Numerical Analysis I

Peter Philip[*]

Lecture Notes

Originally Created for the Class of Winter Semester 2008/2009 at LMU Munich,
Revised and Extended for the Classes of Winter Semesters 2009/2010 and 2014/2015

April 10, 2015

Contents

1 Introduction and Motivation **4**

2 Rounding and Error Analysis **8**

2.1 Floating-Point Numbers Arithmetic, Rounding 8

2.2 Rounding Errors 12

2.3 Landau Symbols 16

2.4 Operator Norms and Matrix Norms 21

2.5 Condition of a Problem 28

3 Interpolation **37**

3.1 Motivation 37

3.2 Polynomial Interpolation 38

3.2.1 Existence and Uniqueness 38

3.2.2 Newton's Divided Difference Interpolation Formula 40

3.2.3 Error Estimates and the Mean Value Theorem for Divided Differences 44

3.3 Hermite Interpolation 47

[*] E-Mail: philip@math.lmu.de

3.4 The Weierstrass Approximation Theorem 55
3.5 Spline Interpolation 58
3.5.1 Introduction 58
3.5.2 Linear Splines 59
3.5.3 Cubic Splines 60

4 Numerical Integration 68
4.1 Introduction 68
4.2 Quadrature Rules Based on Interpolating Polynomials 72
4.3 Newton-Cotes Formulas 75
4.3.1 Definition, Weights, Degree of Accuracy 75
4.3.2 Rectangle Rules ($n = 0$) 78
4.3.3 Trapezoidal Rule ($n = 1$) 79
4.3.4 Simpson's Rule ($n = 2$) 79
4.3.5 Higher Order Newton-Cotes Formulas 81
4.4 Convergence of Quadrature Rules 81
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 ($n = 1$) 86
4.5.4 Composite Simpson's Rules ($n = 2$) 87
4.6 Gaussian Quadrature 88
4.6.1 Introduction 88
4.6.2 Orthogonal Polynomials 88
4.6.3 Gaussian Quadrature Rules 92

5 Numerical Solution of Linear Systems 96
5.1 Motivation 96
5.2 Gaussian Elimination and LU Decomposition 97
5.2.1 Pivot Strategies 97
5.2.2 Gaussian Elimination via Matrix Multiplication and LU Decomposition 100
5.2.3 The Algorithm of LU Decomposition 109

5.3 Cholesky Decomposition 111

5.4 QR Decomposition 114

5.4.1 Definition and Motivation 114

5.4.2 QR Decomposition via Gram-Schmidt Orthogonalization 115

5.4.3 QR Decomposition via Householder Reflections 119

6 Iterative Methods, Solution of Nonlinear Equations 124

6.1 Motivation: Fixed Points and Zeros 124

6.2 Banach Fixed Point Theorem 125

6.3 Newton's Method 126

A b-Adic Expansions of Real Numbers 130

B Operator Norms and Matrix Norms 135

C $\mathbb{R}^m$-Valued Integration 139

D The Vandermonde Determinant 141

E Parameter-Dependent Integrals 143

F Inner Products and Orthogonality 143

G Baire Category and the Uniform Boundedness Principle 145

H Linear Algebra 148

H.1 Gaussian Elimination Algorithm 148

H.2 Orthogonal Matrices 151

H.3 Positive Definite Matrices 152

H.4 Cholesky Decomposition for Positive Semidefinite Matrices 153

1 Introduction and Motivation

The central motivation of Numerical Analysis 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 Analysis 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 Analysis 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

$$x_0 := a_0, \quad x_{n+1} := \frac{1}{2}\left(x_n + \frac{a}{x_n}\right) \text{ for each } n \in \mathbb{N}_0. \tag{1.1}$$

It can be shown 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 (cf. Sec. 6.3 below) for the computation of the zero of the function $f : \mathbb{R}^+ \longrightarrow \mathbb{R}$, $f(x) := x^2 - a$). 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      (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 :  $\epsilon = 10^{-10} * a$   % 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                    (1.3)
                           % contents of the variable $x$ with the computed value
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 Example 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 :  $n = 1$;    $x = a$
2 :  $x = 1/(2 - x)$;    $n = n + 1$
3 :  if $n \leq N$                                   (1.4)
         then goto 2
         else quit
```

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:

