicity.

Proof. See, e.g., [Rud87, Th. 10.43(b)]. □

Theorem K.2. Let $p : \mathbb{C} \to \mathbb{C}$ be a monic complex polynomial function of degree $n$, $n \in \mathbb{N}$, i.e.

$$p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1z + a_0,$$

(K.2a)

$a_0, \ldots, a_{n-1} \in \mathbb{C}$. Then, for each $\epsilon > 0$, there exists $\delta > 0$ such that, for each monic complex polynomial $q : \mathbb{C} \to \mathbb{C}$ of degree $n$, satisfying

$$q(z) = z^n + b_{n-1}z^{n-1} + \cdots + b_1z + b_0,$$

(K.2b)

$b_0, \ldots, b_{n-1} \in \mathbb{C}$, and

$$\forall k \in \{0, \ldots, n-1\} \quad |a_k - b_k| < \delta,$$

(K.2c)

there exist enumerations $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ and $\mu_1, \ldots, \mu_n \in \mathbb{C}$ of the zeros of $p$ and $q$, respectively, listing each zero according to its multiplicity, where

$$\forall k \in \{1, \ldots, n\} \quad |\lambda_k - \mu_k| < \epsilon.$$

(K.3)

Proof. Let $\alpha \in \mathbb{R}^+ \cup \{\infty\}$ denote the minimal distance between distinct zeros of $p$, i.e.

$$\alpha := \begin{cases} \infty & \text{if } p \text{ has only one distinct zero}, \\ \min \{ |\lambda - \tilde{\lambda}| : p(\lambda) = p(\tilde{\lambda}) = 0, \lambda \neq \tilde{\lambda} \} & \text{otherwise}. \end{cases}$$

(K.4)
Without loss of generality, we may assume $0 < \epsilon < \alpha/2$. Let $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_l$, where $1 \leq l \leq n$ be the distinct zeros of $p$. For each $k \in \{1, \ldots, l\}$, define the open disk $D_k := B_\epsilon(\tilde{\lambda}_k)$. Since $\epsilon < \alpha$, $p$ does not have any zeros on $\partial D_k$. Thus,

$$m_k := \min \{ |p(z)| : z \in \partial D_k \} > 0. \quad (K.5a)$$

Also define

$$M_k := \max \left\{ \sum_{j=0}^{n-1} |z|^j : z \in \partial D_k \right\} \quad \text{and note } 1 \leq M_k < \infty. \quad (K.5b)$$

Then, for each $\delta > 0$ satisfying

$$\forall_{k \in \{1, \ldots, l\}} \ M_k \delta < m_k$$

and each $q$ satisfying (K.2b) and (K.2c), the following holds true:

$$\forall_{k \in \{1, \ldots, l\}} \forall_{z \in \partial D_k} \ |q(z) - p(z)| \leq \sum_{j=0}^{n-1} |a_j - b_j| |z|^j \leq \delta \sum_{j=0}^{n-1} |z|^j \leq \delta M_k < m_k \leq |p(z)|, \quad (K.6)$$

showing $p$ and $q$ satisfy hypothesis (K.1) of Rouche’s theorem. Thus, in each $D_k$, the polynomials $p$ and $q$ must have precisely the same number of zeros (counted according to their respective multiplicities). Since $p$ and $q$ both have precisely $n$ zeros (not necessarily distinct), the only (distinct) zero of $p$ in $D_k$ is $\tilde{\lambda}_k$, and the radius of $D_k$ is $\epsilon$, the proof of Th. K.2 is complete.

**Corollary K.3.** Let $A \in \mathcal{M}(n, \mathbb{C})$, $n \in \mathbb{N}$, and let $\| \cdot \|$ denote some arbitrary norm on $\mathcal{M}(n, \mathbb{C})$. Then, for each $\epsilon > 0$, there exists $\delta > 0$ such that, for each matrix $B \in \mathcal{M}(n, \mathbb{C})$, satisfying

$$\|A - B\| < \delta, \quad (K.7)$$

there exist enumerations $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ and $\mu_1, \ldots, \mu_n \in \mathbb{C}$ of the eigenvalues of $A$ and $B$, respectively, listing each eigenvalue according to its (algebraic) multiplicity, where

$$\forall_{k \in \{1, \ldots, n\}} \ |\lambda_k - \mu_k| < \epsilon. \quad (K.8)$$

**Proof.** Recall that each coefficient of the characteristic polynomial $\chi_A$ is a polynomial in the coefficients of $A$ (due to $\chi_A = \det(X \text{Id} - A)$, cf. [Phi19b, Rem. 4.33]), and thus, the function that maps the coefficients of $A$ to the coefficients of $\chi_A$ is continuous from $\mathbb{C}^{n^2}$ to $\mathbb{C}^n$. Thus (and since all norms on $\mathbb{C}^{n^2}$ are equivalent as well as all norms on $\mathbb{C}^n$), given $\gamma > 0$, there exists $\delta = \delta(\gamma) > 0$, such that (K.7) implies

$$\forall_{k \in \{0, \ldots, n-1\}} \ |a_k - b_k| < \gamma,$$

if $a_0, \ldots, a_{n-1}$ and $b_0, \ldots, b_{n-1}$ denote the coefficients of $\chi_A$ and $\chi_B$, respectively. Since the eigenvalues are precisely the zeros of the respective characteristic polynomial, the statement of the corollary is now immediate from the statement of Th. K.2.
K.2 Transformation to Hessenberg Form

In preparation for the application of a numerical eigenvalue approximation algorithm (e.g. the QR algorithm of Sec. 7.4), it is often advisable to transform the matrix under consideration to Hessenberg form (see Def. K.4). The complexity of the transformation to Hessenberg form is $O(n^3)$ (see Th. K.6(c) below). However, each step of the QR algorithm applied to a matrix in Hessenberg form is $O(n^2)$ (see Rem. K.12 below) as compared to $O(n^3)$ for a fully populated matrix (obtaining a QR decomposition via the Householder method of Def. 5.23 requires $O(n)$ matrix multiplications, each matrix multiplication needing $O(n^2)$ multiplications).