```
1 : $n = 1$;    $x = a$
2 : if $x \neq 2$
        then  $x = 1/(2 - x)$;  $n = n + 1$
        else  $x = -5$;         $n = n + 1$          (1.5)
3 : if $n \leq N$
        then goto 2
        else quit
```

—

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 Analysis 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 $p : \mathbb{R} \longrightarrow \mathbb{R}$" 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 $f(x) = ax^2 + c$". 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

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

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

$$S:\, \big\{(a,c)\in\mathbb{R}^2:\, a\neq 0\big\} \longrightarrow \mathbb{C}^2,$$
$$S(a,c) := \begin{cases} \left(\sqrt{\frac{|c|}{|a|}},\, -\sqrt{\frac{|c|}{|a|}}\right) & \text{for } ac<0,\\ \left(i\sqrt{\frac{|c|}{|a|}},\, -i\sqrt{\frac{|c|}{|a|}}\right) & \text{for } ac>0. \end{cases} \tag{1.6b}$$

(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

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

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 Analysis, 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 Rounding and Error Analysis

2.1 Floating-Point Numbers Arithmetic, Rounding

All the numerical problems considered in this class are related to the computation of (approximations of) real numbers. The b-adic representations of real numbers (see Appendix A), 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 2.1. 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

$$(x_1, \dots, x_l) \in \{0, 1, \dots, b-1\}^l, \tag{2.1a}$$

$$N \in \mathbb{Z} \text{ satisfying } N_- \leq N \leq N_+, \tag{2.1b}$$

the strings

$$0\,.\,x_1 x_2 \dots x_l \cdot b^N \quad \text{and} \quad -0\,.\,x_1 x_2 \dots x_l \cdot b^N \tag{2.2}$$

are called *floating-point representations* of the (rational) numbers

$$x := b^N \sum_{\nu=1}^{l} x_\nu b^{-\nu} \quad \text{and} \quad -x, \tag{2.3}$$

respectively, provided that $x_1 \neq 0$ if $x \neq 0$. For floating-point representations, b is called the *radix* (or *base*), $0\,.\,x_1 \dots x_l$ (i.e. $|x|/b^N$) is called the *significand* (or *mantissa* or *coefficient*), $x_1, \dots, x_l$ are called the *significant digits*, and N is called the *exponent*. The number l is sometimes called the *precision*. Let $\operatorname{fl}_l(b, N_-, N_+) \subseteq \mathbb{Q}$ 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 2.2. If one restricts Def. 2.1 to the case $N_- = N_+$, 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 2.3. Given the assumptions of Def. 2.1, for the numbers in (2.3), it always holds that

$$b^{N_- -1} \leq b^{N-1} \leq \sum_{\nu=1}^{l} x_\nu b^{N-\nu} = |x| \leq (b-1)\sum_{\nu=1}^{l} b^{N_+ -\nu} =: \max(l,b,N_+) \overset{\text{Lem. A.2}}{<} b^{N_+}, \tag{2.4}$$

provided that $x \neq 0$. In other words:

$$\mathrm{fl}_l(b,N_-,N_+) \subseteq \mathbb{Q} \cap \big([-\max(l,b,N_+), -b^{N_- -1}] \cup \{0\} \cup [b^{N_- -1}, \max(l,b,N_+)]\big). \tag{2.5}$$

Obviously, $\mathrm{fl}_l(b,N_-,N_+)$ is not closed under arithmetic operations. A result of absolute value bigger than $\max(l,b,N_+)$ 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 2.4. Let $b \in \mathbb{N}$, $b \geq 2$, $l \in \mathbb{N}$, $N \in \mathbb{Z}$, and $\sigma \in \{-1,1\}$. Then, given

$$x = \sigma b^N \sum_{\nu=1}^{\infty} x_\nu b^{-\nu}, \tag{2.6}$$

where $x_\nu \in \{0,1,\dots,b-1\}$ for each $\nu \in \mathbb{N}$ and $x_1 \neq 0$ for $x \neq 0$, define

$$\mathrm{rd}_l(x) := \begin{cases} \sigma b^N \sum_{\nu=1}^{l} x_\nu b^{-\nu} & \text{for } x_{l+1} < b/2, \\ \sigma b^N \big(b^{-l} + \sum_{\nu=1}^{l} x_\nu b^{-\nu}\big) & \text{for } x_{l+1} \geq b/2. \end{cases} \tag{2.7}$$

The number $\mathrm{rd}_l(x)$ is called *x rounded to l digits.*

Remark 2.5. We note that the notation $\mathrm{rd}_l(x)$ of Def. 2.4 is actually not entirely correct, since rd_l is actually a function of the sequence $(\sigma, N, x_1, x_2, \dots)$ rather than of x: It can actually occur that rd_l takes on different values for different representations of the same number x (for example, using decimal notation, consider $x = 0.34\overline{9} = 0.35\overline{0}$, yielding $\mathrm{rd}_1(0.34\overline{9}) = 0.3$ and $\mathrm{rd}_1(0.35\overline{0}) = 0.4$). However, from basic results on b-adic expansions of real numbers (see Th. A.6), we know that x can have at most two different b-adic representations and, for the same x, $\mathrm{rd}_l(x)$ can vary at most $b^N b^{-l}$. Since always writing $\mathrm{rd}_l(\sigma, N, x_1, x_2, \dots)$ does not help with readability, and since writing $\mathrm{rd}_l(x)$ hardly ever causes confusion as to what is meant in a concrete case, the abuse of notation introduced in Def. 2.4 is quite commonly employed.

Lemma 2.6. *Let $b \in \mathbb{N}$, $b \geq 2$, $l \in \mathbb{N}$. If $x \in \mathbb{R}$ is given by* (2.6)*, then* $\mathrm{rd}_l(x) = \sigma b^{N'} \sum_{\nu=1}^{l} x'_\nu b^{-\nu}$*, where $N' \in \{N, N+1\}$ and $x'_\nu \in \{0,1,\dots,b-1\}$ for each $\nu \in \mathbb{N}$ and $x'_1 \neq 0$ for $x \neq 0$. In particular, rd_l maps $\bigcup_{k=1}^{\infty} \mathrm{fl}_k(b,N_-,N_+)$ into $\mathrm{fl}_l(b,N_-,N_++1)$.*

Proof. For $x_{l+1} < b/2$, there is nothing to prove. Thus, assume $x_{l+1} \geq b/2$. Case (i): There exists $\nu \in \{1,\dots,l\}$ such that $x_\nu < b-1$. Then, letting $\nu_0 \in \{1,\dots,l\}$ be the largest index such that $x_\nu < b-1$, one finds $N' = N$ and

$$x'_\nu = \begin{cases} x_\nu & \text{for } 1 \leq \nu < \nu_0, \\ x_{\nu_0} + 1 & \text{for } \nu = \nu_0, \\ 0 & \text{for } \nu_0 < \nu \leq l. \end{cases} \tag{2.8a}$$

Case (ii): $x_\nu = b - 1$ holds for each $\nu \in \{1, \dots, l\}$. Then one obtains $N' = N + 1$,

$$x'_\nu := \begin{cases} 1 & \text{for } \nu = 1, \\ 0 & \text{for } 1 < \nu \leq l, \end{cases} \tag{2.8b}$$

thereby concluding the proof. ■

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 2.7 makes sense for general real numbers x, y, but is intended to be used with numbers from some $\mathrm{fl}_l(b, N_-, N_+)$, i.e. numbers given in floating point representation (in particular, with a finite precision).

Notation 2.7. Let $b \in \mathbb{N}$, $b \geq 2$. Assume $x, y \in \mathbb{R}$ are given in a form analogous to (2.6). We then define, for each $l \in \mathbb{N}$,

$$x \diamond_l y := \mathrm{rd}_l(x \diamond y), \tag{2.9}$$

where $\diamond$ can stand for any of the operations '+', '−', '·', and ':'.

Remark 2.8. It follows from Lem. 2.6 that, given $x, y \in \mathrm{fl}_l(b, N_-, N_+)$, the result of $x \diamond_l y$ as defined in (2.9) is either in $\mathrm{fl}_l(b, N_-, N_+)$ or an overflow or an underflow.

Remark 2.9. The definition of (2.9) 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, it is the property stated in Rem. 2.8 that one expects floating-point operations to satisfy.

Caveat 2.10. 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 $x, y, z \in \mathrm{fl}_l(b, N_-, N_+)$ and one assumes that no overflow or underflow occurs, then there are examples, where $(x +_l y) +_l z \neq x +_l (y +_l z)$, $(x \cdot_l y) \cdot_l z \neq x \cdot_l (y \cdot_l z)$, and $x \cdot_l (y +_l z) \neq x \cdot_l y +_l x \cdot_l z$ (exercise).

Lemma 2.11. *Let $b \in \mathbb{N}$, $b \geq 2$. Suppose*

$$x = b^N \sum_{\nu=1}^{\infty} x_\nu b^{-\nu}, \quad y = b^M \sum_{\nu=1}^{\infty} y_\nu b^{-\nu}, \tag{2.10}$$

where $N, M \in \mathbb{Z}$; $x_\nu, y_\nu \in \{0, 1, \dots, b-1\}$ 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 $\nu \in \{1,\dots,n-1\}$, then $x \geq y$.*

Proof. (a): One estimates

$$\begin{aligned} x - y &= b^N \sum_{\nu=1}^{\infty} x_\nu b^{-\nu} - b^M \sum_{\nu=1}^{\infty} y_\nu b^{-\nu} \geq b^{N-1} - b^{N-1} \sum_{\nu=1}^{\infty} (b-1) b^{-\nu} \\ &= b^{N-1} - b^{N-1}(b-1)\left(\frac{1}{1-b^{-1}} - 1\right) = 0. \end{aligned} \tag{2.11a}$$

(b): One estimates

$$\begin{aligned} x - y &= b^N \sum_{\nu=1}^{\infty} x_\nu b^{-\nu} - b^M \sum_{\nu=1}^{\infty} y_\nu b^{-\nu} = b^N \sum_{\nu=n}^{\infty} x_\nu b^{-\nu} - b^N \sum_{\nu=n}^{\infty} y_\nu b^{-\nu} \\ &\geq b^{N-n} - b^{N-n} \sum_{\nu=1}^{\infty} (b-1) b^{-\nu} \overset{\text{as in (2.11a)}}{=} 0, \end{aligned} \tag{2.11b}$$

concluding the proof of the lemma. ■

Lemma 2.12. *Let $b \in \mathbb{N}$, $b \geq 2$, $l \in \mathbb{N}$. Then, for each $x, y \in \mathbb{R}$:*

(a) $\operatorname{rd}_l(x) = \operatorname{sgn}(x) \operatorname{rd}_l(|x|)$.

(b) $0 \leq x < y$ *implies* $0 \leq \operatorname{rd}_l(x) \leq \operatorname{rd}_l(y)$.

(c) $0 \geq x > y$ *implies* $0 \geq \operatorname{rd}_l(x) \geq \operatorname{rd}_l(y)$.

Proof. (a): If x is given by (2.6), one obtains from (2.7):

$$\operatorname{rd}_l(x) = \sigma \operatorname{rd}_l \left(b^N \sum_{\nu=1}^{\infty} x_\nu b^{-\nu} \right) = \operatorname{sgn}(x) \operatorname{rd}_l(|x|).$$

(b): Suppose x and y are given as in (2.10). Then Lem. 2.11(a) implies $N \leq M$.

Case $x_{l+1} < b/2$: In this case, according to (2.7),

$$\operatorname{rd}_l(x) = b^N \sum_{\nu=1}^{l} x_\nu b^{-\nu}. \tag{2.12}$$

For $N < M$, we estimate

$$\operatorname{rd}_l(y) \geq b^M \sum_{\nu=1}^{l} y_\nu b^{-\nu} \overset{\text{Lem. 2.11(a)}}{\geq} b^N \sum_{\nu=1}^{l} x_\nu b^{-\nu} = \operatorname{rd}_l(x), \tag{2.13}$$

which establishes the case. We claim that (2.13) also holds for $N = M$, now due to Lem. 2.11(b): This is clear if $x_\nu = y_\nu$ for each $\nu \in \{1,\dots,l\}$. Otherwise, define

$$n := \min\big\{\nu \in \{1,\dots,l\} : x_\nu \neq y_\nu\big\}. \tag{2.14}$$

Then $x < y$ and Lem. 2.11(b) yield $x_n < y_n$. Moreover, another application of Lem. 2.11(b) then implies (2.13) for this case.

Case $x_{l+1} \geq b/2$: In this case, according to Lem. 2.6,

$$\mathrm{rd}_l(x) = b^{N'} \sum_{\nu=1}^{l} x'_\nu b^{-\nu}, \tag{2.15}$$

where either $N' = N$ and the x'_ν are given by (2.8a), or $N' = N+1$ and the x'_ν are given by (2.8b). For $N < M$, we obtain

$$\mathrm{rd}_l(y) \geq b^M \sum_{\nu=1}^{l} y_\nu b^{-\nu} \overset{(*)}{\geq} b^{N'} \sum_{\nu=1}^{l} x'_\nu b^{-\nu} = \mathrm{rd}_l(x), \tag{2.16}$$

where, for $N' < M$, $(*)$ holds by Lem. 2.11(a), and, for $N' = N+1 = M$, $(*)$ holds by (2.8b). It remains to consider $N = M$. If $x_\nu = y_\nu$ for each $\nu \in \{1, \dots, l\}$, then $x < y$ and Lem. 2.11(b) yield $b/2 \leq x_{l+1} \leq y_{l+1}$, which, in turn, yields $\mathrm{rd}_l(y) = \mathrm{rd}_l(x)$. Otherwise, once more define n according to (2.14). As before, $x < y$ and Lem. 2.11(b) yield $x_n < y_n$. From Lem. 2.6, we obtain $N' = N$ and that the x'_ν are given by (2.8a) with $\nu_0 \geq n$. Then the values from (2.8a) show that (2.16) holds true once again.

(c) follows by combining (a) and (b). ■

Definition and Remark 2.13. 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 $\epsilon \in \mathrm{fl}_l(b, N_-, N_+)$ such that

$$1 +_l \epsilon \neq 1, \tag{2.17}$$

then it is called the *relative machine precision*. It is an exercise to show that, for $N_+ < -l+1$, one has $1 +_l y = 1$ for every $0 < y \in \mathrm{fl}_l(b, N_-, N_+)$, such that there is no positive number in $\mathrm{fl}_l(b, N_-, N_+)$ satisfying (2.17), whereas, for $N_+ \geq -l+1$:

$$\epsilon = \begin{cases} b^{N_- - 1} & \text{for } -l+1 < N_-, \\ \lceil b/2 \rceil b^{-l} & \text{for } -l+1 \geq N_- \end{cases}$$

(for $x \in \mathbb{R}$, $\lceil x \rceil := \min\{k \in \mathbb{Z} : x \leq k\}$ is called *ceiling of x* or *x rounded up*).

2.2 Rounding Errors

Definition 2.14. Let $v \in \mathbb{R}$ be the exact value and let $a \in \mathbb{R}$ be an approximation for v. The numbers

$$e_\mathrm{a} := |v - a|, \quad e_\mathrm{r} := \frac{e_\mathrm{a}}{|v|} \tag{2.18}$$

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 2.15. *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 (2.6), i.e.*

$$x = \sigma b^N \sum_{\nu=1}^{\infty} x_\nu b^{-\nu}, \tag{2.19}$$

where $x_\nu \in \{0, 1, \dots, b-1\}$ 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*

$$e_{\mathrm{a}}(x) = \big|\operatorname{rd}_l(x) - x\big| \leq \frac{b^{N-l}}{2}.$$

(b) *The relative error of rounding to l digits satisfies, for each $x \neq 0$,*

$$e_{\mathrm{r}}(x) = \frac{\big|\operatorname{rd}_l(x) - x\big|}{|x|} \leq \frac{b^{-l+1}}{2}.$$

(c) *For each $x \neq 0$, one also has the estimate*

$$\frac{\big|\operatorname{rd}_l(x) - x\big|}{|\operatorname{rd}_l(x)|} \leq \frac{b^{-l+1}}{2}.$$

Proof. (a): First, consider the case $x_{l+1} < b/2$: One computes

$$\begin{aligned}
e_{\mathrm{a}}(x) \quad &= \quad \big|\operatorname{rd}_l(x) - x\big| = -\sigma\big(\operatorname{rd}_l(x) - x\big) = b^N \sum_{\nu=l+1}^{\infty} x_\nu b^{-\nu} \\
&= \quad b^{N-l-1} x_{l+1} + b^N \sum_{\nu=l+2}^{\infty} x_\nu b^{-\nu} \\
&\overset{b \text{ even, (A.3)}}{\leq} \quad b^{N-l-1}\left(\frac{b}{2} - 1\right) + b^{N-l-1} = \frac{b^{N-l}}{2}.
\end{aligned} \tag{2.20a}$$

It remains to consider the case $x_{l+1} \geq b/2$. In that case, one obtains

$$\begin{aligned}
\sigma\big(\operatorname{rd}_l(x) - x\big) \quad &= \quad b^{N-l} - b^N x_{l+1} b^{-l-1} - b^N \sum_{\nu=l+2}^{\infty} x_\nu b^{-\nu} \\
&= \quad b^{N-l-1}(b - x_{l+1}) - b^N \sum_{\nu=l+2}^{\infty} x_\nu b^{-\nu} \leq \frac{b^{N-l}}{2}.
\end{aligned} \tag{2.20b}$$

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 (2.20b) implies $\sigma\big(\operatorname{rd}_l(x) - x\big) \geq 0$, i.e. $e_a(x) = \sigma\big(\operatorname{rd}_l(x) - x\big)$.

(b): Since $x \neq 0$, one has $x_1 \geq 1$. Thus, $|x| \geq b^{N-1}$. Then (a) yields

$$e_{\mathrm{r}} = \frac{e_{\mathrm{a}}}{|x|} \leq \frac{b^{N-l}\, b^{-N+1}}{2} = \frac{b^{-l+1}}{2}. \tag{2.21}$$

(c): Again, $x_1 \geq 1$ as $x \neq 0$. Thus, (2.7) implies $|\operatorname{rd}_l(x)| \geq b^{N-1}$. This time, (a) yields

$$\frac{e_{\mathrm{a}}}{|\operatorname{rd}_l(x)|} \leq \frac{b^{N-l}\, b^{-N+1}}{2} = \frac{b^{-l+1}}{2}, \tag{2.22}$$

establishing the case and completing the proof of the proposition. ■

Corollary 2.16. *In the situation of Prop.* 2.15, *let, for* $x \neq 0$*:*

$$\epsilon_l(x) := \frac{\operatorname{rd}_l(x) - x}{x}, \quad \eta_l(x) := \frac{\operatorname{rd}_l(x) - x}{\operatorname{rd}_l(x)}. \tag{2.23}$$

Then

$$\max\big\{|\epsilon_l(x)|, |\eta_l(x)|\big\} \leq \frac{b^{-l+1}}{2} =: \tau_l. \tag{2.24}$$

The number τ_l *is called the* relative computing precision *of floating-point arithmetic with* l *significant digits.* ■

Remark 2.17. From the definitions of $\epsilon_l(x)$ and $\eta_l(x)$ in (2.23), one immediately obtains the relations

$$\operatorname{rd}_l(x) = x\big(1 + \epsilon_l(x)\big) = \frac{x}{1 - \eta_l(x)} \quad \text{for } x \neq 0. \tag{2.25}$$

In particular, according to (2.9), one has for floating-point operations (for $x \diamond y \neq 0$):

$$x \diamond_l y = \operatorname{rd}_l(x \diamond y) = (x \diamond y)\big(1 + \epsilon_l(x \diamond y)\big) = \frac{x \diamond y}{1 - \eta_l(x \diamond y)}. \tag{2.26}$$

—

One can use the formulas of Rem. 2.17 to perform what is sometimes called a *forward analysis* of the rounding error. This technique is illustrated in the next example.

Example 2.18. Let $x := 0.9995 \cdot 10^0$ and $y := -0.9984 \cdot 10^0$. Computing the sum with precision 3 yields

$$\operatorname{rd}_3(x) +_3 \operatorname{rd}_3(y) = 0.100 \cdot 10^1 +_3 (-0.998 \cdot 10^0) = \operatorname{rd}_3(0.2 \cdot 10^{-2}) = 0.2 \cdot 10^{-2}.$$

Letting $\epsilon := \epsilon_3(\operatorname{rd}_3(x) + \operatorname{rd}_3(y))$, applying the formulas of Rem. 2.17 provides:

$$\begin{aligned}
\operatorname{rd}_3(x) +_3 \operatorname{rd}_3(y) &\overset{(2.26)}{=} \big(\operatorname{rd}_3(x) + \operatorname{rd}_3(y)\big)(1 + \epsilon) \\
&\overset{(2.25)}{=} \big(x(1 + \epsilon_3(x)) + y(1 + \epsilon_3(y))\big)(1 + \epsilon) \\
&= (x + y) + e_{\mathrm{a}},
\end{aligned} \tag{2.27}$$

where

$$e_{\mathrm{a}} = x\big(\epsilon + \epsilon_3(x)(1 + \epsilon)\big) + y\big(\epsilon + \epsilon_3(y)(1 + \epsilon)\big). \tag{2.28}$$

Using (2.23) and plugging in the numbers yields:

$$\begin{aligned}
\epsilon &= \frac{\operatorname{rd}_3(x) +_3 \operatorname{rd}_3(y) - \big(\operatorname{rd}_3(x) + \operatorname{rd}_3(y)\big)}{\operatorname{rd}_3(x) + \operatorname{rd}_3(y)} = \frac{0.002 - 0.002}{0.002} = 0,\\
\epsilon_3(x) &= \frac{1 - 0.9995}{0.9995} = \frac{1}{1999} = 0.00050\dots,\\
\epsilon_3(y) &= \frac{-0.998 + 0.9984}{-0.9984} = -\frac{1}{2496} = -0.00040\dots,\\
e_{\mathrm{a}} &= x\epsilon_3(x) + y\epsilon_3(y) = 0.0005 + 0.0004 = 0.0009.
\end{aligned}$$

The corresponding relative error is

$$e_{\mathrm{r}} = \frac{e_{\mathrm{a}}}{|x+y|} = \frac{0.0009}{0.0011} = \frac{9}{11} = 0.\overline{81}.$$

Thus, e_{r} is much larger than both $e_{\mathrm{r}}(x)$ and $e_{\mathrm{r}}(y)$. 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 2.19 and 2.20 below).

Example 2.19. Let us generalize the situation of Example 2.18 to general $x, y \in \mathbb{R}\setminus\{0\}$ and a general precision $l \in \mathbb{N}$. The formulas in (2.27) and (2.28) remain valid if one replaces 3 by l. Moreover, with $b = 10$ and $l \geq 2$, one obtains from (2.24):

$$|\epsilon| \leq 0.5 \cdot 10^{-l+1} \leq 0.5 \cdot 10^{-1} = 0.05.$$

Thus,

$$|e_{\mathrm{a}}| \leq |x|\big(\epsilon + 1.05|\epsilon_l(x)|\big) + |y|\big(\epsilon + 1.05|\epsilon_l(y)|\big),$$

and

$$e_{\mathrm{r}} = \frac{|e_{\mathrm{a}}|}{|x+y|} \leq \frac{|x|}{|x+y|}\big(|\epsilon| + 1.05|\epsilon_l(x)|\big) + \frac{|y|}{|x+y|}\big(|\epsilon| + 1.05|\epsilon_l(y)|\big).$$

One can now distinguish three (not completely disjoint) cases:

(a) If $|x+y| < \max\{|x|, |y|\}$ (in particular, if $\operatorname{sgn}(x) = -\operatorname{sgn}(y)$), then e_{r} is typically larger (potentially much larger, as in Example 2.18) than $|\epsilon_l(x)|$ and $|\epsilon_l(y)|$. Subtractive cancellation falls into this case. If possible, this should be avoided (cf. Example 2.20 below).

(b) If $\operatorname{sgn}(x) = \operatorname{sgn}(y)$, then $|x+y| = |x| + |y|$, implying

$$e_{\mathrm{r}} \leq |\epsilon| + 1.05 \max\big\{|\epsilon_l(x)|, |\epsilon_l(y)|\big\}$$

i.e. the relative error is at most of the same order of magnitude as the number $|\epsilon| + \max\big\{|\epsilon_l(x)|, |\epsilon_l(y)|\big\}$.

(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 *dampening*).

—

As the following Example 2.20 illustrates, subtractive cancellation can often be avoided by rearranging an expression into a mathematically equivalent formula that is numerically more stable.

Example 2.20. 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

$$x_1 = \frac{1}{2a}\left(-b - \operatorname{sgn}(b)\sqrt{b^2 - 4ac}\right), \quad x_2 = \frac{1}{2a}\left(-b + \operatorname{sgn}(b)\sqrt{b^2 - 4ac}\right).$$

If $|4ac| \ll b^2$, then, for the computation of x_2, one is in the situation of Example 2.19(a), i.e. the formula is numerically unstable. However, due to $x_1 x_2 = c/a$, one can use the equivalent formula

$$x_2 = \frac{2c}{-b - \operatorname{sgn}(b)\sqrt{b^2 - 4ac}},$$

where subtractive cancellation can not occur.

2.3 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 2.21. Let $D \subseteq \mathbb{R}$ and consider functions $f, g : D \longrightarrow \mathbb{R}$, where we assume $g(x) \neq 0$ for each $x \in D$. Moreover, let $x_0 \in \mathbb{R} \cup \{-\infty, \infty\}$ 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,

$$\limsup_{x \to x_0} \left| \frac{f(x)}{g(x)} \right| < \infty. \tag{2.29}$$

(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,

$$\lim_{x \to x_0} \left| \frac{f(x)}{g(x)} \right| = 0. \tag{2.30}$$

As mentioned before, O and o are known as *Landau symbols*.

Proposition 2.22. *We consider the setting of Def.* 2.21 *and provide the following equivalent characterizations of the Landau symbols:*

(a) *$f(x) = O(g(x))$ for $x \to x_0$ if, and only if, there exist $C, \delta > 0$ such that*

$$\left|\frac{f(x)}{g(x)}\right| \leq C \quad \textit{for each } x \in D \setminus \{x_0\} \textit{ with} \begin{cases} |x - x_0| < \delta & \textit{for } x_0 \in \mathbb{R}, \\ x > \delta & \textit{for } x_0 = \infty, \\ x < -\delta & \textit{for } x_0 = -\infty. \end{cases} \tag{2.31}$$

(b) *$f(x) = o(g(x))$ for $x \to x_0$ if, and only if, for every $C > 0$, there exists $\delta > 0$ such that* (2.31) *holds.*

Proof. (a): First, suppose that (2.29) holds, and let $M := \limsup_{x \to x_0} \left|\frac{f(x)}{g(x)}\right|$, $C := 1 + M$. Consider the case $x_0 \in \mathbb{R}$. If there were no $\delta > 0$ such that (2.31) holds, then, for each $\delta_n := 1/n$, $n \in \mathbb{N}$, there were some $x_n \in D \setminus \{x_0\}$ with $|x_n - x_0| < \delta_n$ and $y_n := \left|\frac{f(x_n)}{g(x_n)}\right| > C = 1 + M$, implying $\limsup_{n \to \infty} y_n \geq 1 + M > M$, in contradiction to (2.29). The cases $x_0 = \infty$ and $x_0 = -\infty$ are handled analogously (e.g., one can replace $\delta_n := 1/n$ by $\delta_n := n$ and $\delta_n := -n$, respectively). Conversely, if there exist $C, \delta > 0$ such that (2.31) holds true, then, for each sequence $(x_n)_{n \in \mathbb{N}}$ in $D \setminus \{x_0\}$ such that $\lim_{n \to \infty} x_n = x_0$, one has that the sequence $(y_n)_{n \in \mathbb{N}}$ with $y_n := \left|\frac{f(x_n)}{g(x_n)}\right|$ can not have a cluster point larger than C (only finitely many of the x_n are not in the neighborhood of x_0 determined by δ). Thus $\limsup_{n \to \infty} y_n \leq C$, thereby implying (2.29).

(b): There is nothing to prove, since the assertion that, for every $C > 0$, there exists $\delta > 0$ such that (2.31) holds, is precisely the definition of the notation in (2.30). ■

Proposition 2.23. *Again, we consider the setting of Def. 2.21. In addition, for use in some of the following assertions, we introduce functions $f_1, f_2, g_1, g_2 : D \longrightarrow \mathbb{R}$, where we assume $g_1(x) \neq 0$ and $g_2(x) \neq 0$ for each $x \in D$.*

(a) *Suppose $|f| \leq |g|$ in some neighborhood of x_0, i.e. there exists $\epsilon > 0$ such that*

$$|f(x)| \leq |g(x)| \quad \textit{for each } x \in D \setminus \{x_0\} \textit{ with} \begin{cases} |x - x_0| < \epsilon & \textit{for } x_0 \in \mathbb{R}, \\ x > \epsilon & \textit{for } x_0 = \infty, \\ x < -\epsilon & \textit{for } x_0 = -\infty. \end{cases} \tag{2.32}$$

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) *If $f(x) = O(g(x))$ (resp. $f(x) = o(g(x))$) for $x \to x_0$, and if $|f_1| \leq |f|$ and $|g| \leq |g_1|$ in some neighborhood of x_0 (i.e. if there exists $\epsilon > 0$ such that*

$$|f_1(x)| \leq |f(x)|,\ |g(x)| \leq |g_1(x)| \quad \textit{for each } x \in D \setminus \{x_0\}$$
$$\textit{with} \begin{cases} |x - x_0| < \epsilon & \textit{for } x_0 \in \mathbb{R}, \\ x > \epsilon & \textit{for } x_0 = \infty, \\ x < -\epsilon & \textit{for } x_0 = -\infty \end{cases} \Bigg),$$

then $f_1(x) = O(g_1(x))$ (resp. $f_1(x) = o(g_1(x))$) for $x \to x_0$.

(d) *$f(x) = o(g(x))$ for $x \to x_0$ implies $f(x) = O(g(x))$ for $x \to x_0$.*

(e) *If $\alpha \in \mathbb{R} \setminus \{0\}$, then for $x \to x_0$:*

$$\alpha f(x) = O(g(x)) \quad \Rightarrow \quad f(x) = O(g(x)), \tag{2.33a}$$
$$\alpha f(x) = o(g(x)) \quad \Rightarrow \quad f(x) = o(g(x)), \tag{2.33b}$$
$$f(x) = O(\alpha g(x)) \quad \Rightarrow \quad f(x) = O(g(x)), \tag{2.33c}$$
$$f(x) = o(\alpha g(x)) \quad \Rightarrow \quad f(x) = o(g(x)). \tag{2.33d}$$

(f) *For $x \to x_0$:*

$$f(x) = O(g_1(x)) \quad \textit{and} \quad g_1(x) = O(g_2(x)) \quad \textit{implies} \quad f(x) = O(g_2(x)), \tag{2.34a}$$
$$f(x) = o(g_1(x)) \quad \textit{and} \quad g_1(x) = o(g_2(x)) \quad \textit{implies} \quad f(x) = o(g_2(x)). \tag{2.34b}$$

(g) *For $x \to x_0$:*

$$f_1(x) = O(g(x)) \quad \textit{and} \quad f_2(x) = O(g(x)) \quad \textit{implies} \quad f_1(x) + f_2(x) = O(g(x)), \tag{2.35a}$$
$$f_1(x) = o(g(x)) \quad \textit{and} \quad f_2(x) = o(g(x)) \quad \textit{implies} \quad f_1(x) + f_2(x) = o(g(x)). \tag{2.35b}$$

(h) *For $x \to x_0$:*

$$f_1(x) = O(g_1(x)) \quad \textit{and} \quad f_2(x) = O(g_2(x)) \quad \textit{implies} \quad f_1(x)f_2(x) = O(g_1(x)g_2(x)), \tag{2.36a}$$
$$f_1(x) = o(g_1(x)) \quad \textit{and} \quad f_2(x) = o(g_2(x)) \quad \textit{implies} \quad f_1(x)f_2(x) = o(g_1(x)g_2(x)). \tag{2.36b}$$

(i) *For $x \to x_0$:*

$$f(x) = O(g_1(x)g_2(x)) \quad \Rightarrow \quad \frac{f(x)}{g_1(x)} = O(g_2(x)), \tag{2.37a}$$
$$f(x) = o(g_1(x)g_2(x)) \quad \Rightarrow \quad \frac{f(x)}{g_1(x)} = o(g_2(x)). \tag{2.37b}$$

Proof. (a): The hypothesis implies

$$\limsup_{x \to x_0} \left| \frac{f(x)}{g(x)} \right| \leq \limsup_{x \to x_0} \left| \frac{g(x)}{g(x)} \right| = 1 < \infty.$$

(b): If $f = g$, then

$$\limsup_{x \to x_0} \left| \frac{f(x)}{g(x)} \right| = \lim_{x \to x_0} \left| \frac{f(x)}{g(x)} \right| = 1.$$

Since $0 \neq 1 < \infty$, one has $f(x) = O(f(x))$ for $x \to x_0$, but not $f(x) = o(f(x))$ for $x \to x_0$.

(c): The assertions follow, for example, by applying Prop. 2.22: Since $\left|\frac{f_1(x)}{g_1(x)}\right| \leq \left|\frac{f(x)}{g(x)}\right|$ in the ϵ-neighborhood of x_0 as given in the hypothesis, for $f(x) = O(g(x))$ (resp. $f(x) = o(g(x))$), (2.31) is valid for f_1 and g_1 if one replaces δ by $\min\{\delta, \epsilon\}$ (where, for $f(x) = o(g(x))$, δ does, in general, depend on C).

(d): If the limit in (2.30) exists, then it coincides with the limit superior. It is therefore immediate that (2.30) implies (2.29).

(e): Everything follows immediately from the fact that multiplication with α and $1/\alpha$, respectively, commutes with taking the limit and the limit superior in (2.30) and (2.29), respectively.

(f): If, for $x \to x_0$, $f(x) = O(g_1(x))$ and $g_1(x) = O(g_2(x))$, then

$$0 \leq M_1 := \limsup_{x\to x_0} \left|\frac{f(x)}{g_1(x)}\right| < \infty \quad \text{and} \quad 0 \leq M_2 := \limsup_{x\to x_0} \left|\frac{g_1(x)}{g_2(x)}\right| < \infty,$$

implying

$$\limsup_{x\to x_0} \left|\frac{f(x)}{g_2(x)}\right| = \limsup_{x\to x_0} \left|\frac{f(x)}{g_1(x)}\right| \left|\frac{g_1(x)}{g_2(x)}\right| \leq M_1 M_2 < \infty.$$

If, for $x \to x_0$, $f(x) = o(g_1(x))$ and $g_1(x) = o(g_2(x))$, then

$$\lim_{x\to x_0} \left|\frac{f(x)}{g_2(x)}\right| = \lim_{x\to x_0} \left|\frac{f(x)}{g_1(x)}\right| \lim_{x\to x_0} \left|\frac{g_1(x)}{g_2(x)}\right| = 0.$$

(g): If, for $x \to x_0$, $f_1(x) = O(g(x))$ and $f_2(x) = O(g(x))$, then

$$\limsup_{x\to x_0} \left|\frac{f_1(x) + f_2(x)}{g(x)}\right| \leq \limsup_{x\to x_0} \left|\frac{f_1(x)}{g(x)}\right| + \limsup_{x\to x_0} \left|\frac{f_2(x)}{g(x)}\right| < \infty.$$

If, for $x \to x_0$, $f_1(x) = o(g(x))$ and $f_2(x) = o(g(x))$, then

$$\lim_{x\to x_0} \left|\frac{f_1(x) + f_2(x)}{g(x)}\right| = \lim_{x\to x_0} \left|\frac{f_1(x)}{g(x)}\right| + \lim_{x\to x_0} \left|\frac{f_2(x)}{g(x)}\right| = 0.$$

(h): If, for $x \to x_0$, $f_1(x) = O(g_1(x))$ and $f_2(x) = O(g_2(x))$, then

$$\limsup_{x\to x_0} \left|\frac{f_1(x)f_2(x)}{g_1(x)g_2(x)}\right| \leq \limsup_{x\to x_0} \left|\frac{f_1(x)}{g_1(x)}\right| \limsup_{x\to x_0} \left|\frac{f_2(x)}{g_2(x)}\right| < \infty.$$

If, for $x \to x_0$, $f_1(x) = o(g_1(x))$ and $f_2(x) = o(g_2(x))$, then

$$\lim_{x\to x_0} \left|\frac{f_1(x)f_2(x)}{g_1(x)g_2(x)}\right| = \lim_{x\to x_0} \left|\frac{f_1(x)}{g_1(x)}\right| \lim_{x\to x_0} \left|\frac{f_2(x)}{g_2(x)}\right| = 0.$$

(i): Trivial, since $\frac{f_1(x)}{g_1(x)}/g_2(x) = \frac{f_1(x)}{g_1(x)g_2(x)}$. ■

Example 2.24. (a) Consider a polynomial, i.e. $(a_0,\dots,a_n) \in \mathbb{R}^{n+1}$, $n \in \mathbb{N}_0$, and

$$P : \mathbb{R} \longrightarrow \mathbb{R}, \quad P(x) := \sum_{i=0}^{n} a_i x^i, \quad a_n \neq 0. \tag{2.38}$$

Then, for $p \in \mathbb{R}$,

$$P(x) = O(x^p) \qquad \text{for} \qquad x \to \infty \qquad \Leftrightarrow \qquad p \geq n, \tag{2.39a}$$

$$P(x) = o(x^p) \qquad \text{for} \qquad x \to \infty \qquad \Leftrightarrow \qquad p > n : \tag{2.39b}$$

For each $x \neq 0$:

$$\frac{P(x)}{x^p} = \sum_{i=0}^{n} a_i x^{i-p}. \tag{2.40}$$

Since

$$\lim_{x\to\infty} x^{i-p} = \begin{cases} \infty & \text{for } i-p > 0, \\ 1 & \text{for } i = p, \\ 0 & \text{for } i-p < 0, \end{cases} \tag{2.41}$$

(2.40) implies (2.39). Thus, in particular, for $x \to \infty$, each constant function is big O of 1: $a_0 = O(1)$.

(b) In the situation of Ex. 2.19(b), we found that the relative error e_r of rounding could be estimated according to

$$e_\mathrm{r} \leq 0.05 + 1.05 \max\big\{|\epsilon_l(x)|, |\epsilon_l(y)|\big\}.$$

Introducing $\max\big\{|\epsilon_l(x)|, |\epsilon_l(y)|\big\}$ as a new variable $\alpha \in \mathbb{R}_0^+$, one can restate the result as $e_\mathrm{r}(\alpha)$ is at most $O(\alpha)$ (for $\alpha \to \alpha_0$ for every $\alpha_0 > 0$).

Example 2.25. Recall the notion of differentiability: If G is an open subset of $\mathbb{R}^n$, $n \in \mathbb{N}$, $\xi \in G$, then $f : G \longrightarrow \mathbb{R}^m$, $m \in \mathbb{N}$, is called differentiable in ξ if, and only if, there exists a linear map $L : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ and another (not necessarily linear) map $r : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ such that

$$f(\xi + h) - f(\xi) = L(h) + r(h) \tag{2.42a}$$

for each $h \in \mathbb{R}^n$ with sufficiently small $\|h\|_2$, *and*

$$\lim_{h\to 0} \frac{r(h)}{\|h\|_2} = 0. \tag{2.42b}$$

Now, using the Landau symbol o, (2.42b) can be equivalently expressed as

$$\|r(h)\|_2 = o\big(\|h\|_2\big) \quad \text{for } \|h\|_2 \to 0. \tag{2.42c}$$

Example 2.26. 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

$$f(x) = T_m(x,a) + R_m(x,a), \tag{2.43a}$$

where

$$T_m(x,a) := \sum_{k=0}^{m} \frac{f^{(k)}(a)}{k!}(x-a)^k = f(a)+f'(a)(x-a)+\frac{f''(a)}{2!}(x-a)^2+\dots+\frac{f^{(m)}(a)}{m!}(x-a)^m \tag{2.43b}$$

is the mth *Taylor polynomial* and

$$R_m(x,a) := \frac{f^{(m+1)}(\theta)}{(m+1)!}(x-a)^{m+1} \quad \text{with some suitable } \theta \in]x,a[\tag{2.43c}$$

is the *Lagrange form* of the *remainder term*. Since the continuous function $f^{(m+1)}$ is bounded on each compact interval $[a,y]$, $y \in I$, (2.43) imply

$$f(x) - T_m(x,a) = O\big((x-a)^{m+1}\big) \tag{2.44}$$

for $x \to x_0$ for each $x_0 \in I$.

2.4 Operator Norms and Matrix Norms

When working in more general vector spaces than $\mathbb{R}$, errors are measured by more general norms than the absolute value (we already briefly encountered this in Example 2.25). If you are not sufficiently familiar with norms, you might want to consult the relevant subsections of [Phi14, Sec. 1]. A special class of norms of importance to us is the class of norms defined on linear maps between normed vector spaces. In terms of Numerical Analysis, we will mostly be interested in linear maps between $\mathbb{R}^n$ and $\mathbb{R}^m$, i.e. in linear maps given by real matrices. However, introducing the relevant notions for linear maps between general normed vector spaces does not provide much additional difficulty, and, hopefully, even some extra clarity.

Definition 2.27. Let $A : X \longrightarrow Y$ be a linear map between two normed vector spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$. 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 2.28. Let $A : X \longrightarrow Y$ be a linear map between two normed vector spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$. The number

$$\begin{aligned} \|A\| &:= \sup\left\{ \frac{\|Ax\|_Y}{\|x\|_X} : x \in X,\ x \neq 0 \right\} \\ &= \sup\big\{ \|Ax\|_Y : x \in X,\ \|x\|_X = 1 \big\} \in [0,\infty] \end{aligned} \tag{2.45}$$

is called the *operator norm* of A induced by $\|\cdot\|_X$ and $\|\cdot\|_Y$ (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 $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$, and A is given via a real $m \times n$ matrix, the operator norm is also called *matrix norm*.

—

From now on, the space index of a norm will usually be suppressed, i.e. we write just $\|\cdot\|$ instead of both $\|\cdot\|_X$ and $\|\cdot\|_Y$.

Remark 2.29. According to Th. B.1 of the Appendix, for a linear map $A : X \longrightarrow Y$ between two normed vector spaces X and Y, 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 $A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, $(n, m) \in \mathbb{N}^2$, is continuous (this follows, for example, from the fact that each such map is (trivially) differentiable, and every differentiable map is continuous). In particular, each linear map $A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, has all the above equivalent properties.

Theorem 2.30. *Let X and Y be normed vector spaces.*

(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

$$\big\|(\lambda A)x\big\| = \big\|\lambda(Ax)\big\| = |\lambda|\,\big\|Ax\big\|, \tag{2.46}$$

yielding

$$\begin{aligned}\|\lambda A\| &= \sup\big\{\|(\lambda A)x\| : x \in X,\ \|x\| = 1\big\} = \sup\big\{|\lambda|\,\|Ax\| : x \in X,\ \|x\| = 1\big\}\\ &= |\lambda| \sup\big\{\|Ax\| : x \in X,\ \|x\| = 1\big\} = |\lambda|\,\|A\|,\end{aligned} \tag{2.47}$$

showing that the operator norm is homogeneous of degree 1. Finally, if $A, B \in \mathcal{L}(X, Y)$ and $x \in X$, then

$$\|(A+B)x\| = \|Ax + Bx\| \leq \|Ax\| + \|Bx\|, \tag{2.48}$$

yielding

$$\begin{aligned}\|A+B\| &= \sup\big\{\|(A+B)x\| : x \in X,\ \|x\| = 1\big\}\\ &\leq \sup\big\{\|Ax\| + \|Bx\| : x \in X,\ \|x\| = 1\big\}\\ &\leq \sup\big\{\|Ax\| : x \in X,\ \|x\| = 1\big\} + \sup\big\{\|Bx\| : x \in X,\ \|x\| = 1\big\}\\ &= \|A\| + \|B\|,\end{aligned} \tag{2.49}$$

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

$$\|Ax - Ay\| \leq \|A\|\,\|x - y\| \text{ for each } x, y \in X. \tag{2.50}$$

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 (2.50). Now let $L \in \mathbb{R}_0^+$ 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 $\|A\| \leq L$. ■

Lemma 2.31. *If* $\operatorname{Id} : X \longrightarrow X$, $\operatorname{Id}(x) := x$, *is the identity map on a normed vector space* X, *then* $\|\operatorname{Id}\| = 1$ *(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 $\|\operatorname{Id}(x)\| = \|x\| = 1$. ■

Lemma 2.32. *Let* X, Y, Z *be normed vector spaces and consider linear maps* $A \in \mathcal{L}(X, Y)$, $B \in \mathcal{L}(Y, Z)$. *Then*

$$\|BA\| \leq \|B\|\,\|A\|. \tag{2.51}$$

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

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

thereby establishing the case. ■

Example 2.33. Let $m, n \in \mathbb{N}$ and let $A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be the linear map given by the $m \times n$ matrix $(a_{ij})_{(i,j)\in\{1,\dots,m\}\times\{1,\dots,n\}}$. Then

$$\|A\|_\infty := \max\left\{\sum_{j=1}^{n} |a_{ij}| : i \in \{1,\dots,m\}\right\} \tag{2.52a}$$

is called the *row sum norm* of A, and

$$\|A\|_1 := \max\left\{\sum_{i=1}^{m} |a_{ij}| : j \in \{1,\dots,n\}\right\} \tag{2.52b}$$

is called the *column sum norm* of A. It is an exercise to show that $\|A\|_\infty$ is the operator norm induced if $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with the ∞-norm, and $\|A\|_1$ is the operator norm induced if $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with the 1-norm.

Notation 2.34. Let $m, n \in \mathbb{N}$ and let $A = (a_{ij})_{(i,j)\in\{1,\dots,m\}\times\{1,\dots,n\}}$ be a real $m \times n$ matrix.

(a) By A^* we denote the *adjoint* matrix of A, and by $A^{\mathrm{t}} := (a_{ij})_{(j,i)\in\{1,\dots,n\}\times\{1,\dots,m\}}$ (an $n \times m$ matrix) we denote the *transpose* of A. Recall that, for real matrices, A^* and A^{t} are identical, but for complex matrices, one has $A^* = (\overline{a}_{ij})_{(j,i)\in\{1,\dots,n\}\times\{1,\dots,m\}}$, where $\overline{a}_{ij}$ is the complex conjugate of a_{ij}.

(b) If $m = n$, then

$$\operatorname{tr} A := \sum_{i=1}^{n} a_{ii}$$

denotes the *trace* of A.

Remark 2.35. Let $m, n \in \mathbb{N}$ and let $A = (a_{ij})_{(i,j)\in\{1,\dots,m\}\times\{1,\dots,n\}}$ be a real $m \times n$ matrix. Then:

$$\operatorname{tr}(A^*A) = \operatorname{tr}\left(\sum_{k=1}^{m} a_{ki}a_{kj}\right)_{(i,j)\in\{1,\dots,n\}^2} = \sum_{i=1}^{n}\sum_{k=1}^{m} |a_{ki}|^2, \tag{2.53a}$$

$$\operatorname{tr}(AA^*) = \operatorname{tr}\left(\sum_{k=1}^{n} a_{ik}a_{jk}\right)_{(i,j)\in\{1,\dots,m\}^2} = \sum_{i=1}^{m}\sum_{k=1}^{n} |a_{ik}|^2. \tag{2.53b}$$

Since the sums in (2.53a) and (2.53b) are identical, we have shown

$$\operatorname{tr}(A^*A) = \operatorname{tr}(AA^*). \tag{2.54}$$

Definition and Remark 2.36. Let $m, n \in \mathbb{N}$. If A is a real $m \times n$ matrix, then let $\|A\|_2$ denote the operator norm induced by the Euclidean norms on $\mathbb{R}^m$ and $\mathbb{R}^n$. This norm is also known as the *spectral norm* of A (cf. Th. 2.42 below).

Caveat: $\|A\|_2$ is *not(!)* the 2-norm on $\mathbb{R}^{mn}$, which is known as the *Frobenius* norm or the *Hilbert-Schmidt* norm (see Example B.9). As it turns out, for $n, m > 1$, the Frobenius norm is not induced by norms on $\mathbb{R}^m$ and $\mathbb{R}^n$ at all (for a proof see Appendix B).

Unfortunately, there is no formula as simple as the ones in (2.52) for the computation of the operator norm $\|A\|_2$ induced by the Euclidean norms. However, $\|A\|_2$ is often important, which is related to the fact that the Euclidean norm is induced by the Euclidean scalar product. We will see below, that the value of $\|A\|_2$ is related to the eigenvalues of A^*A.

—

We recall three important notions and then two important results from Linear Algebra:

Definition 2.37. Let $n \in \mathbb{N}$ and let $A = (a_{ij})$ be a real $n \times n$ matrix.

(a) A is called *symmetric* if, and only if, $A = A^{\mathrm{t}}$, i.e. if, and only if, $a_{ij} = a_{ji}$ for each $(i, j) \in \{1, \dots, n\}^2$.

(b) A is called *positive semidefinite* if, and only if, $x^{\mathrm{t}}Ax \geq 0$ for each $x \in \mathbb{R}^n$.

(c) A is called *positive definite* if, and only if, A is positive semidefinite and $x^{\mathrm{t}}Ax = 0 \Leftrightarrow x = 0$, i.e. if, and only if, $x^{\mathrm{t}}Ax > 0$ for each $0 \neq x \in \mathbb{R}^n$.

Theorem 2.38. *If $n \in \mathbb{N}$ and A is a real $n \times n$ matrix, then A has precisely n, in general* complex*, eigenvalues $\lambda_i \in \mathbb{C}$ if every eigenvalue is counted with its (algebraic) multiplicity (i.e. the multiplicity of the corresponding zero of the characteristic polynomial $\chi_A(\lambda) = \det(\lambda \operatorname{Id} - A)$ of A – it equals the geometric multiplicity (i.e. the dimension of the eigenspace $\ker(\lambda_i \operatorname{Id} - A)$) if, and only if, A is diagonalizable).* ■

Theorem 2.39. *If $n \in \mathbb{N}$ and A is a real, symmetric, and positive semidefinite $n \times n$ matrix, then all eigenvalues λ_i of A are real and nonnegative: $\lambda_i \in \mathbb{R}_0^+$. Moreover, there is an orthonormal basis $(v_1, \dots, v_n)$ of $\mathbb{R}^n$ such that v_i is an eigenvector for λ_i, i.e., in particular,*

$$Av_i = \lambda_i v_i, \tag{2.55a}$$

$$v_i^{\mathrm{t}} v_j = v_i \cdot v_j = \begin{cases} 0 & \text{for } i \neq j, \\ 1 & \text{for } i = j \end{cases} \tag{2.55b}$$

for each $i, j \in \{1, \dots, n\}$. ■

Lemma 2.40. *Let $m, n \in \mathbb{N}$ and let $A = (a_{ij})$ be a real $m \times n$ matrix. Then A^*A is a symmetric and positive semidefinite $n \times n$ matrix, whereas AA^* is a symmetric $m \times m$ matrix.*

Proof. Since $(A^*)^* = A$, it suffices to consider A^*A. That A^*A is a symmetric $n \times n$ matrix is immediately evident from the representation given in (2.53a) (since $a_{ki}a_{kj} = a_{kj}a_{ki}$). Moreover, if $x \in \mathbb{R}^n$, then $x^{\mathrm{t}}A^*Ax = (Ax)^{\mathrm{t}}(Ax) = \|Ax\|_2^2 \geq 0$, showing that A^*A is positive semidefinite. ■

Definition 2.41. In view of Th. 2.38, we define, for each $n \in \mathbb{N}$ and each real $n \times n$ matrix A:

$$r(A) := \max\big\{|\lambda| : \lambda \in \mathbb{C} \text{ and } \lambda \text{ is eigenvalue of } A\big\}. \tag{2.56}$$

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

Theorem 2.42. *Let $m, n \in \mathbb{N}$. If A is a real $m \times n$ matrix, then its spectral norm $\|A\|_2$ (i.e. the operator norm induced by the Euclidean norms on $\mathbb{R}^m$ and $\mathbb{R}^n$) satisfies*

$$\|A\|_2 = \sqrt{r(A^*A)}. \tag{2.57a}$$

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

$$\|A\|_2 = r(A). \tag{2.57b}$$

Proof. According to Lem. 2.40, A^*A is a symmetric and positive semidefinite $n \times n$ matrix. Then Th. 2.39 yields real nonnegative eigenvalues $\lambda_1, \dots, \lambda_n \geq 0$ and a corresponding orthonormal basis $(v_1, \dots, v_n)$ of eigenvectors satisfying (2.55) with A replaced by A^*A, in particular,

$$A^*Av_i = \lambda_i v_i \quad \text{for each } i \in \{1, \dots, n\}. \tag{2.58}$$

As $(v_1, \dots, v_n)$ is a basis of $\mathbb{R}^n$, for each $x \in \mathbb{R}^n$, there are numbers $x_1, \dots, x_n \in \mathbb{R}$ such that $x = \sum_{i=1}^n x_i v_i$. Thus, one computes

$$\begin{aligned}\|Ax\|_2^2 = (Ax) \cdot (Ax) = x^{\mathrm{t}} A^*Ax &= \left(\sum_{i=1}^n x_i v_i\right) \cdot \left(\sum_{i=1}^n x_i \lambda_i v_i\right) \\ &= \sum_{i=1}^n \lambda_i x_i^2 \leq r(A^*A) \sum_{i=1}^n x_i^2 = r(A^*A)\|x\|_2^2,\end{aligned} \tag{2.59}$$

proving $\|A\|_2 \leq \sqrt{r(A^*A)}$. To verify the remaining inequality, let $\lambda := r(A^*A)$ be the largest of the nonnegative λ_i, and let $v := v_i$ be the corresponding eigenvector from the orthonormal basis. Then $\|v\|_2 = 1$ and choosing $x = v$ in (2.59) yields $\|Av\|_2^2 = r(A^*A)$, thereby proving $\|A\|_2 \geq \sqrt{r(A^*A)}$ and completing the proof of (2.57a). It remains to consider the case where $A = A^*$. Then $A^*A = A^2$ and since

$$Av = \lambda v \quad \Rightarrow \quad A^2 v = \lambda^2 v,$$

Th. 2.38 implies

$$r(A^2) = r(A)^2.$$

Thus,

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

thereby establishing the case. ■

Caveat 2.43. It is *not* admissible to use the simpler formula (2.57b) for nonsymmetric $n \times n$ 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 2.44. *Let $n \in \mathbb{N}$. Then*

$$\|y\|_2 = \max\big\{v \cdot y : \, v \in \mathbb{R}^n \text{ and } \|v\|_2 = 1\big\} \quad \text{for each } y \in \mathbb{R}^n. \tag{2.60}$$

Proof. Let $v \in \mathbb{R}^n$ such that $\|v\|_2 = 1$. One estimates, using the Cauchy-Schwarz inequality [Phi14, (1.81)],

$$v \cdot y \leq |v \cdot y| \leq \|v\|_2 \|y\|_2 = \|y\|_2. \tag{2.61a}$$

Note that (2.60) is trivially true for $y = 0$. If $y \neq 0$, then letting $w := y/\|y\|_2$ yields $\|w\|_2 = 1$ and

$$w \cdot y = y \cdot y/\|y\|_2 = \|y\|_2. \tag{2.61b}$$

Together, (2.61a) and (2.61b) establish (2.60). ■

Proposition 2.45. *Let $m, n \in \mathbb{N}$. If A is a real $m \times n$ matrix, then $\|A\|_2 = \|A^*\|_2$ and, in particular, $r(A^*A) = r(AA^*)$. This allows to use the simpler (smaller) matrix of A^*A and AA^* to compute $\|A\|_2$ (see Example* 2.47 *below).*

Proof. One calculates

$$\begin{aligned}
\|A\|_2 &= \max\big\{\|Ax\|_2 : \, x \in \mathbb{R}^n \text{ and } \|x\|_2 = 1\big\} \\
&\overset{(2.60)}{=} \max\big\{v \cdot Ax : \, v, x \in \mathbb{R}^n \text{ and } \|v\|_2 = \|x\|_2 = 1\big\} \\
&= \max\big\{(A^*v) \cdot x : \, v, x \in \mathbb{R}^n \text{ and } \|v\|_2 = \|x\|_2 = 1\big\} \\
&\overset{(2.60)}{=} \max\big\{\|A^*v\|_2 : \, v \in \mathbb{R}^n \text{ and } \|v\|_2 = 1\big\} \\
&= \|A^*\|_2,
\end{aligned}$$

proving the proposition. ■

Remark 2.46. One can actually show more than we did in Prop. 2.45: 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.

Example 2.47. Consider $A := \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$. One obtains

$$AA^* = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad A^*A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

So one would tend to compute the eigenvalues using AA^*. One finds $\lambda_1 = 1$ and $\lambda_2 = 3$. Thus, $r(AA^*) = 3$ and $\|A\|_2 = \sqrt{3}$.

2.5 Condition of a Problem

We are now ready to exploit 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. $\mathbb{R}^n$) and the output is likewise a vector lying in some normed vector space Y (e.g. $\mathbb{R}^m$). 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 2.48. Let X and Y be normed vector spaces, $U \subseteq X$ open, and $f : U \longrightarrow Y$. Fix $x \in U \setminus \{0\}$ such that $f(x) \neq 0$ and $\delta > 0$ such that $B_\delta(x) = \{\tilde{x} \in X : \|\tilde{x} - x\| < \delta\} \subseteq U$. 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

$$\frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \leq K \, \frac{\|\tilde{x} - x\|}{\|x\|} \quad \text{for every } \tilde{x} \in B_\delta(x). \tag{2.62}$$

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 (2.62) 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 2.49. We remain in the situation of Def. 2.48 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:

$$k_{\mathrm{rel}} := k_{\mathrm{rel}}(f, x) := \lim_{\alpha \to 0} K(\alpha). \tag{2.63}$$

The number $k_{\mathrm{rel}} \in \mathbb{R}_0^+$ is called the *relative condition* of (the problem represented by the solution operator) f at x.

Remark 2.50. The smaller k_{rel}, the more well-behaved is the problem in the vicinity of x, where stability corresponds more or less to $k_{\mathrm{rel}} < 100$.

Remark 2.51. Let us compare the newly introduced notion of f being well-conditioned with some other regularity notions for f. For example, if f is Lipschitz continuous in $B_\delta(x)$, then f is clearly well-conditioned in $B_\delta(x)$. The converse is not true as, for example, $t \mapsto \sqrt{t}$ is well-conditioned in $B_1(1) =]0, 2[$, but not Lipschitz continuous on $B_1(1)$. On the other hand, (2.62) does imply continuity in x. Once again, the converse is not true as, for example, $t \mapsto 1 + \sqrt{|t - 1|}$ is continuous in 1, but it is not well-conditioned in any $B_\delta(1)$ with $\delta > 0$. 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 injective and everywhere well-conditioned, then it is also 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 and we didn't allow $x = 0$ and $f(x) = 0$, whereas well-posedness was defined as a *global* property.

Theorem 2.52. *Let $m, n \in \mathbb{N}$, let $U \subseteq \mathbb{R}^n$ be open, and assume that $f : U \longrightarrow \mathbb{R}^m$ 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*

$$k_{\text{rel}} = \|Df(x)\| \frac{\|x\|}{\|f(x)\|}, \tag{2.64}$$

where $Df(x) : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is the (total) derivative of f in x (a linear map represented by the Jacobian $J_f(x)$). In (2.64), any norm on $\mathbb{R}^n$ and $\mathbb{R}^m$ will work as long as $\|Df(x)\|$ is the corresponding induced operator norm.

Proof. Choose $\delta > 0$ sufficiently small such that $\overline{B}_\delta(x) \subseteq U$. 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. [Phi14, Th. 3.15], which can be applied to the components $f_1, \dots, f_m$ of f):

$$f(x+h) = f(x) + \int_0^1 Df\big(x + th\big)(h)\,\mathrm{d}t\,. \tag{2.65}$$

If $\tilde{x} \in B_\delta(x)$, then we can let $h := \tilde{x} - x$, which permits to restate (2.65) in the form

$$f(\tilde{x}) - f(x) = \int_0^1 Df\big((1-t)x + t\tilde{x}\big)(\tilde{x} - x)\,\mathrm{d}t\,. \tag{2.66}$$

This allows to estimate the norm as follows:

$$\begin{aligned}
\big\|f(\tilde{x}) - f(x)\big\| &= \left\| \int_0^1 Df\big((1-t)x + t\tilde{x}\big)(\tilde{x} - x)\,\mathrm{d}t \right\| \\
&\overset{\text{Th. C.5}}{\leq} \int_0^1 \Big\| Df\big((1-t)x + t\tilde{x}\big)(\tilde{x} - x) \Big\|\,\mathrm{d}t \\
&\leq \int_0^1 \Big\| Df\big((1-t)x + t\tilde{x}\big) \Big\|\, \|\tilde{x} - x\|\,\mathrm{d}t \\
&\leq \sup\Big\{ \big\|Df(y)\big\| : y \in B_\delta(x) \Big\} \|\tilde{x} - x\| \\
&= S(\delta)\|\tilde{x} - x\|
\end{aligned} \tag{2.67}$$

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

$$\frac{\big\|f(\tilde{x}) - f(x)\big\|}{\big\|f(x)\big\|} \leq \frac{\|x\|}{\big\|f(x)\big\|} S(\delta) \frac{\|\tilde{x} - x\|}{\|x\|}$$

and, therefore,

$$K(\delta) \leq \frac{\|x\|}{\big\|f(x)\big\|} S(\delta). \tag{2.68}$$

The hypothesis that f is continuously differentiable implies $\lim_{y \to x} Df(y) = Df(x)$, which implies $\lim_{y \to x} \|Df(y)\| = \|Df(x)\|$ (continuity of the norm, cf. [Phi14, Ex. 1.53]), which, in turn, implies $\lim_{\delta\to 0} S(\delta) = \|Df(x)\|$. Since, also, $\lim_{\delta\to 0} K(\delta) = k_{\text{rel}}$, (2.68) yields

$$k_{\text{rel}} \leq \|Df(x)\| \, \frac{\|x\|}{\|f(x)\|}. \tag{2.69}$$

In still remains to prove the inverse inequality. To that end, choose $v \in \mathbb{R}^n$ such that $\|v\| = 1$ and

$$\big\|Df(x)(v)\big\| = \big\|Df(x)\big\| \tag{2.70}$$

(such a vector v must exist due to the fact that the continuous function $y \mapsto \|Df(x)(y)\|$ must attain its max on the compact unit sphere $S_1(0)$ – note that this argument does not work in infinite-dimensional spaces, where $S_1(0)$ is *not* compact). For $0 < \epsilon < \delta$ consider $\tilde{x} := x + \epsilon v$. Then $\|\tilde{x} - x\| = \epsilon < \delta$, i.e. $\tilde{x} \in B_\delta(x)$. In particular, $\tilde{x}$ is admissible in (2.66), which provides

$$\begin{aligned} f(\tilde{x}) - f(x) &= \epsilon \int_0^1 Df\big((1-t)x + t\tilde{x}\big)(v)\,\mathrm{d}t \\ &= \epsilon Df(x)(v) + \epsilon \int_0^1 \Big(Df\big((1-t)x + t\tilde{x}\big) - Df(x)\Big)(v)\,\mathrm{d}t\,. \end{aligned} \tag{2.71}$$

Once again using $\|\tilde{x} - x\| = \epsilon$ as well as the triangle inequality in the form $\|a + b\| \geq \|a\| - \|b\|$, we obtain from (2.71) and (2.70):

$$\begin{aligned} &\frac{\big\|f(\tilde{x}) - f(x)\big\|}{\big\|f(x)\big\|} \\ &\geq \frac{\|x\|}{\big\|f(x)\big\|} \left(\big\|Df(x)\big\| - \int_0^1 \Big\|Df\big((1-t)x + t\tilde{x}\big) - Df(x)\Big\|\,\mathrm{d}t \right) \frac{\|\tilde{x} - x\|}{\|x\|}. \end{aligned} \tag{2.72}$$

Thus,

$$K(\epsilon) \geq \frac{\|x\|}{\big\|f(x)\big\|} \Big(\big\|Df(x)\big\| - T(\epsilon) \Big), \tag{2.73}$$

where

$$T(\epsilon) := \sup \left\{ \int_0^1 \Big\|Df\big((1-t)x + t\tilde{x}\big) - Df(x)\Big\|\,\mathrm{d}t \;:\; \tilde{x} \in \overline{B}_\epsilon(x) \right\}. \tag{2.74}$$

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 $\tilde{x} \in \overline{B}_\epsilon(x)$ and $t \in [0,1]$, one obtains

$$T(\epsilon) \leq \int_0^1 \alpha\,\mathrm{d}t \, = \alpha,$$

implying

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

In particular, we can take limits in (2.73) to get

$$k_{\text{rel}} \geq \frac{\|x\|}{\big\|f(x)\big\|} \big\|Df(x)\big\|. \tag{2.75}$$

Finally, (2.75) together with (2.69) completes the proof of (2.64). ■

Example 2.53. (a) Let us investigate the condition of the problem of multiplication, by considering

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}, \quad f(x,y) := xy, \quad Df(x,y) = (y,x). \tag{2.76}$$

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

$$k_{\text{rel}}(x,y) = \frac{|x|}{|y|} + \frac{|y|}{|x|}. \tag{2.77}$$

Thus, we see that the relative condition explodes if $|x| \gg |y|$ or $|x| \ll |y|$. Since one can also show (exercise) that f is well-conditioned in $B_\delta(x,y)$ for each $\delta > 0$, 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 division, i.e.

$$g : \mathbb{R} \times \big(\mathbb{R} \setminus \{0\}\big) \longrightarrow \mathbb{R}, \quad g(x,y) := \frac{x}{y}, \quad Dg(x,y) = \left(\frac{1}{y}, -\frac{x}{y^2}\right). \tag{2.78}$$

Once again using the Euclidean norm on $\mathbb{R}^2$, one obtains from (2.64) that, for each $(x,y) \in \mathbb{R}^2$ such that $xy \neq 0$,

$$\begin{aligned} k_{\text{rel}}(x,y) &= \|Dg(x,y)\|_2 \frac{\|(x,y)\|_2}{|g(x,y)|} = \frac{|y|}{|x|} \sqrt{x^2+y^2} \sqrt{\frac{1}{y^2} + \frac{x^2}{y^4}} \\ &= \frac{1}{|x|} \sqrt{x^2+y^2} \sqrt{1 + \frac{x^2}{y^2}} = \frac{x^2+y^2}{|x||y|} = \frac{|x|}{|y|} + \frac{|y|}{|x|}, \end{aligned} \tag{2.79}$$

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 $B_\delta(x,y)$, $\delta > 0$, 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 $B_\delta(x,y)$: 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

$$\left|\frac{x}{y} - \frac{xn}{y}\right| = (n-1)\frac{|x|}{|y|} \to \infty \quad \text{for } n \to \infty.$$

For $0 < \delta < |y|$, the result is different: Suppose $|\tilde{y}| \leq |y| - \delta$. Then

$$\big\|(x,y) - (\tilde{x},\tilde{y})\big\|_2 \geq |\tilde{y} - y| \geq |y| - |\tilde{y}| \geq \delta.$$

Thus, $(\tilde{x},\tilde{y}) \in B_\delta(x,y)$ implies $|\tilde{y}| > |y| - \delta$. In consequence, one estimates, for each

$(\tilde{x}, \tilde{y}) \in B_\delta(x, y)$,

$$\begin{aligned}
\big|g(x,y) - g(\tilde{x},\tilde{y})\big| = \left|\frac{x}{y} - \frac{\tilde{x}}{\tilde{y}}\right| \leq \left|\frac{x}{y} - \frac{x}{\tilde{y}}\right| + \left|\frac{x}{\tilde{y}} - \frac{\tilde{x}}{\tilde{y}}\right| &= \left|\frac{x}{y\tilde{y}}\right| |y - \tilde{y}| + \left|\frac{1}{\tilde{y}}\right| |x - \tilde{x}| \\
&\leq \max\left\{\frac{|x|}{|y|(|y| - \delta)}, \frac{1}{|y| - \delta}\right\} \big\|(x,y) - (\tilde{x},\tilde{y})\big\|_1 \\
&\leq C \max\left\{\frac{|x|}{|y|(|y| - \delta)}, \frac{1}{|y| - \delta}\right\} \big\|(x,y) - (\tilde{x},\tilde{y})\big\|_2
\end{aligned}$$

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 $k_{\mathrm{rel}}(x, y)$ alone is not enough to determine the numerical stability of g at (x, y): Even though $k_{\mathrm{rel}}(x, y)$ 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 $k_{\mathrm{rel}}(x, y)$ can be completely useless and even misleading.

As another important example, we consider the problem of solving the linear system $Ax = b$ with an invertible real $n \times n$ matrix A. 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 2.54. 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 shortly 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:

$$\lambda_1 \approx 0.010, \quad \lambda_2 \approx 0.843, \quad \lambda_3 \approx 3.86, \quad \lambda_4 \approx 30.3, \quad \frac{\lambda_4}{\lambda_1} \approx 2984.$$

For our systematic investigation of this problem, we now apply (2.64) to the general situation:

Example 2.55. Let A be an invertible real $n \times n$ matrix, $n \in \mathbb{N}$ and consider the problem of solving the linear system $Ax = b$. Then the solution operator is

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}^n, \quad f(b) := A^{-1}b, \quad Df(b) = A^{-1}.$$

Fixing some norm on $\mathbb{R}^n$ and using the induced matrix norm, we obtain from (2.64) that, for each $0 \neq b \in \mathbb{R}^n$:

$$k_{\text{rel}}(b) = \|Df(b)\| \frac{\|b\|}{\|f(b)\|} = \|A^{-1}\| \frac{\|b\|}{\|A^{-1}b\|} \overset{x:=A^{-1}b}{=} \|A^{-1}\| \frac{\|Ax\|}{\|x\|} \leq \|A\|\|A^{-1}\|. \tag{2.80}$$

Definition and Remark 2.56. In view of (2.80), we define the *condition number* $\kappa(A)$ (also just called the *condition*) of an invertible real $n \times n$ matrix A, $n \in \mathbb{N}$, by

$$\kappa(A) := \|A\|\|A^{-1}\|, \tag{2.81}$$

where $\|\cdot\|$ denotes a matrix norm induced by some norm on $\mathbb{R}^n$. Then one immediately gets from (2.80) that

$$k_{\text{rel}}(b) \leq \kappa(A) \quad \text{for each } b \in \mathbb{R}^n \setminus \{0\}. \tag{2.82}$$

The condition number clearly depends on the underlying matrix norm. If the matrix norm is the spectral norm, then one calls the condition number the *spectral condition* and one sometimes writes κ_2 instead of κ.

Notation 2.57. For each $n \times n$ matrix A, $n \in \mathbb{N}$, let

$$\lambda(A) := \{\lambda \in \mathbb{C} : \lambda \text{ is eigenvalue of } A\}, \tag{2.83a}$$

$$|\lambda(A)| := \{|\lambda| : \lambda \in \mathbb{C} \text{ is eigenvalue of } A\}, \tag{2.83b}$$

denote the set of eigenvalues of A and the set of absolute values of eigenvalues of A, respectively.

Lemma 2.58. *For the spectral condition of an invertible real $n\times n$ matrix A, one obtains*

$$\kappa_2(A) = \sqrt{\frac{\max\lambda(A^*A)}{\min\lambda(A^*A)}} \tag{2.84}$$

(recall that all eigenvalues of A^*A are real and positive). Moreover, if A is symmetric, then (2.84) *simplifies to*

$$\kappa_2(A) = \frac{\max|\lambda(A)|}{\min|\lambda(A)|} \tag{2.85}$$

(note that $\min|\lambda(A)| > 0$ *for each invertible A).*

Proof. Exercise. ■

Theorem 2.59. *Let A be an invertible real $n \times n$ matrix, $n \in \mathbb{N}$. Assume that $x, b, \Delta x, \Delta b \in \mathbb{R}^n$ satisfy*

$$Ax = b, \quad A(x + \Delta x) = b + \Delta b, \tag{2.86}$$

i.e. Δb can be seen as a perturbation of the input and Δx can be seen as the resulting perturbation of the output. One then has the following estimates for the absolute and relative errors:

$$\|\Delta x\| \le \|A^{-1}\|\|\Delta b\|, \quad \frac{\|\Delta x\|}{\|x\|} \le \kappa(A)\,\frac{\|\Delta b\|}{\|b\|}, \tag{2.87}$$

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

Proof. Since A is linear, (2.86) implies $A(\Delta x) = \Delta b$, i.e. $\Delta x = A^{-1}(\Delta b)$, which already yields the first estimate in (2.87). For $x, b \neq 0$, the first estimate together with $Ax = b$ easily implies the second:

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

thereby establishing the case. ■

Example 2.60. 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 ∞-norm. From (2.87), we know

$$e_{\mathrm{r}} \le \kappa(A)\,\frac{\|\Delta b\|_\infty}{\|b\|_\infty} = \kappa(A)\,\frac{\|\Delta b\|_\infty}{4}.$$

Moreover, since $A^{-1} = \begin{pmatrix} -1 & 2 \\ 1 & -1 \end{pmatrix}$, from (2.81) and (2.52a), we obtain

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

Thus, if the perturbation Δb satisfies $\|\Delta b\|_\infty < 4/900 \approx 0.0044$, then $e_{\mathrm{r}} < \frac{9}{4}\cdot\frac{4}{900} = 10^{-2}$.

Remark 2.61. Note that (2.87) is a much stronger and more useful statement than the $k_{\text{rel}}(b) \leq \kappa(A)$ of (2.80). 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, (2.87) 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 Δb).

—

Let us come back to the instability observed in Example 2.54 and determine its actual cause. Suppose A is any symmetric invertible real $n \times n$ matrix, $\lambda_{\min}, \lambda_{\max} \in \lambda(A)$ such that $|\lambda_{\min}| = \min |\lambda(A)|$, $|\lambda_{\max}| = \max |\lambda(A)|$. 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 Example 2.54.

—

Before concluding this section, we study a generalization of the problem from (2.86). Now, in addition to a perturbation of the right-hand side b, we also allow a perturbation of the matrix A by a matrix ΔA:

$$Ax = b, \quad (A + \Delta A)(x + \Delta x) = b + \Delta b. \tag{2.88}$$

We would now like to control Δx in terms of Δb and ΔA.

Remark 2.62. The set of invertible real $n \times n$ matrices is open in the set of all real $n \times n$ matrices (which is the same as $\mathbb{R}^{n^2}$, where all norms are equivalent) – this follows, for example, from the determinant being a continuous map (even a polynomial) from $\mathbb{R}^{n^2}$ into $\mathbb{R}$, which implies that $\det^{-1}(\mathbb{R} \setminus \{0\})$ must be open. Thus, if A in (2.88) is invertible and $\|\Delta A\|$ is sufficiently small, then $A + \Delta A$ must also be invertible.

Lemma 2.63. *Let $n \in \mathbb{N}$, consider some norm on $\mathbb{R}^n$ and the induced matrix norm on the set of real $n \times n$ matrices. If B is a real $n \times n$ matrix such that $\|B\| < 1$, then* $\mathrm{Id} + B$ *is invertible and*

$$\|(\mathrm{Id} + B)^{-1}\| \leq \frac{1}{1 - \|B\|}. \tag{2.89}$$

Proof. For each $x \in \mathbb{R}^n$:

$$\|(\operatorname{Id}+B)x\| \geq \|x\| - \|Bx\| \geq \|x\| - \|B\|\,\|x\| = (1-\|B\|)\,\|x\|, \tag{2.90}$$

showing that $x \neq 0$ implies $(\operatorname{Id}+B)x \neq 0$, i.e. $\operatorname{Id}+B$ is invertible. Fixing $y \in \mathbb{R}^n$ and applying (2.90) with $x := (\operatorname{Id}+B)^{-1}y$ yields

$$\|(\operatorname{Id}+B)^{-1}y\| \leq \frac{1}{1-\|B\|}\,\|y\|,$$

thereby proving (2.89) and concluding the proof of the lemma. ■

Lemma 2.64. *Let $n \in \mathbb{N}$, consider some norm on $\mathbb{R}^n$ and the induced matrix norm on the set of real $n \times n$ matrices. If A and ΔA are real $n \times n$ matrices such A is invertible and $\|\Delta A\| < \|A^{-1}\|^{-1}$, then $A + \Delta A$ is invertible and*

$$\|(A+\Delta A)^{-1}\| \leq \frac{1}{\|A^{-1}\|^{-1} - \|\Delta A\|}. \tag{2.91}$$

Moreover,

$$\|\Delta A\| \leq \frac{1}{2\|A^{-1}\|} \quad \Rightarrow \quad \|(A+\Delta A)^{-1} - A^{-1}\| \leq C\|\Delta A\|, \tag{2.92}$$

where the constant can be chosen as $C := 2\|A^{-1}\|^2$.

Proof. We can write $A + \Delta A = A(\operatorname{Id}+A^{-1}\Delta A)$. As $\|A^{-1}\Delta A\| \leq \|A^{-1}\|\,\|\Delta A\| < 1$, Lem. 2.63 yields that $\operatorname{Id}+A^{-1}\Delta A$ is invertible. Since A is also invertible, so is $A+\Delta A$. Moreover,

$$\|(A+\Delta A)^{-1}\| = \left\|(\operatorname{Id}+A^{-1}\Delta A)^{-1}A^{-1}\right\| \overset{(2.89)}{\leq} \frac{\|A^{-1}\|}{1-\|A^{-1}\Delta A\|} \leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\,\|\Delta A\|}, \tag{2.93}$$

proving (2.91). One easily verifies the identity

$$(A+\Delta A)^{-1} - A^{-1} = -(A+\Delta A)^{-1}(\Delta A)A^{-1}, \tag{2.94}$$

which yields, for $\|\Delta A\| \leq \frac{1}{2\|A^{-1}\|}$,

$$\begin{aligned}\|(A+\Delta A)^{-1} - A^{-1}\| &\leq \|(A+\Delta A)^{-1}\|\,\|\Delta A\|\,\|A^{-1}\| \overset{(2.93)}{\leq} \frac{\|A^{-1}\|\,\|\Delta A\|\,\|A^{-1}\|}{1-\|A^{-1}\|\,\|\Delta A\|}\\ &\leq 2\,\|A^{-1}\|^2\,\|\Delta A\|,\end{aligned}$$

proving (2.92). ■

Theorem 2.65. *As in Lem. 2.64, let $n \in \mathbb{N}$, consider some norm on $\mathbb{R}^n$ and the induced matrix norm on the set of real $n \times n$ matrices, and let A and ΔA be real $n \times n$ matrices such A is invertible and $\|\Delta A\| < \|A^{-1}\|^{-1}$. Moreover, assume $b, x, \Delta b, \Delta x \in \mathbb{R}^n$ satisfy*

$$Ax = b, \quad (A+\Delta A)(x+\Delta x) = b + \Delta b. \tag{2.95}$$

Then one has the following bound for the relative error:

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1-\|A^{-1}\|\,\|\Delta A\|}\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right). \tag{2.96}$$

Proof. The equations (2.95) imply

$$(A + \Delta A)(\Delta x) = \Delta b - (\Delta A)x. \tag{2.97}$$

In consequence, from (2.91) and $\|b\| \leq \|A\|\,\|x\|$, one obtains

$$\begin{aligned}
\frac{\|\Delta x\|}{\|x\|} &\overset{(2.97)}{=} \frac{\left\|(A+\Delta A)^{-1}\big(\Delta b - (\Delta A)x\big)\right\|}{\|x\|} \\
&\overset{(2.91)}{\leq} \frac{1}{\|A^{-1}\|^{-1} - \|\Delta A\|}\left(\|\Delta A\| + \frac{\|\Delta b\|\|A\|}{\|b\|}\right) \\
&= \frac{\kappa(A)}{1 - \|A^{-1}\|\,\|\Delta A\|}\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right),
\end{aligned}$$

thereby establishing the case. ■

3 Interpolation

3.1 Motivation

Given a finite number of points $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$, 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 $i \in \{0, \dots, n\}$. Then f is called an *interpolate* of the data points. We will restrict ourselves to the case of functions that map (a subset of) $\mathbb{R}$ into $\mathbb{R}$, even though the same problem is also of interest in other contexts, e.g. in higher dimensions. 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 in an arbitrary point.

(b) If f is differentiable, derivatives can be computed in an arbitrary point.

(c) Computation of integrals of the form $\int_a^b f(x)\,\mathrm{d}x$.

(d) Determination of extreme points of f.

(e) Determination of zeros of f.

While a general function f on a nontrivial real interval 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 or that it can at least be approximated by functions from a certain 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, whereas interpolation by rational functions is often desirable if singularities are expected. Due to time constraints, we will only study interpolations related to polynomials in the present lecture, 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

3.2.1 Existence and Uniqueness

Definition and Remark 3.1. For $n \in \mathbb{N}_0$, let $\mathcal{P}_n$ denote the set of all polynomials $p : \mathbb{R} \longrightarrow \mathbb{R}$ of degree at most n. Then $\mathcal{P}_n$ constitutes an $(n+1)$-dimensional real vector space: A linear isomorphism is given by the map

$$f : \mathbb{R}^{n+1} \longrightarrow \mathcal{P}_n, \quad f(a_0, a_1, \dots, a_n) := a_0 + a_1 x + \dots + a_n x^n. \tag{3.1}$$

Theorem 3.2. *Let $n \in \mathbb{N}_0$. Given $n+1$ data points $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$, $x_i \neq x_j$ for $i \neq j$, there is a unique interpolating polynomial $p \in \mathcal{P}_n$ satisfying*

$$p(x_i) = y_i \quad \textit{for each } i \in \{0, 1, \dots, n\}. \tag{3.2}$$

Moreover, one can identify p as the Lagrange interpolating polynomial*, which is given by the explicit formula*

$$p(x) = p_n(x) := \sum_{j=0}^{n} y_j \, L_j(x), \tag{3.3a}$$

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

$$L_j(x) := \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i} \tag{3.3b}$$

are the so-called Lagrange basis polynomials*.*

Proof. If we write p in the form $p(x) = a_0 + a_1 x + \dots + a_n x^n$, then (3.2) is equivalent to the system

$$\begin{aligned} a_0 + a_1 x_0 + \dots + a_n x_0^n &= y_0 \\ a_0 + a_1 x_1 + \dots + a_n x_1^n &= y_1 \\ &\vdots \\ a_0 + a_1 x_n + \dots + a_n x_n^n &= y_n, \end{aligned} \tag{3.4}$$

which constitutes a linear system for the $n+1$ unknowns $a_0, \dots, a_n$. This linear system has a unique solution if, and only if, the determinant

$$D := \begin{vmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{vmatrix} \tag{3.5}$$

does not vanish. One observes that D is the well-known Vandermonde determinant, and its value is given by (see Th. D.1 of the Appendix)

$$D = \prod_{\substack{i,j=0 \\ i>j}}^{n} (x_i - x_j). \tag{3.6}$$

In particular, $D \neq 0$, as we assumed $x_i \neq x_j$ for $i \neq j$, proving both existence and uniqueness of p. It remains to identify p with the Lagrange interpolating polynomial. To that end, denote the right-hand side of (3.3a) by p_n. Since p_n is clearly of degree less than or equal to n, it merely remains to show that p_n satisfies (3.2). Since the x_i are all distinct, we obtain, for each $j, k \in \{0, \dots, n\}$,

$$L_j(x_k) = \delta_{jk} := \begin{cases} 1 & \text{for } k = j, \\ 0 & \text{for } k \neq j, \end{cases} \tag{3.7}$$

which, in turn, yields

$$p_n(x_k) = \sum_{j=0}^{n} y_j \, L_j(x_k) = \sum_{j=0}^{n} y_j \, \delta_{jk} = y_k. \tag{3.8}$$

A second proof of the theorem can be conducted as follows, without making use of the Vandermonde determinant: As above, the direct verification that the Lagrange interpolating polynomial satisfies (3.2) proves existence. Then, for uniqueness, one considers an arbitrary polynomial $q \in \mathcal{P}_n$, satisfying (3.2). Since $q(x_i) = y_i = p_n(x_i)$, the polynomial $r := q - p_n$ is a polynomial of degree at most n having at least $n+1$ zeros. The only such polynomial is $r \equiv 0$, showing $q = p_n$. ■

Example 3.3. We compute the first three Lagrange polynomials. Thus, let

$$(x_0, y_0), (x_1, y_1), (x_2, y_2) \in \mathbb{R}^2$$

be three distinct data points. For $n = 0$, the product in (3.3b) is empty, such that $L_0(x) \equiv 1$ and $p_0 \equiv y_0$. For $n = 1$, (3.3b) yields

$$L_0(x) = \frac{x - x_1}{x_0 - x_1}, \quad L_1(x) = \frac{x - x_0}{x_1 - x_0}, \tag{3.9a}$$

and

$$p_1(x) = y_0 L_0(x) + y_1 L_1(x) = y_0 \, \frac{x - x_1}{x_0 - x_1} + y_1 \, \frac{x - x_0}{x_1 - x_0} = y_0 + \frac{y_1 - y_0}{x_1 - x_0} \, (x - x_0), \tag{3.9b}$$

which is the straight line through (x_0, y_0) and (x_1, y_1). Finally, for $n = 2$, (3.3b) yields

$$\begin{aligned} L_0(x) &= \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}, \quad L_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}, \\ L_2(x) &= \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}, \end{aligned} \tag{3.10a}$$

and

$$\begin{aligned} p_2(x) &= y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x) \\ &= y_0 + \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + \frac{1}{x_2 - x_0}\left(\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0}\right)(x - x_0)(x - x_1). \end{aligned} \tag{3.10b}$$

3.2.2 Newton's Divided Difference Interpolation Formula

As (3.9b) and (3.10b), the interpolating polynomial can be built recursively. 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 $x_0, \dots, x_n$, then, using the recursive formula (3.11a) below, one obtains the interpolating polynomial for $x_0, \dots, x_n, x_{n+1}$ in just $O(n)$ steps.

Looking at the structure of p_2 in (3.10b), one sees that, while the first coefficient is just y_0, the second one is a difference quotient, reminding one of the derivative $y'(x)$, and the third one is a difference quotient of difference quotients, reminding one of the second derivative $y''(x)$. Indeed, the same structure holds for interpolating polynomials of all orders:

Theorem 3.4. *In the situation of Th.* 3.2*, the Lagrange interpolating polynomial $p_n \in \mathcal{P}_n$ satisfying* (3.2) *can be written in the following form known as* Newton's divided difference interpolation formula*:*

$$p_n(x) = \sum_{j=0}^{n} [y_0, \dots, y_j]\, \omega_j(x), \tag{3.11a}$$

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

$$\omega_j(x) := \prod_{i=0}^{j-1} (x - x_i) \tag{3.11b}$$

are the so-called Newton basis polynomials*, and the $[y_0, \dots, y_j]$ are so-called* divided differences*, defined recursively by*

$$\begin{gathered} [y_j] := y_j \quad \textit{for each } j \in \{0, 1, \dots, n\}, \\ [y_j, \dots, y_{j+k}] := \frac{[y_{j+1}, \dots, y_{j+k}] - [y_j, \dots, y_{j+k-1}]}{x_{j+k} - x_j} \\ \textit{for each } j \in \{0, 1, \dots, n\},\ 1 \le k \le n - j \end{gathered} \tag{3.11c}$$

(note that in the recursive part of the definition, one first omits y_j and then y_{j+k}). In particular, (3.11a) *means that the p_n satisfy the recursive relation*

$$p_0(x) = y_0, \tag{3.12a}$$

$$p_j(x) = p_{j-1}(x) + [y_0, \dots, y_j]\, \omega_j(x) \quad \textit{for } 1 \le j \le n. \tag{3.12b}$$

Before we can prove Th. 3.4, we need to study the divided differences in more detail.

Remark 3.5. In terms of practical computations, note that one can obtain the divided differences in $O(n^2)$ steps from the following recursive scheme:

$$\begin{tikzcd}
y_0 & = & {[y_0]} \arrow[dr] & & & & & \\
y_1 & = & {[y_1]} \arrow[r] \arrow[dr] & {[y_0, y_1]} \arrow[dr] & & & & \\
y_2 & = & {[y_2]} \arrow[r] & {[y_1, y_2]} \arrow[r] & {[y_0, y_1, y_2]} & & & \\
\vdots & & & \vdots & \vdots & \ddots & & \\
y_{n-1} & = & {[y_{n-1}]} \arrow[r] \arrow[dr] & {[y_{n-2}, y_{n-1}]} \arrow[r] \arrow[dr] & \dots & \dots & {[y_0, \dots, y_{n-1}]} \arrow[dr] & \\
y_n & = & {[y_n]} \arrow[r] & {[y_{n-1}, y_n]} \arrow[r] & \dots & \dots & {[y_1, \dots, y_n]} \arrow[r] & {[y_0, \dots, y_n]}
\end{tikzcd}$$

Proposition 3.6. *Let $n \in \mathbb{N}_0$ and consider $n+1$ data points*

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2,$$

$x_i \neq x_j$ for $i \neq j$.

(a) *Then, for each pair $(j,k) \in \mathbb{N}_0 \times \mathbb{N}_0$ such that $0 \le j \le j+k \le n$, the divided differences defined in* (3.11c) *satisfy:*

$$[y_j, \dots, y_{j+k}] = \sum_{i=j}^{j+k} y_i \prod_{\substack{m=j \\ m \neq i}}^{j+k} \frac{1}{x_i - x_m}. \tag{3.13}$$

In particular, the value of $[y_j, \dots, y_{j+k}]$ depends only on the data points

$$(x_j, y_j), \dots, (x_{j+k}, y_{j+k}).$$

(b) *The value of $[y_j, \dots, y_{j+k}]$ does* not *depend on the order of the data points.*

Proof. (a): The proof is carried out by induction with respect to k. For $k = 0$ and each $j \in \{0, \dots, n\}$, one has

$$\sum_{i=j}^{j} y_i \prod_{\substack{m=j \\ m \neq i}}^{j} \frac{1}{x_i - x_m} = y_j = [y_j],$$

as desired. Thus, let $k > 0$.

Claim 1. It suffices to show that, for each $0 \leq j \leq n-k$, the following identity holds true:

$$\sum_{i=j}^{j+k} y_i \prod_{\substack{m=j\\m\neq i}}^{j+k} \frac{1}{x_i - x_m} = \frac{1}{x_{j+k}-x_j}\left(\sum_{i=j+1}^{j+k} y_i \prod_{\substack{m=j+1\\m\neq i}}^{j+k} \frac{1}{x_i - x_m} - \sum_{i=j}^{j+k-1} y_i \prod_{\substack{m=j\\m\neq i}}^{j+k-1} \frac{1}{x_i - x_m}\right). \tag{3.14}$$

Proof. One computes as follows (where (3.13) is used, it holds by induction):

$$\begin{aligned}
&\sum_{i=j}^{j+k} y_i \prod_{\substack{m=j\\m\neq i}}^{j+k} \frac{1}{x_i - x_m}\\
&\overset{(3.14)}{=} \quad \frac{1}{x_{j+k}-x_j}\left(\sum_{i=j+1}^{j+k} y_i \prod_{\substack{m=j+1\\m\neq i}}^{j+k} \frac{1}{x_i - x_m} - \sum_{i=j}^{j+k-1} y_i \prod_{\substack{m=j\\m\neq i}}^{j+k-1} \frac{1}{x_i - x_m}\right)\\
&\overset{(3.13)}{=} \quad \frac{1}{x_{j+k}-x_j}\big([y_{j+1},\dots,y_{j+k}] - [y_j,\dots,y_{j+k-1}]\big)\\
&\overset{(3.11c)}{=} \quad [y_j,\dots,y_{j+k}],
\end{aligned}$$

showing that (3.14) does, indeed, suffice. ▲

So it remains to verify (3.14). We have to show that, for each $i \in \{j,\dots,j+k\}$, the coefficients of y_i on both sides of (3.14) are identical.

Case $i = j+k$: Since y_{j+k} occurs only in first sum on the right-hand side of (3.14), one has to show that

$$\prod_{\substack{m=j\\m\neq j+k}}^{j+k} \frac{1}{x_{j+k} - x_m} = \frac{1}{x_{j+k}-x_j} \prod_{\substack{m=j+1\\m\neq j+k}}^{j+k} \frac{1}{x_{j+k} - x_m}. \tag{3.15}$$

However, (3.15) is obviously true.

Case $i = j$: Since y_j occurs only in second sum on the right-hand side of (3.14), one has to show that

$$\prod_{\substack{m=j\\m\neq j}}^{j+k} \frac{1}{x_j - x_m} = -\frac{1}{x_{j+k}-x_j} \prod_{\substack{m=j\\m\neq j}}^{j+k-1} \frac{1}{x_j - x_m}. \tag{3.16}$$

Once again, (3.16) is obviously true.

Case $j < i < j+k$: In this case, one has to show

$$\prod_{\substack{m=j\\m\neq i}}^{j+k} \frac{1}{x_i - x_m} = \frac{1}{x_{j+k}-x_j}\left(\prod_{\substack{m=j+1\\m\neq i}}^{j+k} \frac{1}{x_i - x_m} - \prod_{\substack{m=j\\m\neq i}}^{j+k-1} \frac{1}{x_i - x_m}\right). \tag{3.17}$$

Multiplying (3.17) by

$$\prod_{\substack{m=j\\ m\neq i,j,j+k}}^{j+k} (x_i - x_m)$$

shows that it is equivalent to

$$\frac{1}{x_i - x_j}\,\frac{1}{x_i - x_{j+k}} = \frac{1}{x_{j+k} - x_j}\left(\frac{1}{x_i - x_{j+k}} - \frac{1}{x_i - x_j}\right),$$

which is also clearly valid.

(b): That the value does not depend on the order of the data points can be seen from the representation (3.13), since the value of the right-hand side remains the same when changing the order of factors in each product and when changing the order of the summands (in fact, if suffices to consider the case of switching the places of just *two* data points, as each permutation is the composition of finitely many transpositions). ■

Proof of Th. 3.4. We carry out the proof via induction on n. In Example 3.3, we had already verified that p_n according to (3.11a) agrees with the Lagrange interpolating polynomial for $n = 0, 1, 2$ (of course, $n = 0$ suffices for our induction). Let $n > 0$. Let p be the Lagrange interpolating polynomial according to (3.3a) and let p_n be the polynomial according to (3.11a). We have to show that $p_n = p$. Clearly, each ω_j is of degree $j \leq n$, such that p_n is also of degree at most n. Moreover, $p_n = p_{n-1} + [y_0, \dots, y_n]\,\omega_n$ and, by induction p_{n-1} is the interpolating polynomial for $(x_0, y_0), \dots, (x_{n-1}, y_{n-1})$. Since $\omega_n(x_k) = 0$ for each $k \in \{0, \dots, n-1\}$, we have $p_n(x_k) = p_{n-1}(x_k) = y_k$. This implies that the polynomial $q := p - p_n$ has n distinct zeros, namely $x_0, \dots, x_{n-1}$. Now the coefficient of x^n in p is

$$\sum_{j=0}^{n} y_j \prod_{\substack{i=0\\ i\neq j}}^{n} \frac{1}{x_j - x_i} \tag{3.18a}$$

and the coefficient of x^n in p_n is

$$[y_0, \dots, y_n] = \sum_{i=0}^{n} y_i \prod_{\substack{m=0\\ m\neq i}}^{n} \frac{1}{x_i - x_m}. \tag{3.18b}$$

With the substitutions $i \to m$ and $j \to i$ in (3.18a), one obtains that both coefficients are identical, i.e. q is of degree at most $n - 1$. As it also has n distinct zeros, it must vanish identically, showing $p = p_n$. ■

Corollary 3.7. *Let $n \in \mathbb{N}_0$ and let $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^2$ be data points such that $x_i \neq x_j$ for $i \neq j$. Suppose $A \subseteq \mathbb{R}$ is some set containing $x_0, \dots, x_n$ and $f : A \longrightarrow \mathbb{R}$ satisfies $f(x_i) = y_i$ for each $i \in \{0, \dots, n\}$. If p_n is the interpolating polynomial, then*

$$f(x) = p_n(x) + [f(x), y_0, \dots, y_n]\,\omega_{n+1}(x) \quad \textit{for each } x \in A \setminus \{x_0, \dots, x_n\}. \tag{3.19}$$

Proof. Fix $x \in A \setminus \{x_0, \dots, x_n\}$ and let p_{n+1} be the interpolating polynomial to

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n), (x, f(x)).$$

One then obtains

$$\begin{aligned} p_n(x) + [f(x), y_0, \dots, y_n]\, \omega_{n+1}(x) &\overset{\text{Prop. 3.6(b)}}{=} p_n(x) + [y_0, \dots, y_n, f(x)]\omega_{n+1}(x) \\ &\overset{(3.12b)}{=} p_{n+1}(x) = f(x), \end{aligned}$$

thereby establishing the case. ■

3.2.3 Error Estimates and the Mean Value Theorem for Divided Differences

We will now further investigate the situation where the data points (x_i, y_i) are given by some differentiable real function f, i.e. $y_i = f(x_i)$, which is to be approximated by the interpolating polynomials. The main goal is to prove an estimate that allows to control the error of such approximations. We begin by establishing a result that allows to generalize the well-known mean value theorem to the situation of divided differences.

Theorem 3.8. *Let $a, b \in \mathbb{R}$, $a < b$, and consider $f \in C^{n+1}[a, b]$, $n \in \mathbb{N}_0$. If*

$$(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n) \in [a, b] \times \mathbb{R}$$

are such that $x_i \neq x_j$ for $i \neq j$ and $y_i = f(x_i)$ for each $i \in \{0, \dots, n\}$, then, defining, for each $x \in [a, b]$, the numbers $m := \min\{x, x_0, \dots, x_n\}$, $M := \max\{x, x_0, \dots, x_n\}$, there exists

$$\xi = \xi(x, x_0, \dots, x_n) \in]m, M[,$$

such that

$$f(x) = p_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\, \omega_{n+1}(x), \tag{3.20}$$

where $p_n \in \mathcal{P}_n$ is the interpolating polynomial to the $n + 1$ data points and ω_{n+1} is the Newton basis polynomial to $x_0, \dots, x_n, x$.

Proof. If $x \in \{x_0, \dots, x_n\}$, then $\omega_{n+1}(x) = 0$ and, by the definition of p_n, (3.20) trivially holds for every $\xi \in [a, b]$. Now assume $x \notin \{x_0, \dots, x_n\}$. Then $\omega_{n+1}(x) \neq 0$, and we can set

$$K := \frac{f(x) - p_n(x)}{\omega_{n+1}(x)}. \tag{3.21}$$

Define the auxiliary function

$$\psi : [a, b] \longrightarrow \mathbb{R}, \quad \psi(t) := f(t) - p_n(t) - K\omega_{n+1}(t). \tag{3.22}$$

Then ψ has the $n + 2$ zeros $x, x_0, \dots, x_n$ and Rolle's theorem yields that ψ' has at least $n + 1$ zeros in $]m, M[$, ψ'' has at least n zeros in $]m, M[$, and, inductively, $\psi^{(n+1)}$ still has at least one zero $\xi \in]m, M[$. Computing $\psi^{(n+1)}$ from (3.22), one obtains

$$\psi^{(n+1)} : [a, b] \longrightarrow \mathbb{R}, \quad \psi^{(n+1)}(t) = f^{(n+1)}(t) - K(n+1)!.$$

Thus, $\psi^{(n+1)}(\xi) = 0$ implies

$$K = \frac{f^{(n+1)}(\xi)}{(n+1)!}. \tag{3.23}$$

Combining (3.23) with (3.21) proves (3.20). ■

Recall the mean value theorem that states that, for $f \in C^1[a,b]$, there exists $\xi \in]a,b[$ such that

$$\frac{f(b) - f(a)}{b - a} = f'(\xi).$$

As a simple consequence of Th. 3.8, we are now in a position generalize the mean value theorem to higher derivatives, making use of divided differences.

Corollary 3.9 (Mean Value Theorem for Divided Differences)**.** *Let $a, b \in \mathbb{R}$, $a < b$, $f \in C^{n+1}[a,b]$, $n \in \mathbb{N}_0$. Then, given $n+2$ distinct points $x, x_0, \dots, x_n \in [a,b]$, there exists*

$$\xi = \xi(x, x_0, \dots, x_n) \in]m, M[, \quad m := \min\{x, x_0, \dots, x_n\}, \quad M := \max\{x, x_0, \dots, x_n\}, \tag{3.24}$$

such that, with $y_i := f(x_i)$,

$$[f(x), y_0, \dots, y_n] = \frac{f^{(n+1)}(\xi)}{(n+1)!}. \tag{3.25}$$

Proof. Theorem 3.8 provides $\xi \in]m, M[$ satisfying (3.20). Combining (3.20) with (3.19), results in

$$p_n(x) + [f(x), y_0, \dots, y_n]\omega_{n+1}(x) = p_n(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\,\omega_{n+1}(x),$$

proving (3.25) (note $\omega_{n+1}(x) \neq 0$). ■

One can compute $n![y_0, \dots, y_n]$ as a numerical approximation to $f^{(n)}(x)$. From Cor. 3.9, we obtain the following error estimate:

Corollary 3.10 (Numerical Differentiation)**.** *Let $a, b \in \mathbb{R}$, $a < b$, $f \in C^{n+1}[a,b]$, $n \in \mathbb{N}_0$, $x \in [a,b]$. Then, given $n+1$ distinct points $x_0, \dots, x_n \in [a,b]$ and letting $y_i := f(x_i)$ for each $i \in \{0, \dots, n\}$, the following estimate holds:*

$$\left| f^{(n)}(x) - n![y_0, \dots, y_n] \right| \leq (M - m) \max\Big\{ |f^{(n+1)}(s)| : s \in [m, M] \Big\}, \tag{3.26}$$

where

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

Proof. Note that the max in (3.26) exists, since $f^{(n+1)}$ is continuous on the compact set $[m, M]$. According to the mean value theorem,