Definition K.4. An $n \times n$ matrix $A = (a_{kl})$, $n \in \mathbb{N}$, is said to be in upper (resp. lower) Hessenberg form if, and only if, it is almost in upper (resp. lower) triangular form, except that there are allowed to be nonzero elements in the first subdiagonal (resp. superdiagonal), i.e., if, and only if,

$$a_{kl} = 0 \text{ for } k \geq l + 2 \quad (\text{resp. } a_{kl} = 0 \text{ for } k \leq l - 2).$$

Thus, according to Def. K.4, the $n \times n$ matrix $A = (a_{kl})$ is in upper Hessenberg form if, and only if, it can be written as

$$A = \begin{pmatrix}
\ast & \cdots & \cdots & \cdots & \ast \\
\ast & \ast & \ast & & \\
0 & \ast & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & \ast & \ast
\end{pmatrix}.
$$

How can one transform an arbitrary matrix $A \in \mathcal{M}(n, \mathbb{K})$ into Hessenberg form without changing its eigenvalues? The idea is to use several similarity transformations of the form $A \mapsto HAH$, using Householder matrices $H = H^{-1}$ (cf. Def. 5.19). Householder transformations are also used in Sec. 5.4.3 to obtain QR decompositions.

Algorithm K.5 (Transformation to Hessenberg Form). Given $A \in \mathcal{M}(n, \mathbb{K})$, $n \in \mathbb{N}$, $n > 2$ (for $n = 1$ and $n = 2$, $A$ is always in Hessenberg form), we define the following algorithm: Let $A^{(1)} := A$. For $k = 1, \ldots, n - 2$, $A^{(k)}$ is transformed into $A^{(k+1)}$ by performing precisely one of the following actions:

(a) If $a_{ik}^{(k)} = 0$ for each $i \in \{k + 2, \ldots, n\}$, then

$$A^{(k+1)} := A^{(k)}, \quad H^{(k)} := \text{Id} \in \mathcal{M}(n, \mathbb{K}).$$

(b) Otherwise, it is $x^{(k)} \neq 0$, where

$$x^{(k)} := (a_{k+1,k}^{(k)}, \ldots, a_{n,k}^{(k)})^t \in \mathbb{K}^{n-k},$$

(K.10)
and we set

\[ A^{(k+1)} := H^{(k)} A^{(k)} H^{(k)}, \]

where

\[ H^{(k)} := \begin{pmatrix}
\text{Id}_k & 0 \\
0 & \text{Id}_{n-k} - 2u^{(k)}(u^{(k)}^*)
\end{pmatrix}, \quad u^{(k)} := u(x^{(k)}), \quad (K.11)\]

with \( u(x^{(k)}) \) according to (5.34a).

**Theorem K.6.** Let \( A \in \mathcal{M}(n, \mathbb{K}), n \in \mathbb{N}, n > 2, \) and let \( B := A^{(n-1)} \) be the matrix obtained as a result after having applied Alg. K.5 to \( A \).

(a) Then \( B \) is in upper Hessenberg form and \( \sigma(B) = \sigma(A) \). Moreover, \( m_a(\lambda, B) = m_a(\lambda, A) \) for each \( \lambda \in \sigma(A) \).

(b) If \( A \) is Hermitian, then \( B \) must be Hermitian as well. Thus, \( B \) is a Hermitian matrix in Hessenberg form, i.e., it is tridiagonal (of course, for \( \mathbb{K} = \mathbb{R} \), Hermitian is the same as symmetric).

(c) The transformation of \( A \) into \( B \) requires \( O(n^3) \) arithmetic operations.

**Proof.** (a): Using (5.34c) and blockwise matrix multiplication, a straightforward induction shows that each \( A^{(k+1)} \) has the form

\[ A^{(k+1)} = H^{(k)} A^{(k)} H^{(k)} = \begin{pmatrix}
A_1^{(k)} & A_2^{(k)} H_1^{(k)} \\
0 & \xi_k \xi_1^{(k)}
\end{pmatrix}, \quad (K.12)\]

where \( A_1^{(k)} \) is in upper Hessenberg form. In particular, for \( k + 1 = n - 1 \), we obtain \( B \) to be in upper Hessenberg form. Moreover, since each \( A^{(k+1)} \) is obtained from \( A^{(k)} \) via a similarity transformation, both matrices always have the same eigenvalues with the same multiplicities. In particular, by another induction, the same holds for \( A \) and \( B \).

(b): Since each \( H^{(k)} \) is Hermitian according to Lem. 5.20(a), if \( A^{(k)} \) is Hermitian, then \( (A^{(k+1)})^* = (H^{(k)} A^{(k)} H^{(k)})^* = A^{(k+1)} \), i.e. \( A^{(k+1)} \) is Hermitian as well. Thus, another straightforward induction shows all \( A^{(k)} \) to be Hermitian.
(c): The claim is that the number \(N(n)\) of arithmetic operations needed to compute \(B\) from \(A\) is \(O(n^3)\) for \(n \to \infty\). For simplicity, we do not distinguish between different arithmetic operations, considering addition, multiplication, division, complex conjugation, and even the taking of a square root as one arithmetic operation each, even though the taking of a square root will tend to be orders of magnitude slower than an addition. Still, none of the operations depends on the size \(n\) of the matrix, which justifies treating them as the same, if one is just interested in the asymptotics \(n \to \infty\). First, if \(u, x \in \mathbb{C}^m\), \(H = \text{Id} - 2uu^*\), then computing \(Hx = x - 2u(u^*x)\) needs \(m\) conjugations, \(3m\) multiplications, \(m\) additions, and \(m\) subtractions, yielding

\[
N_1(m) = 6m.
\]

Thus, the number of operations to obtain \(x^*H = (Hx)^*\) is

\[
N_2(m) = 7m
\]

(the same as \(N_1(m)\) plus the additional \(m\) conjugations). The computation of the norm \(\|x\|\) needs \(m\) conjugations, \(m\) multiplications, \(m\) additions, and 1 square root, a scalar multiple \(\xi x\) needs \(m\) multiplications. Thus, the computation of \(u(x)\) according to (5.34a) needs \(8m\) operations (2 norms, 2 scalar multiples) plus a constant number \(C\) of operations (conjugations, multiplications, divisions, square roots) that does not depend on \(m\), the total being

\[
N_3(m) = 8m + C.
\]

Now we are in a position to compute the number \(N^k(n)\) of operations needed to obtain \(A^{(k+1)}\) from \(A^{(k)}\): \(N_3(n - k)\) operations to obtain \(u^{(k)}\) according to (K.11), \(N_1(n - k)\) operations to obtain \(\xi_k e_1\) according to (K.12), \(k \cdot N_2(n - k)\) operations to obtain \(A^{(k)}_2 H_1^{(k)}\) according to (K.12), and \((n - k) \cdot N_1(n - k) + (n - k) \cdot N_2(n - k)\) operations to obtain \(H_1^{(k)} A^{(k)}_3 H_1^{(k)}\) according to (K.12), the total being

\[
N^k(n) = N_3(n - k) + N_1(n - k) + k \cdot N_2(n - k) + (n - k) \cdot N_1(n - k) + (n - k) \cdot N_2(n - k) = C + 8(n - k) + 6(n - k) + 7k(n - k) + 6(n - k)(n - k) + 7(n - k)(n - k) = C + 14(n - k) + 7n(n - k) + 6(n - k)^2.
\]

Finally, we obtain

\[
N(n) = \sum_{k=1}^{n-2} N^k(n) = C(n - 2) + (14 + 7n) \sum_{k=2}^{n-1} k + 6 \sum_{k=2}^{n-1} k^2 = C(n - 2) + (14 + 7n) \left( \frac{(n - 1)n}{2} - 1 \right) + 6 \left( \frac{(n - 1)n(2n - 1)}{6} - 1 \right),
\]

which is \(O(n^3)\) for \(n \to \infty\), completing the proof. ■
K.3 QR Method for Matrices in Hessenberg Form

As mentioned before, it is not advisable to apply the QR method of the previous section directly to a fully populated matrix, since, in this case, the complexity of obtaining the QR decomposition in each step is $O(n^3)$. In the present section, we describe how each step of the QR method can be performed more efficiently for Hessenberg matrices, reducing the complexity of each step to $O(n^2)$. The key idea is to replace the use of Householder matrices by so-called Givens rotations.

Notation K.7. Let $n \in \mathbb{N}$, $n > 1$, $m \in \{1, \ldots, n-1\}$, and $c, s \in \mathbb{K}$. By $G(m, c, s) = (g_{jl}) \in \mathcal{M}(n, \mathbb{K})$ we denote the matrix, where

$$g_{jl} = \begin{cases} 
  c & \text{for } j = l = m, \\
  c & \text{for } j = l = m + 1, \\
  s & \text{for } j = m + 1, l = m, \\
  -s & \text{for } j = m, l = m + 1, \\
  1 & \text{for } j = l \notin \{m, m+1\}, \\
  0 & \text{otherwise,} 
\end{cases} \quad \text{(K.13)}$$

that means

$$G(m, c, s) = \begin{pmatrix} 
  1 & & & & \\
  & \ddots & & & \\
  & & 1 & c & -s \\
  & & s & c & \\
  & & & 1 & \ddots \\
  & & & & 1 
\end{pmatrix} \leftarrow \text{row } m \leftarrow \text{row } m + 1.$$

Lemma K.8. Let $n \in \mathbb{N}$, $n > 1$, $m \in \{1, \ldots, n-1\}$, and $c, s \in \mathbb{K}$. Let $G(m, c, s) = (g_{jl}) \in \mathcal{M}(n, \mathbb{K})$ be as in Not. K.7.

(a) If $|c|^2 + |s|^2 = 1$, then $G(m, c, s)$ is unitary (in this case, $G(m, c, s)$ is a special case of what is known as a Givens rotation).

(b) If $A = (a_{jl}) \in \mathcal{M}(n, \mathbb{K})$ is in Hessenberg form with $a_{21} = \cdots = a_{m,m-1} = 0$, $a_{m+1,m} \neq 0$,

$$c := \frac{a}{\sqrt{|a|^2 + |b|^2}}, \quad s := \frac{b}{\sqrt{|a|^2 + |b|^2}}, \quad a := a_{mm}, \quad b := a_{m+1,m}, \quad \text{(K.14)}$$

and $B = (b_{jl}) := G(m, c, s)^* A$, then $B$ is still in Hessenberg form, but with $b_{21} = \cdots = b_{m,m-1} = b_{m+1,m} = 0$. Moreover, $|c|^2 + |s|^2 = 1$. Rows $m + 2$ through $n$ are identical in $A$ and $B$. 

Proof. (a): \( G(m, c, s) \) is unitary, since
\[
\begin{pmatrix} c & -s \\ s & c \end{pmatrix} \begin{pmatrix} c & s \\ -s & c \end{pmatrix} = \begin{pmatrix} |c|^2 + |s|^2 & c\overline{s} - \overline{c}s \\ s\overline{c} - \overline{s}c & |c|^2 + |s|^2 \end{pmatrix} = \text{Id},
\]

implying \( G(m, c, s)G(m, c, s)^* = \text{Id} \).

(b): Clearly, \(|c|^2 + |s|^2 = 1\). Due to the form of \( G(m, c, s)^* \), \( G(m, c, s)^*A \) can only affect rows \( m \) and \( m+1 \) of \( A \) (all other rows are the same in \( A \) and \( B \)). As \( A \) is in Hessenberg form with \( a_{21} = \cdots = a_{m,m-1} = 0 \), this further implies that columns \( m \) through \( m-1 \) are also the same in \( A \) and \( B \), implying \( B \) to be in Hessenberg form with \( b_{21} = \cdots = b_{m,m-1} = 0 \) as well. Finally,
\[
b_{m+1,m} = -sa_{mm} + ca_{m+1,m} = \frac{-ba + ab}{\sqrt{|a|^2 + |b|^2}} = 0,
\]

thereby establishing the case. \( \blacksquare \)

The idea of each step of the QR method for matrices in Hessenberg form is to transform \( A^{(k)} \) into \( A^{(k+1)} \) by a sequence of \( n-1 \) unitary similarity transformations via Givens rotations \( G_{k,1}, \ldots, G_{k,n-1} \). The \( G_{km} \) are chosen such that the sequence
\[
A^{(k)} =: A^{(k,1)} \mapsto A^{(k,2)} := G_{k,1}^* A^{(k,1)} \mapsto \ldots \mapsto R_k := A^{(k,n)} := G_{k,n-1}^* A^{(k,n-1)}
\]
transforms \( A^{(k)} \) into an upper triangular matrix \( R_k \):

Lemma K.9. If \( A^{(k)} \in \mathcal{M}(n, \mathbb{K}) \) is a matrix in Hessenberg form, \( n \in \mathbb{N}, n \geq 2 \), \( A^{(k,1)} := A^{(k)} \), and, for each \( m \in \{1, \ldots, n-1\} \),
\[
A^{(k,m+1)} := G_{km}^* A^{(k,m)}, \tag{K.15}
\]

\( G_{km} = \text{Id} \) for \( c_{m+1,m}^{(k)} = 0 \), otherwise \( G_{km} = G(m, c, s) \) with \( c, s \) defined as in (K.14), except with \( a := a_{mm}^{(k,m)} \) and \( b := a_{m+1,m}^{(k)} \), then each \( A^{(k,m)} \) is in Hessenberg form, \( a_{21}^{(k,m)} = \cdots = a_{m,m-1}^{(k,m)} = 0 \), and rows \( m+1 \) through \( n \) are identical in \( A^{(k,m)} \) and in \( A^{(k)} \). In particular, \( R_k := A^{(k,n)} \) is an upper triangular matrix.