$$\left| f^{(n)}(x) - f^{(n)}(\xi) \right| \leq |x - \xi| \max\Big\{ |f^{(n+1)}(s)| : s \in [m, M] \Big\} \tag{3.27}$$

for each $\xi \in [m, M]$. For $n = 0$, the left-hand side of (3.26) is $|f(x) - f(x_0)|$ and (3.26) follows from (3.27). If $n \geq 1$ and $x \notin \{x_0, \dots, x_n\}$, then we apply Cor. 3.9 with $n-1$, yielding $n![y_0, \dots, y_n] = f^{(n)}(\xi)$ for some $\xi \in]m, M[$, such that (3.26) follows from (3.27) also in this case. Finally, (3.26) extends to $x \in \{x_0, \dots, x_n\}$ by the continuity of $f^{(n)}$. ■

We would now like to provide a convergence result for the approximation of a (benign) C^∞ function by interpolating polynomials. In preparation, we recall the sup norm:

Definition and Remark 3.11. Let $a, b \in \mathbb{R}$, $a < b$, $f \in C[a, b]$. Then

$$\|f\|_\infty := \sup\big\{|f(x)| : x \in [a, b]\big\}$$

is called the *supremum norm* of f (*sup norm* for short, also known as *uniform norm* (the norm of uniform convergence), *max norm*, or *∞-norm*). It constitutes, indeed, a norm on $C[a, b]$.

Theorem 3.12. *Let $a, b \in \mathbb{R}$, $a < b$, $f \in C^\infty[a, b]$. If there exist $K \in \mathbb{R}_0^+$ and $R < (b-a)^{-1}$ such that*

$$\|f^{(n)}\|_\infty \leq K n! R^n \quad \text{for each } n \in \mathbb{N}_0, \tag{3.28}$$

then, for each sequence of distinct numbers $(x_0, x_1, \dots)$ in $[a, b]$, it holds that

$$\lim_{n \to \infty} \|f - p_n\|_\infty = 0, \tag{3.29}$$

where p_n denotes the interpolating polynomial of degree at most n for the first $n + 1$ points $(x_0, f(x_0)), \dots, (x_n, f(x_n))$.

Proof. Since $\|\omega_{n+1}\|_\infty \leq (b-a)^{n+1}$, we obtain from (3.20):

$$\|f - p_n\|_\infty \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!} \|\omega_{n+1}\|_\infty \leq K R^{n+1} (b-a)^{n+1}. \tag{3.30}$$

As $R < (b-a)^{-1}$, (3.30) establishes (3.29). ■

Example 3.13. Suppose $f \in C^\infty[a, b]$ is such that there exists $C \geq 0$ satisfying

$$\|f^{(n)}\|_\infty \leq C^n \quad \text{for each } n \in \mathbb{N}_0. \tag{3.31}$$

We claim that, in that case, there are $K \in \mathbb{R}_0^+$ and $R < (b-a)^{-1}$ satisfying (3.28) (in particular, this shows that Th. 3.12 applies to all functions of the form $f(x) = e^{kx}$, $f(x) = \sin(kx)$, $f(x) = \cos(kx)$). To verify that (3.31) implies (3.28), 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 $n \in \mathbb{N}_0$,

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

Remark 3.14. When approximating $f \in C^\infty[a,b]$ 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 ω_{n+1} will be much larger in the vicinity of a and b than in the center of the interval. This behavior of ω_{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 $\|\omega_{n+1}\|_\infty$ is minimized. If $a=-1$, $b=1$, then one can show that the optimal choice for the data points is

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

These x_j are precisely the zeros of the so-called *Chebyshev polynomials of the first kind.*

3.3 Hermite Interpolation

So far, we have determined an interpolating polynomial p of degree at most n from values at $n+1$ *distinct* points $x_0,\dots,x_n$. Alternatively, one can prescribe the values of p at less than $n+1$ points and, instead, additionally prescribe the values of derivatives of p at some of the same points, where the total number of pieces of data still needs to be $n+1$. For example, to determine p of degree at most 7, one can prescribe $p(1)$, $p(2)$, $p'(2)$, $p(3)$, $p'(3)$, $p''(3)$, $p'''(3)$, and $p^{(4)}(3)$. 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 interpolation is known as *Hermite interpolation* and is to be studied in the present section. The main tool is a generalized version of the divided differences that no longer requires the underlying x_j to be distinct. One can then still use Newton's interpolation formula (3.11a).

The need to generalize the divided differences to situations of identical x_j emphasizes a certain notational dilemma that has its root in the fact that the divided differences depend both on the x_j and on the prescribed values at the x_j. In the notation, it is common to either suppress the x_j-dependence (as we did in the previous sections, cf. (3.11c)) or to suppress the dependence on the values by writing, e.g., $[x_0,\dots,x_n]$ or $f[x_0,\dots,x_n]$ instead of $[y_0,\dots,y_n]$. In the situation of identical x_j, the y_j represent derivatives of various orders, which is not reflected in the notation $[y_0,\dots,y_n]$. Thus, if there are identical x_j, we opt for the $f[x_0,\dots,x_n]$ notation (which has the disadvantage that it suggests that the values y_j are *a priori* given by an underlying function f, which might or might not be the case). If the x_j are all distinct, then one can write $[y_0,\dots,y_n]$ or $f[x_0,\dots,x_n]$ for the same divided difference.

The essential hint for how to generalize divided differences to identical x_j is given by the mean value theorem for divided differences (3.25), which is restated for the convenience of the reader: If $f \in C^{n+1}[a,b]$, then

$$f[x,x_0,\dots,x_n] := [f(x),y_0,\dots,y_n] = \frac{f^{(n+1)}(\xi)}{(n+1)!}. \tag{3.32}$$

In the limit $x_j \to x$ for each $j \in \{0,\dots,n\}$, the continuity of $f^{(n+1)}$ implies that the expression in (3.32) converges to $f^{(n+1)}(x)/(n+1)!$. Thus, $f[x,\dots,x]$ (x repeated $n+2$

times) needs to represent $f^{(n+1)}(x)/(n+1)!$. Thus, if we want to prescribe the first m ($m \in \mathbb{N}_0$) derivatives $y_j, y_j', \dots, y_j^{(m)}$, of the interpolating polynomial at x_j, then, in generalization of the first part of the definition in (3.11c), we need to set

$$f[x_j] := y_j, \quad f[x_j, x_j] := \frac{y_j'}{1!}, \quad \dots, \quad f[\underbrace{x_j, \dots, x_j}_{m+1}] := \frac{y_j^{(m)}}{m!}. \tag{3.33a}$$

The generalization of the recursive part of the definition in (3.11c) is given by

$$\begin{aligned} f[x_0, \dots, x_m] := \frac{f[x_0, \dots, \widehat{x_j}, \dots, x_m] - f[x_0, \dots, \widehat{x_i}, \dots, x_m]}{x_i - x_j} \\ \text{for each } m \in \mathbb{N}_0 \text{ and } 0 \le i, j \le m \text{ such that } x_i \neq x_j \end{aligned} \tag{3.33b}$$

(the hatted symbols $\widehat{x_j}$ and $\widehat{x_i}$ mean that the corresponding points are omitted).

Remark 3.15. Note that there is a choice involved in the definition of $f[x_0, \dots, x_m]$ in (3.33b): The right-hand side depends on which x_i and x_j are used. We will see in the following, that the definition is actually independent of that choice and the $f[x_0, \dots, x_m]$ are well-defined by (3.33). However, this will still need some work.

Example 3.16. Suppose, we have $x_0 \neq x_1$ and want to compute $f[x_0, x_1, x_1, x_1]$ according to (3.33). First, we obtain $f[x_0]$, $f[x_1]$, $f[x_1, x_1]$, and $f[x_1, x_1, x_1]$ from (3.33a). Then we use (3.33b) to compute, in succession,

$$\begin{aligned} f[x_0, x_1] &= \frac{f[x_1] - f[x_0]}{x_1 - x_0}, \quad f[x_0, x_1, x_1] = \frac{f[x_1, x_1] - f[x_0, x_1]}{x_1 - x_0}, \\ f[x_0, x_1, x_1, x_1] &= \frac{f[x_1, x_1, x_1] - f[x_0, x_1, x_1]}{x_1 - x_0}. \end{aligned}$$

—

Before actually proving something, let us illustrate the algorithm of Hermite interpolation with the following example:

Example 3.17. Given the data

$$\begin{gathered} x_0 = 1, \quad x_1 = 2, \\ y_0 = 1, \quad y_0' = 4, \quad y_1 = 3, \quad y_1' = 1, \quad y_1'' = 2, \end{gathered}$$

we seek the interpolating polynomial p of degree at most 4. Using Newton's interpolation formula (3.11a) with the generalized divided differences according to (3.33) yields

$$\begin{aligned} p(x) = f[1] + f[1,1](x-1) + f[1,1,2](x-1)^2 + f[1,1,2,2](x-1)^2(x-2) \\ + f[1,1,2,2,2](x-1)^2(x-2)^2. \end{aligned} \tag{3.34}$$

The generalized divided differences can be computed from a scheme analogous to the one in Rem. 3.5. For simplicity, it is advisable to order the data points such that identical

x_j are adjacent. In the present case, we get $1, 1, 2, 2, 2$. This results in the scheme

$$\begin{tikzcd}
f[1] = y_0 = 1 \arrow[dr] & & & \\
f[1] = y_0 = 1 \arrow[r] \arrow[dr] & f[1,1] = y_0' = 4 \arrow[dr] & & \\
f[2] = y_1 = 3 \arrow[r] \arrow[dr] & f[1,2] = \frac{3-1}{2-1} = 2 \arrow[r] \arrow[dr] & f[1,1,2] = \frac{2-4}{2-1} = -2 \arrow[dr] & \\
f[2] = y_1 = 3 \arrow[r] \arrow[dr] & f[2,2] = y_1' = 1 \arrow[r] \arrow[dr] & f[1,2,2] = \frac{1-2}{2-1} = -1 \arrow[r] \arrow[dr] & f[1,1,2,2] = \frac{-1-(-2)}{2-1} = 1 \\
f[2] = y_1 = 3 \arrow[r] & f[2,2] = y_1' = 1 \arrow[r] & f[2,2,2] = \frac{y_1''}{2!} = 1 \arrow[r] & f[1,2,2,2] = \frac{1-(-1)}{2-1} = 2
\end{tikzcd}$$

and, finally,

$$\begin{tikzcd}
\dots \quad f[1,1,2,2] = 1 \arrow[dr] & \\
\dots \quad f[1,2,2,2] = 2 \arrow[r] & f[1,1,2,2,2] = \frac{2-1}{2-1} = 1.
\end{tikzcd}$$

Hence, plugging the results of the scheme into (3.34),

$$p(x) = 1 + 4(x-1) - 2(x-1)^2 + (x-1)^2(x-2) + (x-1)^2(x-2)^2.$$

—

To proceed with the general theory, it will be useful to reintroduce the generalized divided differences $f[x_0, \dots, x_n]$ under the assumption that they, indeed, arise from a differentiable function f. A posteriori, we will see that this does not constitute an actual restriction: As it turns out, if one starts out by constructing the $f[x_0, \dots, x_n]$ from (3.33), these divided differences define an interpolating polynomial p via Newton's interpolation formula. If one then uses $f := p$ and computes the $f[x_0, \dots, x_n]$ using the following definition based on f, then one recovers the same $f[x_0, \dots, x_n]$ (defined by (3.33)) one had started out with (cf. the proof of Th. 3.25).

In preparation of the f-based definition of the generalized divided differences, the following notation is introduced:

Notation 3.18. For $n \in \mathbb{N}$, the following notation is introduced:

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

Remark 3.19. The set Σ^n introduced in Not. 3.18 is the convex hull of the set consisting of the origin and the n standard unit vectors, i.e. the convex hull of the $(n-1)$-dimensional standard simplex Δ^{n-1} united with the origin:

$$\Sigma^n = \operatorname{conv}\Big(\{0\} \cup \{e_i : i \in \{1, \dots, n\}\}\Big) = \operatorname{conv}\Big(\{0\} \cup \Delta^{n-1}\Big),$$

where

$$e_i := (0, \dots, \underset{\substack{\uparrow \\ i}}{1}, \dots, 0), \quad \Delta^{n-1} := \operatorname{conv}\{e_i : i \in \{1, \dots, n\}\}. \tag{3.35}$$

Theorem 3.20 (Hermite-Genocchi Formula)**.** *Let $a, b \in \mathbb{R}$, $a < b$, $f \in C^n[a,b]$, $n \in \mathbb{N}$. Then, given $n+1$ distinct points $x_0, \dots, x_n \in [a,b]$, and letting $y_i := f(x_i)$ for each $i \in \{0, \dots, n\}$, the divided differences satisfy the following relation, known as the Hermite-Genocchi formula:*

$$[y_0, \dots, y_n] = \int_{\Sigma^n} f^{(n)}\left(x_0 + \sum_{i=1}^{n} s_i(x_i - x_0)\right) \mathrm{d}s\,, \tag{3.36}$$

where Σ^n is the set according to Not. 3.18.

Proof. First, note that, for each $s \in \Sigma^n$,

$$\begin{aligned} a = x_0 + a - x_0 \le x_0 + \sum_{i=1}^{n} s_i(a - x_0) \le x_s &:= x_0 + \sum_{i=1}^{n} s_i(x_i - x_0) \\ &\le x_0 + \sum_{i=1}^{n} s_i(b - x_0) \le x_0 + b - x_0 = b, \end{aligned}$$

such that $f^{(n)}$ is defined at x_s.

We now proceed to verify (3.36) by showing inductively that the formula actually holds true not only for n, but for each $k = 1, \dots, n$: For $k = 1$, (3.36) is an immediate consequence of the fundamental theorem of calculus (FTC) (note $\Sigma^1 = [0,1]$):

$$\begin{aligned} \int_0^1 f'\big(x_0 + s(x_1 - x_0)\big)\,\mathrm{d}s &= \left[\frac{f\big(x_0 + s(x_1 - x_0)\big)}{x_1 - x_0}\right]_0^1 \\ &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{y_1 - y_0}{x_1 - x_0} = [y_0, y_1]. \end{aligned}$$

Thus, we may assume that (3.36) holds with n replaced by k for some $k \in \{1, \dots, n-1\}$. We must then show that it holds for $k+1$. To that end, we compute

$$\begin{aligned} &\int_{\Sigma^{k+1}} f^{(k+1)}\left(x_0 + \sum_{i=1}^{k+1} s_i(x_i - x_0)\right) \mathrm{d}s \\ &\overset{\text{Fubini}}{=} \int_{\Sigma^k}\left(\int_0^{1-\sum_{i=1}^k s_i} f^{(k+1)}\left(x_0 + \sum_{i=1}^{k} s_i(x_i - x_0) + s_{k+1}(x_{k+1} - x_0)\right) \mathrm{d}s_{k+1}\right) \mathrm{d}(s_1, \dots, s_k) \\ &\overset{\text{FTC}}{=} \quad \frac{1}{x_{k+1} - x_0}\int_{\Sigma^k}\left(f^{(k)}\left(x_{k+1} + \sum_{i=1}^{k} s_i(x_i - x_{k+1})\right)\right. \\ &\qquad\qquad \left. - f^{(k)}\left(x_0 + \sum_{i=1}^{k} s_i(x_i - x_0)\right)\right) \mathrm{d}(s_1, \dots, s_k) \\ &\overset{\text{ind.}}{=} \quad \frac{[y_1, \dots, y_{k+1}] - [y_0, \dots, y_k]}{x_{k+1} - x_0} \\ &\overset{(3.11c)}{=} \quad [y_0, \dots, y_{k+1}], \end{aligned}$$

thereby establishing the case. ■

One now notices that the Hermite-Genocchi formula naturally extends to the situation of nondistinct $x_0, \dots, x_n$, giving raise to the following definition:

Definition 3.21. Let $a, b \in \mathbb{R}$, $a < b$, $f \in C^n[a,b]$, $n \in \mathbb{N}$. Then, given $n+1$ (not necessarily distinct) points $x_0, \dots, x_n \in [a,b]$, define

$$f[x_0, \dots, x_n] := \int_{\Sigma^n} f^{(n)}\left(x_0 + \sum_{i=1}^{n} s_i(x_i - x_0)\right) \mathrm{d}s\,. \tag{3.37a}$$

Moreover, for $x \in [a,b]$, define

$$f[x] := f(x). \tag{3.37b}$$

Remark 3.22. The Hermite-Genocchi formula (3.36) says that, if $x_0, \dots, x_n$ are distinct and $y_i = f(x_i)$, then $f[x_0, \dots, x_n] = [y_0, \dots, y_n]$.

—

Before compiling some important properties of the generalized divided differences, we need to compute the volume of Σ^n.

Lemma 3.23. *For each* $n \in \mathbb{N}$*:*

$$\operatorname{vol}(\Sigma^n) := \int_{\Sigma^n} 1\,\mathrm{d}s = \frac{1}{n!}. \tag{3.38}$$

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

$$\begin{aligned}
t_1 &:= s_1, & s_1 &:= t_1,\\
t_2 &:= s_1 + s_2, & s_2 &:= t_2 - t_1,\\
t_3 &:= s_1 + s_2 + s_3, & s_3 &:= t_3 - t_2,\\
&\ \vdots & &\ \vdots\\
t_n &:= s_1 + \dots + s_n, & s_n &:= t_n - t_{n-1}.
\end{aligned} \tag{3.39}$$

According to (3.39), T is clearly bijective, the change of variables is differentiable, $\det J_{T^{-1}} \equiv 1$, and

$$T(\Sigma^n) = \big\{(t_1, \dots, t_n) \in \mathbb{R}^n : 0 \le t_1 \le t_2 \le \dots \le t_n \le 1\big\}. \tag{3.40}$$

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

$$\begin{aligned}
\operatorname{vol}(\Sigma^n) &= \int_{\Sigma^n} 1\,\mathrm{d}s \overset{\text{(CVT)}}{=} \int_{T(\Sigma^n)} \det J_{T^{-1}}(t)\,\mathrm{d}t = \int_{T(\Sigma^n)} 1\,\mathrm{d}t\\
&\overset{\text{(FT)}}{=} \int_0^1 \cdots \int_0^{t_3} \int_0^{t_2} \mathrm{d}t_1\,\mathrm{d}t_2 \cdots \mathrm{d}t_n = \int_0^1 \cdots \int_0^{t_3} t_2\,\mathrm{d}t_2 \cdots \mathrm{d}t_n\\
&= \int_0^1 \cdots \int_0^{t_4} \frac{t_3^2}{2!}\,\mathrm{d}t_3 \cdots \mathrm{d}t_n = \int_0^1 \frac{t_n^{n-1}}{(n-1)!}\,\mathrm{d}t_n = \frac{1}{n!},
\end{aligned}$$

proving (3.38). ■

Proposition 3.24. *Let $a, b \in \mathbb{R}$, $a < b$, $f \in C^n[a,b]$, $n \in \mathbb{N}$.*

(a) *The function from $[a,b]^{n+1}$ into $\mathbb{R}$,*

$$(x_0, \dots, x_n) \mapsto f[x_0, \dots, x_n],$$

defined by (3.37) is continuous, in particular, continuous in each x_k, $k \in \{0, \dots, n\}$.

(b) *Given $n+1$ (not necessarily distinct) points $x_0, \dots, x_n \in [a,b]$, the value of the generalized divided difference $f[x_0, \dots, x_n]$ does not depend on the order of the x_k.*

(c) *For each $0 \leq k \leq n$ and $x \in [a,b]$:*

$$f[\underbrace{x, \dots, x}_{k+1}] = \frac{f^{(k)}(x)}{k!}.$$

(d) *Suppose $x_0, \dots, x_n \in [a,b]$, not necessarily distinct. Letting $y_j^{(m)} := f^{(m)}(x_j)$ for each $j, m \in \{0, \dots, n\}$, one can use the recursive scheme (3.33) to compute the generalized divided difference $f[x_0, \dots, x_n]$.*

Proof. (a): For $n = 0$, according to (3.37b), the continuity of $x_0 \mapsto f[x_0]$ is merely the assumed continuity of f. Let $n > 0$. Since $f^{(n)}$ is continuous and the continuous function $|f^{(n)}|$ is uniformly bounded on $[a,b]$ (namely by $M := \|f^{(n)}\|_\infty$), the continuity of the divided differences is an immediate consequence of (3.37a) and the continuity of parameter-dependent integrals (see Th. E.1 of the Appendix).

(b): Let $x^\nu := (x_0^\nu, \dots, x_n^\nu) \in [a,b]^{n+1}$ be a sequence such that $\lim_{\nu\to\infty} x^\nu = (x_0, \dots, x_n)$ and such that the x_i^ν are distinct for each ν (see the proof of Th. 3.25 below for an explicit construction of such a sequence). Then, for each permutation π of $\{0, \dots, n\}$:

$$f[x_0, \dots, x_n] \overset{\text{(a)}}{=} \lim_{\nu\to\infty} f[x_0^\nu, \dots, x_n^\nu] \overset{\text{Prop. 3.6(b)}}{=} \lim_{\nu\to\infty} f[x_{\pi(0)}^\nu, \dots, x_{\pi(n)}^\nu] \overset{\text{(a)}}{=} f[x_{\pi(0)}, \dots, x_{\pi(n)}].$$

(c): For $k = 0$, there is nothing to prove. For $k > 0$, we compute

$$f[\underbrace{x, \dots, x}_{k+1}] \overset{\text{(3.37a)}}{=} \int_{\Sigma^k} f^{(k)}\left(x + \sum_{i=1}^{k} s_i(x - x)\right) \mathrm{d}s = \int_{\Sigma^k} f^{(k)}(x)\,\mathrm{d}s \overset{\text{(3.38)}}{=} \frac{f^{(k)}(x)}{k!}.$$

(d): First, the setting $y_j^{(m)} := f^{(m)}(x_j)$ guarantees that (3.33a) agrees with (c). If $n > 0$ and the $x_0, \dots, x_n$ are all distinct, then, for each $i, j \in \{0, \dots, n\}$ with $i \neq j$,

$$\begin{aligned}
f[x_0, \dots, x_n] &= [y_0, \dots, y_n] \overset{\text{Prop. 3.6(b)}}{=} [y_j, y_0, \dots, \widehat{y_j}, \widehat{y_i}, \dots, y_n, y_i] \\
&\overset{\text{(3.11c)}}{=} \frac{[y_0, \dots, \widehat{y_j}, \widehat{y_i}, \dots, y_n, y_i] - [y_j, y_0, \dots, \widehat{y_j}, \widehat{y_i}, \dots, y_n]}{x_i - x_j} \\
&\overset{\text{Prop. 3.6(b)}}{=} \frac{[y_0, \dots, \widehat{y_j}, \dots, y_n] - [y_0, \dots, \widehat{y_i}, \dots, y_n]}{x_i - x_j} \\
&= \frac{f[x_0, \dots, \widehat{x_j}, \dots, x_n] - f[x_0, \dots, \widehat{x_i}, \dots, x_n]}{x_i - x_j},
\end{aligned}$$

proving (3.33b) for the case, where all x_j are distinct. In the general case, where the x_j are not necessarily distinct, one, once again, approximates $(x_0, \dots, x_k)$ by a sequence of distinct points. Then (3.33b) is implied by the distinct case together with (a). ■

The following Th. 3.25 shows that the algorithm of Hermite interpolation employed in Example 3.17 (i.e. using Newton's interpolation formula with the generalized divided differences) is, indeed, valid: It results in the unique polynomial that satisfies the given $n+1$ conditions.

Theorem 3.25. *Let $r, n \in \mathbb{N}_0$. Given $r+1$ distinct numbers $\tilde{x}_0, \dots, \tilde{x}_r \in \mathbb{R}$, natural numbers $m_0, \dots, m_r \in \mathbb{N}$ such that $\sum_{k=0}^{r} m_k = n+1$, and values $y_j^{(m)} \in \mathbb{R}$ for each $j \in \{0, \dots, r\}$, $m \in \{0, \dots, m_j - 1\}$, there exists a unique polynomial $p \in \mathcal{P}_n$ (of degree at most n) such that*

$$p^{(m)}(\tilde{x}_j) = y_j^{(m)} \quad \text{for each } j \in \{0, \dots, r\},\ m \in \{0, \dots, m_j - 1\}. \tag{3.41}$$

Moreover, letting, for each $i \in \{0, \dots, n\}$,

$$x_i := \tilde{x}_j \quad \text{for} \quad \sum_{k=0}^{j-1} m_k < i+1 \leq \sum_{k=0}^{j} m_k, \tag{3.42}$$

p is given by Newton's interpolation formula

$$p(x) = \sum_{j=0}^{n} f[x_0, \dots, x_j]\, \omega_j(x), \tag{3.43a}$$

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

$$\omega_j(x) = \prod_{i=0}^{j-1} (x - x_i), \tag{3.43b}$$

and the $f[x_0, \dots, x_j]$ can be computed via the recursive scheme (3.33).

Proof. Consider the map

$$A : \mathcal{P}_n \longrightarrow \mathbb{R}^{n+1},$$
$$A(p) := \big(p(\tilde{x}_0), \dots, p^{(m_0-1)}(\tilde{x}_0), \dots, p(\tilde{x}_r), \dots, p^{(m_r-1)}(\tilde{x}_r)\big).$$

The existence and uniqueness statement of the theorem is the same as saying that A is bijective. Since A is clearly linear and since $\dim \mathcal{P}_n = \dim \mathbb{R}^{n+1} = n+1$, it suffices to show that A is one-to-one, i.e. $\ker A = 0$. Recall (exercise) that, for $p \in \mathcal{P}_n$, if $m \in \{0, \dots, n-1\}$ and $p(x) = p'(x) = \dots = p^{(m)}(x) = 0$, then x is a zero of multiplicity at least $m+1$ for p. Thus, if $A(p) = (0, \dots, 0)$, then $\tilde{x}_0$ is a zero of p with multiplicity at least m_0, …, $\tilde{x}_r$ is a zero of p with multiplicity at least m_r, such that p has at least $n+1$ zeros. Since p has degree at most n, it follows that $p \equiv 0$, showing $\ker A = 0$, thereby also establishing existence and uniqueness of the interpolating polynomial.

Now let $f = p \in \mathcal{P}_n$ be the unique interpolating polynomial satisfying (3.41). Then the generalized divided differences are given by (3.37), and from Rem. 3.22 we know that $f[x_0, \dots, x_j] = [y_0, \dots, y_j]$ if all the x_k are distinct. In particular, (3.43) holds if all the x_k are distinct. Thus, in the general case, let $x^\nu := (x_0^\nu, \dots, x_n^\nu) \in \mathbb{R}^{n+1}$ be a sequence such that $\lim_{\nu\to\infty} x^\nu = (x_0, \dots, x_n)$ and such that the x_i^ν are distinct for each sufficiently large ν, say for $\nu \geq N \in \mathbb{N}$ (for example, one can let $x_i^\nu := \tilde{x}_j + \frac{\alpha}{\nu}$ if, and only if, $\sum_{k=0}^{j-1} m_k < i+1 \leq \sum_{k=0}^{j} m_k$ and $i+1 = \alpha + \sum_{k=0}^{j-1} m_k$). Then, for each $x \in \mathbb{R}$ and each $\nu \in \mathbb{N}$ with $\nu \geq N$, one has

$$p(x) = \sum_{j=0}^{n} \big[p(x_0^\nu), \dots, p(x_j^\nu)\big] \prod_{i=0}^{j-1} (x - x_i^\nu). \tag{3.44}$$

Taking the limit for $\nu \to \infty$ in (3.44) yields

$$\begin{aligned} p(x) \quad &= \quad \lim_{\nu\to\infty} p(x) = \lim_{\nu\to\infty} \left(\sum_{j=0}^{n} \big[p(x_0^\nu), \dots, p(x_j^\nu)\big] \prod_{i=0}^{j-1} (x - x_i^\nu) \right) \\ &\overset{\text{Prop. 3.24(a)}}{=} \quad \sum_{j=0}^{n} f[x_0, \dots, x_j]\, \omega_j(x), \end{aligned}$$

proving (3.43). Finally, since we have seen that the $f[x_0, \dots, x_j]$ are given by letting $f = p$ be the interpolating polynomial, the validity of the recursive scheme (3.33) follows from Prop. 3.24(c),(d). ■

Theorem 3.26. *Let $a, b \in \mathbb{R}$, $a < b$, $f \in C^{n+1}[a, b]$, $n \in \mathbb{N}$, and let $x_0, \dots, x_n \in [a, b]$ (not necessarily distinct). Let $p \in \mathcal{P}_n$ be the interpolating polynomial given by* (3.43).

(a) *The following holds for each $x \in [a, b]$:*

$$f(x) = p(x) + f[x, x_0, \dots, x_n]\, \omega_{n+1}(x). \tag{3.45}$$

(b) *Let $x \in [a, b]$. Letting $m := \min\{x, x_0, \dots, x_n\}$, $M := \max\{x, x_0, \dots, x_n\}$, there exists $\xi = \xi(x, x_0, \dots, x_n) \in [m, M]$ such that*

$$f(x) = p(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}\, \omega_{n+1}(x). \tag{3.46}$$

(c) (Mean Value Theorem for Generalized Divided Differences) *Given $x \in [a, b]$, let m, M, and ξ be as in* (b). *Then*

$$f[x, x_0, \dots, x_n] = \frac{f^{(n+1)}(\xi)}{(n+1)!}. \tag{3.47}$$

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

$$\begin{aligned} p(x) + f[x, x_0, \dots, x_n]\, \omega_{n+1}(x) \quad &= \quad \sum_{j=0}^{n} f[x_0, \dots, x_j]\, \omega_j(x) + f[x_0, \dots, x_n, x]\, \omega_{n+1}(x) \\ &= \quad f(x). \end{aligned}$$

(b): For $x \in \{x_0, \dots, x_n\}$, (3.46) is clearly true for each $\xi \in [m, M]$. Thus, assume $x \notin \{x_0, \dots, x_n\}$. We know from Th. 3.8 that (3.46) holds if the $x_0, \dots, x_n$ are all distinct. In the general case, we once again apply the technique of approximating $(x_0, \dots, x_n)$ by a sequence $x^\nu := (x_0^\nu, \dots, x_n^\nu) \in [m, M]^{n+1}$ such that $\lim_{\nu\to\infty} x^\nu = (x_0, \dots, x_n)$ and such that the x_i^ν are distinct for each ν. Then Th. 3.8 provides a sequence $\xi(x^\nu) \in]m, M[$ such that, for each ν, (3.46) holds with $\xi = \xi(x^\nu)$. Since $(\xi(x^\nu))_{\nu\in\mathbb{N}}$ is a sequence in the compact interval $[m, M]$, there exists a subsequence $(\tilde{x}^\nu)_{\nu\in\mathbb{N}}$ of $(x^\nu)_{\nu\in\mathbb{N}}$ such that $\xi(\tilde{x}^\nu)$ converges to some $\xi \in [m, M]$. Finally, the continuity of $f^{(n+1)}$ yields

$$\begin{aligned}\frac{f^{(n+1)}(\xi)}{(n+1)!} &= \lim_{\nu\to\infty} \frac{f^{(n+1)}\big(\xi(\tilde{x}^\nu)\big)}{(n+1)!} \overset{(3.25)}{=} \lim_{\nu\to\infty} f[x, \tilde{x}_0^\nu, \dots, \tilde{x}_n^\nu] = f[x, x_0, \dots, x_n] \\ &\overset{(3.45)}{=} \frac{f(x) - p(x)}{\omega_{n+1}(x)},\end{aligned}$$

proving (3.46).

(c): Analogous to the proof of Cor. 3.9, one merely has to combine (3.45) and (3.46). ■

3.4 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. 3.27. Even though related, this topic is not strictly part of the theory of interpolation.

Theorem 3.27 (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 $p : \mathbb{R} \longrightarrow \mathbb{R}$ such that $\|f - p\restriction_{[a,b]}\|_\infty < \epsilon$, where $p\restriction_{[a,b]}$ denotes the restriction of p to $[a, b]$.*

Theorem 3.27 will be a corollary of the fact that the Bernstein polynomials corresponding to $f \in C[0, 1]$ (see Def. 3.28) converge uniformly to f on $[0, 1]$ (see Th. 3.29 below).

Definition 3.28. Given $f : [0, 1] \longrightarrow \mathbb{R}$, define the *Bernstein polynomials $B_n f$* corresponding to f by

$$B_n f : \mathbb{R} \longrightarrow \mathbb{R}, \quad (B_n f)(x) := \sum_{\nu=0}^{n} f\left(\frac{\nu}{n}\right) \binom{n}{\nu} x^\nu (1 - x)^{n-\nu} \quad \text{for each } n \in \mathbb{N}. \tag{3.48}$$

Theorem 3.29. *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. 3.28 converges uniformly to f on $[0, 1]$, i.e.*

$$\lim_{n\to\infty} \|f - (B_n f)\restriction_{[0,1]}\|_\infty = 0. \tag{3.49}$$

Proof. We begin by noting

$$(B_n f)(0) = f(0) \quad \text{and} \quad (B_n f)(1) = f(1) \quad \text{for each } n \in \mathbb{N}. \tag{3.50}$$

For each $n \in \mathbb{N}$ and $\nu \in \{0, \dots, n\}$, we introduce the abbreviation

$$q_{n\nu}(x) := \binom{n}{\nu} x^\nu (1-x)^{n-\nu}. \tag{3.51}$$

Then

$$1 = \big(x + (1-x)\big)^n = \sum_{\nu=0}^{n} q_{n\nu}(x) \quad \text{for each } n \in \mathbb{N} \tag{3.52}$$

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

$$\big|f(x) - (B_n f)(x)\big| \le \sum_{\nu=0}^{n} \left| f(x) - f\left(\frac{\nu}{n}\right) \right| q_{n\nu}(x) \quad \text{for each } x \in [0,1],\ n \in \mathbb{N}. \tag{3.53}$$

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 $x \in [0,1]$, $n \in \mathbb{N}$, $\nu \in \{0, \dots, n\}$:

$$\left| x - \frac{\nu}{n} \right| < \delta \quad \Rightarrow \quad \left| f(x) - f\left(\frac{\nu}{n}\right) \right| < \frac{\epsilon}{2}. \tag{3.54}$$

For the moment, we fix $x \in [0,1]$ and $n \in \mathbb{N}$ and define

$$N_1 := \left\{ \nu \in \{0, \dots, n\} : \left| x - \frac{\nu}{n} \right| < \delta \right\},$$
$$N_2 := \left\{ \nu \in \{0, \dots, n\} : \left| x - \frac{\nu}{n} \right| \ge \delta \right\}.$$

Then

$$\sum_{\nu \in N_1} \left| f(x) - f\left(\frac{\nu}{n}\right) \right| q_{n\nu}(x) \le \frac{\epsilon}{2} \sum_{\nu \in N_1} q_{n\nu}(x) \le \frac{\epsilon}{2} \sum_{\nu=0}^{n} q_{n\nu}(x) = \frac{\epsilon}{2}, \tag{3.55}$$

and with $M := \|f\|_\infty$,

$$\begin{aligned} \sum_{\nu \in N_2} \left| f(x) - f\left(\frac{\nu}{n}\right) \right| q_{n\nu}(x) &\le \sum_{\nu \in N_2} \left| f(x) - f\left(\frac{\nu}{n}\right) \right| q_{n\nu}(x) \frac{\left(x - \frac{\nu}{n}\right)^2}{\delta^2} \\ &\le \frac{2M}{\delta^2} \sum_{\nu=0}^{n} q_{n\nu}(x) \left(x - \frac{\nu}{n} \right)^2. \end{aligned} \tag{3.56}$$

To compute the sum on the right-hand side of (3.56), observe

$$\left(x - \frac{\nu}{n} \right)^2 = x^2 - 2x\frac{\nu}{n} + \left(\frac{\nu}{n}\right)^2 \tag{3.57}$$

and

$$\sum_{\nu=0}^{n} \binom{n}{\nu} x^\nu (1-x)^{n-\nu} \frac{\nu}{n} = x \sum_{\nu=1}^{n} \binom{n-1}{\nu-1} x^{\nu-1} (1-x)^{(n-1)-(\nu-1)} \overset{(3.52)}{=} x \tag{3.58}$$

as well as

$$\begin{aligned}
&\sum_{\nu=0}^{n}\binom{n}{\nu}x^{\nu}(1-x)^{n-\nu}\left(\frac{\nu}{n}\right)^2=\frac{x}{n}\sum_{\nu=1}^{n}(\nu-1)\binom{n-1}{\nu-1}x^{\nu-1}(1-x)^{(n-1)-(\nu-1)}+\frac{x}{n}\\
&=\frac{x^2}{n}(n-1)\sum_{\nu=2}^{n}\binom{n-2}{\nu-2}x^{\nu-2}(1-x)^{(n-2)-(\nu-2)}+\frac{x}{n}\\
&=x^2\left(1-\frac{1}{n}\right)+\frac{x}{n}=x^2+\frac{x}{n}(1-x). \qquad (3.59)
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
\sum_{\nu=0}^{n} q_{n\nu}(x)\left(x-\frac{\nu}{n}\right)^2 &\overset{(3.57),(3.52),(3.58),(3.59)}{=} x^2\cdot 1-2x\cdot x+x^2+\frac{x}{n}(1-x)\\
&\leq \frac{1}{4n} \quad \text{for each } x\in[0,1],\ n\in\mathbb{N},
\end{aligned}$$

and together with (3.56):

$$\sum_{\nu\in N_2}\left|f(x)-f\left(\frac{\nu}{n}\right)\right|q_{n\nu}(x)\leq\frac{2M}{\delta^2}\frac{1}{4n}<\frac{\epsilon}{2}\quad\text{for each } x\in[0,1],\ n>\frac{M}{\delta^2\epsilon}. \tag{3.60}$$

Combining (3.53), (3.55), and (3.60) yields

$$\big|f(x)-(B_nf)(x)\big|<\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon\quad\text{for each } x\in[0,1],\ n>\frac{M}{\delta^2\epsilon},$$

proving the claimed uniform convergence. ■

Proof of Th. 3.27. Define

$$\begin{aligned}
\phi:\ [a,b]\longrightarrow[0,1],\quad \phi(x):=\frac{x-a}{b-a},\\
\phi^{-1}:\ [0,1]\longrightarrow[a,b],\quad \phi^{-1}(x):=(b-a)x+a.
\end{aligned}$$

Given $\epsilon>0$, and letting

$$g:\ [0,1]\longrightarrow\mathbb{R},\quad g(x):=f\big(\phi^{-1}(x)\big),$$

Th. 3.29 provides a polynomial $q:\mathbb{R}\longrightarrow\mathbb{R}$ such that $\|g-q{\restriction}_{[0,1]}\|_\infty<\epsilon$. Defining

$$p:\mathbb{R}\longrightarrow\mathbb{R},\quad p(x):=q\big(\phi(x)\big)=q\left(\frac{x-a}{b-a}\right)$$

(having extended ϕ in the obvious way) yields a polynomial p such that

$$|f(x)-p(x)|=\big|g(\phi(x))-q(\phi(x))\big|<\epsilon\quad\text{for each } x\in[a,b]$$

as needed. ■

3.5 Spline Interpolation

3.5.1 Introduction

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. 3.14).

The advantage of the approach of the previous sections is that the interpolating polynomial is C^∞ (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 interpolations follow 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^∞, but merely of class C 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 the sup-norm $\|\cdot\|_\infty$, then spline interpolation is usually a good option.

Definition 3.30. Let $a, b \in \mathbb{R}$, $a < b$. Then, given $N \in \mathbb{N}$,

$$\Delta := (x_0, \dots, x_N) \in \mathbb{R}^{N+1} \tag{3.61}$$

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

$$S_{\Delta,l} := \left\{ s \in C^{l-1}[a, b] : \underset{k\in\{0,\dots,N-1\}}{\forall} \; \underset{p_k\in\mathcal{P}_l}{\exists} \; p_k\restriction_{[x_k,x_{k+1}]} = s\restriction_{[x_k,x_{k+1}]} \right\}. \tag{3.62}$$

Of particular importance are the elements of $S_{\Delta,1}$ (*linear splines*) and of $S_{\Delta,3}$ (*cubic splines*).

Remark 3.31. $S_{\Delta,l}$ is a real vector space of dimension $N+l$. Without any continuity restrictions, one had the freedom to choose N arbitrary polynomials of degree at most l, giving dimension $N(l+1)$. However, the continuity requirements for the derivatives of order $0,\dots,l-1$ yield l conditions at each of the $N-1$ interior knots, such that $\dim S_{\Delta,l} = N(l+1)-(N-1)l = N+l$.

3.5.2 Linear Splines

Given a node vector as in (3.61), and values $y_0,\dots,y_N \in \mathbb{R}$, 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 3.32. *Let $a,b \in \mathbb{R}$, $a<b$. Then, given $N \in \mathbb{N}$, a knot vector $\Delta := (x_0,\dots,x_N) \in \mathbb{R}^{N+1}$ for $[a,b]$, and values $(y_0,\dots,y_N) \in \mathbb{R}^{N+1}$, there exists a unique interpolating linear spline $s \in S_{\Delta,1}$ satisfying*

$$s(x_k) = y_k \quad \textit{for each } k \in \{0,\dots,N\}. \tag{3.63}$$

Moreover s is explicitly given by

$$s(x) = y_k + \frac{y_{k+1}-y_k}{x_{k+1}-x_k}\,(x-x_k) \quad \textit{for each } x \in [x_k,x_{k+1}]. \tag{3.64}$$

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

$$\|f-s\|_\infty \le \frac{1}{8}\|f''\|_\infty\, h^2, \tag{3.65}$$

where

$$h := h(\Delta) := \max\bigl\{x_{k+1}-x_k : k \in \{0,\dots,N-1\}\bigr\}$$

is the mesh size *of the partition Δ.*

Proof. Since s defined by (3.64) clearly satisfies (3.63), is continuous on $[a,b]$, and, for each $k \in \{0,\dots,N\}$, the definition of (3.64) extends to a unique polynomial in $\mathcal{P}_1$, 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 $\mathcal{P}_1$ and, thus, s. For the error estimate, consider $x \in [x_k,x_{k+1}]$, $k \in \{0,\dots,N-1\}$. Note that, due to the inequality $\alpha\,\beta \le \bigl(\frac{\alpha+\beta}{2}\bigr)^2$ valid for all $\alpha,\beta \in \mathbb{R}$, one has

$$(x-x_k)(x_{k+1}-x) \le \left(\frac{(x-x_k)+(x_{k+1}-x)}{2}\right)^2 \le \frac{h^2}{4}. \tag{3.66}$$

Moreover, on $[x_k,x_{k+1}]$, s coincides with the interpolating polynomial $p \in \mathcal{P}_1$ with $p(x_k)=f(x_k)$ and $p(x_{k+1})=f(x_{k+1})$. Thus, we can use (3.20) to compute

$$\bigl|f(x)-s(x)\bigr| = \bigl|f(x)-p(x)\bigr| \le \frac{\|f''\|_\infty}{2!}\,(x-x_k)(x_{k+1}-x) \overset{(3.66)}{\le} \frac{1}{8}\|f''\|_\infty\, h^2,$$

thereby proving (3.65). ■

3.5.3 Cubic Splines

Given a knot vector

$$\Delta = (x_0, \dots, x_N) \in \mathbb{R}^{N+1} \quad \text{and values} \quad (y_0, \dots, y_N) \in \mathbb{R}^{N+1}, \tag{3.67}$$

the task is to find an interpolating cubic spline, i.e. an element s of $S_{\Delta,3}$ satisfying

$$s(x_k) = y_k \quad \text{for each } k \in \{0, \dots, N\}. \tag{3.68}$$

As we require $s \in S_{\Delta,3}$, according to (3.62), for each $k \in \{0, \dots, N-1\}$, we must find $p_k \in \mathcal{P}_3$ such that $p_k\restriction_{[x_k,x_{k+1}]} = s\restriction_{[x_k,x_{k+1}]}$. To that end, for each $k \in \{0, \dots, N-1\}$, we employ the ansatz

$$p_k(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3 \tag{3.69a}$$

with suitable

$$(a_k, b_k, c_k, d_k) \in \mathbb{R}^4; \tag{3.69b}$$

to then define s by

$$s(x) = p_k(x) \quad \text{for each } x \in [x_k, x_{k+1}]. \tag{3.69c}$$

It is immediate that s is well-defined by (3.69c) and an element of $S_{\Delta,3}$ (cf. (3.62)) satisfying (3.68) if, and only if,

$$\begin{aligned}
p_0(x_0) &= y_0, && \text{(3.70a)}\\
p_{k-1}(x_k) &= y_k = p_k(x_k) \quad \text{for each } k \in \{1, \dots, N-1\}, && \text{(3.70b)}\\
p_{N-1}(x_N) &= y_N, && \text{(3.70c)}\\
p'_{k-1}(x_k) &= p'_k(x_k) \quad \text{for each } k \in \{1, \dots, N-1\}, && \text{(3.70d)}\\
p''_{k-1}(x_k) &= p''_k(x_k) \quad \text{for each } k \in \{1, \dots, N-1\}, && \text{(3.70e)}
\end{aligned}$$

using the convention $\{1, \dots, N-1\} = \emptyset$ for $N - 1 < 1$. Note that (3.70) constitutes a system of $2 + 4(N - 1) = 4N - 2$ conditions, while (3.69a) 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 (3.81) 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 (3.70) are satisfied, the strategy is to transform (3.70) 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

$$s''_k := s''(x_k), \quad k \in \{0, \dots, N\}, \tag{3.71}$$

as new variables. As we will see in the following lemma, (3.70) is satisfied if, and only if, the s''_k satisfy the linear system (3.73):

Lemma 3.33. *Given a knot vector and values according to* (3.67)*, and introducing the abbreviations*

$$h_k := x_{k+1} - x_k, \qquad \textit{for each } k \in \{0,\dots,N-1\}, \tag{3.72a}$$

$$g_k := 6 \underbrace{\left(\frac{y_{k+1}-y_k}{h_k} - \frac{y_k - y_{k-1}}{h_{k-1}} \right)}_{(h_k+h_{k-1})[y_{k-1},y_k,y_{k+1}]} \qquad \textit{for each } k \in \{1,\dots,N-1\}, \tag{3.72b}$$

the $p_k \in \mathcal{P}_3$ of (3.69a) *satisfy* (3.70) *(and that means s given by* (3.69c) *is in $S_{\Delta,3}$ and satisfies* (3.68)*) if, and only if, the $N+1$ numbers $s''_0,\dots,s''_N$ defined by* (3.71) *solve the linear system consisting of the $N-1$ equations*

$$h_k s''_k + 2(h_k + h_{k+1}) s''_{k+1} + h_{k+1} s''_{k+2} = g_{k+1}, \quad k \in \{0,\dots,N-2\}, \tag{3.73}$$

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

$$a_k = y_k \qquad \textit{for each } k \in \{0,\dots,N-1\}, \tag{3.74a}$$

$$b_k = \frac{y_{k+1}-y_k}{h_k} - \frac{h_k}{6}\,(s''_{k+1} + 2s''_k) \qquad \textit{for each } k \in \{0,\dots,N-1\}, \tag{3.74b}$$

$$c_k = \frac{s''_k}{2} \qquad \textit{for each } k \in \{0,\dots,N-1\}, \tag{3.74c}$$

$$d_k = \frac{s''_{k+1} - s''_k}{6h_k} \qquad \textit{for each } k \in \{0,\dots,N-1\}. \tag{3.74d}$$

Proof. First, for subsequent use, from (3.69a), we obtain the first two derivatives of the p_k for each $k \in \{0,\dots,N-1\}$:

$$p'_k(x) = b_k + 2c_k(x - x_k) + 3d_k(x - x_k)^2, \tag{3.75a}$$

$$p''_k(x) = 2c_k + 6d_k(x - x_k). \tag{3.75b}$$

As usual, for the claimed equivalence, we prove two implications.

(3.70) implies (3.73) and (3.74):

Plugging $x = x_k$ into (3.69a) yields

$$p_k(x_k) = a_k \quad \text{for each } k \in \{0,\dots,N-1\}. \tag{3.76}$$

Thus, (3.74a) is a consequence of (3.70a) and (3.70b).

Plugging $x = x_k$ into (3.75b) yields

$$s''_k = p''_k(x_k) = 2c_k \quad \text{for each } k \in \{0,\dots,N-1\}, \tag{3.77}$$

i.e. (3.74c).

Plugging (3.70e), i.e. $p''_{k-1}(x_k) = p''_k(x_k)$, into (3.75b) yields

$$2c_{k-1} + 6d_{k-1}(x_k - x_{k-1}) = 2c_k \quad \text{for each } k \in \{1,\dots,N-1\}. \tag{3.78a}$$

When using (3.74c) and solving for d_{k-1}, one obtains

$$d_{k-1} = \frac{s''_k - s''_{k-1}}{6\,h_{k-1}} \quad \text{for each } k \in \{1,\dots,N-1\}, \tag{3.78b}$$

i.e. the relation of (3.74d) for each $k \in \{0,\dots,N-2\}$. Moreover, (3.71) and (3.75b) imply

$$s''_N = s''(x_N) = p''_{N-1}(x_N) = 2c_{N-1} + 6d_{N-1}h_{N-1}, \tag{3.78c}$$

which, after solving for d_{N-1}, shows that the relation of (3.74d) also holds for $k = N-1$.

Plugging the first part of (3.70b) and (3.70c), i.e. $p_{k-1}(x_k) = y_k$ into (3.69a) yields

$$a_{k-1} + b_{k-1}\,h_{k-1} + c_{k-1}\,h^2_{k-1} + d_{k-1}\,h^3_{k-1} = y_k \quad \text{for each } k \in \{1,\dots,N\}. \tag{3.79a}$$

When using (3.74a), (3.74c), (3.74d), and solving for b_{k-1}, one obtains

$$b_{k-1} = \frac{y_k - y_{k-1}}{h_{k-1}} - \frac{s''_{k-1}\,h_{k-1}}{2} - \frac{s''_k - s''_{k-1}}{6\,h_{k-1}}\,h^2_{k-1} = \frac{y_k - y_{k-1}}{h_{k-1}} - \frac{h_{k-1}}{6}\,(s''_k + 2s''_{k-1}) \tag{3.79b}$$

for each $k \in \{1,\dots,N\}$, i.e. (3.74b).

Plugging (3.70d), i.e. $p'_{k-1}(x_k) = p'_k(x_k)$, into (3.75a) yields

$$b_{k-1} + 2c_{k-1}\,h_{k-1} + 3d_{k-1}\,h^2_{k-1} = b_k \quad \text{for each } k \in \{1,\dots,N-1\}. \tag{3.80a}$$

Finally, applying (3.74b), (3.74c), and (3.74d), one obtains

$$\frac{y_k - y_{k-1}}{h_{k-1}} - \frac{h_{k-1}}{6}\,(s''_k + 2s''_{k-1}) + s''_{k-1}\,h_{k-1} + \frac{s''_k - s''_{k-1}}{2}\,h_{k-1} = \frac{y_{k+1} - y_k}{h_k} - \frac{h_k}{6}\,(s''_{k+1} + 2s''_k), \tag{3.80b}$$

or, rearranged,

$$h_{k-1}\,s''_{k-1} + 2(h_{k-1} + h_k)s''_k + h_k\,s''_{k+1} = g_k, \quad k \in \{1,\dots,N-1\}, \tag{3.80c}$$

which is (3.73).

(3.73) and (3.74) imply (3.70):

First, note that plugging $x = x_k$ into (3.75b) again yields $s''_k = p''_k(x_k)$. Using (3.74a) in (3.69a) yields (3.70a) and the $p_k(x_k) = y_k$ part of (3.70b). Since (3.74b) is equivalent to (3.79b), and (3.79b) (when using (3.74a), (3.74c), and (3.74d)) implies (3.79a), which, in turn, is equivalent to $p_{k-1}(x_k) = y_k$ for each $k \in \{1,\dots,N\}$, we see that (3.74) implies (3.70c) and the $p_{k-1}(x_k) = y_k$ part of (3.70b). Since (3.73) is equivalent to (3.80c) and (3.80b), and (3.80b) (when using (3.74a) – (3.74d)) implies (3.80a), which, in turn, is equivalent to $p'_{k-1}(x_k) = p'_k(x_k)$ for each $k \in \{1,\dots,N-1\}$, we see that (3.73) and (3.74) implies (3.70d). Finally, (3.74d) is equivalent to (3.78b), which (when using (3.74c)) implies (3.78a). Since (3.78a) is equivalent to $p''_{k-1}(x_k) = p''_k(x_k)$ for each $k \in \{1,\dots,N-1\}$, (3.74) also implies (3.70d), concluding the proof of the lemma. ■

As already mentioned after formulating (3.70), (3.70) yields two conditions less than there are variables. Not surprisingly, the same is true in the linear system (3.73), 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:

$$s''(x_0) = s''(x_N) = 0. \tag{3.81a}$$

Periodic boundary conditions:

$$s'(x_0) = s'(x_N) \quad \text{and} \quad s''(x_0) = s''(x_N). \tag{3.81b}$$

Dirichlet boundary conditions for the first derivative:

$$s'(x_0) = y_0' \in \mathbb{R} \quad \text{and} \quad s'(x_N) = y_N' \in \mathbb{R}. \tag{3.81c}$$

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

Since (3.81a) fixes the values of s_0'' and s_N'' as 0, (3.73) 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

$$As'' = g \tag{3.82a}$$

by introducing the matrix