Proof. This is just Lem. K.8(b) combined with a straightforward induction on \( m \). \( \blacksquare \)

Algorithm K.10 (QR Method for Matrices in Hessenberg Form). Given \( A \in \mathcal{M}(n, \mathbb{K}) \) in Hessenberg form, \( n \in \mathbb{N}, n \geq 2 \), we define the following algorithm for the computation of a sequence \( (A^{(k)})_{k \in \mathbb{N}} \) of matrices in \( \mathcal{M}(n, \mathbb{K}) \): Let \( A^{(1)} := A \). For each \( k \in \mathbb{N} \), \( A^{(k+1)} \) is obtained from \( A^{(k)} \) by a sequence of \( n-1 \) unitary similarity transformations:
\[
A^{(k)} =: B^{(k,1)} \mapsto B^{(k,2)} := G_{k,1}^* B^{(k,1)} G_{k,1} \mapsto \ldots \mapsto A^{(k+1)} := B^{(k,n)} := G_{k,n-1}^* B^{(k,n-1)} G_{k,n-1}, \tag{K.16a}
\]
where the $G_{km}$, $m \in \{1, \ldots, n - 1\}$, are defined as in Lem. K.9. To obtain $G_{km}$, one needs to know $a_{km}$ and $a_{m+1,m}$. However, as it turns out, one does not actually have to compute the matrices $A^{(k,m)}$ explicitly, since, letting

$$C^{(k,m)} := G_{km}^* B^{(k,m)}, \quad (K.16b)$$

one has (see Th. K.11(c) below)

$$C^{(k,m)}_{m+1,m+1} = a_{m+1,m+1}, \quad C^{(k,m)}_{m+2,m+1} = a_{m+2,m+1}. \quad (K.16c)$$

Thus, when computing $B^{(k,m+1)}$ for (K.16a), one computes $C^{(k,m)}$ first and stores the elements in (K.16c) to be available to obtain $G_{k,m+1}$.

**Theorem K.11.** Let $A \in \mathcal{M}(n, \mathbb{F})$ be in Hessenberg form, $n \in \mathbb{N}$, $n \geq 2$. Let the matrices $A^{(k)}$, $A^{(k,m)}$, $B^{(k,m)}$, and $C^{(k,m)}$ be defined as in Alg. K.10.

(a) The QR method for matrices in Hessenberg form as defined in Alg. K.10 is, indeed, a QR method as defined in Alg. 7.18, i.e. (7.37) holds for the $A^{(k)}$. In particular, Alg. K.10 preserves the eigenvalues (with multiplicities) of $A$ and Th. 7.22 applies (also cf. Rem. 7.23(a)).

(b) Each $C^{(k,m)}$ and each $A^{(k)}$ is in Hessenberg form.

(c) Let $m \in \{1, \ldots, n - 1\}$. Then all columns with indices $m + 1$ through $n$ of the matrices $C^{(k,m)}$ and $A^{(k,m+1)}$ are identical (in particular, (K.16c) holds true).

**Proof.** Step 1: Assuming $A^{(k)}$ to be in Hessenberg form, we show that each $C^{(k,m)}$, $m \in \{1, \ldots, n - 1\}$, is in Hessenberg form as well, (c) holds, and $A^{(k+1)}$ is in Hessenberg form: Let $m \in \{1, \ldots, n - 1\}$. According to the definition in Lem. K.9,

$$A^{(k,m+1)} = G_{km}^* \cdots G_{k1}^* A^{(k)}. \quad (K.16a)$$

Thus, according to (K.16a),

$$C^{(k,m)} = G_{km}^* B^{(k,m)} = G_{km}^* G_{k,m-1}^* \cdots G_{k1}^* A^{(k)} G_{k1} \cdots G_{k,m-1} = A^{(k,m+1)} G_{k1} \cdots G_{k,m-1}. \quad (K.17)$$

If a matrix $Z \in \mathcal{M}(n, \mathbb{F})$ is in Hessenberg form with $z_{l+1,l} = z_{l+2,l+1} = 0$, then right multiplication by $G_{kl}$ only affects columns $l, l + 1$ of $Z$ and rows $1, \ldots, l + 1$ of $Z$. As $A^{(k,m+1)}$ is in Hessenberg form with $a_{21}^{(k,m+1)} = \cdots = a_{m+1,m}^{(k,m+1)} = 0$ by Lem. K.9, (K.17) implies $C^{(k,m)}$ is in Hessenberg form with columns $m, \ldots, n$ of $C^{(k,m)}$ and $A^{(k,m+1)}$ being identical. As $A^{(k+1)} = C^{(k,n-1)} G_{k,n-1}$, and $G_{k,n-1}$ only affects columns $n - 1$ and $n$ of $C^{(k,n-1)}$, $A^{(k+1)}$ must also be in Hessenberg form, concluding Step 1.

Step 2: Assuming $A^{(k)}$ to be in Hessenberg form, we show (7.37) to hold for $A^{(k)}$ and $A^{(k+1)}$: According to (K.16a),

$$A^{(k+1)} = G_{k,n-1}^* \cdots G_{k1}^* A^{(k)} G_{k1} \cdots G_{k,n-1} = R_k Q_k. \quad (K.18a)$$
where
\[ R_k = G_{k,n-1}^* \cdots G_{k1}^* A^{(k)}, \quad Q_k := G_{k1} \cdots G_{k,n-1}, \quad A^{(k)} = Q_k R_k. \] (K.18b)

As \( R_k \) is upper triangular by Lem. K.9 and \( Q_k \) is unitary (as it is the product of unitary matrices), we have proved that (7.37) holds and, thereby, completed Step 2.

Finally, since \( A^{(1)} \) is in Hessenberg form by hypothesis, using Steps 1 and 2 in an induction on \( k \) concludes the proof of the theorem. ■

\textbf{Remark K.12.} In the situation of Alg. K.10 it takes \( O(n^2) \) arithmetic operations to transform \( A^{(k)} \) into \( A^{(k+1)} \) (see the proof of Th. K.6(c) for a list of admissible arithmetic operations): One has to obtain the \( n - 1 \) Givens rotations \( G_{k1}, \ldots, G_{k,n-1} \) using (K.14), and one also needs the adjoint \( n - 1 \) Givens rotations. As the number of nontrivial entries in the \( G_{km} \) is always 4 and, thus, independent of \( n \), the total of arithmetic operations to obtain the \( G_{km} \) is \( N_G(n) = O(n) \). Since each \( C^{(k,m)} \) is in Hessenberg form, right multiplication with \( G_{km} \) only affects at most an \( (m+2) \times 2 \) submatrix, needing at most \( C(m+2) \) arithmetic operations, \( C \in \mathbb{N} \). The result differs from a Hessenberg matrix at most in one element, namely the element with index \( (m+2,m) \), i.e. left multiplication of the result with \( G_{km}^* \) only affects at most an \( 2 \times (n-m) \) submatrix, needing at most \( C(n-m) \) arithmetic operations (this is also consistent with the very first step, where one computes \( G_{k1}^* A^{(k)} \) with \( A^{(k)} \) in Hessenberg form). Thus, the total is at most
\[ N_G(n) + (n-1)C(n-m+1+m+2) = N_G(n) + (n-1)C(n+3), \]
which is \( O(n^2) \) for \( n \to \infty \).

\section{L Minimization}

\subsection{L.1 Golden Section Search}

The goal of this section is to provide a derivative-free algorithm for one-dimensional minimization problems
\[ \min f(t), \quad f : I \rightarrow \mathbb{R}, \quad I \subseteq \mathbb{R}, \] (L.1)
that is worst-case optimal in the same way the bisection method is worst-case optimal for root-finding problems. Let us emphasize that one-dimensional minimization problems also often occur as subproblems, so-called line searches in multi-dimensional minimization: One might use a derivative-based method to determine a descent direction \( u \) at some point \( x \) for some objective function \( J : X \rightarrow \mathbb{R} \), that means some \( u \in X \) such that the auxiliary objective function
\[ f : [0,a] \rightarrow \mathbb{R}, \quad f(t) := J(x + tu) \] (L.2)
is known to be decreasing in some neighborhood of 0. The line search then aims to minimize \( J \) on the line segment \( \{x + tu : t \in [0,a]\} \), i.e. to perform the (one-dimensional) minimization of \( f \).
Recall the bisection algorithm for finding a zero of a one-dimensional continuous function $f : [a, b] \to \mathbb{R}$, defined on some interval $[a, b] \subseteq \mathbb{R}$, $a < b$, with $f(a) < 0 < f(b)$: In each step, one has an interval $[a_n, b_n]$, $a_n < b_n$, with $f(a_n) < 0 < f(b_n)$, and one evaluates $f$ at the midpoint $c_n := (a_n + b_n)/2$, choosing $[a_n, c_n]$ as the next interval if $f(c_n) > 0$ and $[c_n, b_n]$ if $f(c_n) < 0$. If $f$ is continuous, then the boundary points of the intervals converge to a zero of $f$. Choosing the midpoint is worst-case optimal in the sense that, without further knowledge on $f$, any other choice for $c_n$ might force one to choose the larger of the two subintervals for the next step, which would then have length bigger than $(b_n - a_n)/2$.

As we will see below, golden section search is the analogue of the bisection algorithm for minimization problems, where the situation is slightly more subtle. First, we have to address the problem of \textit{bracketing a minimum}: For the bisection method, by the intermediate value theorem, a zero of the continuous function $f$ is bracketed if $f(a) < f(b)$ (or $f(a) > f(b)$). However, to bracket the min of a continuous function $f$, we need \textit{three} points $a < b < c$, where

$$f(b) = \min\{f(a), f(b), f(c)\}. \quad (L.3)$$

\textbf{Definition L.1.} Let $I \subseteq \mathbb{R}$ be an interval, $f : I \to \mathbb{R}$. Then $(a, b, c) \in I^3$ is called a \textit{minimization triplet} for $f$ if, and only if, $a < b < c$ and (L.3) holds.

Clearly, if $f$ is continuous, then (L.3) implies $f$ to have (at least one) global min in $[a, b]$. Thus, the problem of bracketing a minimum is to determine a minimization triplet for $f$, where one has to be aware that the problem has no solution if $f$ is strictly increasing or strictly decreasing. Still, we can formulate the following algorithm:

\textbf{Algorithm L.2} (Bracketing a Minimum). Let $f : I \to \mathbb{R}$ be defined on a nontrivial interval $I \subseteq \mathbb{R}$, $m := \inf I$, $M := \sup I$. We define a finite sequence $(a_1, a_2, \ldots, a_N)$, $N \geq 3$, of distinct points in $I$ via the following recursion: For the first two points choose $a < b$ in the interior of $I$. Set

$$a_1 := a, \quad a_2 := b, \quad \sigma := 1 \quad \text{if } f(b) \leq f(a),$$

$$a_1 := b, \quad a_2 := a, \quad \sigma := -1 \quad \text{if } f(b) > f(a). \quad (L.4)$$

If the algorithm has not been stopped in the previous step (i.e. if $2 \leq n < N$), then set $t_n := a_n + 2\sigma|a_n - a_{n-1}|$ and

$$a_{n+1} := \begin{cases} t_n & \text{if } t_n \in I, \\ M & \text{if } t_n > M, M \in I, \\ (a_n + M)/2 & \text{if } t_n > M, M \notin I, \\ m & \text{if } t_n < m, m \in I, \\ (m + a_n)/2 & \text{if } t_n < m, m \notin I. \end{cases} \quad (L.5)$$

Stop the iteration (and set $N := n + 1$) if at least one of the following conditions is satisfied: (i) $f(a_{n+1}) \geq f(a_n)$, (ii) $a_{n+1} \in \{m, M\}$, (iii) $a_{n+1} \notin [a_{\min}, a_{\max}]$ (where
\( a_{\text{min}}, a_{\text{max}} \in I \) are some given bounds), (iv) \( n + 1 = n_{\text{max}} \) (where \( n_{\text{max}} \in \mathbb{N} \) is some given bound).

**Remark L.3.** (a) Since (L.4) guarantees \( f(a_1) \geq f(a_2) \), condition (i) plus an induction shows
\[
f(a_1) \geq f(a_2) > \cdots > f(a_{N-1}).
\]
(b) If Alg. L.2 terminates due to condition (i), then (a) implies \((a_{N-2}, a_{N-1}, a_N)\) (for \( \sigma = 1 \)) or \((a_N, a_{N-1}, a_{N-2})\) (for \( \sigma = -1 \)) to be a minimization triplet for \( f \).
(c) If Alg. L.2 terminates due to conditions (ii) – (iv) (and (i) is not satisfied), then the method has failed to find a minimization triplet for \( f \). One then has three options: (a) abort the search (after all, a minimization triplet for \( f \) might not exist and \( a_N \) is a good candidate for a local min), (b) search on the other side of \( a_1 \), (c) continue the original search with adjusted bounds (if possible).