$$A := \begin{pmatrix} 2(h_0+h_1) & h_1 & & & & \\ h_1 & 2(h_1+h_2) & h_2 & & & \\ & h_2 & 2(h_2+h_3) & h_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots \\ & & & h_{N-3} & 2(h_{N-3}+h_{N-2}) & h_{N-2} \\ & & & & h_{N-2} & 2(h_{N-2}+h_{N-1}) \end{pmatrix} \tag{3.82b}$$

and the vectors

$$s'' := \begin{pmatrix} s_1'' \\ \vdots \\ s_{N-1}'' \end{pmatrix}, \quad g := \begin{pmatrix} g_1 \\ \vdots \\ g_{N-1} \end{pmatrix}. \tag{3.82c}$$

The matrix A from (3.82b) belongs to an especially benign class of matrices, so-called strictly diagonally dominant matrices:

Definition 3.34. Let A be a real $N \times N$ matrix, $N \in \mathbb{N}$. Then A is called *strictly diagonally dominant* if, and only if,

$$\sum_{\substack{j=1 \\ j \neq k}}^{N} |a_{kj}| < |a_{kk}| \quad \text{for each } k \in \{1, \dots, N\}. \tag{3.83}$$

Lemma 3.35. *Let A be a real $N \times N$ matrix, $N \in \mathbb{N}$. If A is strictly diagonally dominant, then, for each $x \in \mathbb{R}^N$,*

$$\|x\|_\infty \leq \max\left\{ \frac{1}{|a_{kk}| - \sum_{\substack{j=1\\ j\neq k}}^N |a_{kj}|} : k \in \{1,\dots,N\} \right\} \|Ax\|_\infty. \tag{3.84}$$

In particular, A is invertible with

$$\|A^{-1}\|_\infty \leq \max\left\{ \frac{1}{|a_{kk}| - \sum_{\substack{j=1\\ j\neq k}}^N |a_{kj}|} : k \in \{1,\dots,N\} \right\}, \tag{3.85}$$

where $\|A^{-1}\|_\infty$ denotes the row sum norm according to (2.52a).

Proof. Exercise. ■

Theorem 3.36. *Given a knot vector and values according to* (3.67)*, there is a unique cubic spline $s \in S_{\Delta,3}$ that satisfies* (3.68) *and the natural boundary conditions* (3.81a)*. Moreover, s can be computed by first solving the linear system* (3.82) *for the s''_k (the h_k and the g_k are given by* (3.72)*), then determining the a_k, b_k, c_k, and d_k according to* (3.74)*, and, finally, using* (3.69) *to get s.*

Proof. Lemma 3.33 shows that the existence and uniqueness of s is equivalent to (3.82a) having a unique solution. That (3.82a) does, indeed, have a unique solution follows from Lem. 3.35, as A as defined in (3.82b) is clearly strictly diagonally dominant. It is also part of the statement of Lem. 3.33 that s can be computed in the described way. ■

Theorem 3.37. *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 $\Delta := (x_0,\dots,x_N) \in \mathbb{R}^{N+1}$ for $[a,b]$, define*

$$h_k := x_{k+1} - x_k \quad \text{for each } k \in \{0,\dots,N-1\}, \tag{3.86a}$$

$$h_{\min} := \min\big\{h_k : k \in \{0,\dots,N-1\}\big\}, \tag{3.86b}$$

$$h_{\max} := \max\big\{h_k : k \in \{0,\dots,N-1\}\big\}. \tag{3.86c}$$

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

$$\max\Big\{\big|s''(x_k) - f''(x_k)\big| : k \in \{0,\dots,N\}\Big\} \leq C\,\|f^{(4)}\|_\infty\, h_{\max}^2, \tag{3.87}$$

then the following error estimates hold:

$$\|f - s\|_\infty \leq c\,\|f^{(4)}\|_\infty\, h_{\max}^4, \tag{3.88a}$$

$$\|f' - s'\|_\infty \leq 2c\,\|f^{(4)}\|_\infty\, h_{\max}^3, \tag{3.88b}$$

$$\|f'' - s''\|_\infty \leq 2c\,\|f^{(4)}\|_\infty\, h_{\max}^2, \tag{3.88c}$$

$$\|f''' - s'''\|_\infty \leq 2c\,\|f^{(4)}\|_\infty\, h_{\max}^1, \tag{3.88d}$$

where

$$c := \frac{h_{\max}}{h_{\min}}\left(C + \frac{1}{4}\right). \tag{3.89}$$

At the x_k, the third derivative s''' does not need to exist. In that case, (3.88d) *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 3.38. *Once again, consider the situation of Th.* 3.37*, but, instead of* (3.87)*, 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* (3.87) *with $C = 3/4$, and, in consequence, the error estimates* (3.88) *with $c = h_{\max}/h_{\min}$.*

Proof. We merely need to show that (3.87) holds with $C = 3/4$. Since s is the interpolating cubic spline satisfying natural boundary conditions, s'' satisfies the linear system (3.82a). Deviding, for each $k \in \{1, \dots, N-1\}$, the $(k-1)$th equation of that system by $3(h_{k-1} + h_k)$ yields the form

$$Bs'' = \hat{g}, \tag{3.90a}$$

with

$$B := \begin{pmatrix} \frac{2}{3} & \frac{h_1}{3(h_0+h_1)} & & & & \\ \frac{h_1}{3(h_1+h_2)} & \frac{2}{3} & \frac{h_2}{3(h_1+h_2)} & & & \\ & \frac{h_2}{3(h_2+h_3)} & \frac{2}{3} & \frac{h_3}{3(h_2+h_3)} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots \\ & & & & \frac{h_{N-3}}{3(h_{N-3}+h_{N-2})} & \frac{2}{3} & \frac{h_{N-2}}{3(h_{N-3}+h_{N-2})} \\ & & & & & \frac{h_{N-2}}{3(h_{N-2}+h_{N-1})} & \frac{2}{3} \end{pmatrix} \tag{3.90b}$$

and

$$\hat{g} := \begin{pmatrix} \hat{g}_1 \\ \vdots \\ \hat{g}_{N-1} \end{pmatrix} := \begin{pmatrix} \frac{g_1}{3(h_0+h_1)} \\ \vdots \\ \frac{g_{N-1}}{3(h_{N-2}+h_{N-1})} \end{pmatrix}. \tag{3.90c}$$

Taylor's theorem applied to f'' provides, for each $k \in \{1, \dots, N-1\}$, $\alpha_k \in]x_{k-1}, x_k[$ and $\beta_k \in]x_k, x_{k+1}[$ such that the following equations (3.91) hold, where (3.91a) has been multiplied by $\frac{h_{k-1}}{3(h_{k-1}+h_k)}$ and (3.91b) has been multiplied by $\frac{h_k}{3(h_{k-1}+h_k)}$:

$$\frac{h_{k-1}\, f''(x_{k-1})}{3(h_{k-1}+h_k)} = \frac{h_{k-1}\, f''(x_k)}{3(h_{k-1}+h_k)} - \frac{h_{k-1}^2\, f'''(x_k)}{3(h_{k-1}+h_k)} + \frac{h_{k-1}^3 f^{(4)}(\alpha_k)}{6(h_{k-1}+h_k)}, \tag{3.91a}$$

$$\frac{h_k\, f''(x_{k+1})}{3(h_{k-1}+h_k)} = \frac{h_k\, f''(x_k)}{3(h_{k-1}+h_k)} + \frac{h_k^2\, f'''(x_k)}{3(h_{k-1}+h_k)} + \frac{h_k^3 f^{(4)}(\beta_k)}{6(h_{k-1}+h_k)}. \tag{3.91b}$$

Adding (3.91a) and (3.91b) results in

$$\frac{h_{k-1}}{3(h_{k-1}+h_k)}\,f''(x_{k-1})+\frac{2}{3}\,f''(x_k)+\frac{h_k}{3(h_{k-1}+h_k)}\,f''(x_{k+1})=f''(x_k)+R_k+\delta_k, \tag{3.92a}$$

where

$$R_k := \frac{1}{3}\,(h_k-h_{k-1})\,f'''(x_k), \tag{3.92b}$$

$$\delta_k := \frac{1}{6(h_{k-1}+h_k)}\big(h_{k-1}^3 f^{(4)}(\alpha_k)+h_k^3 f^{(4)}(\beta_k)\big). \tag{3.92c}$$

Using the matrix B from (3.90b), (3.92a) takes the form

$$B\begin{pmatrix} f''(x_1)\\ \vdots\\ f''(x_{N-1})\end{pmatrix}=\begin{pmatrix} f''(x_1)\\ \vdots\\ f''(x_{N-1})\end{pmatrix}+\begin{pmatrix} R_1\\ \vdots\\ R_{N-1}\end{pmatrix}+\begin{pmatrix} \delta_1\\ \vdots\\ \delta_{N-1}\end{pmatrix}. \tag{3.93}$$

In the same way we applied Taylor's theorem to f'' to obtain (3.91), we now apply Taylor's theorem to f to obtain for each $k\in\{1,\dots,N-1\}$, $\xi_k\in]x_k,x_{k+1}[$ and $\eta_k\in]x_{k-1},x_k[$ such that

$$f(x_{k+1})=f(x_k)+h_k\,f'(x_k)+\frac{h_k^2}{2}\,f''(x_k)+\frac{h_k^3}{6}\,f'''(x_k)+\frac{h_k^4}{24}\,f^{(4)}(\xi_k), \tag{3.94a}$$

$$f(x_{k-1})=f(x_k)-h_{k-1}\,f'(x_k)+\frac{h_{k-1}^2}{2}\,f''(x_k)-\frac{h_{k-1}^3}{6}\,f'''(x_k)+\frac{h_{k-1}^4}{24}\,f^{(4)}(\eta_k). \tag{3.94b}$$

Multiplying (3.94a) and (3.94b) by $2/h_k$ and $2/h_{k-1}$, respectively, leads to

$$2\,\frac{f(x_{k+1})-f(x_k)}{h_k}=2\,f'(x_k)+h_k\,f''(x_k)+\frac{h_k^2}{3}\,f'''(x_k)+\frac{h_k^3}{12}\,f^{(4)}(\xi_k), \tag{3.95a}$$

$$-2\,\frac{f(x_k)-f(x_{k-1})}{h_{k-1}}=-2\,f'(x_k)+h_{k-1}\,f''(x_k)-\frac{h_{k-1}^2}{3}\,f'''(x_k)+\frac{h_{k-1}^3}{12}\,f^{(4)}(\eta_k). \tag{3.95b}$$

Adding (3.95a) and (3.95b) followed by a division by $h_{k-1}+h_k$ yields

$$2\,\frac{f(x_{k+1})-f(x_k)}{h_k(h_{k-1}+h_k)}-2\,\frac{f(x_k)-f(x_{k-1})}{h_{k-1}(h_{k-1}+h_k)}=f''(x_k)+R_k+\hat{\delta}_k, \tag{3.96}$$

where R_k is as in (3.92b) and

$$\hat{\delta}_k := \frac{1}{12(h_{k-1}+h_k)}\big(h_{k-1}^3 f^{(4)}(\eta_k)+h_k^3 f^{(4)}(\xi_k)\big).$$

Observing that

$$\hat{g}_k \overset{(3.90c)}{=} \frac{g_k}{3(h_{k-1}+h_k)} \overset{(3.72b),\,y_k=f(x_k)}{=} 2\,\frac{f(x_{k+1})-f(x_k)}{h_k(h_{k-1}+h_k)}-2\,\frac{f(x_k)-f(x_{k-1})}{h_{k-1}(h_{k-1}+h_k)},$$

we can write (3.96) as

$$\begin{pmatrix} f''(x_1) \\ \vdots \\ f''(x_{N-1}) \end{pmatrix} = \begin{pmatrix} \hat{g}_1 \\ \vdots \\ \hat{g}_{N-1} \end{pmatrix} - \begin{pmatrix} R_1 \\ \vdots \\ R_{N-1} \end{pmatrix} - \begin{pmatrix} \hat{\delta}_1 \\ \vdots \\ \hat{\delta}_{N-1} \end{pmatrix}. \tag{3.97}$$

Using (3.97) in the right-hand side of (3.93) and then subtracting (3.90a) 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

$$\frac{2}{3} - \frac{h_{k-1}}{3(h_{k-1}+h_k)} - \frac{h_k}{3(h_{k-1}+h_k)} = \frac{1}{3} \quad \text{for } k \in \{1,\dots,N-1\},$$

we conclude that B is strictly diagonally dominant, such that Lem. 3.35 allows us to estimate

$$\begin{aligned} &\max\left\{\left|f''(x_k) - s''(x_k)\right| : k \in \{0,\dots,N\}\right\} \leq 3\max\left\{|\delta_k| + |\hat{\delta}_k| : k \in \{1,\dots,N-1\}\right\} \\ &\overset{(*)}{\leq} \frac{3}{4}\, h_{\max}^2 \, \|f^{(4)}\|_\infty, \end{aligned} \tag{3.98}$$

where

$$|\delta_k| + |\hat{\delta}_k| \leq \left(\frac{1}{6} + \frac{1}{12}\right) \frac{h_{k-1}^3 + h_k^3}{h_{k-1}+h_k} \, \|f^{(4)}\|_\infty \leq \frac{1}{4}\, h_{\max}^2 \, \|f^{(4)}\|_\infty$$

was used for the inequality at $(*)$. The estimate (3.98) shows that s satisfies (3.87) with $C = 3/4$, thereby concluding the proof of the theorem. ■

Note that (3.88) is, in more than one way, much better than (3.65): From (3.88) we get convergence of the first three derivatives, and the convergence of $\|f - s\|_\infty$ is also much faster in (3.88) (proportional to h^4 rather than to h^2 in (3.65)). The disadvantage of (3.88), however, is that we needed to assume the existence of a continuous fourth derivative of f.

Remark 3.39. 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

$$f'(x_k) = s'(x_k). \tag{3.99}$$

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 (3.99). Thus, one will prefer one or the other form of interpolation, depending on either (3.99) (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 1 and 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.

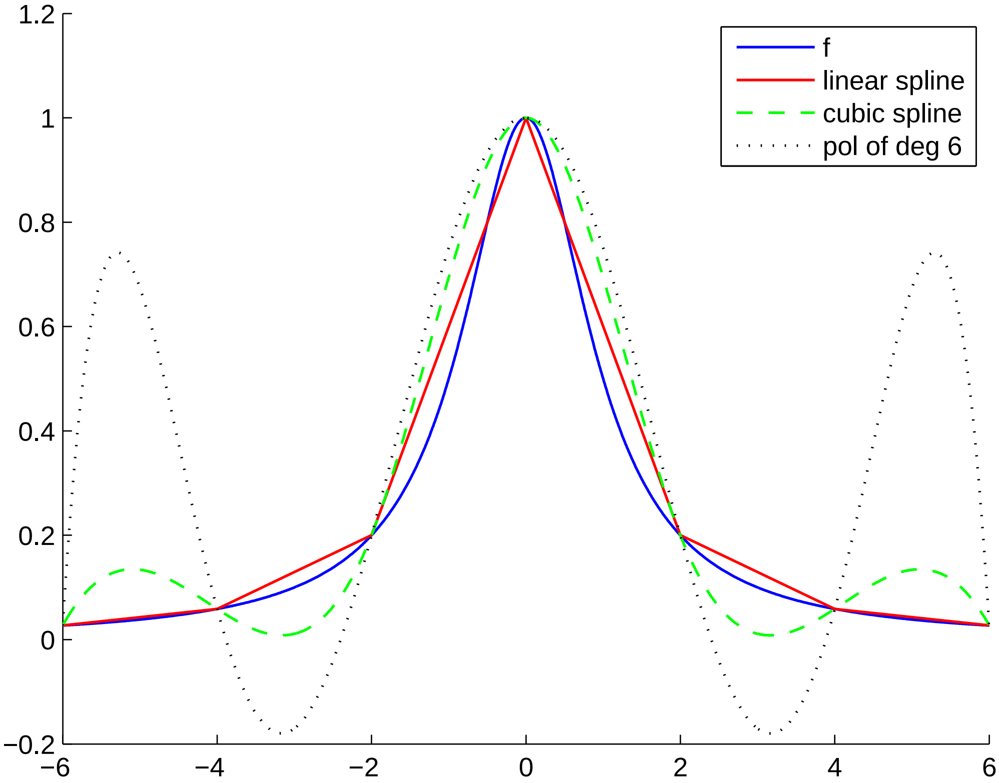

Figure 1: Different interpolations of $f(x) = 1/(x^2 + 1)$ with data points at $(x_0, \ldots, x_6) = (-6, -4, \ldots, 4, 6)$.

4 Numerical Integration

4.1 Introduction

The goal is the computation of (an approximation of) the value of definite integrals

$$\int_a^b f(x)\,\mathrm{d}x\,. \tag{4.1}$$

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

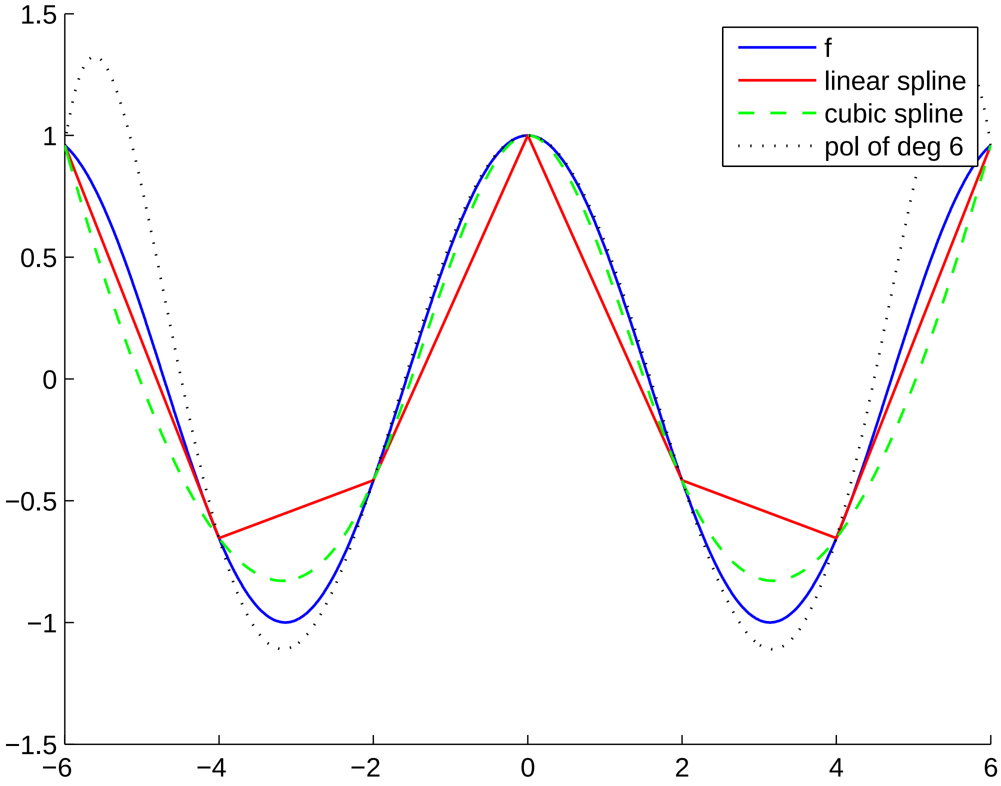

Figure 2: Different interpolations of $f(x) = \cos(x)$ with data points at $(x_0, \dots, x_6) = (-6, -4, \dots, 4, 6)$.

elementary functions (such as polynomials, exponential functions, and trigonometric functions). An example is the important function

$$\Phi(y) := \int_0^y \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, \mathrm{d}x\,, \tag{4.2}$$

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 $[a, b]$, $a < b$, $(x_0, \dots, x_n) \in \mathbb{R}^{n+1}$, $n \in \mathbb{N}$, is called a *partition* of $[a, b]$ if, and only if, $a = x_0 < x_1 < \dots < x_n = b$ (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 $(t_1, \dots, t_n) \in \mathbb{R}^n$ such that $t_i \in [x_{i-1}, x_i]$ for each $i \in \{1, \dots, n\}$.

Given a partition Δ (with or without tags) of $[a,b]$ as above, the number

$$h(\Delta) := \max\big\{h_i : i \in \{1,\dots,n\}\big\}, \quad h_i := x_i - x_{i-1}, \tag{4.3}$$

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

$$\sum_{i=1}^{n} f(t_i)h_i \tag{4.4}$$

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 $f : [a,b] \longrightarrow \mathbb{R}$ by $R[a,b]$.

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

$$\Delta_k = \big((x_0^k,\dots,x_{n_k}^k),\,(t_1^k,\dots,t_{n_k}^k)\big), \quad k \in \mathbb{N}, \tag{4.5}$$

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

$$I(f) := \int_a^b f(x)\,\mathrm{d}x := \lim_{k\to\infty} \sum_{i=1}^{n_k} f(t_i^k)h_i^k \tag{4.6}$$

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., [Phi13, 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 Φ is the function defined in (4.2), one can use the tagged partitions Δ_n of $[0,1]$ given by $x_i = t_i = i/n$ for $i \in \{0,\dots,n\}$, $n \in \mathbb{N}$. 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:

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:

n	I_n
1	0.2426
2	0.2978
10	0.3342
100	0.3415
1000	0.3422
5000	0.3422

—

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

$$\int_a^b \rho(x)\,\mathrm{d}x > 0 \tag{4.7}$$

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

$$I_\rho(f) := \int_a^b f(x)\rho(x)\,\mathrm{d}x\,, \tag{4.8}$$

that means the usual integral $I(f)$ corresponds to $\rho \equiv 1$ (note that $f\rho \in R[a, b]$, cf. [Phi13, Th. 10.17(c)]).

—

In the sequel, we will first study methods designed for the numerical solution of nonweighted 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

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

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

Remark 4.7. Note that every quadrature rule constitutes a linear functional on $R[a, b]$.

—

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 $\rho:\,[a,b]\longrightarrow\mathbb{R}_0^+$, a quadrature rule I_n is defined to have the *degree of accuracy* $r\in\mathbb{N}_0$ if, and only if,

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

but,

$$I_n(x^{r+1})\neq\int_a^b x^{r+1}\,\rho(x)\,\mathrm{d}x\,.\tag{4.10b}$$

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 $p\in\mathcal{P}_r$.

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

$$p:\,\mathbb{R}\longrightarrow\mathbb{R},\quad p(x):=\prod_{i=0}^{n}(x-x_i)^2,\tag{4.11}$$

since

$$I_n(p)=0\neq\int_a^b p(x)\rho(x)\,\mathrm{d}x\,.\tag{4.12}$$

In particular, I_n can have degree of accuracy at most $r=2n+1$.

(c) 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,

$$\int_a^b\rho(x)\,\mathrm{d}x\,=I_n(1)=(b-a)\sum_{i=0}^{n}\sigma_i,\tag{4.13}$$

which simplifies to

$$\sum_{i=0}^{n}\sigma_i=1\quad\text{for }\rho\equiv1.\tag{4.14}$$

4.2 Quadrature Rules Based on Interpolating Polynomials

In this section, we consider $\rho\equiv1$, 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 $x_0,\dots,x_n\in[a,b]$, $n\in\mathbb{N}_0$, is defined by

$$I_n:\,R[a,b]\longrightarrow\mathbb{R},\quad I_n(f):=\int_a^b p_n(x)\,\mathrm{d}x\,,\tag{4.15}$$

where $p_n=p_n(f)\in\mathcal{P}_n$ is the interpolating polynomial corresponding to the data points $(x_0,f(x_0)),\dots,(x_n,f(x_n))$.

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 $x_0,\dots,x_n \in [a,b]$, $n \in \mathbb{N}_0$, as defined in* (4.15)*.*

(a) *I_n is, indeed, a quadrature rule in the sense of Def.* 4.6*, where the weights σ_i, $i \in \{0,\dots,n\}$, are given in terms of the Lagrange basis polynomials L_i of* (3.3b) *by*

$$\begin{aligned}\sigma_i &= \frac{1}{b-a}\int_a^b L_i(x)\,\mathrm{d}x = \frac{1}{b-a}\int_a^b \prod_{\substack{k=0\\k\neq i}}^n \frac{x-x_k}{x_i-x_k}\,\mathrm{d}x\\ &\overset{(*)}{=} \int_0^1 \prod_{\substack{k=0\\k\neq i}}^n \frac{t-t_k}{t_i-t_k}\,\mathrm{d}t\,, \quad \textit{where} \quad t_i := \frac{x_i-a}{b-a}.\end{aligned} \tag{4.16}$$

(b) *I_n has at least degree of accuracy n, i.e. it is exact for each $q \in \mathcal{P}_n$.*

(c) *For each $f \in C^{n+1}[a,b]$, one has the error estimate*

$$\left|\int_a^b f(x)\,\mathrm{d}x - I_n(f)\right| \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!}\int_a^b |\omega_{n+1}(x)|\,\mathrm{d}x\,, \tag{4.17}$$

where $\omega_{n+1}(x) = \prod_{i=0}^n (x-x_i)$ is the Newton basis polynomial.

Proof. (a): (4.16) immediately follows by plugging (3.3) (with $y_j = f(x_j)$) into (4.15) and comparing with (4.9). The equality labeled $(*)$ is due to

$$\frac{x-x_k}{x_i-x_k} = \frac{\frac{x-a}{b-a} - \frac{x_k-a}{b-a}}{\frac{x_i-a}{b-a} - \frac{x_k-a}{b-a}}$$

and the change of variables $x \mapsto t := \frac{x-a}{b-a}$.

(b): For $q \in \mathcal{P}_n$, we have $p_n(q) = q$ such that $I_n(q) = \int_a^b q(x)\,\mathrm{d}x$ is clear from (4.15).

(c): This is also immediate, namely from (4.15) and (3.20). Actually, there is a subtle point here, since what we claimed is only true given the (Riemann) integrability of the function $x \mapsto f^{(n+1)}(\xi(x))$. Due to the dependence of ξ on x, this is not trivial. However, we will show in the proof of Prop. 4.14, that $x \mapsto f^{(n+1)}(\xi(x))$ is, indeed, even piecewise continuous on $[a,b]$. ■

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 (3.28) holds, i.e.

$$\|f^{(n)}\|_\infty \leq K n! R^n \quad \text{for each } n \in \mathbb{N}_0, \tag{4.18}$$

then, for each sequence of distinct numbers $(x_0, x_1, \dots)$ in $[a,b]$, the corresponding quadrature rules based on interpolating polynomials converge, i.e.

$$\lim_{n\to\infty}\left|\int_a^b f(x)\,\mathrm{d}x - I_n(f)\right| = 0\,: \tag{4.19}$$

Indeed, $|\int_a^b f(x)\,\mathrm{d}x - I_n(f)| \leq \int_a^b \|f-p_n\|_\infty\,\mathrm{d}x = (b-a)\,\|f-p_n\|_\infty$ and $\lim_{n\to\infty}\|f-p_n\|_\infty = 0$ according to (3.29).

—

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 $g:[a,b]\longrightarrow\mathbb{R}$ 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*

$$\int_a^b h(t)g(t)\,\mathrm{d}t = h(\tau)\int_a^b g(t)\,\mathrm{d}t\,. \tag{4.20}$$

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

$$m_h\int_a^b g(t)\,\mathrm{d}t \leq \int_a^b h(t)g(t)\,\mathrm{d}t \leq M_h\int_a^b g(t)\,\mathrm{d}t\,.$$

Thus, if we define

$$F:[a,b]\longrightarrow\mathbb{R},\quad F(s):=h(s)\int_a^b g(t)\,\mathrm{d}t - \int_a^b h(t)g(t)\,\mathrm{d}t\,,$$

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. τ satisfies (4.20). 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 $x_0,\dots,x_n\in[a,b]$, $n\in\mathbb{N}_0$, 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*

$$\int_a^b f(x)\,\mathrm{d}x - I_n(f) = \frac{f^{(n+1)}(\tau)}{(n+1)!}\int_a^b \omega_{n+1}(x)\,\mathrm{d}x\,. \tag{4.21}$$

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 (4.21)).

Proof. From Th. 3.26, we know that, for each $x\in[a,b]$, there exists $\xi(x)\in[a,b]$ such that (cf. (3.46))

$$f(x) = p_n(x) + \frac{f^{(n+1)}\big(\xi(x)\big)}{(n+1)!}\,\omega_{n+1}(x). \tag{4.22}$$

Combining (4.22) with (3.45) yields

$$f^{(n+1)}\big(\xi(x)\big) = f[x,x_0,\dots,x_n]\,(n+1)! \tag{4.23}$$

for each $x \in [a,b]$ such that $\omega_{n+1}(x) \neq 0$, i.e. for each $x \in [a,b] \setminus \{x_0, \dots, x_n\}$. Since $x \mapsto f[x, x_0, \dots, x_n]$ is continuous on $[a,b]$ by Prop. 3.24(a), the function $x \mapsto f^{(n+1)}\big(\xi(x)\big)$ is also continuous in each $x \in [a,b] \setminus \{x_0, \dots, x_n\}$, i.e. piecewise continuous on $[a,b]$. Moreover, an argument as in the proof of Th. 3.26(b) shows that the $\xi(x_0), \dots, \xi(x_n)$ can be chosen such that (4.23) holds for each $x \in [a,b]$ and $x \mapsto f^{(n+1)}\big(\xi(x)\big)$ is continuous on $[a,b]$. Thus, integrating (4.22) and applying Lem. 4.13 yields $\tau \in [a,b]$ satisfying

$$\int_a^b f(x)\,\mathrm{d}x \, - I_n(f) = \frac{1}{(n+1)!} \int_a^b f^{(n+1)}\big(\xi(x)\big)\omega_{n+1}(x)\,\mathrm{d}x \, = \frac{f^{(n+1)}(\tau)}{(n+1)!} \int_a^b \omega_{n+1}(x)\,\mathrm{d}x\,, \tag{4.24}$$

proving (4.21). ■

Remark 4.15. There exist precisely 3 examples where ω_{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 ω_{n+1}, ω_{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 $x_0, \dots, x_n \in [a,b]$, $n \in \mathbb{N}$, are equally-spaced. Moreover, a Newton-Cotes formula is called *closed* or a *Newton-Cotes closed quadrature rule* if, and only if,

$$x_i := a + i\,h, \quad h := \frac{b-a}{n}, \quad i \in \{0, \dots, n\}, \quad n \in \mathbb{N}. \tag{4.25}$$

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

$$\sigma_i = \sigma_{i,n} = \frac{1}{n} \int_0^n \prod_{\substack{k=0\\k\neq i}}^{n} \frac{s-k}{i-k}\,\mathrm{d}s\,. \tag{4.26}$$

Given n, one can compute the σ_i explicitly and store or tabulate them for further use. The next lemma shows that one actually does not need to compute *all* σ_i:

Lemma 4.17. *The weights σ_i of the Newton-Cotes closed quadrature rules (see* (4.26)*) are symmetric, i.e.*

$$\sigma_i = \sigma_{n-i} \quad \text{for each } i \in \{0, \dots, n\},\ n \in \mathbb{N}. \tag{4.27}$$

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 $p \in \mathcal{P}_n$, but even for $p \in \mathcal{P}_{n+1}$. 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, $i \in \{0,\dots,n\}$, are defined according to (4.25), then the antiderivative of the Newton basis polynomial $\omega := \omega_{n+1}$, i.e. the function*

$$F : \mathbb{R} \longrightarrow \mathbb{R}, \quad F(x) := \int_a^x \prod_{i=0}^{n} (y - x_i)\,\mathrm{d}y\,, \tag{4.28}$$

satisfies

$$F(x) \begin{cases} = 0 & \text{for } x = a, \\ > 0 & \text{for } a < x < b, \\ = 0 & \text{for } x = b. \end{cases} \tag{4.29}$$

Proof. $F(a) = 0$ is clear. The remaining assertions of (4.29) are shown in several steps.

Claim 1. With $h \in]0, b-a[$ defined according to (4.25), one has

$$\begin{matrix} \omega(x_{2j} + \tau) > 0, \\ \omega(x_{2j+1} + \tau) < 0, \end{matrix} \quad \text{for each } \tau \in]0,h[\text{ and each } j \in \Big\{0,\dots,\frac{n}{2} - 1\Big\}. \tag{4.30}$$

Proof. As ω has degree precisely $n+1$ and a zero at each of the $n+1$ distinct points $x_0,\dots,x_n$, each of these zeros must be simple. In particular, ω 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 ω is odd. This implies that ω 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 (4.30) now follows by induction. ▲

Claim 2. On the left half of the interval $[a,b]$, $|\omega|$ decays in the following sense:

$$\big|\omega(x+h)\big| < \big|\omega(x)\big| \quad \text{for each } x \in \left[a, \frac{a+b}{2} - h\right] \setminus \{x_0,\dots,x_{n/2-1}\}. \tag{4.31}$$

Proof. For each x that is not a zero of ω, we compute

$$\begin{aligned} \frac{\omega(x+h)}{\omega(x)} &= \frac{\prod_{i=0}^{n}(x+h-x_i)}{\prod_{i=0}^{n}(x-x_i)} = \frac{(x+h-a)\prod_{i=1}^{n}(x+h-x_i)}{(x-b)\prod_{i=0}^{n-1}(x-x_i)} \\ &\overset{x_i - h = x_{i-1}}{=} \frac{(x+h-a)\prod_{i=0}^{n-1}(x-x_i)}{(x-b)\prod_{i=0}^{n-1}(x-x_i)} = \frac{x+h-a}{x-b}. \end{aligned}$$

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 (4.31). ▲

Claim 3. $F(x) > 0$ for each $x \in \left]a, \frac{a+b}{2}\right]$.

Proof. There exists $\tau \in]0,h]$ and $i \in \{0,\dots,n/2-1\}$ such that $x = x_i + \tau$. Thus,

$$F(x) = \int_{x_i}^{x} \omega(y)\,\mathrm{d}y + \sum_{k=0}^{i-1} \int_{x_k}^{x_{k+1}} \omega(y)\,\mathrm{d}y\,, \tag{4.32}$$

and it suffices to show that, for each $\tau \in]0,h]$,

$$\int_{x_{2j}}^{x_{2j}+\tau} \omega(y)\,\mathrm{d}y > 0 \qquad \text{if } j \in \left\{k \in \mathbb{N}_0 : 0 \le k \le \frac{n/2-1}{2}\right\}, \tag{4.33a}$$

$$\int_{x_{2j}}^{x_{2j+1}+\tau} \omega(y)\,\mathrm{d}y > 0 \qquad \text{if } j \in \left\{k \in \mathbb{N}_0 : 0 \le k \le \frac{n/2-2}{2}\right\}. \tag{4.33b}$$

While (4.33a) is immediate from (4.30), for (4.33b), one calculates

$$\int_{x_{2j}}^{x_{2j+1}+\tau} \omega(y)\,\mathrm{d}y = \int_{x_{2j}}^{x_{2j}+\tau} \overbrace{\omega(y) + \underbrace{\omega(y+h)}_{=-|\omega(y+h)|}}^{>0 \text{ by } (4.31)} \mathrm{d}y + \overbrace{\int_{x_{2j}+\tau}^{x_{2j+1}} \omega(y)\,\mathrm{d}y}^{\ge 0 \text{ by } (4.30)} > 0$$

for each $\tau \in]0,h]$ and each $0 \le j \le \frac{n/2-2}{2}$, thereby establishing the case. ▲

Claim 4. ω is antisymmetric with respect to the midpoint of $[a,b]$, i.e.

$$\omega\left(\frac{a+b}{2} + \tau\right) = -\omega\left(\frac{a+b}{2} - \tau\right) \quad \text{for each } \tau \in \mathbb{R}. \tag{4.34}$$

Proof. From $(a+b)/2 - x_i = -\big((a+b)/2 - x_{n-i}\big) = (n/2-i)h$ for each $i \in \{0,\dots,n\}$, one obtains

$$\begin{aligned}\omega\left(\frac{a+b}{2} + \tau\right) &= \prod_{i=0}^{n}\left(\frac{a+b}{2} + \tau - x_i\right) = -\prod_{i=0}^{n}\left(\frac{a+b}{2} - \tau - x_{n-i}\right) \\ &= -\prod_{i=0}^{n}\left(\frac{a+b}{2} - \tau - x_i\right) = -\omega\left(\frac{a+b}{2} - \tau\right),\end{aligned}$$

which is (4.34). ▲

Claim 5. F is symmetric with respect to the midpoint of $[a,b]$, i.e.

$$F\left(\frac{a+b}{2} + \tau\right) = F\left(\frac{a+b}{2} - \tau\right) \quad \text{for each } \tau \in \mathbb{R}. \tag{4.35}$$

Proof. For each $\tau \in \left[0, \frac{b-a}{2}\right]$, we obtain

$$F\left(\frac{a+b}{2} + \tau\right) = \int_a^{(a+b)/2-\tau} \omega(x)\,\mathrm{d}x + \int_{(a+b)/2-\tau}^{(a+b)/2+\tau} \omega(x)\,\mathrm{d}x = F\left(\frac{a+b}{2} - \tau\right),$$

since the second integral vanishes due to (4.34). ▲

Finally, (4.29) 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 $I_n : R[a, b] \longrightarrow \mathbb{R}$ 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

$$I(x^{n+2}) \neq I_n(x^{n+2}). \tag{4.36}$$

To that end, let $x_{n+1} \in [a, b] \setminus \{x_0, \dots, x_n\}$, and let $q \in \mathcal{P}_{n+1}$ be the interpolating polynomial for $f(x) = x^{n+2}$ and $x_0, \dots, x_{n+1}$, whereas $p_n \in \mathcal{P}_n$ is the corresponding interpolating polynomial for $x_0, \dots, x_n$ (note that p_n interpolates both f and q). As we already know that I_n has at least degree of accuracy $n + 1$, we obtain

$$\int_a^b p_n(x)\,\mathrm{d}x = I_n(f) = I_n(q) = \int_a^b q(x)\,\mathrm{d}x\,. \tag{4.37}$$

By applying (3.20) with $f(x) = x^{n+2}$ and using $f^{(n+2)}(x) = (n + 2)!$, one obtains

$$f(x)-q(x) = \frac{f^{(n+2)}\big(\xi(x)\big)}{(n+2)!}\,\omega_{n+2}(x) = \omega_{n+2}(x) = \prod_{i=0}^{n+1}(x-x_i) \quad \text{for each } x \in [a,b]. \tag{4.38}$$

Integrating (4.38) while taking into account (4.37) as well as using F from (4.28) yields

$$\begin{aligned} I(f) - I_n(f) &= \int_a^b \prod_{i=0}^{n+1}(x - x_i)\,\mathrm{d}x = \int_a^b F'(x)(x - x_{n+1})\,\mathrm{d}x \\ &= \big[F(x)(x - x_{n+1})\big]_a^b - \int_a^b F(x)\,\mathrm{d}x \overset{(4.29)}{=} -\int_a^b F(x)\,\mathrm{d}x \overset{(4.29)}{<} 0, \end{aligned}$$

proving (4.36). ■

4.3.2 Rectangle Rules ($n = 0$)

Note that, for $n = 0$, there is no closed Newton-Cotes formula, as (4.25) 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 $\int_a^b f(x)\,\mathrm{d}x$ 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 are usually not exact for polynomials of degree 1 (i.e. for non-constant affine functions); however, the midpoint rule has degree of accuracy $r = 1$ (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*

$$\int_a^b f(x)\,\mathrm{d}x - (b-a)f(a) = \frac{(b-a)^2}{2}\, f'(\tau). \tag{4.39}$$

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

$$\int_a^b f(x)\,\mathrm{d}x - (b-a)f(a) = f'(\tau) \int_a^b (x-a)\,\mathrm{d}x = \frac{(b-a)^2}{2}\, f'(\tau),$$

proving (4.39). ■

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):

$$\begin{aligned} I_1(f) :&= \int_a^b \left(f(a) + \frac{f(b)-f(a)}{b-a}\,(x-a) \right) \mathrm{d}x = \left[f(a)x + \frac{1}{2}\frac{f(b)-f(a)}{b-a}\,(x-a)^2 \right]_a^b \\ &= (b-a)\left(\frac{1}{2}\, f(a) + \frac{1}{2}\, f(b) \right) \end{aligned} \tag{4.40}$$

for each $f \in C[a,b]$. Note that (4.40) 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*

$$\int_a^b f(x)\,\mathrm{d}x - I_1(f) = -\frac{(b-a)^3}{12}\, f''(\tau) = -\frac{h^3}{12}\, f''(\tau). \tag{4.41}$$

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 (4.26), one finds

$$\sigma_0 = \frac{1}{2} \int_0^2 \frac{s-1}{-1}\,\frac{s-2}{-2}\,\mathrm{d}s = \frac{1}{4}\left[\frac{s^3}{3} - \frac{3s^2}{2} + 2s \right]_0^2 = \frac{1}{6}. \tag{4.42}$$

Next, one obtains $\sigma_2 = \sigma_0 = \frac{1}{6}$ from (4.27) and $\sigma_1 = 1 - \sigma_0 - \sigma_2 = \frac{2}{3}$ from Rem. 4.9. Plugging this into (4.9) yields

$$I_2(f) = (b-a)\left(\frac{1}{6}f(a) + \frac{2}{3}f\left(\frac{a+b}{2}\right) + \frac{1}{6}f(b)\right) \tag{4.43}$$

for each $f \in C[a,b]$.

Lemma 4.25. *Simpson's rule 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*

$$\int_a^b f(x)\,\mathrm{d}x - I_2(f) = -\frac{(b-a)^5}{2880}\,f^{(4)}(\tau) = -\frac{h^5}{90}\,f^{(4)}(\tau). \tag{4.44}$$

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 $q \in \mathcal{P}_3$ 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 $p_2 \in \mathcal{P}_2$ 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

$$\int_a^b p_2(x)\,\mathrm{d}x = I_2(f) = I_2(q) = \int_a^b q(x)\,\mathrm{d}x\,. \tag{4.45}$$

By applying (3.46) with $n = 3$, one obtains

$$f(x) = q(x) + \frac{f^{(4)}\big(\xi(x)\big)}{4!}\,\omega_4(x) \quad \text{for each } x \in [a,b], \tag{4.46}$$

where

$$\omega_4(x) = (x-a)\left(x - \frac{a+b}{2}\right)^2 (x-b). \tag{4.47}$$

As in the proof of Prop. 4.14, it follows that the function $x \mapsto f^{(4)}\big(\xi(x)\big)$ can be chosen continuous. Moreover, according to (4.47), $\omega_4(x) \le 0$ for each $x \in [a,b]$. Thus, integrating (4.46) and applying (4.45) as well as Lem. 4.13 yields $\tau \in [a,b]$ satisfying

$$\begin{aligned}\int_a^b f(x)\,\mathrm{d}x - I_2(f) &= \frac{f^{(4)}(\tau)}{4!}\int_a^b \omega_4(x)\,\mathrm{d}x = -\frac{f^{(4)}(\tau)}{24}\,\frac{(b-a)^5}{120}\\ &= -\frac{(b-a)^5}{2880}\,f^{(4)}(\tau) = -\frac{h^5}{90}\,f^{(4)}(\tau),\end{aligned} \tag{4.48}$$

where it was used that

$$\begin{aligned}\int_a^b \omega_4(x)\,\mathrm{d}x &= -\int_a^b \frac{(x-a)^2}{2}\left(2\left(x-\frac{a+b}{2}\right)(x-b)+\left(x-\frac{a+b}{2}\right)^2\right)\mathrm{d}x\\ &= -\left(\frac{(b-a)^5}{24}-\int_a^b \frac{(x-a)^3}{6}\left(2(x-b)+4\left(x-\frac{a+b}{2}\right)\right)\mathrm{d}x\right)\\ &= -\left(\frac{(b-a)^5}{24}-2\frac{(b-a)^5}{24}+\int_a^b \frac{(x-a)^4}{24}\,6\,\mathrm{d}x\right)\\ &= -\frac{(b-a)^5}{120}. \end{aligned}\tag{4.49}$$

This completes the proof of (4.44). ■

4.3.5 Higher Order Newton-Cotes Formulas

As before, let $[a,b]$ be a real interval, $a<b$, and let $I_n : R[a,b] \longrightarrow \mathbb{R}$ be the Newton-Cotes closed quadrature rule based on interpolating polynomials $p \in \mathcal{P}_n$. 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 $f \in R[a,b]$ (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

$$\lim_{n\to\infty}\sum_{i=0}^{n}|\sigma_{i,n}| = \infty. \tag{4.50}$$

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 (4.50), 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*

$$I_n : C[a,b] \longrightarrow \mathbb{R}, \quad I_n(f) = (b-a)\sum_{i=0}^{n}\sigma_i\, f(x_i), \tag{4.51}$$

with distinct $x_0, \dots, x_n \in [a,b]$ and weights $\sigma_0, \dots, \sigma_n \in \mathbb{R}$, $n \in \mathbb{N}$, the operator norm of I_n induced by $\|\cdot\|_\infty$ on $C[a,b]$ is

$$\|I_n\| = (b-a) \sum_{i=0}^{n} |\sigma_i|. \tag{4.52}$$

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 $\phi : \{0,\dots,n\} \longrightarrow \{0,\dots,n\}$ to be a reordering such that $x_{\phi(0)} < x_{\phi(1)} < \dots < x_{\phi(n)}$, let $y_i := \operatorname{sgn}(\sigma_{\phi(i)})$ for each $i \in \{0,\dots,n\}$, and

$$f(x) := \begin{cases} y_0 & \text{for } x \in [a, x_{\phi(0)}], \\ y_i + \frac{y_{i+1}-y_i}{x_{\phi(i+1)}-x_{\phi(i)}} (x - x_{\phi(i)}) & \text{for } x \in [x_{\phi(i)}, x_{\phi(i+1)}],\ i \in \{0,\dots,n-1\}, \\ y_n & \text{for } x \in [x_{\phi(n)}, b]. \end{cases}$$

Note that f is a linear spline on $[x_{\phi(0)}, x_{\phi(n)}]$ (cf. (3.64)), 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

$$I_n(f) = (b-a) \sum_{i=0}^{n} \sigma_i \, f(x_i) = (b-a) \sum_{i=0}^{n} \sigma_i \operatorname{sgn}(\sigma_i) = (b-a) \sum_{i=0}^{n} |\sigma_i|,$$

showing the remaining inequality $\|I_n\| \geq (b-a) \sum_{i=0}^{n} |\sigma_i|$ and, thus, (4.52). ■

Theorem 4.28 (Polya)**.** *Given a real interval $[a,b]$, $a < b$, and a weight function $\rho : [a,b] \longrightarrow \mathbb{R}_0^+$, consider a sequence of quadrature rules*

$$I_n : C[a,b] \longrightarrow \mathbb{R}, \quad I_n(f) = (b-a) \sum_{i=0}^{n} \sigma_{i,n} \, f(x_{i,n}), \tag{4.53}$$

with distinct points $x_{0,n}, \dots, x_{n,n} \in [a,b]$ and weights $\sigma_{0,n}, \dots, \sigma_{n,n} \in \mathbb{R}$, $n \in \mathbb{N}$. The sequence converges in the sense that

$$\lim_{n\to\infty} I_n(f) = \int_a^b f(x)\rho(x) \, \mathrm{d}x \quad \textit{for each } f \in C[a,b] \tag{4.54}$$

if, and only if, the following two conditions are satisfied:

(i) $\lim_{n\to\infty} I_n(p) = \int_a^b p(x)\rho(x)\,\mathrm{d}x$ *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*

$$\sum_{i=0}^{n} |\sigma_{i,n}| \leq C \quad \textit{for each } n \in \mathbb{N}.$$

Proof. Suppose (i) and (ii) hold true. Let $W := \int_a^b \rho(x)\,\mathrm{d}x$. Given $f \in C[a,b]$ and $\epsilon > 0$, according to the Weierstrass Approximation Theorem 3.27, there exists a polynomial p such that

$$\|f - p{\restriction_{[a,b]}}\|_\infty < \tilde{\epsilon} := \frac{\epsilon}{C(b-a) + 1 + W}. \tag{4.55}$$

Due to (i), there exists $N \in \mathbb{N}$ such that, for each $n \geq N$,

$$\left| I_n(p) - \int_a^b p(x)\rho(x)\,\mathrm{d}x \right| < \tilde{\epsilon}.$$

Thus, for each $n \geq N$,

$$\begin{aligned}
&\left| I_n(f) - \int_a^b f(x)\rho(x)\,\mathrm{d}x \right| \\
&\leq \big|I_n(f) - I_n(p)\big| + \left| I_n(p) - \int_a^b p(x)\rho(x)\,\mathrm{d}x \right| + \left| \int_a^b p(x)\rho(x)\,\mathrm{d}x - \int_a^b f(x)\rho(x)\,\mathrm{d}x \right| \\
&< (b-a)\sum_{i=0}^n |\sigma_{i,n}|\,\big|f(x_{i,n}) - p(x_{i,n})\big| + \tilde{\epsilon} + \tilde{\epsilon}W < \tilde{\epsilon}\big((b-a)C + 1 + W\big) = \epsilon,
\end{aligned}$$

proving (4.54).

Conversely, (4.54) 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 (4.54) implies

$$\sup\big\{|T(f)| : T \in \mathcal{T}\big\} = \sup\big\{|I_n(f)| : n \in \mathbb{N}\big\} < \infty \quad \text{for each } f \in C[a,b]. \tag{4.56}$$

Then the Banach-Steinhaus Th. G.4 yields

$$\sup\big\{\|I_n\| : n \in \mathbb{N}\big\} = \sup\big\{\|T\| : T \in \mathcal{T}\big\} < \infty, \tag{4.57}$$

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 G. 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 $q \in \mathcal{P}_n$, $I_m(q)$ is exact for each $m \geq n$, Condition (ii) of Th. 4.28 is violated due to (4.50). 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. 3.12), 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 $f : [0, 1] \longrightarrow \mathbb{R}$, I_n, $n \in \mathbb{N}$, has the form

$$I_n(f) = \sum_{l=0}^{n} \sigma_l f(\xi_l) \tag{4.58}$$

with $0 = \xi_0 < \xi_1 < \cdots < \xi_n = 1$ and weights $\sigma_0, \dots, \sigma_n \in \mathbb{R}$. Given a real interval $[a, b]$, $a < b$, and $N \in \mathbb{N}$, set

$$x_k := a + k\,h, \quad h := \frac{b-a}{N}, \qquad \text{for each } k \in \{0, \dots, N\}, \tag{4.59a}$$

$$x_{kl} := x_{k-1} + h\,\xi_l, \qquad \text{for each } k \in \{1, \dots, N\}, \quad l \in \{0, \dots, n\}. \tag{4.59b}$$

Then

$$I_{n,N}(f) := h \sum_{k=1}^{N} \sum_{l=0}^{n} \sigma_l f(x_{kl}), \quad f : [a, b] \longrightarrow \mathbb{R}, \tag{4.60}$$

are called the *composite quadrature rules* based on I_n.

Remark 4.30. The heuristics behind (4.60) is

$$\int_{x_{k-1}}^{x_k} f(x)\,\mathrm{d}x = h \int_0^1 f(x_{k-1} + \xi h)\,\mathrm{d}\xi \approx h \sum_{l=0}^{n} \sigma_l f(x_{kl}). \tag{4.61}$$

Theorem 4.31. *Let $[a,b]$, $a<b$, be a real interval, $n \in \mathbb{N}_0$. If I_n has the form* (4.58) *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* (4.60) *based on I_n converge for each Riemann integrable $f : [a,b] \longrightarrow \mathbb{R}$, i.e.*

$$\forall_{f \in R[a,b]} \quad \lim_{N\to\infty} I_{n,N}(f) = \int_a^b f(x)\,\mathrm{d}x\,. \tag{4.62}$$

Proof. Let $f \in R[a,b]$. Introducing, for each $l \in \{0,\dots,n\}$, the abbreviation

$$S_l(N) := \sum_{k=1}^{N} \frac{b-a}{N} f(x_{kl}), \tag{4.63}$$

(4.60) yields

$$I_{n,N}(f) = \sum_{l=0}^{n} \sigma_l S_l(N). \tag{4.64}$$

Note that, for each $l \in \{0,\dots,n\}$, $S_l(N)$ is a Riemann sum for the tagged partition $\Delta_{Nl} := \big((x_0^N,\dots,x_N^N), (x_{1l}^N,\dots,x_{Nl}^N)\big)$ and $\lim_{N\to\infty} h(\Delta_{Nl}) = \lim_{N\to\infty} \frac{b-a}{N} = 0$. Thus, applying Th. 4.3,

$$\lim_{N\to\infty} S_l(N) = \int_a^b f(x)\,\mathrm{d}x \quad \text{for each } l \in \{0,\dots,n\}.$$

This, in turn, implies

$$\lim_{N\to\infty} I_{n,N}(f) = \lim_{N\to\infty} \sum_{l=0}^{n} \sigma_l S_l(N) = \int_a^b f(x)\,\mathrm{d}x \sum_{l=0}^{n} \sigma_l = \int_a^b f(x)\,\mathrm{d}x$$

as, due to hypothesis,

$$\sum_{l=0}^{n} \sigma_l = I_n(1) = \int_0^1 \mathrm{d}x\, = 1,$$

proving (4.62). ■

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 (4.60) 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 $K = K(f) \in \mathbb{R}_0^+$ such that

$$\left| I_{n,N}(f) - \int_a^b f(x)\,\mathrm{d}x \right| \le K\,h^r \quad \text{for each } N \in \mathbb{N} \tag{4.65}$$

($h = (b-a)/N$ as before).

4.5.2 Composite Rectangle Rules ($n = 0$)

Definition and Remark 4.33. Using the rectangle rule $\int_0^1 f(x)\,\mathrm{d}x \approx f(0)$ of Sec. 4.3.2 as I_0 on $[0,1]$, the formula (4.60) yields, for $f : [a,b] \longrightarrow \mathbb{R}$, the corresponding *composite rectangle rules*

$$I_{0,N}(f) = h \sum_{k=1}^{N} f(x_{k0}) \overset{x_{k0}=x_{k-1}}{=} h\big(f(x_0) + f(x_1) + \cdots + f(x_{N-1})\big). \tag{4.66}$$

Theorem 4.34. *For each $f \in C^1[a,b]$, there exists $\tau \in [a,b]$ satisfying*

$$\int_a^b f(x)\,\mathrm{d}x - I_{0,N}(f) = \frac{(b-a)^2}{2N} f'(\tau) = \frac{b-a}{2}\, h\, f'(\tau). \tag{4.67}$$

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 $k \in \{1,\dots,N\}$, a point $\tau_k \in [x_{k-1}, x_k]$ such that

$$\int_{x_{k-1}}^{x_k} f(x)\,\mathrm{d}x - \frac{b-a}{N} f(x_{k-1}) = \frac{(b-a)^2}{2N^2} f'(\tau_k) = \frac{h^2}{2} f'(\tau_k). \tag{4.68}$$

Summing (4.68) from $k = 1$ through N yields

$$\int_a^b f(x)\,\mathrm{d}x - I_{0,N}(f) = \frac{b-a}{2}\, h\, \frac{1}{N} \sum_{k=1}^{N} f'(\tau_k).$$

As f' is continuous and

$$\min\big\{f'(x) : x \in [a,b]\big\} \le \alpha := \frac{1}{N} \sum_{k=1}^{N} f'(\tau_k) \le \max\big\{f'(x) : x \in [a,b]\big\},$$

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 (4.65) holds with $r = 1$ and $K := (b-a)\|f'\|_\infty/2$. ■

4.5.3 Composite Trapezoidal Rules ($n = 1$)

Definition and Remark 4.35. Using the trapezoidal rule (4.40) as I_1 on $[0,1]$, the formula (4.60) yields, for $f : [a,b] \longrightarrow \mathbb{R}$, the corresponding *composite trapezoidal rules*

$$\begin{aligned} I_{1,N}(f) &= h \sum_{k=1}^{N} \frac{1}{2}\big(f(x_{k0}) + f(x_{k1})\big) \overset{\substack{x_{k0}=x_{k-1},\\ x_{k1}=x_k}}{=} h \sum_{k=1}^{N} \frac{1}{2}\big(f(x_{k-1}) + f(x_k)\big) \\ &= \frac{h}{2}\Big(f(a) + 2\big(f(x_1) + \cdots + f(x_{N-1})\big) + f(b)\Big). \end{aligned} \tag{4.69}$$

Theorem 4.36. *For each $f \in C^2[a,b]$, there exists $\tau \in [a,b]$ satisfying*

$$\int_a^b f(x)\,\mathrm{d}x - I_{1,N}(f) = -\frac{(b-a)^3}{12N^2}\,f''(\tau) = -\frac{b-a}{12}\,h^2\,f''(\tau). \tag{4.70}$$

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 $k \in \{1,\dots,N\}$, a point $\tau_k \in [x_{k-1}, x_k]$ such that

$$\int_{x_{k-1}}^{x_k} f(x)\,\mathrm{d}x - \frac{h}{2}\big(f(x_{k-1}) + f(x_k)\big) = -\frac{h^3}{12}\,f''(\tau_k).$$

Summing the above equation from $k = 1$ through N yields

$$\int_a^b f(x)\,\mathrm{d}x - I_{1,N}(f) = -\frac{b-a}{12}\,h^2\,\frac{1}{N}\sum_{k=1}^N f''(\tau_k).$$

As f'' is continuous and

$$\min\big\{f''(x) : x \in [a,b]\big\} \le \alpha := \frac{1}{N}\sum_{k=1}^N f''(\tau_k) \le \max\big\{f''(x) : x \in [a,b]\big\},$$

the intermediate value theorem provides $\tau \in [a,b]$ satisfying $\alpha = f''(\tau)$, proving (4.70). The claimed order of convergence now follows as well, since (4.70) implies that (4.65) 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 (4.43) as I_2 on $[0,1]$, the formula (4.60) yields, for $f : [a,b] \longrightarrow \mathbb{R}$, the corresponding *composite Simpson's rules $I_{2,N}(f)$*. It is an exercise to check that

$$I_{2,N}(f) = \frac{h}{6}\left(f(a) + 4\sum_{k=1}^N f(x_{k-\frac{1}{2}}) + 2\sum_{k=1}^{N-1} f(x_k) + f(b)\right), \tag{4.71}$$

where $x_{k-\frac{1}{2}} := (x_{k-1} + x_k)/2$ for each $k \in \{1,\dots,N\}$.

Theorem 4.38. *For each $f \in C^4[a,b]$, there exists $\tau \in [a,b]$ satisfying*

$$\int_a^b f(x)\,\mathrm{d}x - I_{2,N}(f) = -\frac{(b-a)^5}{2880N^4}\,f^{(4)}(\tau) = -\frac{b-a}{2880}\,h^4\,f^{(4)}(\tau). \tag{4.72}$$

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. 3.14. 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. 4.52 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 quadrature, it is common to use the notation

$$I_n(f) = \sum_{i=1}^{n} \sigma_i \, f(\lambda_i), \tag{4.73}$$

i.e. one writes λ_i instead of x_i and the factor $(b-a)$ is incorporated into the weights σ_i.

4.6.2 Orthogonal Polynomials

Notation 4.39. Let $\mathcal{P}$ denote the real vector space of polynomials $p: \mathbb{R} \longrightarrow \mathbb{R}$.

—

During Gaussian quadrature, the λ_i are determined as the zeros of polynomials that are orthogonal with respect to certain inner products (also known as scalar products) on the (infinite-dimensional) real vector space $\mathcal{P}$ (see Appendix F).

Notation 4.40. Given a real interval $[a,b]$, $a<b$, and a weight function $\rho: [a,b] \longrightarrow \mathbb{R}_0^+$ as defined in Def. 4.5, let

$$\langle \cdot, \cdot \rangle_\rho : \mathcal{P} \times \mathcal{P} \longrightarrow \mathbb{R}, \qquad \langle p, q \rangle_\rho := \int_a^b p(x)q(x)\rho(x)\,\mathrm{d}x\,, \tag{4.74a}$$