Given a minimization triplet \((a, b, c)\) for some function \( f \), we now want to evaluate \( f \) at \( x \in [a, b] \) or at \( x \in [b, c] \) to obtain a new minimization triplet \((a', b', c')\), where \( c' - a' < c - a \). Moreover, the strategy for choosing \( x \) should be worst-case optimal.

**Lemma L.4.** Let \( I \subseteq \mathbb{R} \) be an interval, \( f : I \rightarrow \mathbb{R} \). If \((a, b, c) \in I^3\) is a minimization triplet for \( f \), then so are the following triples:

\[
\begin{align*}
(a, x, b) & \quad \text{for } a < x < b \text{ and } f(b) \geq f(x), & (L.6a) \\
(x, b, c) & \quad \text{for } a < x < b \text{ and } f(b) \leq f(x), & (L.6b) \\
(a, b, x) & \quad \text{for } b < x < c \text{ and } f(b) \leq f(x), & (L.6c) \\
(b, x, c) & \quad \text{for } b < x < c \text{ and } f(b) \geq f(x). & (L.6d)
\end{align*}
\]

**Proof.** All four cases are immediate. \( \blacksquare \)

While Lem. L.4 shows the forming of new, smaller triplets to be easy, the question remains of how to choose \( x \). Given a minimization triplet \((a, b, c)\), the location of \( b \) is a fraction \( w \) of the way between \( a \) and \( c \), i.e.

\[
w := \frac{b - a}{c - a}, \quad 1 - w = \frac{c - b}{c - a}. \quad (L.7)
\]

We will now assume \( w \leq \frac{1}{2} \) – if \( w \geq \frac{1}{2} \), then one can make the analogous (symmetric) argument, starting from \( c \) rather than from \( a \) (we will come back to this point below). The new point \( x \) will be an additional fraction \( z \) toward \( c \),

\[
z := \frac{x - b}{c - a}. \quad (L.8)
\]
If one chooses the new minimization triplet according to (L.6), then its length $l$ is

$$l = \begin{cases} 
  w(c-a) & \text{for } a < x < b \text{ and } f(b) \geq f(x), \\
  (1-(w+z))(c-a) & \text{for } a < x < b \text{ and } f(b) \leq f(x), \\
  (w+z)(c-a) & \text{for } b < x < c \text{ and } f(b) \leq f(x), \\
  (1-w)(c-a) & \text{for } b < x < c \text{ and } f(b) \geq f(x).
\end{cases} \quad (L.9)$$

For the moment, consider the case $b < x < c$. As we assume no prior knowledge about $f$, we have no control over which of the two choices we will have to make to obtain the new minimization triplet, i.e., if possible, we should make the last two factors in front of $c-a$ in (L.9) equal, leading to $w + z = 1 - w$ or

$$z = 1 - 2w. \quad (L.10)$$

If we were to choose $x$ between $a$ and $b$, then it could occur that $l = (1-(w+z)) > 1-w$ (note $z < 0$ in this case), that means this choice is not worst-case optimal. Together with the symmetric argument for $w > \frac{1}{2}$, this means we always need to put $x$ in the larger of the two subintervals. While (L.10) provides a formula for determining $z$ in terms of $w$, we still have the freedom to choose $w$. If there is an optimal value for $w$, then this value should remain constant throughout the algorithm – it should be the same for the old triplet and for the new triplet. This leads to

$$w = \frac{x-b}{c-b} (L.8) = \frac{z(c-a)}{c-b} = \frac{z}{1-w}, \quad (L.11)$$

which, combined with (L.10), yields

$$w^2 - 3w + 1 = 0 \iff w = \frac{3 \pm \sqrt{5}}{2}. \quad (L.12)$$

Fortunately, precisely one of the values of $w$ in (L.12) satisfies $0 < w < \frac{1}{2}$, namely

$$w = \frac{3 - \sqrt{5}}{2} \approx 0.38, \quad 1 - w = \frac{\sqrt{5} - 1}{2} \approx 0.62. \quad (L.13)$$

Then $\frac{1-w}{w} = \frac{1+\sqrt{5}}{2}$ is the value of the golden ratio or golden section, which gives the following Alg. L.5 its name. Together with the symmetric argument for $w > \frac{1}{2}$, we can summarize our findings by stating that the new point $x$ should be a fraction of $w = \frac{3-\sqrt{5}}{2}$ into the larger of the two subintervals (where the fraction is measured from the central point of the triplet):

**Algorithm L.5 (Golden Section Search).** Let $I \subseteq \mathbb{R}$ be a nontrivial interval, $f : I \rightarrow \mathbb{R}$. Assume that $(a,b,c) \in I^3$ is a minimization triplet for $f$. We define the following recursive algorithm for the computation of a sequence of minimization triplets for $f$: Set $(a_1,b_1,c_1) := (a,b,c)$. Given $n \in \mathbb{N}$ and a minimization triplet $(a_n,b_n,c_n)$, set

$$x_n := \begin{cases} 
  b_n + w(c_n - b_n) & \text{if } c_n - b_n \geq b_n - a_n, \\
  b_n - w(b_n - a_n) & \text{if } c_n - b_n < b_n - a_n,
\end{cases} \quad w := \frac{3 - \sqrt{5}}{2}, \quad (L.14)$$

and choose $(a_{n+1},b_{n+1},c_{n+1})$ according to (L.6) (in practice, one will stop the iteration if $c_n - a_n < \epsilon$, where $\epsilon > 0$ is some given bound).

(a) The value of $f(b_n)$ decreases monotonically, i.e. $f(b_{n+1}) \leq f(b_n)$ for each $n \in \mathbb{N}$.

(b) With the possible exception of one step, one always has

$$c_{n+1} - a_{n+1} \leq q(c_n - a_n), \quad \text{where} \quad q := \frac{1}{2} + \frac{\alpha}{2} \approx 0.69, \quad \alpha := \frac{3 - \sqrt{5}}{2}. \quad (L.15)$$

(c) One has

$$x := \lim_{n \to \infty} a_n = \lim_{n \to \infty} b_n = \lim_{n \to \infty} c_n, \quad \{x\} = \bigcap_{n \in \mathbb{N}} [a_n, c_n]. \quad (L.16)$$

Proof. (a) clearly holds, since each new minimization triplet is chosen according to (L.6).

(b): As, according to (L.14), $x_n$ is always chosen in the larger subinterval, it suffices to consider the case $w := \frac{b_n - a_n}{c_n - a_n} \leq \frac{1}{2}$. If $w = \alpha$, then (L.9) and (L.10) guarantee

$$c_{n+1} - a_{n+1} = (w + z)(c_n - a_n) = (1 - w)(c_n - a_n) < q(c_n - a_n).$$

Moreover, due to (L.11), if $w = \alpha$ (or $w = 1 - \alpha$) has occurred once, then $w = \alpha$ or $w = 1 - \alpha$ will hold for all following steps. Let $w \neq \alpha$, $w \leq \frac{1}{2}$. If $(a_{n+1}, b_{n+1}, c_{n+1}) = (b_n, x_n, c_n)$, then (L.15) might not hold, but $w = \alpha$ holds in the next step, i.e. (L.15) does hold for all following steps. Thus, it only remains to consider the case $(a_{n+1}, b_{n+1}, c_{n+1}) = (a_n, b_n, x_n)$. In this case,

$$\frac{c_{n+1} - a_{n+1}}{c_n - a_n} = \frac{x_n - a_n}{c_n - a_n} = \frac{b_n - a_n}{c_n - a_n} + \frac{x_n - b_n}{c_n - a_n} = w + \frac{x_n - b_n}{c_n - b_n} \frac{c_n - b_n}{c_n - a_n} = w + \alpha(1 - w) = \alpha + (1 - \alpha)w \leq \alpha + (1 - \alpha)\frac{1}{2} = q,$$

establishing the case.

(c): Clearly, $[a_{n+1}, c_{n+1}] \subseteq [a_n, c_n]$ for each $n \in \mathbb{N}$. Applying an induction to (L.15) yields

$$\forall n \in \mathbb{N}, \quad c_{n+1} - a_{n+1} \leq q^{n-1}(c_1 - a_1),$$

where the exponent on the right-hand side is $n - 1$ instead of $n$ due to the possible exception in one step. Since $q < 1$, this means $\lim_{n \to \infty}(c_n - a_n) = 0$, proving (L.16).

Definition L.7. Let $I \subseteq \mathbb{R}$ be an interval, $f : I \to \mathbb{R}$. We call $f$ locally monotone if, and only if, for each $x \in I$, there exists $\epsilon > 0$ such that $f$ is monotone (increasing or decreasing) on $I \cap [x - \epsilon, x]$ and monotone on $I \cap [x, x + \epsilon]$.

Example L.8. The function

$$f : \mathbb{R} \to \mathbb{R}, \quad f(x) := \begin{cases} x \sin \frac{1}{x} & \text{for } x \neq 0, \\ 0 & \text{for } x = 0, \end{cases}$$

is an example of a continuous function that is not locally monotone (clearly, $f$ is not monotone in any interval $[-\epsilon, 0]$ or $[0, \epsilon]$, $\epsilon > 0$).
Theorem L.9. Let \( I \subseteq \mathbb{R} \) be a nontrivial interval and assume \( f : I \rightarrow \mathbb{R} \) to be continuous and locally monotone in the sense of Def. L.7. Consider Alg. L.5 and let \( x := \lim_{n \to \infty} b_n \) (cf. Lem. L.6(c)). Then \( f \) has a local min at \( x \).

Proof. The continuity of \( f \) implies \( f(x) = \lim_{n \to \infty} f(b_n) \). Then Lem. L.6(a) implies

\[
f(x) = \inf \{ f(b_n) : n \in \mathbb{N} \}.
\]

If \( f \) does not have a local min at \( x \), then

\[
\forall n \in \mathbb{N} \exists y_n \in I \left( y_n \in ]a_n, c_n[ \text{ and } f(y_n) < f(x) \right).
\]

If \( a_n < y_n < x \), then \( f(a_n) > f(y_n) < f(x) \), showing \( f \) is not monotone in \([a_n, x]\). Likewise, if \( x < y_n < c_n \), then \( f(x) > f(y_n) < f(c_n) \), showing \( f \) is not monotone in \([x, c_n]\). Thus, there is no \( \epsilon > 0 \) such that \( f \) is monotone on both \( I \cap [x - \epsilon, x] \) and \( I \cap [x, x + \epsilon] \), in contradiction to the assumption that \( f \) be locally monotone. \( \blacksquare \)

L.2 Nelder-Mead Method

The Nelder-Mead method of [NM65] is a method for minimization in \( \mathbb{R}^n \), i.e. for an objective function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \). It is also known as the downhill simplex method or the amoeba method. The method has nothing to do with the simplex method of linear programming (except that both methods make use of the geometrical structure called simplex).

The Nelder-Mead method is a derivative-free method that is simple to implement. It is an example of a method that seems to perform quite well in practise, even though there are known examples, where it will fail, and, in general, the related convergence theory available in the literature remains quite limited. Here, we will merely present the algorithm together with its heuristics, and provide some relevant references to the literature.

Recall that a (nondegenerate) simplex in \( \mathbb{R}^n \) is the convex hull of \( n+1 \) points \( x_1, \ldots, x_{n+1} \in \mathbb{R}^n \) such that the points \( x_2 - x_1, \ldots, x_{n+1} - x_1 \) are linearly independent. If \( \Delta = \text{conv}\{x_1, \ldots, x_{n+1}\} \), then \( x_1, \ldots, x_{n+1} \) are called the vertices of \( \Delta \).

The idea of the Nelder-Mead method is to construct a sequence \((\Delta^k)_{k \in \mathbb{N}}\),

\[
\Delta^k = \text{conv}\{x^k_1, \ldots, x^k_{n+1}\} \subseteq \mathbb{R}^n,
\]

of simplices such that the size of the \( \Delta^k \) converges to 0 and the value of \( f \) at the vertices of the \( \Delta^k \) decreases, at least on average. To this end, in each step, one finds the vertex \( v \), where \( f \) is largest and either replaces \( v \) by a new vertex (by moving it in the direction of the centroid of the remaining vertices, see below) or one shrinks the simplex.

Algorithm L.10 (Nelder-Mead). Given \( f : \mathbb{R}^n \rightarrow \mathbb{R}, n \in \mathbb{N} \), we provide a recursion to obtain a sequence of simplices \((\Delta^k)_{k \in \mathbb{N}}\) as in (L.19):
Initialization: Choose \( x_1^1 \in \mathbb{R}^n \setminus \{e_1, \ldots, e_n\} \) (the initial guess for the desired \( \min \) of \( f \)), and set
\[
\forall_{i \in \{1, \ldots, n\}} x_{i+1}^1 := x_1^1 + \delta_i e_i, \tag{L.20}
\]
where the \( e_i \) denote the standard unit vectors, and the \( \delta_i > 0 \) are initialization parameters that one should try to adapt to the (conjectured) characteristic length scale (or scales) of the considered problem.