$$\| \cdot \|_\rho : \mathcal{P} \longrightarrow \mathbb{R}_0^+, \qquad \|p\|_\rho := \sqrt{\langle p, p \rangle_\rho}. \tag{4.74b}$$

Lemma 4.41. *Given a real interval $[a,b]$, $a<b$, and a weight function $\rho: [a,b] \longrightarrow \mathbb{R}_0^+$, (4.74a) defines an inner product on $\mathcal{P}$.*

Proof. As $\rho \in R[a,b]$, $\rho \geq 0$, and $\int_a^b \rho(x)\,\mathrm{d}x > 0$, there must be some nontrivial interval $J \subseteq [a,b]$ and $\epsilon > 0$ such that $\rho \geq \epsilon$ on J. Thus, if $0 \neq p \in \mathcal{P}$, then there must be some nontrivial interval $J_p \subseteq J$ and $\epsilon_p > 0$ such that $p^2\rho \geq \epsilon_p$ on J_p, implying $\langle p,p\rangle_\rho = \int_a^b p^2(x)\rho(x)\,\mathrm{d}x \ \geq \epsilon_p|J_p| > 0$ and proving $\langle\cdot,\cdot\rangle_\rho$ satisfies F.1(i). Given $p,q,r \in \mathcal{P}$ and $\lambda,\mu \in \mathbb{R}$, one computes

$$\begin{aligned}\langle \lambda p + \mu q, r\rangle_\rho &= \int_a^b \big(\lambda p(x) + \mu q(x)\big) r(x)\rho(x)\,\mathrm{d}x \\ &= \lambda \int_a^b q(x)r(x)\rho(x)\,\mathrm{d}x \ + \mu \int_a^b q(x)r(x)\rho(x)\,\mathrm{d}x \\ &= \lambda\langle p,r\rangle_\rho + \mu\langle q,r\rangle_\rho,\end{aligned}$$

proving F.1(ii). Since, for $p,q \in \mathcal{P}$, one also has

$$\langle p,q\rangle_\rho = \int_a^b p(x)q(x)\rho(x)\,\mathrm{d}x \ = \int_a^b q(x)p(x)\rho(x)\,\mathrm{d}x \ = \langle q,p\rangle_\rho,$$

such that F.1(iii) holds as well, the proof of the lemma is complete. ■

Remark 4.42. Given a real interval $[a,b]$, $a<b$, and a weight function $\rho:\ [a,b] \longrightarrow \mathbb{R}_0^+$, consider the inner product space $\big(\mathcal{P}, \langle\cdot,\cdot\rangle_\rho\big)$, where $\langle\cdot,\cdot\rangle_\rho$ is given by (4.74a). Then, for each $n \in \mathbb{N}_0$, $q \in \mathcal{P}_n^\perp$ (cf. (F.2)) if, and only if,

$$\int_a^b q(x)p(x)\rho(x) = 0 \quad \text{for each } p \in \mathcal{P}_n. \tag{4.75}$$

At first glance, it might not be obvious that there are always $q \in \mathcal{P}\setminus\{0\}$ that satisfy (4.75). However, such $q \neq 0$ can always be found by the general procedure known as Gram-Schmidt orthogonalization which is reviewed in the following theorem.

Theorem 4.43 (Gram-Schmidt Orthogonalization)**.** *Let $\big(X,\langle\cdot,\cdot\rangle\big)$ be an inner product space with induced norm $\|\cdot\|$ according to* (F.1)*. Let $x_0,x_1,\dots$ be a finite or infinite sequence of vectors in X. Define $v_0,v_1,\dots$ recursively as follows:*

$$v_0 := x_0, \quad v_n := x_n - \sum_{\substack{k=0,\\ v_k\neq 0}}^{n-1} \frac{\langle x_n, v_k\rangle}{\|v_k\|^2}\, v_k \tag{4.76}$$

for each $n \in \mathbb{N}$, additionally assuming that n is less than or equal to the max index of the sequence $x_0,x_1,\dots$ if the sequence is finite. Then the sequence $v_0,v_1,\dots$ constitutes an orthogonal system (see Def. F.2(a)*). Of course, by omitting the $v_k = 0$ and by dividing each $v_k \neq 0$ by its norm, one can also obtain an orthonormal system (nonempty if at least one $v_k \neq 0$). Moreover, $v_n = 0$ if, and only if, $x_n \in \operatorname{span}\{x_0,\dots,x_{n-1}\}$. In particular, if the $x_0,x_1,\dots$ are all linearly independent, then so are the $v_0,v_1,\dots$.*

Proof. We show by induction on n, that, for each $0 \leq m < n$, $v_n \perp v_m$. For $n=0$, there is nothing to show. Thus, let $n>0$ and $0 \leq m < n$. By induction, $\langle v_k,v_m\rangle = 0$ for each $0 \leq k,m < n$ such that $k \neq m$. For $v_m = 0$, $\langle v_n,v_m\rangle = 0$ is clear. Otherwise,

$$\langle v_n,v_m\rangle = \left\langle x_n - \sum_{\substack{k=0,\\ v_k\neq 0}}^{n-1} \frac{\langle x_n, v_k\rangle}{\|v_k\|^2}\, v_k,\ v_m \right\rangle = \langle x_n,v_m\rangle - \frac{\langle x_n,v_m\rangle}{\|v_m\|^2}\langle v_m,v_m\rangle = 0,$$

thereby establishing the case. So we know that $v_0, v_1, \dots$ constitutes an orthogonal system. Next, by induction, for each n, we obtain $v_n \in \operatorname{span}\{x_0, \dots, x_n\}$ directly from (4.76). Thus, $v_n = 0$ implies $x_n = \sum_{\substack{k=0,\\ v_k \neq 0}}^{n-1} \frac{\langle x_n, v_k \rangle}{\|v_k\|^2} v_k \in \operatorname{span}\{x_0, \dots, x_{n-1}\}$. Conversely, if $x_n \in \operatorname{span}\{x_0, \dots, x_{n-1}\}$, then

$$\begin{aligned}\dim \operatorname{span}\{v_0, \dots, v_{n-1}, v_n\} &= \dim \operatorname{span}\{x_0, \dots, x_{n-1}, x_n\} = \dim \operatorname{span}\{x_0, \dots, x_{n-1}\} \\ &= \dim \operatorname{span}\{v_0, \dots, v_{n-1}\},\end{aligned}$$

which, due to Lem. F.3(c), implies $v_n = 0$. Finally, if all $x_0, x_1, \dots$ are linearly independent, then all $v_k \neq 0$, $k = 0, 1, \dots$, such that the $v_0, v_1, \dots$ are linearly independent by Lem. F.3(c). ■

While Gram-Schmidt orthogonalization works in every inner product space, we will see in the next theorem that, for polynomials, there is another recursive relation that is theoretically useful as well as more efficient.

Theorem 4.44. *Given an inner product $\langle \cdot, \cdot \rangle$ on $\mathcal{P}$ that satisfies*

$$\langle xp, q \rangle = \langle p, xq \rangle \quad \text{for each } p, q \in \mathcal{P} \tag{4.77}$$

(clearly, the inner product from (4.74a) *satisfies* (4.77)*), let $\|\cdot\|$ denote the induced norm. If, for each $n \in \mathbb{N}_0$, $x_n \in \mathcal{P}$ is defined by $x_n(x) := x^n$ and $p_n := v_n$ is given by Gram-Schmidt orthogonalization according to* (4.76)*, then the p_n satisfy the following recursive relation (sometimes called* three-term recursion*):*

$$p_0 = x_0 \equiv 1, \quad p_1 = x - \beta_0, \tag{4.78a}$$

$$p_{n+1} = (x - \beta_n) p_n - \gamma_n^2 p_{n-1} \quad \text{for each } n \in \mathbb{N}, \tag{4.78b}$$

where

$$\beta_n := \frac{\langle x p_n, p_n \rangle}{\|p_n\|^2} \quad \text{for each } n \in \mathbb{N}_0, \qquad \gamma_n := \frac{\|p_n\|}{\|p_{n-1}\|} \quad \text{for each } n \in \mathbb{N}. \tag{4.78c}$$

Proof. Recall that, according to Th. 4.43, the linear independence of the x_n guarantees the linear independence of the $p_n = v_n$. We observe that $p_0 = x_0 \equiv 1$ is clear from (4.76) and the definition of x_0. Next, form (4.76), we compute

$$p_1 = v_1 = x_1 - \frac{\langle x_1, p_0 \rangle}{\|p_0\|^2} p_0 = x - \frac{\langle x p_0, p_0 \rangle}{\|p_0\|^2} = x - \beta_0.$$

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,

$$r := p_{n+1} - q_{n+1} \in \mathcal{P}_n,$$

as the coefficient of x^{n+1} in both p_{n+1} and q_{n+1} is 1. According to Lem. F.3(b), it now suffices to show $r \in \mathcal{P}_n^\perp$. Moreover, we know $p_{n+1} \in \mathcal{P}_n^\perp$, since $p_{n+1} \perp p_k$ for each $k \in \{0,\dots,n\}$ according to Th. 4.43 and $\mathcal{P}_n = \operatorname{span}\{p_0,\dots,p_n\}$. Thus, by Lem. F.3(a), to prove $r \in \mathcal{P}_n^\perp$, it suffices to show

$$q_{n+1} \in \mathcal{P}_n^\perp. \tag{4.79}$$

To this end, we start by using $\langle p_{n-1}, p_n\rangle = 0$ and the definition of β_n to compute

$$\langle q_{n+1}, p_n\rangle = \big\langle (x-\beta_n)p_n - \gamma_n^2 p_{n-1}, p_n\big\rangle = \big\langle x p_n, p_n\big\rangle - \beta_n\big\langle p_n, p_n\big\rangle = 0. \tag{4.80a}$$

Similarly, also using (4.77) as well as the definition of γ_n, we obtain

$$\begin{aligned}\langle q_{n+1}, p_{n-1}\rangle &= \langle x p_n, p_{n-1}\rangle - \gamma_n^2\|p_{n-1}\|^2 = \langle p_n, x p_{n-1}\rangle - \|p_n\|^2\\ &= \big\langle p_n, x p_{n-1} - p_n\big\rangle \overset{(*)}{=} 0,\end{aligned} \tag{4.80b}$$

where, at $(*)$, it was used that $x p_{n-1} - p_n \in \mathcal{P}_{n-1}$ due to the fact that x^n is canceled. Finally, for an arbitrary $q \in \mathcal{P}_{n-2}$ (setting $\mathcal{P}_{-1} := \{0\}$), it holds that

$$\langle q_{n+1}, q\rangle = \langle p_n, xq\rangle - \beta_n\langle p_n, q\rangle - \gamma_n^2\langle p_{n-1}, q\rangle = 0, \tag{4.80c}$$

due to $p_n \in \mathcal{P}_{n-1}^\perp$ and $p_{n-1} \in \mathcal{P}_{n-2}^\perp$. Combining the equations (4.80) yields $q_{n+1} \perp q$ for each $q \in \operatorname{span}\big(\{p_n, p_{n-1}\} \cup \mathcal{P}_{n-2}\big) = \mathcal{P}_n$, establishing (4.79) and completing the proof. ■

Remark 4.45. The reason that (4.78) is more efficient than (4.76) lies in the fact that the computation of p_{n+1} according to (4.76) requires the computation of the $n+1$ inner products $\langle x_{n+1}, p_0\rangle, \dots, \langle x_{n+1}, p_n\rangle$, whereas the computation of p_{n+1} according to (4.78) merely requires the computation of the 2 inner products $\langle p_n, p_n\rangle$ and $\langle x p_n, p_n\rangle$ occurring in β_n and γ_n.

Definition 4.46. In the situation of Th. 4.44, we call the polynomials p_n, $n \in \mathbb{N}_0$, given by (4.78), the *orthogonal polynomials* with respect to $\langle\cdot,\cdot\rangle$ (p_n is called the nth orthogonal polynomial).

Example 4.47. Considering $\rho \equiv 1$ on $[-1,1]$, one obtains $\langle p, q\rangle_\rho = \int_{-1}^1 p(x)q(x)\,\mathrm{d}x$ 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 4.48. *Given a real interval* $[a,b]$, $a<b$, *and a weight function* $\rho:\,[a,b]\longrightarrow\mathbb{R}_0^+$, *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* $\lambda_1,\dots,\lambda_n$, *which all are simple and lie in the open interval* $]a,b[$.

Proof. Let $a<\lambda_1<\lambda_2<\dots<\lambda_k<b$ 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 λ_j). From $p_n\in\mathcal{P}_n$, we know $k\leq n$. The goal is to show $k=n$. Seeking a contradiction, we assume $k<n$. This implies

$$q(x):=\prod_{i=1}^{k}(x-\lambda_i)\in\mathcal{P}_{n-1}.$$

Thus, as $p_n\in\mathcal{P}_{n-1}^{\perp}$, we get

$$\langle p_n,q\rangle_\rho=0. \tag{4.81}$$

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

$$\langle p_n,q\rangle_\rho=\int_a^b p_n(x)q(x)\rho(x)\,\mathrm{d}x\neq 0, \tag{4.82}$$

in contradiction to (4.81). This shows $k=n$, and, in particular, that all zeros of p_n are simple. ■

4.6.3 Gaussian Quadrature Rules

Definition 4.49. Given a real interval $[a,b]$, $a<b$, a weight function $\rho:\,[a,b]\longrightarrow\mathbb{R}_0^+$, and $n\in\mathbb{N}$, let $\lambda_1,\dots,\lambda_n\in[a,b]$ be the zeros of the nth orthogonal polynomial p_n with respect to $\langle\cdot,\cdot\rangle_\rho$ and let $L_1,\dots,L_n$ be the corresponding Lagrange basis polynomials (cf. (3.3b)), i.e.

$$L_j(x):=\prod_{\substack{i=1\\ i\neq j}}^{n}\frac{x-\lambda_i}{\lambda_j-\lambda_i}\quad\text{for each } j\in\{1,\dots,n\}, \tag{4.83}$$

then the quadrature rule

$$I_n:\,R[a,b]\longrightarrow\mathbb{R},\quad I_n(f):=\sum_{j=1}^{n}\sigma_j\,f(\lambda_j), \tag{4.84a}$$

where

$$\sigma_j:=\langle L_j,1\rangle_\rho=\int_a^b L_j(x)\rho(x)\,\mathrm{d}x\quad\text{for each } j\in\{1,\dots,n\}, \tag{4.84b}$$

is called the nth order *Gaussian quadrature rule* with respect to ρ.

Theorem 4.50. *Consider the situation of Def.* 4.49; *in particular let* I_n *be the Gaussian quadrature rule defined according to* (4.84).

(a) *I_n is exact for each polynomial of degree at most $2n-1$. This can be stated in the form*

$$\langle p, 1\rangle_\rho = \sum_{j=1}^{n} \sigma_j \, p(\lambda_j) \quad \textit{for each } p \in \mathcal{P}_{2n-1}. \tag{4.85}$$

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 σ_j, $j \in \{1, \dots, n\}$, are positive: $\sigma_j > 0$.*

(c) *For $\rho \equiv 1$, I_n is based on interpolating polynomials for $\lambda_1, \dots, \lambda_n$ in the sense of Def.* 4.10.

Proof. (a): Recall from Linear Algebra that the remainder theorem holds in the ring of polynomials $\mathcal{P}$. In particular, since p_n has degree precisely n, given an arbitrary $p \in \mathcal{P}_{2n-1}$, there exist $q, r \in \mathcal{P}_{n-1}$ such that

$$p = q p_n + r. \tag{4.86}$$

Then, the relation $p_n(\lambda_j) = 0$ implies

$$p(\lambda_j) = r(\lambda_j) \quad \text{for each } j \in \{1, \dots, n\}. \tag{4.87}$$

Since $r \in \mathcal{P}_{n-1}$, it must be the unique interpolating polynomial for the data

$$\big(\lambda_1, r(\lambda_1)\big), \dots, \big(\lambda_n, r(\lambda_n)\big).$$

Thus, according to the Lagrange interpolating polynomial formula (3.3a):

$$r(x) = \sum_{j=1}^{n} r(\lambda_j) L_j(x) \overset{(4.87)}{=} \sum_{j=1}^{n} p(\lambda_j) L_j(x). \tag{4.88}$$

This allows to compute

$$\begin{aligned} \langle p, 1\rangle_\rho &\overset{(4.86)}{=} \int_a^b \big(q(x) p_n(x) + r(x)\big)\rho(x)\,\mathrm{d}x = \langle q, p_n\rangle_\rho + \langle r, 1\rangle_\rho \\ &\overset{\substack{p_n \in \mathcal{P}_{n-1}^\perp, \\ (4.88)}}{=} \sum_{j=1}^{n} p(\lambda_j) \langle L_j, 1\rangle_\rho \overset{(4.84b)}{=} \sum_{j=1}^{n} \sigma_j \, p(\lambda_j), \end{aligned} \tag{4.89}$$

proving (4.85).

(b): For each $j \in \{1, \dots, n\}$, we apply (4.85) to $L_j^2 \in \mathcal{P}_{2n-2}$ to obtain

$$0 < \|L_j\|_\rho^2 = \langle L_j^2, 1\rangle_\rho \overset{(4.85)}{=} \sum_{k=1}^{n} \sigma_k L_j^2(\lambda_k) \overset{(3.7)}{=} \sigma_j.$$

(c): This follows from (a): If $f : [a,b] \longrightarrow \mathbb{R}$ is given and $p \in \mathcal{P}_{n-1}$ is the interpolating polynomial for the data

$$\big(\lambda_1, f(\lambda_1)\big), \dots, \big(\lambda_n, f(\lambda_n)\big),$$

then

$$I_n(f) = \sum_{j=1}^{n} \sigma_j\, f(\lambda_j) = \sum_{j=1}^{n} \sigma_j\, p(\lambda_j) \overset{(4.85)}{=} \langle p, 1\rangle_1 = \int_a^b p(x)\,\mathrm{d}x\,,$$

which establishes the case. ■

Theorem 4.51. *The converse of Th.* 4.50(a) *is also true: If, for $n \in \mathbb{N}$, $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ and $\sigma_1, \dots, \sigma_n \in \mathbb{R}$ are such that* (4.85) *holds, then $\lambda_1, \dots, \lambda_n$ must be the zeros of the nth orthogonal polynomial p_n and $\sigma_1, \dots, \sigma_n$ must satisfy* (4.84b)*.*

Proof. We first verify $q = p_n$ for

$$q : \mathbb{R} \longrightarrow \mathbb{R}, \quad q(x) := \prod_{j=1}^{n} (x - \lambda_j). \tag{4.90}$$

To that end, let $m \in \{0, \dots, n-1\}$ and apply (4.85) to the polynomial $p(x) := x^m\, q(x)$ (note $p \in \mathcal{P}_{n+m} \subseteq \mathcal{P}_{2n-1}$) 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 $q \in \mathcal{P}_{n-1}^\perp$ and $q - p_n \in \mathcal{P}_{n-1}^\perp$. 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 $q - p_n \in \mathcal{P}_{n-1}$. Since $\mathcal{P}_{n-1} \cap \mathcal{P}_{n-1}^\perp = \{0\}$ according to Lem. F.3(b), we have established $q = p_n$, thereby also identifying the λ_j as the zeros of p_n as claimed. Finally, applying (4.85) 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 σ_j, indeed, satisfy (4.84b). ■

Theorem 4.52. *For each $f \in C[a,b]$, the Gaussian quadrature rules $I_n(f)$ of Def.* 4.49 *converge:*

$$\lim_{n\to\infty} I_n(f) = \int_a^b f(x)\rho(x)\,\mathrm{d}x \quad \textit{for each } f \in C[a,b]. \tag{4.91}$$

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 $p \in \mathcal{P}_{2n-1}$ and Condition (ii) of Th. 4.28 is satisfied since the positivity of the weights yields

$$\sum_{j=1}^{n} |\sigma_{j,n}| = \sum_{j=1}^{n} \sigma_{j,n} = I_n(1) = \int_a^b \rho(x)\,\mathrm{d}x \in \mathbb{R}^+ \tag{4.92}$$

for each $n \in \mathbb{N}$. ■

Remark 4.53. **(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 $[-1,1]$:

$$\int_a^b f(u)\,\mathrm{d}u = \frac{b-a}{2}\int_{-1}^1 f\big(u(x)\big)\,\mathrm{d}x = \frac{b-a}{2}\int_{-1}^1 f\left(\frac{b-a}{2}\,x + \frac{a+b}{2}\right)\mathrm{d}x\,. \quad (4.93)$$

This is useful in the context of Gaussian quadrature rules, since the computation of the zeros λ_j of the orthogonal polynomials is, in general, not easy. However, due to (4.93), 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 ρ.

(b) If one is given the λ_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 4.54. *Consider the situation of Def.* 4.49*; in particular let I_n be the Gaussian quadrature rule defined according to* (4.84)*. Then, for each $f \in C^{2n}[a,b]$, there exists $\tau \in [a,b]$ such that*

$$\int_a^b f(x)\rho(x)\,\mathrm{d}x - I_n(f) = \frac{f^{(2n)}(\tau)}{(2n)!}\int_a^b p_n^2(x)\rho(x)\,\mathrm{d}x\,, \quad (4.94)$$

where p_n is the nth orthogonal polynomial, which can be further estimated by

$$\big|p_n(x)\big| \le (b-a)^n \quad \textit{for each } x \in [a,b]. \quad (4.95)$$

Proof. As before, let $\lambda_1, \dots, \lambda_n$ be the zeros of the nth orthogonal polynomial p_n. Given $f \in C^{2n}[a,b]$, let $q \in \mathcal{P}_{2n-1}$ be the Hermite interpolating polynomial for the $2n$ points

$$\lambda_1, \lambda_1, \lambda_2, \lambda_2, \dots, \lambda_n, \lambda_n.$$

The identity (3.45) yields

$$f(x) = q(x) + f[x, \lambda_1, \lambda_1, \lambda_2, \lambda_2, \dots, \lambda_n, \lambda_n]\underbrace{\prod_{j=1}^n (x-\lambda_j)^2}_{p_n^2(x)}\,. \quad (4.96)$$

Now we multiply (4.96) by ρ, replace the generalized divided difference by using the mean value theorem for generalized divided differences (3.47), and integrate over $[a,b]$:

$$\int_a^b f(x)\rho(x)\,\mathrm{d}x = \int_a^b q(x)\rho(x)\,\mathrm{d}x + \frac{1}{(2n)!}\int_a^b f^{(2n)}\big(\xi(x)\big)p_n^2(x)\rho(x)\,\mathrm{d}x\,. \quad (4.97)$$

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

$$\int_a^b f(x)\rho(x)\,\mathrm{d}x = \int_a^b q(x)\rho(x)\,\mathrm{d}x + \frac{f^{(2n)}(\tau)}{(2n)!} \int_a^b p_n^2(x)\rho(x)\,\mathrm{d}x\,. \tag{4.98}$$

Since $q \in \mathcal{P}_{2n-1}$, I_n is exact for q, i.e.

$$\int_a^b q(x)\rho(x)\,\mathrm{d}x = I_n(q) = \sum_{j=1}^n \sigma_j\, q(\lambda_j) = \sum_{j=1}^n \sigma_j\, f(\lambda_j) = I_n(f). \tag{4.99}$$

Combining (4.99) and (4.98) proves (4.94). Finally, (4.95) is clear from the representation $p_n(x) = \prod_{j=1}^n (x - \lambda_j)$ and $\lambda_j \in [a,b]$ for each $j \in \{1,\dots,n\}$. ■

Remark 4.55. 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 Motivation

For simplicity, we will restrict ourselves to the case of systems over the real numbers. In principle, systems over other fields are of interest as well.

Definition 5.1. Given a real $n \times m$ matrix A, $m, n \in \mathbb{N}$, and $b \in \mathbb{R}^n$, the equation

$$Ax = b \tag{5.1}$$

is called a *linear system* for the unknown $x \in \mathbb{R}^m$. The matrix one obtains by adding b as the $(m+1)$th column to A is called the *augmented matrix* of the linear system. It is denoted by $(A|b)$.

—

The goal of this section is to determine solutions to linear systems (5.1), provided such solutions exist. If the linear system does not have any solutions, than one is interested in solving the least squares minimization problem, i.e. one aims at finding x such that $\|Ax - b\|_2$ is minimized. Since one often needs to solve (5.1) with the same A and many different b, it is of particular interest to decompose A in a way that facilitates the efficient computation of solutions when varying b (if A is invertible, then such decompositions can be used to obtain A^{-1}, but it is often not necessary to compute A^{-1} explicitly).

5.2 Gaussian Elimination and LU Decomposition

5.2.1 Pivot Strategies

The first method for the solution of linear systems one usually encounters is the Gaussian elimination algorithm. It is assumed the Gaussian elimination algorithm is known. However, a review is provided in Appendix H.1 (see, in particular, Def. H.5).

The Gaussian elimination algorithm suffers from the fact that it can be numerically unstable even for matrices with small condition number, as demonstrated in the following example.

Example 5.2. Consider the linear system $Ax = b$ with

$$A := \begin{pmatrix} 0.0001 & 1 \\ 1 & 1 \end{pmatrix}, \quad b := \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \tag{5.2}$$

The exact solution is

$$x = \begin{pmatrix} 1.00010001 \\ 0.99989999 \end{pmatrix}. \tag{5.3}$$

We now examine what occurs if we solve the system approximately using the Gaussian elimination algorithm according to Def. H.5 and a floating point arithmetic, rounding to 3 significant digits:

When applying the Gaussian elimination algorithm to $(A|b)$, there is only one step, and it is carried out according to Def. H.5(a), i.e. the second row is replaced by the sum of the second row and the first row multiplied by -10^4. The exact result is

$$\begin{pmatrix} 10^{-4} & 1 & | & 1 \\ 0 & -9999 & | & -9998 \end{pmatrix}. \tag{5.4a}$$

However, when rounding to 3 digits, it is replaced by

$$\begin{pmatrix} 10^{-4} & 1 & | & 1 \\ 0 & -10^4 & | & -10^4 \end{pmatrix}, \tag{5.4b}$$

yielding the approximate solution

$$x = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{5.4c}$$

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 $(A|b)$:

$$\begin{pmatrix} 1 & 1 & | & 2 \\ 10^{-4} & 1 & | & 1 \end{pmatrix}. \tag{5.5a}$$

Now Gaussian elimination and rounding yields

$$\begin{pmatrix} 1 & 1 & | & 2 \\ 0 & 1 & | & 1 \end{pmatrix} \tag{5.5b}$$

with the approximate solution

$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{5.5c}$$

which is much better than (5.4c).

—

Example 5.2 shows that it can be advantageous to switch rows even if one could just apply Def. H.5(a) in the kth step of the Gaussian elimination algorithm. Similarly, when applying Def. H.5(b), 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 kth step, one should first switch the kth 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 5.2 is the column maximum strategy (cf. Def. 5.4 below): In the kth step of Gaussian elimination, 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. Def. H.5), where the pivot element (cf. Def. H.1) 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 (5.5c). However, consider what happens in the following example:

Example 5.3. Consider the linear system $Ax = b$ with

$$A := \begin{pmatrix} 1 & 10000 \\ 1 & 1 \end{pmatrix}, \quad b := \begin{pmatrix} 10000 \\ 2 \end{pmatrix}. \tag{5.6}$$

The exact solution is the same as in Example 5.2, i.e. given by (5.3) (the first row of Example 5.2 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 fist step we obtain

$$\left(\begin{array}{cc|c} 1 & 10000 & 10000 \\ 0 & -9999 & -9998 \end{array}\right), \tag{5.7a}$$

which, after rounding to 3 digits, is replaced by

$$\left(\begin{array}{cc|c} 1 & 10000 & 10000 \\ 0 & -10000 & -10000 \end{array}\right), \tag{5.7b}$$

yielding the unsatisfactory approximate solution

$$x = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{5.7c}$$

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. 5.4 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:

$$\begin{pmatrix} 1 & 1 & | & 2 \\ 1 & 10000 & | & 10000 \end{pmatrix}. \tag{5.8a}$$

Now Gaussian elimination and rounding yields

$$\begin{pmatrix} 1 & 1 & | & 2 \\ 0 & 10000 & | & 10000 \end{pmatrix}, \tag{5.8b}$$

once again with the acceptable approximate solution

$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{5.8c}$$

—

Both pivot strategies discussed above are summarized in the following definition:

Definition 5.4. Let A be a real $n \times m$ matrix A, $m, n \in \mathbb{N}$, and $b \in \mathbb{R}^n$. We define two modified versions of the Gaussian elimination algorithm of Def. H.5 by adding one of the following actions (0) or (0′) at the beginning of the algorithm's kth step (for each $k \geq 1$ such that $r(k) < n$ and $k \leq m+1$).

As in Def. H.5, let $(A^{(1)}|b^{(1)}) := (A|b)$, $r(1) := 1$, and, for each $k \geq 1$ such that $r(k) < n$ and $k \leq m+1$, transform $(A^{(k)}|b^{(k)})$ into $(A^{(k+1)}|b^{(k+1)})$ and $r(k)$ into $r(k+1)$. In contrast to Def. H.5, to achieve the transformation, first perform either (0) or (0′):

(0) *Column Maximum Strategy:* Determine $i \in \{r(k), \dots, n\}$ such that

$$\left|a_{ik}^{(k)}\right| = \max\left\{\left|a_{\alpha k}^{(k)}\right| : \alpha \in \{r(k), \dots, n\}\right\}. \tag{5.9}$$

(0′) *Relative Column Maximum Strategy:* For each $\alpha \in \{r(k), \dots, n\}$, define

$$S_\alpha^{(k)} := \sum_{\beta=k}^{m+1} \left|a_{\alpha\beta}^{(k)}\right|. \tag{5.10a}$$

If $\max\left\{S_\alpha^{(k)} : \alpha \in \{r(k), \dots, n\}\right\} = 0$, then $(A^{(k)}|b^{(k)})$ is already in echelon form and the algorithm is halted. Otherwise, determine $i \in \{r(k), \dots, n\}$ such that

$$\frac{\left|a_{ik}^{(k)}\right|}{S_i^{(k)}} = \max\left\{\frac{\left|a_{\alpha k}^{(k)}\right|}{S_\alpha^{(k)}} : \alpha \in \{r(k), \dots, n\},\ S_\alpha^{(k)} \neq 0\right\}. \tag{5.10b}$$

If $a_{ik}^{(k)} = 0$, then nothing needs to be done (Def. H.5(c)). If $a_{ik}^{(k)} \neq 0$ and $i = r(k)$, then no switching is needed, i.e. one proceeds by Def. H.5(a). If $a_{ik}^{(k)} \neq 0$ and $i \neq r(k)$, then one proceeds by Def. H.5(b) (using the found row number i for switching).

Remark 5.5. (a) Clearly, Th. H.6 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 $(\tilde{A}|\tilde{b})$ in echelon form, such that the linear system $\tilde{A}x = \tilde{b}$ 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.

5.2.2 Gaussian Elimination via Matrix Multiplication and LU Decomposition

The Gaussian elimination algorithm can not only be used to solve one particular linear system, but it can also be used, with virtually no extra effort, to decompose the matrix A into simpler matrices L and U that then facilitate the simple solution of $Ax = b$ when varying b (i.e. without having to reapply Gaussian elimination for each new b).

The decomposition is easily obtained from the fact that each elementary row operation occurring during Gaussian elimination (cf. Def. H.3) can be obtained via multiplication with a suitable matrix. Namely, adding the multiple of one row to another row is obtained by multiplication by a so-called Frobenius matrix (see Def. 5.6(b)), while the switching of rows is obtained by a permutation matrix (see Def. and Rem. 5.10).

We start by reviewing definitions and properties of Frobenius and related matrices.

Definition 5.6. (a) An $n \times n$ matrix A, $n \in \mathbb{N}$, is called *upper triangular* or *right triangular* (respectively *lower triangular* or *left triangular*) if, and only if, $a_{ij} = 0$ for each $i, j \in \{1, \dots, n\}$ such that $i > j$ (respectively $i < j$). A triangular matrix A is called *unipotent* if, and only if, $a_{ii} = 1$ for each $i \in \{1, \dots, n\}$.

(b) A unipotent lower triangular $n \times n$ matrix A, $n \in \mathbb{N}$, is called a *Frobenius matrix* of index k, $k \in \{1, \dots, n\}$, if, and only if, $a_{ij} = 0$ for each $j \neq k$, i.e. if, and only if, it has the following form:

$$A = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & a_{k+1,k} & 1 & & & \\ & & a_{k+2,k} & & 1 & & \\ & & \vdots & & & \ddots & \\ & & a_{n,k} & & & & 1 \end{pmatrix}. \tag{5.11}$$

The following Rem. 5.7 recalls relevant properties of general triangular matrices, while the subsequent Rem. 5.8 deals with Frobenius matrices.

Remark 5.7. The following assertions are easily verified from the definition of matrix multiplication:

(a) The set $\mathcal{L}$ of lower triangular $n \times n$ matrices is closed under matrix multiplication. Since matrix multiplication is always associative, in algebraic terms, this means that $\mathcal{L}$ forms a semigroup. The same is true for the set $\mathcal{R}$ of upper triangular $n \times n$ matrices.

(b) The inverse of an invertible lower triangular $n \times n$ matrix is again a lower triangular $n \times n$ matrix. Together with (a), this means that the set of all invertible lower triangular $n \times n$ matrices forms a group. The unipotent lower triangular $n \times n$ matrices form a subgroup of that group. Analogous statements hold true for the set of upper triangular $n \times n$ matrices.

Remark 5.8. The following assertions are also easily verified from the definition of matrix multiplication:

(a) The elementary row operation of row addition can be achieved by multiplication by a suitable Frobenius matrix from the left. In particular, consider the Gaussian elimination algorithm of Def. H.5 applied to an $n \times m$ matrix A (note that the algorithm can be applied to *any* matrix, so it does not really matter if A represents an augmented matrix of a linear system or not). If $A^{(k)} \mapsto A^{(k+1)}$ is the matrix transformation occurring in the kth step of the Gaussian elimination algorithm of Def. H.5 and the kth step is done by performing Def. H.5(a), i.e. by, for each $i \in \{r(k)+1, \dots, n\}$, replacing the ith row by the ith row plus $-a_{ik}^{(k)}/a_{r(k),k}^{(k)}$ times the $r(k)$th row, then

$$A^{(k+1)} = L_k\, A^{(k)}, \tag{5.12a}$$

where

$$L_k = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & -l_{r(k)+1,r(k)} & 1 & & & \\ & & -l_{r(k)+2,r(k)} & & 1 & & \\ & & \vdots & & & \ddots & \\ & & -l_{n,r(k)} & & & & 1 \end{pmatrix}, \quad l_{i,r(k)} := a_{ik}^{(k)}/a_{r(k),k}^{(k)}, \tag{5.12b}$$

is an $n \times n$ Frobenius matrix of index $r(k)$.

(b) If the $n \times n$ matrix L is a Frobenius matrix,

$$L = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & l_{k+1,k} & 1 & & & \\ & & l_{k+2,k} & & 1 & & \\ & & \vdots & & & \ddots & \\ & & l_{n,k} & & & & 1 \end{pmatrix}, \tag{5.13a}$$

then L is invertible with

$$L^{-1} = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & -l_{k+1,k} & 1 & & & \\ & & -l_{k+2,k} & & 1 & & \\ & & \vdots & & & \ddots & \\ & & -l_{n,k} & & & & 1 \end{pmatrix}. \tag{5.13b}$$

In particular, if L is a Frobenius matrix of index k, then L^{-1} is also a Frobenius matrix of index k.

(c) If

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ l_{21} & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \dots & 1 \end{pmatrix} \tag{5.14a}$$

is an arbitrary unipotent lower triangular $n \times n$ matrix, then it is the product of the following $n-1$ Frobenius matrices:

$$L = \tilde{L}_1 \cdots \tilde{L}_{n-1}, \tag{5.14b}$$

where

$$\tilde{L}_k := \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & l_{k+1,k} & 1 & & & \\ & & l_{k+2,k} & & 1 & & \\ & & \vdots & & & \ddots & \\ & & l_{n,k} & & & & 1 \end{pmatrix} \quad \text{for each } k \in \{1, \dots, n-1\} \tag{5.14c}$$

is a Frobenius matrix of index k.

Theorem 5.9. *Let A be a real $n \times m$ matrix, $m, n \in \mathbb{N}$. Moreover, let $\tilde{A}$ be the $n \times m$ matrix in echelon form resulting at the end of the Gaussian elimination algorithm of Def.* H.5. *Moreover, assume that* no row switching occurred *during the application of the Gaussian elimination algorithm, i.e., in each step* (a) *or* (c) *of Def.* H.5 *was used, but never* (b)*.*

(a) *There exists a unipotent lower triangular $n \times n$ matrix L such that*

$$A = L\tilde{A}. \tag{5.15a}$$

This is the version of the LU decomposition without row switching for general $n \times m$ matrices. Obviously, this is not an LU decomposition in the strict sense, since $\tilde{A}$ is in echelon form, but not necessarily upper triangular. For LU decompositions in the strict sense, see (b)*.*

(b) *If A is an $n \times n$ matrix, then $\tilde{A}$ is an upper triangular matrix, which is emphasized by writing $U := \tilde{A}$. Thus, there exist a unipotent lower triangular $n \times n$ matrix L and an upper triangular matrix U such that*

$$A = LU = L\tilde{A}. \tag{5.15b}$$

This is called an LU decomposition of A. If A is invertible, then $U = \tilde{A}$ is invertible and the LU decomposition is unique.

Proof. (a): Let $N \in \{0, \dots, m\}$ be the final number of steps that is needed to perform the Gaussian elimination algorithm according to Def. H.5 (note that we now have m where we had $m+1$ in Def. H.5, which is due to the fact that our current matrix A is not augmented by a vector b). If $N = 0$, then A consists of precisely one row and there is nothing to prove (set $L := (1)$). If $N \geq 1$, then let $A^{(1)} := A$ and, recursively, for each $k \in \{1, \dots, N\}$, let $A^{(k+1)}$ be the matrix obtained in the kth step of the Gaussian elimination algorithm. If Def. H.5(a) is used in the kth step, then, according to Rem. 5.8(a),

$$A^{(k+1)} = L_k \, A^{(k)}, \tag{5.16}$$

where L_k is the $n \times n$ Frobenius matrix of index $r(k)$ given by (5.12b). If Def. H.5(c) is used in the kth step, then (5.16) also holds, namely with $L_k := \mathrm{Id}$. By induction, (5.16) implies

$$\tilde{A} = A^{(N+1)} = L_N \cdots L_1 A. \tag{5.17}$$

From Rem. 5.8(b), we know that each L_k is invertible with L_k^{-1} also being a Frobenius matrix of index $r(k)$. Thus,

$$L_1^{-1} \cdots L_N^{-1} \tilde{A} = A. \tag{5.18}$$

Comparing with (5.14) shows that $L := L_1^{-1} \cdots L_N^{-1}$ is a unipotent lower triangular matrix, thereby establishing (a). Note: From (5.14) and the definition of the L_k, we actually know all entries of L explicitly: Every nonzero entry of the $r(k)$th column is given by (5.12b). This will be used when formulating the LU decomposition algorithm in Sec. 5.2.3 below.

(b): Everything follows easily from (a): Existence is clear since a quadratic matrix in echelon form is upper triangular. If A is invertible, then so is $U = \tilde{A} = L^{-1}A$. If A is invertible, and we have LU decompositions

$$L_1 U_1 = A = L_2 U_2, \tag{5.19}$$

then we obtain $U_1 U_2^{-1} = L_1^{-1} L_2 =: E$. As both upper triangular and unipotent lower triangular matrices are closed under matrix multiplication, the matrix E is unipotent and both upper and lower triangular, showing $E = \mathrm{Id}$. This, in turn, yields $U_1 = U_2$ as well as $L_1 = L_2$, i.e. the uniqueness claimed in (b). ■

We now proceed to review the definition and properties of permutation matrices which occur in connection with row switching:

Definition and Remark 5.10. Let $n \in \mathbb{N}$. Each bijective map $\pi : \{1, \dots, n\} \longrightarrow \{1, \dots, n\}$ is called a *permutation* of $\{1, \dots, n\}$. The set of permutations of $\{1, \dots, n\}$ forms a group with respect to the composition of maps, the so-called *symmetric group* S_n. For each $\pi \in S_n$, we define an $n \times n$ matrix

$$P_\pi := \begin{pmatrix} e^{\mathrm{t}}_{\pi(1)} & e^{\mathrm{t}}_{\pi(2)} & \dots & e^{\mathrm{t}}_{\pi(n)} \end{pmatrix}, \tag{5.20}$$

where the e_i denote the standard unit (row) vectors of $\mathbb{R}^n$ (cf. (3.35)). The matrix P_π is called the *permutation matrix* associated with π.

Remark 5.11. (a) A real $n \times n$ matrix P is a permutation matrix if, and only if, each row and each column of P have precisely one entry 1 and all other entries of P are 0.

(b) If $\pi \in S_n$ and the columns of P_π are given according to (5.20), then the rows of P_π are given according to the inverse permutation π^{-1},

$$P_\pi = \begin{pmatrix} e_{\pi^{-1}(1)} \\ e_{\pi^{-1}(2)} \\ \dots \\ e_{\pi^{-1}(n)} \end{pmatrix} : \tag{5.21}$$

Consider the ith row. Then $p_{ij} = 1$ if, and only if, $i = \pi(j)$ (since p_{ij} also belongs to the jth column). Thus, $p_{ij} = 1$ if, and only if, $j = \pi^{-1}(i)$, which means that the ith row is $e_{\pi^{-1}(i)}$.

(c) Left multiplication of a matrix A by a permutation matrix P_π permutes the rows of A according to π,

$$P_\pi \begin{pmatrix} — & r_1 & — \\ — & r_2 & — \\ — & \vdots & — \\ — & r_n & — \end{pmatrix} = \begin{pmatrix} — & r_{\pi(1)} & — \\ — & r_{\pi(2)} & — \\ — & \vdots & — \\ — & r_{\pi(n)} & — \end{pmatrix},$$

which follows from the special case

$$P_\pi\, e_i^{\mathrm{t}} = \begin{pmatrix} e_{\pi^{-1}(1)} \\ e_{\pi^{-1}(2)} \\ \dots \\ e_{\pi^{-1}(n)} \end{pmatrix} e_i^{\mathrm{t}} = \begin{pmatrix} e_{\pi^{-1}(1)} \cdot e_i \\ e_{\pi^{-1}(2)} \cdot e_i \\ \dots \\ e_{\pi^{-1}(n)} \cdot e_i \end{pmatrix} = e_{\pi(i)}^{\mathrm{t}},$$

that holds for each $i \in \{1, \dots, n\}$.

(d) Right multiplication of a matrix A by a permutation matrix P_π permutes the columns of A according to π^{-1},

$$\begin{pmatrix} | & | & \dots & | \\ c_1 & c_2 & \dots & c_n \\ | & | & \dots & | \end{pmatrix} P_\pi = \begin{pmatrix} | & | & \dots & | \\ c_{\pi^{-1}(1)} & c_{\pi^{-1}(2)} & \dots & c_{\pi^{-1}(n)} \\ | & | & \dots & | \end{pmatrix},$$

which follows from the special case

$$\begin{aligned} e_i\, P_\pi &= e_i \begin{pmatrix} e_{\pi(1)}^{\mathrm{t}} & e_{\pi(2)}^{\mathrm{t}} & \dots & e_{\pi(n)}^{\mathrm{t}} \end{pmatrix} = \begin{pmatrix} e_i \cdot e_{\pi(1)} & e_i \cdot e_{\pi(2)} & \dots & e_i \cdot e_{\pi(n)} \end{pmatrix} \\ &= e_{\pi^{-1}(i)}, \end{aligned}$$

that holds for each $i \in \{1, \dots, n\}$.

(e) For each $\pi, \sigma \in S_n$, one has

$$P_\pi P_\sigma = P_{\pi\circ\sigma} \tag{5.22}$$

which, in algebraic terms, means that the map $\pi \mapsto P_\pi$ constitutes a group homomorphism and that the permutation matrices form a representation of S_n.

(f) Combining (e) with (5.21) and (5.20) yields

$$P^{-1} = P^{\mathrm{t}} \tag{5.23}$$

for each permutation matrix P.

For obvious reasons, for the Gaussian elimination algorithm, permutation matrices that switch precisely two rows are especially important.

Definition and Remark 5.12. A permutation matrix P_τ corresponding to a transposition $\tau = (ij) \in S_n$ (τ permutes i and j and leaves all other elements fixed) is called a *transposition matrix* and is denoted by $P_{ij} := P_\tau$. The transposition matrix P_{ij} has the form

$$P_{ij} = \begin{pmatrix}
1 & & & & & & & & & & \\
& \ddots & & & & & & & & & \\
& & 1 & & & & & & & & \\
& & & 0 & & & & 1 & & & \\
& & & & 1 & & & & & & \\
& & & & & \ddots & & & & & \\
& & & & & & 1 & & & & \\
& & & 1 & & & & 0 & & & \\
& & & & & & & & 1 & & \\
& & & & & & & & & \ddots & \\
& & & & & & & & & & 1
\end{pmatrix}. \tag{5.24}$$

Since every permutation is the composition of a finite number of transpositions, it is implied by Rem. 5.11(e) that every permutation matrix is the product of a finite number of transposition matrices.

Remark 5.13. It is immediate from Rem. 5.11(c),(d) that left multiplication of a matrix A by P_{ij} switches the ith and jth row of A, whereas right multiplication of A by P_{ij} switches the ith and jth column of A. In particular, the elementary row operation of row switching can be achieved by multiplication by a suitable transposition matrix from the left. In particular, consider the Gaussian elimination algorithm of Def. H.5 applied to an $n \times m$ matrix A. If $A^{(k)} \mapsto A^{(k+1)}$ is the matrix transformation occurring in the kth step of the Gaussian elimination algorithm of Def. H.5 and the kth step is done by performing Def. H.5(b), then

$$A^{(k+1)} = L_k \, P_{i,r(k)} \, A^{(k)}, \tag{5.25}$$

where L_k is the Frobenius matrix of index $r(k)$ defined in (5.12b).

—

We are now in a position to prove the existence of LU decompositions without the restriction that no row switching is used in the Gaussian elimination algorithm.

Theorem 5.14. *Let A be a real $n \times m$ matrix, $m, n \in \mathbb{N}$. Moreover, let $\tilde{A}$ be the $n \times m$ matrix in echelon form resulting at the end of the Gaussian elimination algorithm of Def.* H.5.

(a) *There exists a unipotent lower triangular $n \times n$ matrix L and an $n \times n$ permutation matrix P such that*

$$PA = L\tilde{A}. \tag{5.26a}$$

As in Th. 5.9, *this is not an LU decomposition in the strict sense, since $\tilde{A}$ is in echelon form, but not necessarily upper triangular. One obtains LU decompositions in the strict sense if A is quadratic, which is the case considered in* (b).

(b) *If A is an $n \times n$ matrix, then $\tilde{A}$ is an upper triangular matrix, which is emphasized by writing $U := \tilde{A}$. Thus, there exist a unipotent lower triangular $n \times n$ matrix L, an $n \times n$ permutation matrix P, and an upper triangular matrix U such that*

$$PA = LU = L\tilde{A}. \tag{5.26b}$$

This is called an LU decomposition of PA.

In addition, in each case, one can choose P and L such that $|l_{ij}| \leq 1$ for each entry l_{ij} of L.

Proof. (a): Let $N \in \{0, \dots, m\}$ be the final number of steps that is needed to perform the Gaussian elimination algorithm according to Def. H.5 (note that we now have m where we had $m+1$ in Def. H.5, which is due to the fact that our current matrix A is not augmented by a vector b). If $N = 0$, then A consists of precisely one row and there is nothing to prove (set $L := P := (1)$). If $N \geq 1$, then let $A^{(1)} := A$ and, recursively, for each $k \in \{1, \dots, N\}$, let $A^{(k+1)}$ be the matrix obtained in the kth step of the Gaussian elimination algorithm. If Def. H.5(b) is used in the kth step, then, according to Rem. 5.13,

$$A^{(k+1)} = L_k \, P_k \, A^{(k)}, \tag{5.27}$$

where $P_k := P_{i,r(k)}$ is the $n \times n$ permutation matrix that switches rows $r(k)$ and i, while L_k is the $n \times n$ Frobenius matrix of index $r(k)$ given by (5.12b). If (a) or (c) of Def. H.5 is used in the kth step, then (5.27) also holds, namely with $P_k := \mathrm{Id}$ for (a) and with $L_k := P_k := \mathrm{Id}$ for (c). By induction, (5.27) implies

$$\tilde{A} = A^{(N+1)} = L_N P_N \cdots L_1 P_1 A. \tag{5.28}$$

To show that we can transform (5.28) into (5.26a), we first rewrite the right-hand side of (5.28) taking into account $P_k^{-1} = P_k$ for each of the transposition matrices P_k:

$$\tilde{A} = L_N (P_N L_{N-1} \underbrace{P_N)(P_N}_{\mathrm{Id}} P_{N-1} L_{N-2} \underbrace{P_{N-1} P_N)(P_N P_{N-1}}_{\mathrm{Id}} P_{N-2} \cdots L_1 \underbrace{P_2 \cdots P_N) P_N P_{N-1} \cdots P_2}_{\mathrm{Id}} P_1 A. \tag{5.29}$$

Defining

$$L'_N := L_N, \quad L'_k := P_N P_{N-1} \cdots P_{k+1} L_k P_{k+1} \cdots P_{N-1} P_N \quad \text{for each } k \in \{1, \dots, N-1\}, \tag{5.30}$$

(5.29) takes the form

$$\tilde{A} = L'_N \cdots L'_1 P_N P_{N-1} \cdots P_2 P_1 A. \tag{5.31}$$

We now observe that the L'_k are still Frobenius matrices of index $r(k)$, except that the entries $l_{i,r(k)}$ of L_k with $i > r(k)$ have been permuted according to $P_N P_{N-1} \cdots P_{k+1}$:

This follows since

$$P_{ij}\begin{pmatrix} 1 &&&&&&& \\ & \ddots &&&&&& \\ && 1 &&&&& \\ && \vdots & \ddots &&&& \\ && l_{i,r(k)} && 1 &&& \\ && \vdots &&& \ddots && \\ && l_{j,r(k)} &&&& 1 & \\ && \vdots &&&&& \ddots \end{pmatrix} P_{ij} = \begin{pmatrix} 1 &&&&&&& \\ & \ddots &&&&&& \\ && 1 &&&&& \\ && \vdots & \ddots &&&& \\ && l_{j,r(k)} && 1 &&& \\ && \vdots &&& \ddots && \\ && l_{i,r(k)} &&&& 1 & \\ && \vdots &&&&& \ddots \end{pmatrix}, \tag{5.32}$$

as left multiplication by P_{ij} switches the ith and jth row, switching $l_{i,r(k)}$ and $l_{j,r(k)}$ and moving the corresponding 1's off the diagonal, whereas right multiplication by P_{ij} switches ith and jth column, moving both 1's back onto the diagonal while leaving the $r(k)$th column unchanged.

Finally, using that each $(L'_k)^{-1}$, according to Rem. 5.8(b), exists and is also a Frobenius matrix, (5.31) becomes

$$L\tilde{A} = PA \tag{5.33}$$

with

$$P := P_N P_{N-1} \cdots P_2 P_1, \quad L := (L'_1)^{-1} \cdots (L'_N)^{-1}. \tag{5.34}$$

As in the proof of Th. 5.9, it follows from (5.14) that L is a unipotent lower triangular matrix, thereby completing the proof of (a).

For (b), we once again only remark that a quadratic matrix in echelon form is upper triangular.

To see that one can always choose P such that $|l_{ij}| \leq 1$ for each entry l_{ij} of L, note that when using the Gaussian elimination algorithm with column maximum strategy according to Def. 5.4, (5.9) guarantees that $|l_{i,r(k)}| \leq 1$ for $l_{i,r(k)}$ defined as in (5.12b). ■

Remark 5.15. Note that, even for invertible A, the triple (P, L, U) of (5.26b) is, in general, not unique. For example

$$\underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{P_1} \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}}_{L_1} \underbrace{\begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix}}_{U_1},$$

$$\underbrace{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}}_{P_2} \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}}_{L_2} \underbrace{\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}}_{U_2}.$$

Remark 5.16. (a) Suppose the goal is to solve linear systems $Ax = b$ with fixed A and varying b. Moreover, suppose the LU decomposition $PA = LU$ is at hand. Solving $Ax = b$ is equivalent to solving $PAx = Pb$, i.e. $LUx = Pb$, which is equivalent to solving $Lz = Pb$ for z and then solving $Ux = z$ for x. One can show that the number of steps in finding the LU decomposition of an $n \times n$ matrix A is $O(n^3)$, whereas solving $Lz = Pb$ and $Ux = z$ needs only $O(n^2)$ steps.

(b) One can use (a) to determine A^{-1} of an invertible $n \times n$ matrix A by solving the n linear systems

$$Av_1 = e_1, \ \ldots, \ Av_n = e_n, \tag{5.35}$$

where $e_1, \ldots, e_n$ are the standard (column) unit vectors. Then the v_k are obviously the column vectors of A^{-1}.

(c) Even though the stategy of (a) works fine in many situations, it can fail numerically due to the following stability issue: While the condition numbers, for invertible A, always satisfy

$$\kappa(PA) \leq \kappa(L)\kappa(U) \tag{5.36}$$

due to (2.81) and (2.51), the right-hand side of (5.36) can be much larger than the left-hand side. In that case, it is numerically ill-advised to solve $Lz = Pb$ and $Ux = z$ instead of solving $Ax = b$. There exist more involved decomposition algorithms that are more stable, for example, the QR decomposition (where A is decomposed into an orthogonal matrix Q and an upper triangular matrix R) via the Householder method (see Sec. 5.4.3 below).

5.2.3 The Algorithm of LU Decomposition

Let us crystallize the proof of Th. 5.14 into an algorithm to actually compute the matrices L, P, and $\tilde{A}$ occurring in the respective decompositions (5.26a) and (5.26b) of a given $n \times m$ matrix A. Once again, let $N \in \{0, \ldots, m\}$ be the final number of steps that is needed to perform the Gaussian elimination algorithm according to Def. H.5.

(a) *Algorithm for $\tilde{A}$ (i.e. for U for quadratic A)*: Starting with $A^{(1)} := A$, the kth step of the Gaussian elimination algorithm of Def. H.5 (possibly using a pivot strategy according to Def. 5.4) yields a matrix $A^{(k+1)}$ and $A^{(N+1)}$ is $\tilde{A}$.

(b) *Algorithm for P*: Starting with $P^{(1)} := \mathrm{Id}$, in the kth step of the Gaussian elimination algorithm of Def. H.5, define $P^{(k+1)} := P_k P^{(k)}$, where $P_k := P_{i,r(k)}$ if rows $r(k)$ and i are switched according to (b) of Def. H.5, and $P_k := \mathrm{Id}$, otherwise. In the last step, one obtains $P = P^{(N+1)}$.

(c) *Algorithm for L*: Starting with $L^{(1)} := 0$ (zero matrix), we obtain $L^{(k+1)}$ from $L^{(k)}$ in the kth step of the Gaussian elimination algorithm of Def. H.5 as follows: For Def. H.5(c), set $L^{(k+1)} := L^{(k)}$. If rows $r(k)$ and i are switched according to Def. H.5(b), then we first switch rows $r(k)$ and i in $L^{(k)}$ to obtain some $\tilde{L}^{(k)}$ (this conforms to the definition of the L'_k in (5.30), we will come back to this point below). For Def. H.5(a) and for the elimination step of Def. H.5(b), we first copy all elements of $L^{(k)}$ (resp. $\tilde{L}^{(k)}$) to $L^{(k+1)}$, but then change the elements of the $r(k)$th column according to (5.12b): Set $l^{(k+1)}_{i,r(k)} := a^{(k)}_{ik} / a^{(k)}_{r(k),k}$ for each $i \in \{r(k)+1, \ldots, n\}$. In the last step, we obtain L from $L^{(N+1)}$ by setting all elements on the diagonal to 1 (postponing this step to the end avoids worrying about the diagonal elements when switching rows earlier).

To see that this procedure does, indeed, provide the correct L, we go back to the proof of Th. 5.14(a): From (5.33), (5.34), and (5.30), it follows that the $r(k)$th column of L is precisely the $r(k)$th column of L_k^{-1} permuted according to $P_N \cdots P_{k+1}$. This is precisely what is described in (c) above: We start with the $r(k)$th column of L_k^{-1} by setting $l_{i,r(k)}^{(k+1)} := a_{ik}^{(k)}/a_{r(k),k}^{(k)}$ and then apply $P_{k+1}, \dots, P_N$ during the remaining steps.

Example 5.17. Let us determine the LU decomposition (5.26b) for the matrix

$$A := \begin{pmatrix} 1 & 4 & 2 & 3 \\ 1 & 2 & 1 & 0 \\ 2 & 6 & 3 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix},$$

using the column maximum strategy, which will guarantee $|l_{ij}| \leq 1$ for all components of L. According to the algorithm described above, we start by initializing

$$A^{(1)} := A, \quad P^{(1)} := \mathrm{Id}, \quad L^{(1)} := 0, \quad r(1) := 1.$$

The column maximum strategy requires us to switch rows 1 and 3 using P_{13}:

$$P^{(2)} = P_{13}P^{(1)} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad P_{13}A = \begin{pmatrix} 2 & 6 & 3 & 1 \\ 1 & 2 & 1 & 0 \\ 1 & 4 & 2 & 3 \\ 0 & 0 & 1 & 4 \end{pmatrix}.$$

Eliminating the first column yields

$$L^{(2)} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad A^{(2)} = \begin{pmatrix} 2 & 6 & 3 & 1 \\ 0 & -1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 1 & \frac{1}{2} & \frac{5}{2} \\ 0 & 0 & 1 & 4 \end{pmatrix}, \quad r(2) = r(1) + 1 = 2.$$

In the next step, the column maximum strategy does not require any row switching, so we proceed by eliminating the second column:

$$P^{(3)} = P^{(2)}, \qquad L^{(3)} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$A^{(3)} = \begin{pmatrix} 2 & 6 & 3 & 1 \\ 0 & -1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 1 & 4 \end{pmatrix}, \qquad r(3) = r(2) + 1 = 3.$$

Now we need to switch rows 3 and 4 using P_{34}:

$$P = P^{(4)} = P_{34}P^{(3)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

$$P_{34}L^{(3)} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{2} & -1 & 0 & 0 \end{pmatrix}, \quad P_{34}A^{(3)} = \begin{pmatrix} 2 & 6 & 3 & 1 \\ 0 & -1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 2 \end{pmatrix}.$$

Accidentally, elimination of the third column does not require any additional work and we have

$$L^{(4)} = P_{34}L^{(3)}, \qquad L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{1}{2} & -1 & 0 & 1 \end{pmatrix},$$

$$U = \tilde{A} = A^{(4)} = P_{34}A^{(3)}, \qquad r(4) = r(3) + 1 = 4.$$

Recall that L is obtained from $L^{(4)}$ by setting the diagonal values to 1. One checks that, indeed, $PA = LU$.

Remark 5.18. Note that in the kth step of the algorithm for $\tilde{A}$, we eliminate all elements of the kth column below the row with number $r(k)$, while in the kth step of the algorithm for L, we populate elements of the $r(k)$th column below the diagonal for the first time. Thus, when implementing the algorithm in practice, one can save memory capacity by storing the new elements for L at the locations of the previously eliminated elements of A. This strategy is sometimes called *compact storage*.

5.3 Cholesky Decomposition

Symmetric positive definite matrices A (and also symmetric positive semidefinite matrices, see Appendix H.4) can be decoposed in the particularly simple form $A = LL^{\mathrm{t}}$, where L is lower triangular:

Theorem 5.19. *Let A be a real $n \times n$ matrix, $n \in \mathbb{N}$. If A is symmetric and positive definite (cf. Def.* 2.37(a),(c)*), then there exists a unique lower triangular real $n \times n$ matrix L with positive diagonal entries (i.e. with $l_{jj} > 0$ for each $j \in \{1, \dots, n\}$) and*

$$A = LL^{\mathrm{t}}. \tag{5.37}$$

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

$$A = \left(\begin{array}{c|c} B & v \\ \hline v^{\mathrm{t}} & a_{nn} \end{array}\right), \tag{5.38}$$

where B is the $(n-1)\times(n-1)$ matrix with $b_{ij} = a_{ij}$ for each $i, j \in \{1, \dots, n-1\}$ and $v \in \mathbb{R}^{n-1}$ is the column vector with $v_i = a_{in}$ for each $i \in \{1, \dots, n-1\}$. According to Prop. H.16, B is positive definite (and, clearly, B is also symmetric) and $a_{nn} > 0$. According to the induction hypothesis, there exists a unique lower triangular real $(n-1)\times(n-1)$ matrix L' with positive diagonal entries and $B = L'(L')^{\mathrm{t}}$. Then L' is invertible and we can define

$$w := (L')^{-1}\, v. \tag{5.39}$$

Moreover, let

$$\alpha := a_{nn} - w^{\mathrm{t}}\, w. \tag{5.40}$$

Then $\alpha \in \mathbb{R}$ and there exists $\beta \in \mathbb{C}$ such that $\beta^2 = \alpha$. We set

$$L := \left(\begin{array}{c|c} L' & 0 \\ \hline w^{\mathrm{t}} & \beta \end{array}\right). \tag{5.41}$$

Then L is lower triangular,

$$L^{\mathrm{t}} = \left(\begin{array}{c|c} (L')^{\mathrm{t}} & w \\ \hline 0 & \beta \end{array}\right),$$

and

$$LL^{\mathrm{t}} = \left(\begin{array}{c|c} L'(L')^{\mathrm{t}} & L'w \\ \hline w^{\mathrm{t}}(L')^{\mathrm{t}} & w^{\mathrm{t}}w + \beta^2 \end{array}\right) = \left(\begin{array}{c|c} B & v \\ \hline v^{\mathrm{t}} & a_{nn} \end{array}\right) = A. \tag{5.42}$$

It only remains to be shown one can choose $\beta \in \mathbb{R}^+$ and that this choice determines β uniquely. Since L' is a triangular matrix with positive entries on its diagonal, it is $\det L' \in \mathbb{R}^+$ and

$$\det A = (\det L')^2\, \beta^2 \overset{\text{Prop. H.17(b)}}{>} 0. \tag{5.43}$$

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 α. ∎

Remark 5.20. From the proof of Th. 5.19, it is not hard to extract a recursive algorithm to actually compute the Cholesky decomposition of a symmetric positive definite $n \times n$ matrix A: One starts by setting

$$l_{11} := \sqrt{a_{11}}, \tag{5.44a}$$

where the proof of Th. 5.19 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. 5.19), and one needs to construct $l_{r1}, \dots, l_{rr}$. Using (5.39), one has to solve $L'w = v$ for

$$w = \begin{pmatrix} l_{r1} \\ \vdots \\ l_{r,r-1} \end{pmatrix}, \quad \text{where} \quad v = \begin{pmatrix} a_{1r} \\ \vdots \\ a_{r-1,r} \end{pmatrix}.$$

As L' has lower triangular form, we obtain $l_{r1}, \dots, l_{r,r-1}$ successively, using forward substitution, where

$$l_{rj} := \frac{a_{jr} - \sum_{k=1}^{j-1} l_{jk} l_{rk}}{l_{jj}} \quad \text{for each } j \in \{1, \dots, r-1\} \tag{5.44b}$$

and the proof of Th. 5.19 guarantees $l_{jj} > 0$. Finally, one uses (5.40) (with n replaced by r) to obtain

$$l_{rr} = \sqrt{\alpha} = \sqrt{a_{rr} - w^{\mathrm{t}} w} = \sqrt{a_{rr} - \sum_{k=1}^{r-1} l_{rk}^2}, \tag{5.44c}$$

where the proof of Th. 5.19 guarantees $\alpha > 0$.

Example 5.21. 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. 5.20, we compute

$$\begin{aligned}
l_{11} &= \sqrt{a_{11}} = \sqrt{1} = 1, \\
l_{21} &= \frac{a_{12}}{l_{11}} = \frac{-1}{1} = -1, \\
l_{22} &= \sqrt{a_{22} - l_{21}^2} = \sqrt{2-1} = 1, \\
l_{31} &= \frac{a_{13}}{l_{11}} = \frac{-2}{1} = -2, \\
l_{32} &= \frac{a_{23} - l_{21} l_{31}}{l_{22}} = \frac{3 - (-1)(-2)}{1} = 1, \\
l_{33} &= \sqrt{a_{33} - l_{31}^2 - l_{32}^2} = \sqrt{9-4-1} = 2.
\end{aligned}$$

Thus, the Cholesky decomposition for A is

$$LL^{\mathrm{t}} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -2 & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & -1 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & -1 & -2 \\ -1 & 2 & 3 \\ -2 & 3 & 9 \end{pmatrix} = A.$$

—

We conclude the section by showing that a Cholesky decomposition can not exist if A is not symmetric positive semidefinite:

Proposition 5.22. *For $n \in \mathbb{N}$, let A and L be real $n \times n$ matrices, where $A = LL^{\mathrm{t}}$. Then A is symmetric and positive semidefinite.*

Proof. A is symmetric, since

$$A^{\mathrm{t}} = (LL^{\mathrm{t}})^{\mathrm{t}} = LL^{\mathrm{t}} = A.$$

A is positive semidefinite, since

$$\underset{x \in \mathbb{R}^n}{\forall} \quad x^{\mathrm{t}} A x = x^{\mathrm{t}} L L^{\mathrm{t}} x = (L^{\mathrm{t}} x)^{\mathrm{t}} (L^{\mathrm{t}} x) \geq 0,$$

which completes the proof. ■

5.4 QR Decomposition

5.4.1 Definition and Motivation

For a short review of orthogonal matrices see Appendix H.2.

Definition 5.23. Let A be a real $n \times m$ matrix, $m, n \in \mathbb{N}$.

(a) A decomposition

$$A = Q\tilde{A} \tag{5.45a}$$

is called a *QR decomposition* of A if, and only if, Q is an orthogonal $n \times n$ matrix and $\tilde{A}$ is an $n \times m$ matrix in echelon form. If A is an $n \times n$ matrix, then $\tilde{A}$ is a right or upper triangular matrix, i.e. (5.45a) is a QR decomposition in the strict sense, which is emphasized by writing $R := \tilde{A}$:

$$A = QR. \tag{5.45b}$$

(b) Let $r := \operatorname{rk}(A)$ be the rank of A (see Def. and Rem. H.8). A decomposition

$$A = Q\tilde{A} \tag{5.46a}$$

is called an *economy size QR decomposition* of A if, and only if, Q is an $n \times r$ matrix such that its columns form an orthonormal system in $\mathbb{R}^n$ and $\tilde{A}$ is an $r \times m$ matrix in echelon form. If $r = m$, then $\tilde{A}$ is a right or upper triangular matrix, i.e. (5.46a) is an economy size QR decomposition in the strict sense, which is emphasized by writing $R := \tilde{A}$:

$$A = QR. \tag{5.46b}$$

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.

Lemma 5.24. *Let A and Q be real $n \times n$ matrices, $n \in \mathbb{N}$, A being invertible and Q being orthogonal. With respect to the Euclidean norm on $\mathbb{R}^n$, the following holds (recall* Def. 2.28 *of a matrix norm and* Def. *and* Rem. 2.56 *for the condition of a matrix):*

$$\|QA\| = \|AQ\| = \|A\|, \tag{5.47a}$$

$$\kappa_2(Q) = 1, \tag{5.47b}$$

$$\kappa_2(QA) = \kappa_2(AQ) = \kappa_2(A). \tag{5.47c}$$

Proof. Exercise. ■

Remark 5.25. Suppose, as in Rem. 5.16(a), the goal is to solve linear systems $Ax = b$ with fixed A and varying b. In Rem. 5.16(a), we described how to make use of a given LU decomposition $PA = LU$ in this situation. An analogous procedure can be used given a QR decomposition $A = QR$. Solving $Ax = QRx = b$ is equivalent to solving $Qz = b$ for z and then solving $Rx = z$ for x. Due to $Q^{-1} = Q^{\mathrm{t}}$, solving for $z = Q^{\mathrm{t}}b$ is particularly simple. In Rem. 5.16(c), we noted that the condition of using the LU decomposition to solve $Ax = b$ can be much worse than the condition of A. However, for the QR decomposition, the situation is *much better*: Due to Lem. 5.24, 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! Of course, one can also use the QR decomposition to determine A^{-1} of an invertible $n \times n$ matrix A as described in Rem. 5.16(b).

5.4.2 QR Decomposition via Gram-Schmidt Orthogonalization

As just noted in Rem. 5.25, it is numerically advisable to make use of a QR decomposition if it is available. However, so far we have not adressed the question of how to obtain such a decomposition. The simplest way, but not the most numerically stable, is to use the Gram-Schmidt orthogonalization of Th. 4.43. More precisely, Gram-Schmidt orthogonalization provides the economy size QR decomposition of Def. 5.23(b) as described in Th. 5.26(a). 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 5.26. *Let A be a real $n \times m$ matrix, $m, n \in \mathbb{N}$, with columns $x_1, \dots, x_m$. If $r := \operatorname{rk}(A) > 0$, then define increasing functions*

$$\rho : \{1, \dots, m\} \longrightarrow \{0, \dots, r\}, \quad \rho(k) := \dim(\operatorname{span}\{x_1, \dots, x_k\}), \tag{5.48a}$$

$$\mu : \{1, \dots, r\} \longrightarrow \{1, \dots, m\}, \quad \mu(k) := \min\big\{j \in \{1, \dots, m\} : \rho(j) = k\big\}. \tag{5.48b}$$

(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 an orthogonal $n \times n$ matrix Q, an $n \times m$ matrix $\tilde{A}$ in echelon form, an $n \times r$ matrix Q_{e} (for $r > 0$) such that its columns form an orthonormal system in $\mathbb{R}^n$, and (for $r > 0$) an $r \times m$ matrix $\tilde{A}_{\mathrm{e}}$ in echelon form such that*

$$A = Q\tilde{A}, \tag{5.49a}$$

$$A = Q_{\mathrm{e}}\tilde{A}_{\mathrm{e}} \quad \textit{for } r > 0. \tag{5.49b}$$

(i) *For $r > 0$, the columns $q_1, \dots, q_r \in \mathbb{R}^n$ of Q_e can be computed from the columns $x_1, \dots, x_m \in \mathbb{R}^n$ of A, obtaining $v_1, \dots, v_m \in \mathbb{R}^n$ from* (4.76) *(the first index is now* 1 *instead of* 0 *in* (4.76)*), letting, for each $k \in \{1, \dots, r\}$, $\tilde{v}_k := v_{\mu(k)}$ (i.e. the $\tilde{v}_1, \dots, \tilde{v}_r$ are the $v_1, \dots, v_n$ with each $v_k = 0$ omitted) and, finally, letting $q_k := \tilde{v}_k / \|\tilde{v}_k\|_2$ for each $k = 1, \dots, r$.*

(ii) *For $r > 0$, the entries r_{ik} of $\tilde{A}_\mathrm{e}$ can be defined by*

$$r_{ik} := \begin{cases} \langle x_k, q_i \rangle & \text{for } k \in \{1, \dots, m\},\ i < \rho(k), \\ \langle x_k, q_i \rangle & \text{for } k \in \{1, \dots, m\},\ i = \rho(k),\ k > \mu\big(\rho(k)\big), \\ \|\tilde{v}_{\rho(k)}\|_2 & \text{for } k \in \{1, \dots, m\},\ i = \rho(k),\ k = \mu\big(\rho(k)\big), \\ 0 & \text{otherwise.} \end{cases}$$

(iii) *For the columns $q_1, \dots, q_n$ of Q, one can complete the orthonormal system $q_1, \dots, q_r$ of* (i) *by $q_{r+1}, \dots, q_n \in \mathbb{R}^n$ to an orthonormal basis of $\mathbb{R}^n$.*

(iv) *For $\tilde{A}$, one can use $\tilde{A}_\mathrm{e}$ as in* (ii)*, completed by $n - r$ 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 $\tilde{A}_\mathrm{e}$ 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 $q_{r+1}, \dots, q_n$ of Q.*

Proof. (a): We start by showing that, for $r > 1$, if Q_e and $\tilde{A}_\mathrm{e}$ are defined by (i) and (ii), respectively, then they provide the desired economy size QR decomposition of A. From Th. 4.43, we know that $v_k = 0$ if, and only if, $x_k \in \operatorname{span}\{x_1, \dots, x_{k-1}\}$. Thus, in (i), we indeed obtain r linearly independent $\tilde{v}_k$ and Q_e is an $n \times r$ matrix such that its columns form an orthonormal system in $\mathbb{R}^n$. From (ii), we see that $\tilde{A}_\mathrm{e}$ is an $r \times m$ matrix, as $r = \rho(m)$. To see that $\tilde{A}_\mathrm{e}$ is in echelon form, consider its ith row, $i \in \{1, \dots, r\}$. 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 ith row, and as μ is strictly increasing, $\tilde{A}_\mathrm{e}$ is in echelon form. To show $A = Q_\mathrm{e}\tilde{A}_\mathrm{e}$, note that, for each $k \in \{1, \dots, m\}$,

$$x_k = \left\{ \begin{matrix} \sum_{i=1}^{\rho(k)} \langle x_k, q_i \rangle \, q_i & \text{for } k > \mu\big(\rho(k)\big) \\ \sum_{i=1}^{\rho(k)-1} \langle x_k, q_i \rangle \, q_i + \|\tilde{v}_{\rho(k)}\|_2 \, q_{\rho(k)} & \text{for } k = \mu\big(\rho(k)\big) \end{matrix} \right\} = \sum_{i=1}^{r} r_{ik} q_i \tag{5.50}$$

as a consequence of (4.76), 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 $\tilde{A}_\mathrm{e}$ 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 QR decomposition, as it clearly implies the uniqueness statement for the economy size QR decomposition. Suppose

$$A = QR = \tilde{Q}\tilde{R}, \tag{5.51}$$

where $Q, \tilde{Q}$ are orthogonal $n \times n$ matrices, and $R, \tilde{R}$ are $n \times m$ matrices in echelon form such that all pivot elements are positive. We write (5.51) in the equivalent form

$$x_k = \sum_{i=1}^{r} r_{ik} q_i = \sum_{i=1}^{r} \tilde{r}_{ik} \tilde{q}_i \quad \text{for each } k \in \{1, \dots, m\}, \tag{5.52}$$

and show, by induction on $\alpha \in \{1, \dots, r\}$, that $q_\alpha = \tilde{q}_\alpha$ and $r_{ik} = \tilde{r}_{ik}$ for each $i \in \{1, \dots, r\}$, $k \in \{1, \dots, \mu(\alpha)\}$, as well as

$$\operatorname{span}\{x_1, \dots, x_{\mu(\alpha)}\} = \operatorname{span}\{q_1, \dots, q_{\alpha-1}, q_\alpha\}. \tag{5.53}$$

Consider $\alpha = 1$. For $1 \le k < \mu(1)$, 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 $i \in \{1, \dots, r\}$, $1 \le k < \mu(1)$ (if any). Next,

$$0 \ne x_{\mu(1)} = r_{1,\mu(1)} q_1 = \tilde{r}_{1,\mu(1)} \tilde{q}_1, \tag{5.54}$$

since the echelon forms of R and $\tilde{R}$ imply $r_{i,\mu(1)} = \tilde{r}_{i,\mu(1)} = 0$ for each $i > 1$. As $\|q_1\|_2 = \|\tilde{q}_1\|_2 = 1$, (5.54) implies $q_1 = \pm\tilde{q}_1$. As we also know that both pivot elements $r_{1,\mu(1)}$ and $\tilde{r}_{1,\mu(1)}$ are positive, $q_1 = \tilde{q}_1$ follows. This, in turn, lets one conclude $r_{1,\mu(1)} = \tilde{r}_{1,\mu(1)}$ (again using (5.54)). Note that (5.54) also implies $\operatorname{span}\{x_1, \dots, x_{\mu(1)}\} = \operatorname{span}\{x_{\mu(1)}\} = \operatorname{span}\{q_1\}$. Now let $\alpha > 1$ and, by induction, assume $q_\beta = \tilde{q}_\beta$ for $1 \le \beta < \alpha$ and $r_{ik} = \tilde{r}_{ik}$ for each $i \in \{1, \dots, r\}$, $k \in \{1, \dots, \mu(\alpha - 1)\}$, as well as

$$\operatorname{span}\{x_1, \dots, x_{\mu(\beta)}\} = \operatorname{span}\{q_1, \dots, q_\beta\}. \tag{5.55}$$

We have to show $r_{ik} = \tilde{r}_{ik}$ for each $i \in \{1, \dots, r\}$, $k \in \{\mu(\alpha - 1) + 1, \dots, \mu(\alpha)\}$, and $q_\alpha = \tilde{q}_\alpha$, as well as (5.53). For each $k \in \{\mu(\alpha - 1) + 1, \dots, \mu(\alpha) - 1\}$, from (5.55) with $\beta = \alpha - 1$, we know

$$\operatorname{span}\{x_1, \dots, x_k\} = \operatorname{span}\{x_1, \dots, x_{\mu(\alpha-1)}\} = \operatorname{span}\{q_1, \dots, q_{\alpha-1}\}, \tag{5.56}$$

such that (5.52) implies $r_{ik} = \tilde{r}_{ik} = 0$ for each $i > \alpha - 1$. Then, for $1 \le i \le \alpha - 1$, multiplying (5.52) by q_i and using the orthonormality of the q_i provides $r_{ik} = \tilde{r}_{ik} = \langle x_k, q_i \rangle$. A similar argument also works for $k = \mu(\alpha)$: As we had just seen, $r_{\alpha l} = \tilde{r}_{\alpha l} = 0$ for each $1 \le l < k$. So the echelon forms of R and $\tilde{R}$ imply $r_{ik} = \tilde{r}_{ik} = 0$ for each $i > \alpha$. Then, as before, $r_{ik} = \tilde{r}_{ik} = \langle x_k, q_i \rangle$ for each $i < \alpha$ by multiplying (5.52) by q_i, and, finally,

$$0 \ne x_k - \sum_{i=1}^{\alpha-1} r_{ik} q_i = x_{\mu(\alpha)} - \sum_{i=1}^{\alpha-1} r_{ik} q_i = r_{\alpha k} q_\alpha = \tilde{r}_{\alpha k} \tilde{q}_\alpha. \tag{5.57}$$

Analogous to the case $\alpha = 1$, the positivity of $r_{\alpha k}$ and $\tilde{r}_{\alpha k}$ implies first $q_\alpha = \tilde{q}_\alpha$ followed by $r_{\alpha k} = \tilde{r}_{\alpha k}$. To conclude the induction, we note that combining (5.57) with (5.56) verifies (5.53). In particular, we have shown that (5.52) holds as well as $x_k = \sum_{i=1}^{r} r_{ik} q_i$, i.e. $0 = \sum_{i=r+1}^{n} r_{ik} q_i = \sum_{i=r+1}^{n} \tilde{r}_{ik} \tilde{q}_i$, which implies $r_{ik} = 0$ for each $i > r$ due to the linear independence of the q_i, and $\tilde{r}_{ik} = 0$ for each $i > r$ due to the linear independence of the $\tilde{q}_i$. ■

Remark 5.27. Gram-Schmidt orthogonalization according to (4.76) becomes numerically unstable in cases where x_k is "almost" in $\operatorname{span}\{x_1, \dots, x_{k-1}\}$, resulting in a very small $0 \neq v_k$, in which case $\|v_k\|^{-1}$ can become arbitrarily large. In the following section, we will study an alternative method to compute the QR decomposition that avoids this issue.

Example 5.28. We once again consider the matrix from Ex. 5.17, i.e.

$$A := \begin{pmatrix} 1 & 4 & 2 & 3 \\ 1 & 2 & 1 & 0 \\ 2 & 6 & 3 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix}.$$

Applying Gram-Schmidt orthogonalization (4.76) 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

$$v_1 = x_1 = \begin{pmatrix} 1 \\ 1 \\ 2 \\ 0 \end{pmatrix},$$

$$v_2 = x_2 - \frac{\langle x_2, v_1 \rangle}{\|v_1\|_2^2} v_1 = x_2 - \frac{18}{6} v_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix},$$

$$v_3 = x_3 - \frac{\langle x_3, v_1 \rangle}{\|v_1\|_2^2} v_1 - \frac{\langle x_3, v_2 \rangle}{\|v_2\|_2^2} v_2 = x_3 - \frac{9}{6} v_1 - \frac{1}{2} v_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix},$$

$$v_4 = x_4 - \frac{\langle x_4, v_1 \rangle}{\|v_1\|_2^2} v_1 - \frac{\langle x_4, v_2 \rangle}{\|v_2\|_2^2} v_2 - \frac{\langle x_4, v_3 \rangle}{\|v_2\|_2^2} v_3 = x_4 - \frac{5}{6} v_1 - \frac{3}{2} v_2 - \frac{4}{1} v_3 = \begin{pmatrix} 2/3 \\ 2/3 \\ -2/3 \\ 0 \end{pmatrix}.$$

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

$$Q = \begin{pmatrix} 1/\sqrt{6} & 1/\sqrt{2} & 0 & 1/\sqrt{3} \\ 1/\sqrt{6} & -1/\sqrt{2} & 0 & 1/\sqrt{3} \\ 2/\sqrt{6} & 0 & 0 & -1/\sqrt{3} \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Next, we obtain

$$\begin{array}{llll}
r_{11} = \|v_1\|_2 = \sqrt{6}, & r_{12} = \langle x_2, q_1 \rangle = 3\sqrt{6}, & r_{13} = \langle x_3, q_1 \rangle = 9/\sqrt{6}, & r_{14} = \langle x_4, q_1 \rangle = 5/\sqrt{6}, \\
r_{21} = 0, & r_{22} = \|v_2\|_2 = \sqrt{2}, & r_{23} = \langle x_3, q_2 \rangle = 1/\sqrt{2}, & r_{24} = \langle x_4, q_2 \rangle = 3/\sqrt{2}, \\
r_{31} = 0, & r_{32} = 0, & r_{33} = \|v_3\|_2 = 1, & r_{34} = \langle x_4, q_3 \rangle = 4, \\
r_{41} = 0, & r_{42} = 0, & r_{43} = 0, & r_{44} = \|v_4\|_2 = 2/\sqrt{3},
\end{array}$$

that means

$$R = \tilde{A} = \begin{pmatrix} \sqrt{6} & 3\sqrt{6} & 9/\sqrt{6} & 5/\sqrt{6} \\ 0 & \sqrt{2} & 1/\sqrt{2} & 3/\sqrt{2} \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 2/\sqrt{3} \end{pmatrix}.$$

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. 5.27. On the other hand, the Gram-Schmidt method provides k orthogonal vectors in k steps, whereas, for the Householder method, one has to complete all n steps to obtain n orthogonal vectors. We begin with some preparatory definitions and remarks:

Definition 5.29. A real $n \times n$ matrix H, $n \in \mathbb{N}$, is called a *Householder matrix* or *Householder reflection* or *Householder transformation* if, and only if, there exists $u \in \mathbb{R}^n$ such that $\|u\|_2 = 1$ and

$$H = \operatorname{Id} - 2\,uu^{\mathrm{t}}, \tag{5.58}$$

where, here and in the following, u is treated as a *column* vector. If H is a Householder matrix, then the linear map $H : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is also called a *Householder reflection* or *Householder transformation*.

Lemma 5.30. *Let the real $n \times n$ matrix H, $n \in \mathbb{N}$, be a Householder matrix. Then the following holds:*

(a) *H is symmetric: $H^{\mathrm{t}} = H$.*

(b) *H is involutary: $H^2 = \operatorname{Id}$.*

(c) *H is orthogonal: $H^{\mathrm{t}} H = \operatorname{Id}$.*

Proof. As H is a Householder matrix, there exists $u \in \mathbb{R}^n$ such that $\|u\|_2 = 1$ and (5.58) is satisfied.

(a): One computes

$$H^{\mathrm{t}} = \big(\operatorname{Id} - 2\,uu^{\mathrm{t}}\big)^{\mathrm{t}} = \operatorname{Id} - 2(u^{\mathrm{t}})^{\mathrm{t}}u^{\mathrm{t}} = H. \tag{5.59}$$

(b): Another easy calculation:

$$H^2 = \big(\operatorname{Id} - 2\,uu^{\mathrm{t}}\big)\big(\operatorname{Id} - 2\,uu^{\mathrm{t}}\big) = \operatorname{Id} - 4\,uu^{\mathrm{t}} + 4uu^{\mathrm{t}}uu^{\mathrm{t}} = \operatorname{Id}, \tag{5.60}$$

where $u^{\mathrm{t}}u = \|u\|^2 = 1$ was used in the last equality.

(c) is immediate from (a) and (b). ■

Remark 5.31. Let the real $n \times n$ matrix H, $n \in \mathbb{N}$, be a Householder matrix, with $H = \operatorname{Id} - 2\,uu^{\mathrm{t}}$, $u \in \mathbb{R}^n$, $\|u\|_2 = 1$. Then the map $H : \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $x \mapsto Hx$ constitutes the reflection through the hyperplane

$$V := \operatorname{span}\{u\}^{\perp} = \{z \in \mathbb{R}^n : z^{\mathrm{t}}u = 0\} : \tag{5.61}$$

Note that, for each $x \in \mathbb{R}^n$, $x - uu^{\mathrm{t}}x \in V$:

$$(x - uu^{\mathrm{t}}x)^{\mathrm{t}}u = x^{\mathrm{t}}u - (x^{\mathrm{t}}u)u^{\mathrm{t}}u \overset{u^{\mathrm{t}}u=1}{=} 0. \tag{5.62}$$

In particular, $\operatorname{Id} - uu^{\mathrm{t}}$ 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 $e_1 \in \mathbb{R}^k$. This task then has to be performed for varying dimensions k. That such Householder reflections exist is the contents of the following lemma:

Lemma 5.32. *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 $x \notin \operatorname{span}\{e_1\}$, then*

$$u := \frac{x + \sigma\, e_1}{\|x + \sigma\, e_1\|_2} \quad \textit{for} \quad \sigma = \pm\|x\|_2, \tag{5.63}$$

satisfies

$$\|u\|_2 = 1 \quad \textit{and} \tag{5.64}$$

$$(\operatorname{Id} - 2uu^{\mathrm{t}})x = -\sigma\, e_1. \tag{5.65}$$

Proof. First, note that $x \notin \operatorname{span}\{e_1\}$ implies $\|x + \sigma\, e_1\|_2 \neq 0$ such that u is well-defined. Moreover, (5.64) is immediate from the definition of u. To verify (5.65), note

$$\|x + \sigma\, e_1\|_2^2 = \|x\|_2^2 + 2\sigma\, e_1^{\mathrm{t}}x + \sigma^2 = 2(x + \sigma\, e_1)^{\mathrm{t}}x \tag{5.66a}$$

due to the definition of σ. Together with the definition of u, this yields

$$2u^{\mathrm{t}}x = \frac{2(x + \sigma\, e_1)^{\mathrm{t}}x}{\|x + \sigma\, e_1\|_2} = \|x + \sigma\, e_1\|_2 \tag{5.66b}$$

and

$$2uu^{\mathrm{t}}x = x + \sigma\, e_1, \tag{5.66c}$$

proving (5.65). ■

Remark 5.33. In (5.63), one has the freedom to choose the sign of σ. It is numerically advisable to choose

$$\sigma = \sigma(x) := \begin{cases} \|x\|_2 & \text{for } x_1 \geq 0, \\ -\|x\|_2 & \text{for } x_1 < 0, \end{cases} \tag{5.67}$$

to avoid subtractive cancellation of digits.

Definition 5.34. Given a real $n \times m$ matrix A, $m, n \in \mathbb{N}$, the *Householder method* is the following procedure: Let $A^{(1)} := A$, $Q^{(1)} := \mathrm{Id}$ (identity on $\mathbb{R}^n$), $r(1) := 1$. For $k \geq 1$, as long as $r(k) < n$ and $k \leq m$, 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 $a^{(k)}_{ik} = 0$ for each $i \in \{r(k), \dots, n\}$, then

$$A^{(k+1)} := A^{(k)}, \quad Q^{(k+1)} := Q^{(k)}, \quad r(k+1) := r(k).$$

(b) If $a^{(k)}_{r(k),k} \neq 0$, but $a^{(k)}_{ik} = 0$ for each $i \in \{r(k)+1, \dots, n\}$,

$$A^{(k+1)} := A^{(k)}, \quad Q^{(k+1)} := Q^{(k)}, \quad r(k+1) := r(k) + 1.$$

(c) Otherwise, i.e. if $x^{(k)} \notin \mathrm{span}\{e_1\}$, where

$$x^{(k)} := \left(a^{(k)}_{r(k),k}, \dots, a^{(k)}_{n,k}\right)^{\mathrm{t}} \in \mathbb{R}^{n-r(k)+1} \tag{5.68}$$

and e_1 is the standard unit vector in $\mathbb{R}^{n-r(k)+1}$, then

$$A^{(k+1)} := H^{(k)} A^{(k)}, \quad Q^{(k+1)} := Q^{(k)} H^{(k)}, \quad r(k+1) := r(k) + 1,$$

where

$$H^{(k)} := \left(\begin{array}{c|c} \mathrm{Id} & 0 \\ \hline 0 & \mathrm{Id} - 2u^{(k)}(u^{(k)})^{\mathrm{t}} \end{array}\right), \quad u^{(k)} := \frac{x^{(k)} + \sigma\, e_1}{\|x^{(k)} + \sigma\, e_1\|_2}, \tag{5.69}$$

with $\sigma = \sigma(x^{(k)})$ chosen as in Rem. 5.33, with the upper Id being the identity on $\mathbb{R}^{r(k)-1}$, and with the lower Id being the identity on $\mathbb{R}^{n-r(k)+1}$.

Theorem 5.35. *Given a real $n \times m$ matrix A, $m, n \in \mathbb{N}$, the Householder method as defined in Def.* 5.34 *yields a QR decomposition of A. More precisely, if $r(k) = n$ or $k = m + 1$, then $Q := Q^{(k)}$ is an orthogonal $n \times n$ matrix and $\tilde{A} := A^{(k)}$ is an $n \times m$ matrix in echelon form satisfying $A = Q\tilde{A}$.*

Proof. First note that the Householder method terminates after at most m steps. If $1 \leq N \leq m+1$ is the maximal k occurring during the Householder method, and one lets $H^{(k)} := \mathrm{Id}$ for each k such that Def. 5.34(a) or Def. 5.34(b) was used, then

$$Q = H^{(1)} \cdots H^{(N-1)}, \quad \tilde{A} = H^{(N-1)} \cdots H^{(1)} A. \tag{5.70}$$

Since, according to Lem. 5.30(b), $(H^{(k)})^2 = \mathrm{Id}$ for each k, $Q\tilde{A} = A$ is an immediate consequence of (5.70). Moreover, it follows from Lem. 5.30(c) that the $H^{(k)}$ are all orthogonal, implying that Q is orthogonal. 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 $(i,j) \in \{r(k), \dots, n\} \times \{1, \dots, k-1\}$. 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 don't assume anything). Observe that multiplication of $A^{(k)}$ with $H^{(k)}$ as defined in (5.69) 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 $(i,j) \in \{r(k), \dots, n\} \times \{1, \dots, k-1\}$. Moreover, if we are in the case of Def. 5.34(c), then the choice of $u^{(k)}$ in (5.69) and (5.65) of Lem. 5.32 guarantee

$$\begin{pmatrix} a_{r(k),k}^{(k+1)} \\ a_{r(k)+1,k}^{(k+1)} \\ \vdots \\ a_{n,k}^{(k+1)} \end{pmatrix} \in \operatorname{span} \left\{ \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \right\}. \tag{5.71}$$

In each case, after the application of (a), (b), or (c) of Def. 5.34, $a_{ij}^{(k+1)} = 0$ for each $(i,j) \in \{r(k+1), \dots, n\} \times \{1, \dots, k\}$. We have to show that the first $r(k+1)$ rows of $A^{(k+1)}$ are in echelon form. For Def. 5.34(a), it is $r(k+1) = r(k)$ and there is nothing to prove. For Def. 5.34(b),(c), we know $a_{r(k+1),j}^{(k+1)} = 0$ for each $j \in \{1, \dots, k\}$, 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, $r(k) = n$ or $k = m+1$, we have shown that $\tilde{A}$ is in echelon form. ■

Example 5.36. We apply the Householder method of Def. 5.34 to the matrix A of the previous Examples 5.17 and 5.28, i.e. to

$$A := \begin{pmatrix} 1 & 4 & 2 & 3 \\ 1 & 2 & 1 & 0 \\ 2 & 6 & 3 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix}. \tag{5.72a}$$

We start with

$$A^{(1)} := A, \quad Q^{(1)} := \mathrm{Id}, \quad r(1) := 1. \tag{5.72b}$$

Since $x^{(1)} := (1,1,2,0)^{\mathrm{t}} \notin \operatorname{span}\{(1,0,0,0)^{\mathrm{t}}\}$, we compute

$$u^{(1)} := \frac{x^{(1)} + \sigma(x^{(1)})\,e_1}{\|x^{(1)} + \sigma(x^{(1)})\,e_1\|_2} = \frac{1}{\|x^{(1)} + \sqrt{6}\,e_1\|_2} \begin{pmatrix} 1+\sqrt{6} \\ 1 \\ 2 \\ 0 \end{pmatrix}, \tag{5.72c}$$

$$\begin{aligned} H^{(1)} &:= \operatorname{Id} - 2u^{(1)}(u^{(1)})^{\mathrm{t}} = \operatorname{Id} - \frac{1}{6+\sqrt{6}} \begin{pmatrix} 7+2\sqrt{6} & 1+\sqrt{6} & 2+2\sqrt{6} & 0 \\ 1+\sqrt{6} & 1 & 2 & 0 \\ 2+2\sqrt{6} & 2 & 4 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ &\approx \begin{pmatrix} -0.4082 & -0.4082 & -0.8165 & 0 \\ -0.4082 & 0.8816 & -0.2367 & 0 \\ -0.8165 & -0.2367 & 0.5266 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \end{aligned} \tag{5.72d}$$

$$A^{(2)} := H^{(1)} A^{(1)} \approx \begin{pmatrix} -2.4495 & -7.3485 & -3.6742 & -2.0412 \\ 0 & -1.2899 & -0.6449 & -1.4614 \\ 0 & -0.5798 & -0.2899 & -1.9229 \\ 0 & 0 & 1 & 4 \end{pmatrix}, \tag{5.72e}$$

$$Q^{(2)} := Q^{(1)} H^{(1)} = H^{(1)}, \quad r(2) := r(1) + 1 = 2. \tag{5.72f}$$

Since $x^{(2)} := (-1.2899, -0.5798, 0)^{\mathrm{t}} \notin \operatorname{span}\{(1,0,0)^{\mathrm{t}}\}$, we compute

$$u^{(2)} := \frac{x^{(2)} + \sigma(x^{(2)})\,e_1}{\|x^{(2)} + \sigma(x^{(2)})\,e_1\|_2} = \frac{x^{(2)} - \|x^{(2)}\|_2\,e_1}{\left\|x^{(2)} - \|x^{(2)}\|_2\,e_1\right\|_2} \approx \begin{pmatrix} -0.9778 \\ -0.2096 \\ 0 \end{pmatrix}, \tag{5.72g}$$

$$H^{(2)} := \begin{pmatrix} 1 & 0 \\ 0 & \operatorname{Id} - 2u^{(2)}(u^{(2)})^{\mathrm{t}} \end{pmatrix} \approx \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.9121 & -0.4100 & 0 \\ 0 & -0.4100 & 0.9121 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{5.72h}$$

$$A^{(3)} = H^{(2)} A^{(2)} \approx \begin{pmatrix} -2.4495 & -7.3485 & -3.6742 & -2.0412 \\ 0 & 1.4142 & 0.7071 & 2.1213 \\ 0 & 0 & 0 & -1.1547 \\ 0 & 0 & 1 & 4 \end{pmatrix}, \tag{5.72i}$$

$$Q^{(3)} = Q^{(2)} H^{(2)} \approx \begin{pmatrix} -0.4082 & 0.7071 & -0.5774 & 0 \\ -0.4082 & -0.7071 & -0.5774 & 0 \\ -0.8165 & -0.0000 & 0.5774 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad r(3) = r(2) + 1 = 3. \tag{5.72j}$$

Since $x^{(3)} := (0,1)^{\mathrm{t}} \notin \operatorname{span}\{(1,0)^{\mathrm{t}}\}$, we compute

$$u^{(3)} := \frac{x^{(3)} + \sigma(x^{(3)})\, e_1}{\|x^{(3)} + \sigma(x^{(3)})\, e_1\|_2} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}, \tag{5.72k}$$

$$H^{(3)} := \left(\begin{array}{c|c} \mathrm{Id} & 0 \\ \hline 0 & \mathrm{Id} - 2u^{(3)}(u^{(3)})^{\mathrm{t}} \end{array}\right) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{pmatrix}, \tag{5.72l}$$

$$R = \tilde{A} = A^{(4)} = H^{(3)} A^{(3)} \approx \begin{pmatrix} -2.4495 & -7.3485 & -3.6742 & -2.0412 \\ 0 & 1.4142 & 0.7071 & 2.1213 \\ 0 & 0 & -1 & -4 \\ 0 & 0 & 0 & 1.1547 \end{pmatrix}, \tag{5.72m}$$

$$Q = Q^{(4)} = Q^{(3)} H^{(3)} \approx \begin{pmatrix} -0.4082 & 0.7071 & 0 & 0.5774 \\ -0.4082 & -0.7071 & 0 & 0.5774 \\ -0.8165 & 0 & 0 & -0.5774 \\ 0 & 0 & -1 & 0 \end{pmatrix}. \tag{5.72n}$$

One verifies $A = QR$.

6 Short Introduction to Iterative Methods, Solution of Nonlinear Equations

6.1 Motivation: Fixed Points and Zeros

Definition 6.1. **(a)** Given a function $\varphi : A \longrightarrow B$, where A, B can be arbitrary sets, $x \in A$ is called a *fixed point* of φ if, and only if, $\varphi(x) = x$.

(b) If $f : A \longrightarrow Y$, 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 present section, we will study methods for finding roots and 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 theorem of Sec. 6.2, can often also be used to find zeros, while methods formulated for zeros, such as Newton's method of Sec. 6.3, can often also be used to find fixed points.

Remark 6.2. **(a)** If $\varphi : A \longrightarrow B$, where A, B are subsets of a common vector space X, then $x \in A$ is a fixed point of φ if, and only if, $\varphi(x) - x = 0$, i.e. if, and only if, x is a zero of the function $f : A \longrightarrow X$, $f(x) := \varphi(x) - x$ (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 $f : A \longrightarrow Y$, 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 $\varphi : A \longrightarrow Z$, $\varphi(x) := f(x) + x$.