Iteration Step: Given \( \Delta^k, k \in \mathbb{N} \), sort the vertices such that (L.19) holds with
\[
f(x_1^k) \leq f(x_2^k) \cdots \leq f(x_{n+1}^k). \tag{L.21}
\]
According to the following rules, we will either replace (only) \( x_{n+1}^k \) or shrink the simplex.
In preparation, we compute the centroid of the first \( n \) vertices, i.e.
\[
\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i^k, \tag{L.22a}
\]
and the search direction \( u \) for the new vertex,
\[
u := \bar{x} - x_{n+1}^k. \tag{L.22b}
\]
One will now apply precisely one of the following rules (1) – (4) in decreasing priority (i.e. rule \( (j) \) can only be used if none of the conditions for rules (1) – \((j-1)\) applied):

1. Reflection: Compute the reflected point
   \[
x_r := \bar{x} + \mu_r u = x_{n+1}^k + (\mu_r + 1) u \tag{L.23a}
   \]
   and the corresponding value
   \[
f_r := f(x_r), \tag{L.23b}
   \]
   where \( \mu_r > 0 \) is the reflection parameter (\( \mu_r = 1 \) being a common setting). If
   \[
f(x_r^k) \leq f_r < f(x_n^k), \tag{L.23c}
   \]
   then replace \( x_{n+1}^k \) with \( x_r \) to obtain the new simplex \( \Delta^{k+1} \). Heuristics: As \( f \) is largest at \( x_{n+1}^k \), one reflects the point through the hyperplane spanned by the remaining vertices hoping to find smaller values on the other side. If one, indeed, finds a sufficiently smaller value (better than the second worst), then one takes the new point, except if the new point has a smaller value than all other vertices, in which case, one tries to go even further in that direction by applying rule (2) below.

2. Expansion: If
   \[
f(x_r) < f(x_r^k) \tag{L.24a}
   \]
   (i.e. (L.23c) does not hold and \( x_r \) is better than all \( x_i^k \)), then compute the expanded point
   \[
x_e := \bar{x} + \mu_e u = x_{n+1}^k + (\mu_e + 1) u, \tag{L.24b}
   \]
where $\mu_e > \mu_r$ is the expansion parameter ($\mu_e = 2$ being a common setting). If

$$f(x_e) < f_r,$$

(L.24c)

then replace $x_{n+1}^k$ with $x_e$ to obtain the new simplex $\Delta^{k+1}$. If (L.24c) does not hold, then, as in (1), replace $x_{n+1}^k$ with $x_i$ to obtain $\Delta^{k+1}$. If neither (L.23c) nor (L.24a) holds, then continue with (3) below. Heuristics: If $x_r$ has the smallest value so far, then one hopes to find even smaller values by going further in the same direction.

(3) Contraction: If neither (L.23c) nor (L.24a) holds, then

$$f_r \geq f(x_{n+1}^k).$$

(L.25a)

In this cases, compute the contracted point

$$x_c := \bar{x} + \mu_c u = x_{n+1}^k + (\mu_c + 1) u,$$

(L.25b)

where $-1 < \mu_c < 0$ is the contraction parameter ($\mu_c = -1/2$ being a common setting). If

$$f(x_c) < f(x_{n+1}^k),$$

(L.25c)

then replace $x_{n+1}^k$ with $x_c$ to obtain the new simplex $\Delta^{k+1}$. Heuristics: If (L.25a) holds, then $x_r$ would still be the worst vertex of the new simplex. If the new vertex is to be still the worst, then it should at least be such that it reduces the volume of the simplex. Thus, the contracted point is chosen between $x_{n+1}^k$ and $\bar{x}$.

(4) Reduction: If (L.25a) holds and (L.25c) does not hold, then we have failed to find a better vertex in the direction $u$. We then shrink the simplex around the best point $x_1^k$ by setting $x_1^{k+1} := x_1^k$ and

$$\forall \ i \in \{2, \ldots, n+1\} \quad x_i^{k+1} := x_i^k + \mu_s (x_i^k - x_1^k),$$

(L.26)

where $0 < \mu_s < 1$ is the shrink parameter ($\mu_s = 1/2$ being a common setting). Heuristics: If $f$ increases by moving from the worst point in the direction of the better points, then that indicates the simplex being too large, possibly containing several local mins. Thus, we shrink the simplex around the best point, hoping to thereby improve the situation.


(a) Clearly, all simplices $\Delta^k$, $k \in \mathbb{N}$, are nondegenerate: For $\Delta^1$ this is guaranteed by choosing $x_1^1 \notin \{e_1, \ldots, e_n\}$; for $k > 1$, this is guaranteed since (1) – (3) choose the new vertex not in $\text{conv}\{x_1^k, \ldots, x_n^k\}$ (note $\{x_{n+1}^k + \mu(\bar{x} - x_{n+1}^k) : \mu \in \mathbb{R}\} \cap \text{conv}\{x_1^k, \ldots, x_n^k\} = \{\bar{x}\}$), and $x_2^k - x_1^k, \ldots, x_{n+1}^k - x_1^k$ are linearly independent in (4).

(b) (1) – (3) are designed such that the average value of $f$ at the vertices decreases:

$$\frac{1}{n+1} \sum_{i=1}^{n+1} f(x_i^{k+1}) < \frac{1}{n+1} \sum_{i=1}^{n+1} f(x_i^k).$$

(L.27)

However, if (4) occurs, then (L.27) can fail to hold.
As mentioned above, even though Alg. L.10 seems to perform well in practice, one can construct examples, where it will converge to points that are not local minima (see, e.g., [Kel99, Sec. 8.1.3]). Some positive results regarding the convergence of Alg. L.10 can be found in [Kel99, Sec. 8.1.2] and [LRWW98]. In general, multidimensional minimization is not an easy task, and it is not uncommon for algorithms to converge to wrong points. It is therefore recommended in [PTVF07, pp. 503-504] to restart the algorithm at a point that it claimed to be a minimum (also see [Kel99, Sec. 8.1.4] regarding restarts of Alg. L.10).

References


[DH08] Peter Deuflhard and Andreas Hohmann. Numerische Mathematik 1, 4th ed. Walter de Gruyter, Berlin, Germany, 2008 (German).


REFERENCES


REFERENCES