Given a function g, an *iterative method* defines a sequence $x_0, x_1, \dots,$ where, for $n \in \mathbb{N}_0$, x_{n+1} is computed from x_n (or possibly $x_0, \dots, x_n$) by using g. For example "using g" can mean applying g to x_n (as in the case of the Banach fixed point method of Th. 6.5 below) or it can mean applying g and g' (as in the case of Newton's method according to (6.9) below). We will just investigate two particular iterative methods below, namely the mentioned Banach fixed point method that is designed to provide sequences that converge to a fixed point of g, and Newton's method that is designed to provide sequences that converge to a zero of g.

6.2 Banach Fixed Point Theorem A.K.A. Contraction Mapping Principle

The generic setting for this section are metric spaces (see [Phi14, Sec. 1.2]).

Definition 6.3. Let $\emptyset \neq A$ be a subset of a metric space (X, d). A map $\varphi : A \longrightarrow A$ is called a *contraction* if, and only if, there exists $0 \leq L < 1$ satisfying

$$d\big(\varphi(x), \varphi(y)\big) \leq L\, d(x, y) \quad \text{for each } x, y \in A. \tag{6.1}$$

Remark 6.4. According to Def. 6.3, $\varphi : A \longrightarrow A$ is a contraction if, and only if, φ is Lipschitz continuous with Lipschitz constant $L < 1$.

The following Th. 6.5 constitutes the Banach fixed point theorem. It is also known as the contraction mapping principle. Its proof is surprisingly simple, e.g. about an order of magnitude easier than the proof of the Brouwer fixed point theorem.

Theorem 6.5 (Banach Fixed Point Theorem)**.** *Let $\emptyset \neq A$ be a closed subset of a complete metric space (X, d) (for example, a Banach space). If $\varphi : A \longrightarrow A$ is a contraction with Lipschitz constant $0 \leq L < 1$, then φ has a unique fixed point $x_* \in A$. Moreover, for each initial value $x_0 \in A$, the sequence $(x_n)_{n \in \mathbb{N}_0}$, defined by*

$$x_{n+1} := \varphi(x_n) \quad \textit{for each } n \in \mathbb{N}_0, \tag{6.2}$$

converges to x_*:

$$\lim_{n \to \infty} \varphi^n(x_0) = x_*. \tag{6.3}$$

Furthermore, for each such sequence, we have the error estimate

$$d(x_n, x_*) \leq \frac{L}{1 - L}\, d(x_n, x_{n-1}) \leq \frac{L^n}{1 - L}\, d(x_1, x_0) \tag{6.4}$$

for each $n \in \mathbb{N}$.

Proof. We start with uniqueness: Let $x_*, x_{**} \in A$ be fixed points of φ. Then

$$d(x_*, x_{**}) = d\big(\varphi(x_*), \varphi(x_{**})\big) \leq L\, d(x_*, x_{**}), \tag{6.5}$$

which implies $1 \leq L$ for $d(x_*, x_{**}) > 0$. Thus, $L < 1$ implies $d(x_*, x_{**}) = 0$ and $x_* = x_{**}$. Next, we turn to existence. A simple induction on $m - n$ shows

$$\begin{gathered} d(x_{m+1}, x_m) \leq L\, d(x_m, x_{m-1}) \leq L^{m-n}\, d(x_{n+1}, x_n) \\ \text{for each } m, n \in \mathbb{N}_0,\ m > n. \end{gathered} \tag{6.6}$$

This, in turn, allows us to estimate, for each $n, k \in \mathbb{N}_0$:

$$\begin{aligned} d(x_{n+k}, x_n) &\leq \sum_{m=n}^{n+k-1} d(x_{m+1}, x_m) \overset{(6.6)}{\leq} \sum_{m=n}^{n+k-1} L^{m-n}\, d(x_{n+1}, x_n) \\ &\leq \frac{1}{1-L}\, d(x_{n+1}, x_n) \overset{(6.6)}{\leq} \frac{L^n}{1-L}\, d(x_1, x_0) \to 0 \quad \text{for } n \to \infty, \end{aligned} \tag{6.7}$$

establishing that $(x_n)_{n\in\mathbb{N}_0}$ constitutes a Cauchy sequence. Since X is complete, this Cauchy sequence must have a limit $x_* \in X$, and since the sequence is in A and A is closed, $x_* \in A$. The continuity of φ allows to take limits in (6.2), resulting in $x_* = \varphi(x_*)$, showing that x_* is a fixed point and proving existence.

Finally, the error estimate (6.4) follows from (6.7) by fixing n and taking the limit for $k \to \infty$. ■

Example 6.6. Suppose, we are looking for a fixed point of the map $\varphi(x) = \cos x$ (or, equivalently, for a zero of $f(x) = \cos x - x$). To apply the Banach fixed point Th. 6.5, we need to restrict φ to a set A such that $\varphi(A) \subseteq A$. This is the case for $A := [0, 1]$. Moreover, $\varphi : A \longrightarrow A$ is a contraction, due to $\sin 1 < 1$ and the mean value theorem providing $\tau \in]0, 1[$, satisfying

$$\big|\varphi(x) - \varphi(y)\big| = \big|\varphi'(\tau)\big|\, |x - y| < (\sin 1)|x - y| \tag{6.8}$$

for each $x, y \in A$. Since $\mathbb{R}$ is complete and A is closed in $\mathbb{R}$, Th. 6.5 yields the existence of a unique fixed point $x_* \in [0, 1]$ and $\lim_{n\to\infty} \varphi^n(x_0) = x_*$ for each $x_0 \in [0, 1]$.

6.3 Newton's Method

As mentioned above, Newton's method is an iterative method designed to provide sequences $(x_n)_{n\in\mathbb{N}_0}$ that converge to a zero of a given fuction f. We will see that, if Newton's method converges, than the convergence is faster than for the Banach fixed point method of the previous section. However, one also needs to assume more regularity of f.

If $A \subseteq \mathbb{R}$ and $f : A \longrightarrow \mathbb{R}$ is differentiable, then Newton's method is defined by the recursion

$$x_0 \in A, \quad x_{n+1} := x_n - \frac{f(x_n)}{f'(x_n)} \quad \text{for each } n \in \mathbb{N}_0. \tag{6.9a}$$

Analogously, Newton's method can also be defined for differentiable $f : A \longrightarrow \mathbb{R}^m$, $A \subseteq \mathbb{R}^m$:

$$x_0 \in A, \quad x_{n+1} := x_n - \big(Df(x_n)\big)^{-1} f(x_n) \quad \text{for each } n \in \mathbb{N}_0. \tag{6.9b}$$

And, actually, the same can be done 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 lecture.

In practice, in each step of Newton's method (6.9b), one will determine x_{n+1} as the solution to the linear system

$$Df(x_n)x_{n+1} = Df(x_n)x_n - f(x_n). \tag{6.10}$$

There are clearly several issues with Newton's method as formulated in (6.9):

(a) Is the derivative of f in x_n invertible?

(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. An example of a set of sufficient conditions is provided by the following theorem:

Theorem 6.7. *Let $m \in \mathbb{N}$ and fix a norm $\|\cdot\|$ on $\mathbb{R}^m$. Let $A \subseteq \mathbb{R}^m$ be open. Moreover let $f : A \longrightarrow A$ be differentiable, let $x_* \in A$ be a zero of f, and let $r > 0$ be (sufficiently small) such that*

$$B_r(x_*) = \big\{x \in \mathbb{R}^m : \|x - x_*\| < r\big\} \subseteq A. \tag{6.11}$$

Assume that $Df(x_*)$ is invertible and that $\beta > 0$ is such that

$$\Big\|\big(Df(x_*)\big)^{-1}\Big\| \leq \beta \tag{6.12}$$

(here and in the following, we consider the induced operator norm on real $n\times n$ matrices). Finally, assume that the map $Df : B_r(x_*) \longrightarrow \mathcal{L}(\mathbb{R}^m, \mathbb{R}^m)$ is Lipschitz continuous with Lipschitz constant $L \in \mathbb{R}^+$, i.e.

$$\big\|(Df)(x) - (Df)(y)\big\| \leq L\,\|x - y\| \quad \textit{for each } x, y \in B_r(x_*). \tag{6.13}$$

Then, letting

$$\delta := \min\left\{r, \frac{1}{2\beta L}\right\}, \tag{6.14}$$

Newton's method (6.9b) 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

$$\lim_{n\to\infty} x_n = x_* \tag{6.15}$$

with the error estimates

$$\|x_{n+1} - x_*\| \leq \beta\, L\, \|x_n - x_*\|^2 \leq \frac{1}{2}\, \|x_n - x_*\|, \tag{6.16}$$

$$\|x_n - x_*\| \leq \left(\frac{1}{2}\right)^{2^n - 1} \|x_0 - x_*\| \tag{6.17}$$

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

$$\left\| \big(Df(x)\big)^{-1} \right\| \leq 2\beta. \tag{6.18}$$

Proof. Fix $x \in B_\delta(x_*)$. We will apply Lem. 2.64 with

$$A := Df(x_*), \quad \Delta A := Df(x) - Df(x_*), \quad Df(x) = A + \Delta A. \tag{6.19}$$

To apply Lem. 2.64, we must estimate ΔA. We obtain

$$\|\Delta A\| \overset{(6.13)}{\leq} L\, \|x - x_*\| < L\, \delta \leq \frac{1}{2\beta} \overset{(6.12)}{\leq} \frac{1}{2} \|A^{-1}\|^{-1}, \tag{6.20}$$

such that we can, indeed, apply Lem. 2.64. In consequence, $Df(x)$ is invertible with

$$\left\| \big(Df(x)\big)^{-1} \right\| \overset{(2.91)}{\leq} \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\, \|\Delta A\|} \overset{(6.20)}{\leq} 2\|A^{-1}\| \leq 2\beta, \tag{6.21}$$

thereby establishing the case. ▲

Claim 2. For each $x, y \in B_r(x_*)$, the following holds:

$$\big\| f(x) - f(y) - Df(y)(x - y) \big\| \leq \frac{L}{2}\, \|x - y\|^2. \tag{6.22}$$

Proof. We apply the chain rule to the auxiliary function

$$\phi : [0,1] \longrightarrow \mathbb{R}^m, \quad \phi(t) := f\big(y + t(x - y)\big) \tag{6.23}$$

to obtain

$$\phi'(t) := \big(Df(y + t(x - y))\big)(x - y) \quad \text{for each } t \in [0,1]. \tag{6.24}$$

This allows us to write

$$f(x) - f(y) - Df(y)(x - y) = \phi(1) - \phi(0) - \phi'(0) = \int_0^1 \big(\phi'(t) - \phi'(0)\big)\, \mathrm{d}t\,. \tag{6.25}$$

Next, we estimate the norm of the integrand:

$$\big\|\phi'(t) - \phi'(0)\big\| \leq \big\|Df(y + t(x - y)) - Df(y)\big\|\, \|x - y\| \leq L\, t\, \|x - y\|^2. \tag{6.26}$$

Thus,

$$\left\|\int_0^1 \big(\phi'(t)-\phi'(0)\big)\,\mathrm{d}t\right\| \le \int_0^1 \big\|\phi'(t)-\phi'(0)\big\|\,\mathrm{d}t \le \int_0^1 L\,t\,\|x-y\|^2\,\mathrm{d}t = \frac{L}{2}\,\|x-y\|^2. \tag{6.27}$$

Combining (6.27) with (6.25) proves (6.22). ▲

We will now show that

$$x_n \in B_\delta(x_*) \tag{6.28}$$

and the error estimate (6.16) holds for each $n \in \mathbb{N}_0$. We proceed by induction. For $n = 0$, (6.28) holds by hypothesis. Next, we assume that $n \in \mathbb{N}_0$ and (6.28) holds. Using $f(x_*) = 0$ and (6.9b), we write

$$\begin{aligned} x_{n+1} - x_* &= x_n - x_* - \big(Df(x_n)\big)^{-1}\big(f(x_n) - f(x_*)\big) \\ &= -\big(Df(x_n)\big)^{-1}\big(f(x_n) - f(x_*) - Df(x_n)(x_n - x_*)\big). \end{aligned} \tag{6.29}$$

Applying the norm to (6.29) and using (6.18) and (6.22) implies

$$\|x_{n+1}-x_*\| \le 2\,\beta\,\frac{L}{2}\,\|x_n-x_*\|^2 = \beta\,L\,\|x_n-x_*\|^2 \le \beta\,L\,\delta\|x_n-x_*\| \le \frac{1}{2}\,\|x_n-x_*\|, \tag{6.30}$$

which proves $x_{n+1} \in B_\delta(x_*)$ as well as (6.16).

Another induction shows that (6.16) implies, for each $n \in \mathbb{N}$,

$$\|x_n - x_*\| \le (\beta L)^{1+2+\dots+2^{n-1}}\|x_0 - x_*\|^{2^n} \le (\beta L\delta)^{2^n-1}\|x_0 - x_*\|, \tag{6.31}$$

where the geometric sum formula $1+2+\dots+2^{n-1} = \frac{2^n-1}{2-1} = 2^n - 1$ was used for the last inequality. Moreover, $\beta L\delta \le \frac{1}{2}$ together with (6.31) proves (6.17), and, in particular, (6.15).

Finally, we show that x_* is the unique zero in $B_\delta(x_*)$: Suppose $x_*, x_{**} \in B_\delta(x_*)$ with $f(x_*) = f(x_{**}) = 0$. Then

$$\|x_{**} - x_*\| = \left\|\big(Df(x_*)\big)^{-1}\big(f(x_{**}) - f(x_*) - Df(x_*)(x_{**} - x_*)\big)\right\|. \tag{6.32}$$

Applying (6.18) and (6.22) yields

$$\|x_{**} - x_*\| \le 2\beta\,\frac{L}{2}\,\|x_* - x_{**}\|^2 \le \beta\,L\,\delta\,\|x_* - x_{**}\| \le \frac{1}{2}\|x_* - x_{**}\|, \tag{6.33}$$

implying $\|x_* - x_{**}\| = 0$ and completing the proof of the theorem. ■

The main drawback of Th. 6.7 is the assumption of the existence and location of the zero x_*, which can be very difficult to verify in cases where one would like to apply Newton's method. There exist related theorems that do provide the existence of a zero, but here we will not have time to pursue this issue further.

A b-Adic Expansions of Real Numbers

We are mostly used to representing real numbers in the decimal system. For example, we write

$$x = \frac{395}{3} = 131.\overline{6} = 1 \cdot 10^2 + 3 \cdot 10^1 + 1 \cdot 10^0 + \sum_{n=1}^{\infty} 6 \cdot 10^{-n}. \tag{A.1a}$$

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

$$x = 10000011.\overline{10} = 2^7 + 2^1 + 2^0 + \sum_{n=0}^{\infty} 2^{-(2n+1)}. \tag{A.1b}$$

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.

Definition A.1. Given a natural number $b \geq 2$, an integer $N \in \mathbb{Z}$, and a sequence $(d_N, d_{N-1}, d_{N-2}, \dots)$ of nonnegative integers such that $d_n \in \{0, \dots, b-1\}$ for each $n \in \{N, N-1, N-2, \dots\}$, the expression

$$\sum_{\nu=0}^{\infty} d_{N-\nu}\, b^{N-\nu} \tag{A.2}$$

is called a *b-adic series*. The number b is called the *base* or the *radix*, and the numbers d_ν are called *digits*.

Lemma A.2. *Given a natural number $b \geq 2$, consider the b-adic series given by* (A.2). *Then*

$$\sum_{\nu=0}^{\infty} d_{N-\nu}\, b^{N-\nu} \leq b^{N+1}, \tag{A.3}$$

and, in particular, the b-adic series converges to some $x \in \mathbb{R}_0^+$. Moreover, equality in (A.3) *holds if, and only if, $d_n = b-1$ for every $n \in \{N, N-1, N-2, \dots\}$.*

Proof. Exercise. ■

Definition A.3. 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.2), than one calls the b-adic series a *b-adic expansion* of x.

—

In the following Th. A.6, we will show that each nonnegative real number is represented by a b-adic series. First, another preparatory lemma.

Lemma A.4. *Given a natural number $b \geq 2$, consider two b-adic series*

$$x := \sum_{\nu=0}^{\infty} d_{N-\nu}\, b^{N-\nu} = \sum_{\nu=0}^{\infty} e_{N-\nu}\, b^{N-\nu}, \tag{A.4}$$

$N \in \mathbb{Z}$ and $d_n, e_n \in \{0, \dots, b-1\}$ for each $n \in \{N, N-1, N-2, \dots\}$. If $d_N < e_N$, then $e_N = d_N + 1$, $d_n = b-1$ for each $n < N$ and $e_n = 0$ for each $n < N$.

Proof. By subtracting $d_N b^N$ from both series, one can assume $d_N = 0$ without loss of generality. From Lem. A.2, we know

$$x = \sum_{\nu=0}^{\infty} d_{N-\nu}\, b^{N-\nu} = \sum_{\nu=0}^{\infty} d_{N-1-\nu}\, b^{N-1-\nu} \leq b^N. \tag{A.5a}$$

On the other hand:

$$x = \sum_{\nu=0}^{\infty} e_{N-\nu}\, b^{N-\nu} \geq b^N. \tag{A.5b}$$

Combining (A.5a) and (A.5b) yields $x = b^N$. Once again employing Lem. A.2, (A.5a) also shows that $d_n = b-1$ for each $n \leq N-1$ as claimed. Since $e_N > 0$ and $e_n \geq 0$ for each n, equality in (A.5b) can only occur for $e_N = 1$ and $e_n = 0$ for each $n < N$, thereby completing the proof of the lemma. ■

Notation A.5. For each $x \in \mathbb{R}$, we let

$$\lfloor x \rfloor := \max\{k \in \mathbb{Z} : \, k \leq x\} \tag{A.6}$$

denote the *integral part* of x (also called *floor* of x or x *rounded down*).

Theorem A.6. *Given a natural number $b \geq 2$ and a nonnegative real number $x \in \mathbb{R}_0^+$, there exists a b-adic series representing x, i.e. there is $N \in \mathbb{Z}$ and a sequence $(d_N, d_{N-1}, d_{N-2}, \dots)$ of nonnegative integers satisfying $d_n \in \{0, \dots, b-1\}$ for each $n \in \{N, N-1, N-2, \dots\}$ and*

$$x = \sum_{\nu=0}^{\infty} d_{N-\nu}\, b^{N-\nu}. \tag{A.7}$$

If one introduces the additional requirement that $0 \neq d_N$, then each $x > 0$ has either a unique b-adic expansion or precisely two b-adic expansions. More precisely, for $0 \neq d_N$ and $x > 0$, the following statements are equivalent:

(i) The b-adic expansion of x is not unique.

(ii) There are precisely two b-adic expansions of x.

(iii) There exists a b-adic expansion of x such that $d_n = 0$ for each $n \leq n_0$ for some $n_0 \leq N$.

(iv) There exists a b-adic expansion of x such that $d_n = b-1$ for each $n \leq n_0$ for some $n_0 \leq N$.

Proof. We start by constructing numbers N and d_n, $n \in \{N, N-1, N-2, \dots\}$, such that (A.7) holds. For $x = 0$, one chooses an arbitrary $N \in \mathbb{Z}$ and $d_n = 0$ for each $n \in \{N, N-1, N-2, \dots\}$. Thus, for the remainder of the proof, fix $x > 0$. Let

$$N := \max\{n \in \mathbb{Z} : b^n \leq x\}. \tag{A.8}$$

The numbers $d_{N-n} \in \{0, \dots, b-1\}$ and $x_n \in \mathbb{R}^+$, $n \in \mathbb{N}_0$, are defined inductively by letting

$$d_N := \left\lfloor \frac{x}{b^N} \right\rfloor, \qquad x_0 := d_N b^N, \tag{A.9a}$$

$$d_{N-n} := \left\lfloor \frac{x - x_{n-1}}{b^{N-n}} \right\rfloor, \qquad x_n := x_{n-1} + d_{N-n}\, b^{N-n} \quad \text{for } n \geq 1. \tag{A.9b}$$

Claim 1. One can verify by induction on n that the numbers d_{N-n} and x_n enjoy the following properties for each $n \in \mathbb{N}_0$:

$$d_{N-n} \in \{0, \dots, b-1\}, \tag{A.10a}$$

$$0 < x_n = \sum_{\nu=0}^{n} d_{N-\nu}\, b^{N-\nu} \leq x, \tag{A.10b}$$

$$x - x_n < b^{N-n}. \tag{A.10c}$$

Proof. The induction is carried out for all three statements of (A.10) simultaneously. From (A.8), we know $b^N \leq x < b^{N+1}$, i.e. $1 \leq \frac{x}{b^N} < b$. Using (A.9a), this yields $d_N \in \{1, \dots, b-1\}$ and $0 < x_0 = d_N b^N = b^N d_N \leq b^N \frac{x}{b^N} = x$ as well as $x - x_0 = x - d_N b^N = b^N(\frac{x}{b^N} - d_N) < b^N$. For $n \geq 1$, by induction, one obtains $0 \leq x - x_{n-1} < b^{1+N-n}$, i.e. $0 \leq \frac{x - x_{n-1}}{b^{N-n}} < b$. Using (A.9b), this yields $d_{N-n} \in \{0, \dots, b-1\}$ and $x_n = x_{n-1} + d_{N-n} b^{N-n} \leq x_{n-1} + b^{N-n} \frac{x - x_{n-1}}{b^{N-n}} = x$. Moreover, by induction, $0 < x_{n-1} = \sum_{\nu=0}^{n-1} d_{N-\nu}\, b^{N-\nu}$, such that (A.9b) implies $x_n = x_{n-1} + d_{N-n}\, b^{N-n} \geq x_{n-1} > 0$ and $x_n = x_{n-1} + d_{N-n}\, b^{N-n} = d_{N-n}\, b^{N-n} + \sum_{\nu=0}^{n-1} d_{N-\nu}\, b^{N-\nu} = \sum_{\nu=0}^{n} d_{N-\nu}\, b^{N-\nu}$. Finally, $x - x_n = x - x_{n-1} - d_{N-n}\, b^{N-n} = b^{N-n}(\frac{x - x_{n-1}}{b^{N-n}} - d_{N-n}) \leq b^{N-n}$, completing the proof of the claim. ▲

Since, for each $n \in \mathbb{N}_0$,

$$0 \overset{\text{(A.10b)}}{\leq} x - x_n = b^{N-n-1}\, \frac{x - x_n}{b^{N-n-1}} \overset{\text{(A.9b)}}{\leq} b^{N-n-1}\, (d_{N-n-1} + 1) \leq b^{N-n}, \tag{A.11}$$

and $\lim_{n\to\infty} b^{N-n} = 0$, we have $\lim_{n\to\infty} x_n = x$, thereby establishing (A.7).

It remains to verify the equivalence of (i) – (iv).

(ii) ⇒ (i) is trivial.

"(iii) ⇒ (i)": Assume (iii) holds. Without loss of generality, we can assume that n_0 is the largest index such that $d_n = 0$ for each $n \leq n_0$. We distinguish two cases. If $n_0 < N - 1$ or $d_N \neq 1$, then

$$\sum_{\nu=0}^{N-n_0-2} d_{N-\nu}\, b^{N-\nu} + (d_{n_0+1} - 1)b^{n_0+1} + \sum_{\nu=N-n_0}^{\infty} (b-1)\, b^{N-\nu}$$

is a different b-adic expansion of x and its first coefficient is nonzero. If $n_0 = N - 1$ and $d_N = 1$, then

$$\sum_{\nu=1}^{\infty}(b-1)\, b^{N-\nu} = \sum_{\nu=0}^{\infty}(b-1)\, b^{N-1-\nu}$$

is a different b-adic expansion of x and its first coefficient is nonzero.

"(iv) ⇒ (i)": Assume (iv) holds. Without loss of generality, we can assume that n_0 is the largest index such that $d_n = b - 1$ for each $n \leq n_0$. Then

$$\sum_{\nu=0}^{N-n_0-2} d_{N-\nu}\, b^{N-\nu} + (d_{n_0+1} + 1)b^{n_0+1} + \sum_{\nu=N-n_0}^{\infty} 0\, b^{N-\nu}$$

is a different b-adic expansion of x and its first coefficient is nonzero.

We will now show that, conversely, (i) implies (ii), (iii), and (iv). To that end, let $x > 0$ and suppose that x has two different b-adic expansions

$$x = \sum_{\nu=0}^{\infty} d_{N_1-\nu}\, b^{N_1-\nu} = \sum_{\nu=0}^{\infty} e_{N_2-\nu}\, b^{N_2-\nu} \tag{A.12}$$

with $N_1, N_2 \in \mathbb{Z}$; $d_n, e_n \in \{0, \dots, b-1\}$; and $d_{N_1}, e_{N_2} > 0$. This implies

$$x \geq b^{N_1}, \quad x \geq b^{N_2}. \tag{A.13a}$$

Moreover, Lem. A.2 yields

$$x \leq b^{N_1+1}, \quad x \leq b^{N_2+1}. \tag{A.13b}$$

If $N_2 > N_1$, then (A.13) imply $N_2 = N_1 + 1$ and $b^{N_2} \leq x \leq b^{N_1+1} = b^{N_2}$, i.e. $x = b^{N_2} = b^{N_1+1}$. Since $e_{N_2} > 0$, one must have $e_{N_2} = 1$, and, in turn, $e_n = 0$ for each $n < N_2$. Moreover, $x = b^{N_1+1}$ and Lem. A.2 imply that $d_n = b-1$ for each $n \in \{N_1, N_1 - 1, \dots\}$. Thus, for $N_2 > N_1$, the value of N_1 is determined by N_2 and the values of all d_n and e_n are also completely determined, showing that there are precisely two b-adic expansions of x. Moreover, the d_n have the property required in (iv) and the e_n have the property required in (iii). The argument also shows that, for $N_1 > N_2$, one must have $N_1 = N_2+1$ with the e_n taking the values of the d_n and vice versa. Once again, there are precisely two b-adic expansions of x; now the d_n have the property required in (iii) and the e_n have the property required in (iv).

It remains to consider the case $N := N_1 = N_2$. Since, by hypothesis, the two b-adic expansions of x in (A.12) are not identical, there must be a largest index $n \leq N$ such that $d_n \neq e_n$. Thus, (A.12) implies

$$y := \sum_{\nu=0}^{\infty} d_{n-\nu}\, b^{n-\nu} = \sum_{\nu=0}^{\infty} e_{n-\nu}\, b^{n-\nu}. \tag{A.14}$$

Now Lem. A.4 shows that there are precisely two b-adic expansions of x, one having the property required in (iii) and the other having property required in (iv).

Thus, in each case ($N_2 > N_1$, $N_1 > N_2$, and $N_1 = N_2$), we find that (i) implies (ii), (iii), and (iv), thereby concluding the proof of the theorem. ■

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.7. Given a natural number $b \geq 2$ and finite sequences

$$(d_{N_1}, d_{N_1-1}, \dots, d_0) \in \{0, \dots, b-1\}^{N_1+1}, \tag{A.15a}$$

$$(e_1, e_2, \dots, e_{N_2}) \in \{0, \dots, b-1\}^{N_2}, \tag{A.15b}$$

$$(p_1, p_2, \dots, p_{N_3}) \in \{0, \dots, b-1\}^{N_3}, \tag{A.15c}$$

$N_1, N_2, N_3 \in \mathbb{N}_0$ (where $N_2 = 0$ or $N_3 = 0$ is supposed to mean that the corresponding sequence is empty), the respective string

$$\begin{array}{ll} (d_{N_1} d_{N_1-1} \dots d_0)_b & \text{for } N_2 = N_3 = 0, \\ (d_{N_1} d_{N_1-1} \dots d_0 \,.\, e_1 \dots e_{N_2} \overline{p_1 \dots p_{N_3}})_b & \text{for } N_2 + N_3 > 0 \end{array} \tag{A.16}$$

represents the number

$$\sum_{\nu=0}^{N_1} d_\nu\, b^\nu + \sum_{\nu=1}^{N_2} e_\nu\, b^{-\nu} + \sum_{\alpha=0}^{\infty} \sum_{\nu=1}^{N_3} p_\nu\, b^{-N_2-\alpha N_3-\nu}. \tag{A.17}$$

Example A.8. For the number from (A.1), we get

$$x = (131.\overline{6})_{10} = (10000011.\overline{10})_2 = (83.\overline{\mathrm{A}})_{16} \tag{A.18}$$

(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).

—

One frequently needs to convert representations with respect to one base into representations with respect to another base. When working with digital computers, conversions between bases 10 and 2 and vice versa are the most obvious ones that come up. Converting representations is related to the following elementary remainder theorem and the well-known long division algorithm.

Theorem A.9. *For each pair of numbers $(a, b) \in \mathbb{N}^2$, there exists a unique pair of numbers $(q, r) \in \mathbb{N}_0^2$ satisfying the two conditions $a = qb + r$ and $0 \leq r < b$.*

Proof. Existence: Define

$$q := \max\{n \in \mathbb{N}_0 : nb \leq a\}, \tag{A.19a}$$

$$r := a - qb. \tag{A.19b}$$

Then $q \in \mathbb{N}_0$ by definition and (A.19b) immediately yields $a = qb + r$ as well as $r \in \mathbb{Z}$. Moreover, from (A.19a), $qb \leq a = qb + r$, i.e. $0 \leq r$, in particular, $r \in \mathbb{N}_0$. Since (A.19a) also implies $(q+1)b > a = qb + r$, we also have $b > r$ as required.

Uniqueness: Suppose $(q_1, r_1) \in \mathbb{N}_0$, satisfying the two conditions $a = q_1 b + r_1$ and $0 \leq r_1 < b$. Then $q_1 b = a - r_1 \leq a$ and $(q_1 + 1)b = a - r_1 + b > a$, showing $q_1 = \max\{n \in \mathbb{N}_0 : nb \leq a\} = q$. This, in turn, implies $r_1 = a - q_1 b = a - qb = r$, thereby establishing the case. ■

B Operator Norms and Matrix Norms

Theorem B.1. *For a linear map $A : X \longrightarrow Y$ between two normed vector spaces $(X, \|\cdot\|)$ and $(Y, \|\cdot\|)$, 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) ⇒ (d) ⇒ (e)" is clear.

"(e) ⇒ (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 $L \in \mathbb{R}_0^+$ 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

$$\|Ax - Ay\| \leq \|A\|\,\|x - y\| \text{ for each } x, y \in X. \tag{B.2}$$

For $x = y$, there is nothing to prove. Thus, let $x \neq y$. One computes

$$\frac{\|Ax - Ay\|}{\|x - y\|} = \left\| A\left(\frac{x - y}{\|x - y\|}\right)\right\| \leq \|A\| \tag{B.3}$$

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 $0 \neq x \in M$:

$$\frac{\|Ax\|}{\|x\|} = \left\| A\left(\frac{x}{\|x\|}\right)\right\| \leq \|A\| \tag{B.4}$$

as $\left\|\frac{x}{\|x\|}\right\| = 1$. Thus $\|Ax\| \leq \|A\|\|x\| \leq r\|A\|$, showing that $A(M) \subseteq B_{r\|A\|}(0)$. Thus, $A(M)$ is bounded, thereby establishing the case.

"(a) $\Rightarrow$ (b)": Since A is bounded, it maps the bounded set $B_1(0) \subseteq X$ into some bounded subset of Y. Thus, there is $r > 0$ such that $A(B_1(0)) \subseteq B_r(0) \subseteq Y$. In particular, $\|Ax\| < r$ for each $x \in X$ satisfying $\|x\| = 1$, showing $\|A\| \leq r < \infty$. ■

One can ask the question if every norm on $\mathcal{L}(X, Y)$ is an operator norm induced by norms on X and Y. The following Example B.9 shows that this is not necessarily the case. This is not part of the core material of this lecture; it is just meant as an aside for the interested reader. We need some preparatory notions and results:

Theorem B.2. *$\mathcal{L}(X, Y)$ with the operator norm is a Banach space (i.e. a complete normed vector space) provided that Y is a Banach space (even if X is not a Banach space).*

Proof. See, e.g., [RF10, Ch. 10, Prop. 3] or [Alt06, Th. 3.3]. ■

Definition and Remark B.3. If X is a normed vector space, then the space $X^* := \mathcal{L}(X, \mathbb{R})$ is called the *dual* of X (the elements of X^* are called *bounded linear functionals*). According to (2.45), one has, for each $f \in X^*$,

$$\|f\|_{X^*} = \sup\big\{|f(x)| : x \in X,\ \|x\|_X = 1\big\}. \tag{B.5}$$

Moreover, according to Th. B.2, the fact that $\mathbb{R}$ is a Banach space implies that X^* is always a Banach space.

Theorem B.4. *Let X be a normed vector space. If $0 \neq x \in X$, then there exists $f \in X^*$ (i.e. a bounded linear functional f) such that*

$$\|f\| = 1 \quad \textit{and} \quad f(x) = \|x\|. \tag{B.6}$$

Proof. For a proof of the general case stated in the theorem, see, e.g., [RF10, Prop. 10.6]. It is based on the Hahn-Banach theorem of functional analysis, which, in turn, relies on the axiom of choice. Here, however, we are only interested in the case $X = \mathbb{R}^n$, $n \in \mathbb{N}$, and we provide a direct proof for this special case:

For $X = \mathbb{R}^n$, we can construct f by induction on n. For $n = 1$, $f : X \longrightarrow \mathbb{R}$, $f(y) := \operatorname{sgn}(y)\,\|y\|$ will do the job. For $n > 1$, let $\{b_1, \dots, b_n\}$ be a basis of X such that $x \in V := \operatorname{span}\{b_1, \dots, b_{n-1}\}$. By induction, there is a map $g \in V^*$ such that $\|g\| = 1$ and $g(x) = \|x\|$. The goal is to define

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}, \quad f(v + \lambda b_n) := g(v) + \lambda f(b_n) \tag{B.7}$$

with a suitable choice for $f(b_n)$. To find a suitable $f(b_n)$, note that, for each $v_1, v_2 \in V$:

$$g(v_1) + g(v_2) = g(v_1 + v_2) \leq \|v_1 + v_2\| \leq \|v_1 - b_n\| + \|v_2 + b_n\|, \tag{B.8}$$

i.e.

$$-\|v_1 - b_n\| + g(v_1) \leq \|v_2 + b_n\| - g(v_2), \tag{B.9}$$

implying

$$\alpha := \sup\big\{ -\|v - b_n\| + g(v) : v \in V \big\} \leq \inf\big\{ \|v + b_n\| - g(v) : v \in V \big\} =: \beta. \tag{B.10}$$

Choose an arbitrary $\gamma \in [\alpha, \beta]$ and set $f(b_n) := \gamma$. It remains to show

$$|f(v + \lambda b_n)| \leq \|v + \lambda b_n\| \tag{B.11}$$

for each $v \in V$ and each $\lambda \in \mathbb{R}$. To that end, fix $v \in V$. For $\lambda = 0$, (B.11) holds by induction. For $\lambda > 0$:

$$\begin{aligned} f(v + \lambda b_n) &= g(v) + \lambda\gamma = \lambda\big(g(v/\lambda) + \gamma\big) \\ &\leq \lambda\big(g(v/\lambda) + \|v/\lambda + b_n\| - g(v/\lambda)\big) = \|v + \lambda b_n\|. \end{aligned} \tag{B.12}$$

For $\lambda < 0$:

$$\begin{aligned} f(v + \lambda b_n) &= g(v) + \lambda\gamma = -\lambda\big(g(v/(-\lambda)) - \gamma\big) \\ &\leq -\lambda\big(g(v/(-\lambda)) + \|v/(-\lambda) - b_n\| - g(v/(-\lambda))\big) = \|v + \lambda b_n\|. \end{aligned} \tag{B.13}$$

So we obtain that (B.11) holds for each $v \in V$ and each $\lambda \in \mathbb{R}$ without the absolute value on the left-hand side. However, replacing v by $-v$ and λ by $-\lambda$ then shows that it remains true with the absolute value as well. ■

Definition B.5. Let X, Y be normed vector spaces, and let $A : X \longrightarrow Y$ be a bounded and linear. The map

$$A^* : Y^* \longrightarrow X^*, \quad A^*(f) := f \circ A, \tag{B.14}$$

is called the *adjoint* or *dual* of A.

Example B.6. Let $m,n \in \mathbb{N}$ and let $A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be the linear map given by the $m \times n$ matrix (a_{ij}). We claim that the adjoint map $A^* : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ is given by the adjoint matrix (a_{ji}), showing that both notions are consistently named:

Note that we can indentify $\mathbb{R}^n$ with $(\mathbb{R}^n)^*$, where $x^* = (x_1^*, \dots, x_n^*) \in (\mathbb{R}^n)^*$ acts on $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ via $x^*(x) = \sum_{i=1}^n x_i^* x_i$ and analogously for $\mathbb{R}^m$ and $(\mathbb{R}^m)^*$. Thus, for each $y^* = (y_1^*, \dots, y_m^*) \in (\mathbb{R}^m)^*$ and each $x = (x_1, \dots, x_n) \in \mathbb{R}^n$:

$$\begin{aligned}(A^* y^*)(x) = y^*(Ax) &= \sum_{i=1}^m y_i^*(Ax)_i = \sum_{i=1}^m y_i^* \sum_{j=1}^n a_{ij} x_j = \sum_{j=1}^m y_j^* \sum_{i=1}^n a_{ji} x_i \\ &= \sum_{i=1}^n \sum_{j=1}^m a_{ji} x_i y_j^* = \sum_{i=1}^n x_i \sum_{j=1}^m a_{ji} y_j^*, \end{aligned} \tag{B.15}$$

identifying $(A^* y^*)_i = \sum_{j=1}^m a_{ji} y_j^*$ for each $i = 1, \dots, n$, establishing that the adjoint map A^* is, indeed, given by the adjoint matrix (a_{ji}).

Theorem B.7. *Let X, Y be Banach spaces, and let $A : X \longrightarrow Y$ be a bounded linear operator.*

(a) *The adjoint operator A^* of A according to Def.* B.5 *is well-defined, i.e. $f \circ A \in X^*$ for each $f \in Y^*$.*

(b) *A^* is a bounded linear operator and $\|A^*\| = \|A\|$.*

Proof. (a): As a composition of linear maps, $f \circ A$ is linear; as a composition of continuous maps, $f \circ A$ is continuous.

(b): If $\lambda, \mu \in \mathbb{R}$ and $f, g \in Y^*$, then

$$A^*(\lambda f + \mu g) = (\lambda f + \mu g) \circ A = \lambda (f \circ A) + \mu (g \circ A) = \lambda\, A^*(f) + \mu\, A^*(g), \tag{B.16}$$

showing that A^* is linear. Moreover, if $f \in Y^*$ and $x \in X$, then

$$\big|(A^* f)x\big| = \big|f(A(x))\big| \le \|f\|\,\|A(x)\| \le \|f\|\,\|A\|\,\|x\|, \tag{B.17}$$

showing $\|A^* f\| \le \|f\|\,\|A\|$ and $\|A^*\| \le \|A\|$.

The proof of $\|A^*\| \ge \|A\|$ is based on Th. B.4: For each $x \in X$ such that $Ax \ne 0$, Th. B.4 provides $f_x \in Y^*$ with $\|f_x\| = 1$ and $f_x(Ax) = \|Ax\|$. Now let $(x_n)_{n \in \mathbb{N}}$ be a sequence in X such that $\|x_n\| = 1$ for each $n \in \mathbb{N}$ and $\lim_{n\to\infty} \|Ax_n\| = \|A\|$. Then, for each $n \in \mathbb{N}$,

$$\big|(A^* f_{x_n}) x_n\big| = \big|f_{x_n}(Ax_n)\big| = \|Ax_n\| = \|A\| \tag{B.18}$$

implies $\|A^* f_{x_n}\| \ge \|Ax_n\|$, showing $\|A^*\| \ge \|A\|$. ■

Lemma B.8. *Let $(X, \|\cdot\|)$ and $(Y, \|\cdot\|)$ be normed vector spaces and, for $A \in \mathcal{L}(X,Y)$, let $\|A\|$ denote the corresponding induced operator norm. Let $\alpha \in \mathbb{R}^+$. Then the norm $\|\cdot\|'$ on $\mathcal{L}(X,Y)$ defined by $\|A\|' := \alpha\|A\|$ is the induced operator norm corresponding to $(X, \|\cdot\|)$ and $(Y, \|\cdot\|')$, where, for each $y \in Y$, $\|y\|' := \alpha\|y\|$. In particular, a norm N on $\mathcal{L}(X,Y)$ is an operator norm induced by norms on X and Y if, and only if, all positive multiples αN are operator norms induced by norms on X and Y.*

Proof. We only need to show that $\|A\|'$ is the induced operator norm corresponding to $(X, \|\cdot\|')$ and $(Y, \|\cdot\|)$. One computes

$$\begin{aligned}\|A\|' = \alpha\,\|A\| = \alpha\, \sup\big\{\|Ax\| : x \in X,\ \|x\| = 1\big\} = \sup\big\{\alpha\,\|Ax\| : x \in X,\ \|x\| = 1\big\}\\ = \sup\big\{\|Ax\|' : x \in X,\ \|x\| = 1\big\}, \qquad \text{(B.19)}\end{aligned}$$

thereby establishing the case. ■

Example B.9. Let $m, n \in \mathbb{N}$ and let $A : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be the linear map given by the $m \times n$ matrix (a_{ij}). In view of Rem. 2.35, we define

$$\|A\|_{\text{Frob}} := \|A\|_{\text{HS}} := \sqrt{\operatorname{tr}(A^*A)} = \sqrt{\operatorname{tr}(AA^*)} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}, \qquad \text{(B.20)}$$

called the *Frobenius* norm or the *Hilbert-Schmidt* norm of A. The claim is now that, for $m, n > 1$, there do not exist norms on $\mathbb{R}^m$ and $\mathbb{R}^n$ such that the Frobenius norm is the resulting induced operator norm. In view of Lem. B.8, it suffices to show that the modified Frobenius norm $\|A\|'_{\text{Frob}} := \|A\|_{\text{Frob}}/\sqrt{\max\{m,n\}}$ is not induced by norms on $\mathbb{R}^m$ and $\mathbb{R}^n$. This follows from the fact that, for each $m, n \in \mathbb{N}$ with $m, n > 1$, there exists an $m \times n$ matrix A such that

$$\|A^*A\|'_{\text{Frob}} > \|A^*\|'_{\text{Frob}} \|A\|'_{\text{Frob}} \qquad \text{(B.21)}$$

(it is an exercise to find such matrices A): If the modified Frobenius norm on A were induced by norms on $\mathbb{R}^m$ and $\mathbb{R}^n$, then, according to Example B.6 and Th. B.7(b), the modified Frobenius norm on A^* were induced by (the corresponding operator) norms on $(\mathbb{R}^m)^* \cong \mathbb{R}^m$ and $(\mathbb{R}^n)^* \cong \mathbb{R}^n$, implying that (B.21) could not occur, as induced norms must satisfy (2.51). Finally, Lem. B.8 yields that no positive multiple of the Frobenius norm is induced by norms on $\mathbb{R}^m$ and $\mathbb{R}^n$.

Caveat: It is underlined again that, as seen in the previous example, even though the Frobenius norm can be interpreted as the 2-norm on $\mathbb{R}^{mn}$, it is *not* induced by the Euclidean norms on $\mathbb{R}^m$ and $\mathbb{R}^n$.

C $\mathbb{R}^m$-Valued Integration

In the proof of the well-conditionedness of a continuously differentiable function $f : U \longrightarrow \mathbb{R}^m$, $U \subseteq \mathbb{R}^n$, in Th. 2.52, we make use of $\mathbb{R}^m$-valued integrals. In particular, in the estimate (2.67), we use that $\|\int_A f\| \leq \int_A \|f\|$ 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. For a general treatment of vector-valued integrals, see, for example, [Alt06, Sec. A1] or [Yos80, Sec. V.5].

Definition C.1. Let $A \subseteq \mathbb{R}^n$ be measurable, $m, n \in \mathbb{N}$.

(a) A function $f : A \longrightarrow \mathbb{R}^m$ is called *measurable* (respectively, *integrable*) if, and only if, each coordinate function $f_i = \pi_i \circ f : A \longrightarrow \mathbb{R}$, $i = 1, \dots, m$, is measurable (respectively, integrable).

(b) If $f : A \longrightarrow \mathbb{R}^m$ is integrable, then

$$\int_A f := \left(\int_A f_1, \dots, \int_A f_m \right) \in \mathbb{R}^m \tag{C.1}$$

is the ($\mathbb{R}^m$-valued) *integral* of f over A.

Remark C.2. The linearity of the $\mathbb{R}$-valued integral implies the linearity of the $\mathbb{R}^m$-valued integral.

Theorem C.3. *Let $A \subseteq \mathbb{R}^n$ be measurable, $m, n \in \mathbb{N}$. Then $f : A \longrightarrow \mathbb{R}^m$ is measurable in the sense of Def.* C.1(a) *if, and only if, $f^{-1}(O)$ is measurable for each open subset O of $\mathbb{R}^m$.*

Proof. Assume $f^{-1}(O)$ is measurable for each open subset O of $\mathbb{R}^m$. Let $i \in \{1, \dots, m\}$. If $O_i \subseteq \mathbb{R}$ is open in $\mathbb{R}$, then $O := \pi_i^{-1}(O_i) = \{x \in \mathbb{R}^m : x_i \in O_i\}$ is open in $\mathbb{R}^m$. Thus, $f_i^{-1}(O_i) = f^{-1}(O)$ is measurable, showing that each f_i is measurable, i.e. f is measurable. Now assume f is measurable, i.e. each f_i is measurable. Since every open $O \subseteq \mathbb{R}^m$ is a countable union of open sets of the form $O = O_1 \times \dots \times O_m$ with each O_i being an open subset of $\mathbb{R}$, it suffices to show that the preimages of such open sets are measurable. So let O be as above. Then $f^{-1}(O) = \bigcap_{i=1}^m f_i^{-1}(O_i)$, showing that $f^{-1}(O)$ is measurable. ■

Corollary C.4. *Let $A \subseteq \mathbb{R}^n$ be measurable, $m, n \in \mathbb{N}$. If $f : A \longrightarrow \mathbb{R}^m$ is measurable, then $\|f\| : A \longrightarrow \mathbb{R}$ is measurable.*

Proof. If $O \subseteq \mathbb{R}$ is open, then $\|\cdot\|^{-1}(O)$ is an open subset of $\mathbb{R}^m$ by the continuity of the norm. In consequence, $\|f\|^{-1}(O) = f^{-1}\big(\|\cdot\|^{-1}(O)\big)$ is measurable. ■

Theorem C.5. *Let $A \subseteq \mathbb{R}^n$ be measurable, $m, n \in \mathbb{N}$. For each norm $\|\cdot\|$ on $\mathbb{R}^m$ and each integrable $f : A \longrightarrow \mathbb{R}^m$, the following holds:*

$$\left\| \int_A f \right\| \leq \int_A \|f\|. \tag{C.2}$$

Proof. First assume that $B \subseteq A$ is measurable, $y \in \mathbb{R}^m$, and $f = y\,\chi_B$, where χ_B is the characteristic function of B (i.e. the f_i are y_i on B and 0 on $A \setminus B$). Then

$$\left\| \int_A f \right\| = \Big\| \big(y_1 \lambda_m(B), \dots, y_m \lambda_m(B)\big) \Big\| = \lambda_m(B) \|y\| = \int_A \|f\|, \tag{C.3}$$

where λ_m denotes m-dimensional Lebesgue measure. Next, consider the case that f is a so-called *simple function*, that means f takes only finitely many values $y_1, \dots, y_N$, $N \in \mathbb{N}$, and each preimage $B_i := f^{-1}\{y_i\} \subseteq \mathbb{R}^n$ is measurable. Then $f = \sum_{i=1}^N y_i \chi_{B_i}$, where, without loss of generality, we may assume that the B_i are pairwise disjoint. We obtain

$$\left\| \int_A f \right\| \leq \sum_{i=1}^{N} \left\| \int_A y_i \chi_{B_i} \right\| = \sum_{i=1}^{N} \int_A \|y_i \chi_{B_i}\| = \int_A \left\| \sum_{i=1}^{N} y_i \chi_{B_i} \right\| = \int_A \|f\| \tag{C.4}$$

(note that, as the B_i are disjoint, the integrands of the last two integrals are, indeed, equal at each $x \in A$).

Now, if f is integrable, then each f_i is integrable (i.e. $f_i \in L^1(A)$) and there exist sequences of simple functions $\phi_{i,k} : A \longrightarrow \mathbb{R}$ such that $\lim_{k\to\infty} \|\phi_{i,k} - f_i\|_{L^1(A)} = 0$. In particular,

$$0 \leq \lim_{k\to\infty} \left| \int_A \phi_{i,k} - \int_A f_i \right| \leq \lim_{k\to\infty} \|\phi_{i,k} - f_i\|_{L^1(A)} = 0. \tag{C.5}$$

Thus, we obtain

$$\begin{aligned} \left\| \int_A f \right\| &= \left\| \left(\int_A f_1, \dots, \int_A f_m \right) \right\| = \left\| \left(\lim_{k\to\infty} \int_A \phi_{1,k}, \dots, \lim_{k\to\infty} \int_A \phi_{m,k} \right) \right\| = \lim_{k\to\infty} \left\| \int_A \phi_k \right\| \\ &\leq \lim_{k\to\infty} \int_A \|\phi_k\| \overset{(*)}{=} \int_A \|f\|, \end{aligned} \tag{C.6}$$

where the equality at $(*)$ holds due to $\lim_{k\to\infty} \big\| \|\phi_k\| - \|f\| \big\|_{L^1(A)} = 0$, which, in turn, is verified by

$$\begin{aligned} 0 \leq \int_A \big| \|\phi_k\| - \|f\| \big| &\leq \int_A \|\phi_k - f\| \leq C \int_A \|\phi_k - f\|_1 \\ &= C \int_A \sum_{i=1}^{m} |\phi_{i,k} - f_i| \to 0 \quad \text{for } k \to \infty, \end{aligned} \tag{C.7}$$

with $C \in \mathbb{R}^+$ since the norms $\|\cdot\|$ and $\|\cdot\|_1$ are equivalent on $\mathbb{R}^m$. ■

D The Vandermonde Determinant

Theorem D.1. *Let $n \in \mathbb{N}$ and $\lambda_0, \lambda_1, \dots, \lambda_n \in \mathbb{C}$. Moreover, let*

$$V := \begin{pmatrix} 1 & \lambda_0 & \dots & \lambda_0^n \\ 1 & \lambda_1 & \dots & \lambda_1^n \\ \vdots & & & \vdots \\ 1 & \lambda_n & \dots & \lambda_n^n \end{pmatrix} \tag{D.1}$$

be the corresponding Vandermonde matrix*. Then its determinant, the so-called* Vandermonde determinant *is given by*

$$\det(V) = \prod_{\substack{k,l=0 \\ k>l}}^{n} (\lambda_k - \lambda_l). \tag{D.2}$$

Proof. The proof can be conducted by induction with respect to n: For $n = 1$, we have

$$\det(V) = \begin{vmatrix} 1 & \lambda_0 \\ 1 & \lambda_1 \end{vmatrix} = \lambda_1 - \lambda_0 = \prod_{\substack{k,l=0\\k>l}}^{1} (\lambda_k - \lambda_l), \tag{D.3}$$

showing (D.2) holds for $n = 1$. Now let $n > 1$. We know from Linear Algebra that the value of a determinant does not change if we add a multiple of a column to a different column. Adding the $(-\lambda_0)$-fold of the nth column to the $(n+1)$st column, we obtain in the $(n+1)$st column

$$\begin{pmatrix} 0 \\ \lambda_1^n - \lambda_1^{n-1}\lambda_0 \\ \vdots \\ \lambda_n^n - \lambda_n^{n-1}\lambda_0 \end{pmatrix}. \tag{D.4}$$

Next, one adds the $(-\lambda_0)$-fold of the $(n-1)$st column to the nth column, and, successively, the $(-\lambda_0)$-fold of the mth column to the $(m+1)$st column. One finishes, in the nth step, by adding the $(-\lambda_0)$-fold of the first column to the second column, obtaining

$$\det(V) = \begin{vmatrix} 1 & \lambda_0 & \dots & \lambda_0^n \\ 1 & \lambda_1 & \dots & \lambda_1^n \\ \vdots & & & \vdots \\ 1 & \lambda_n & \dots & \lambda_n^n \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & \lambda_1 - \lambda_0 & \lambda_1^2 - \lambda_1\lambda_0 & \dots & \lambda_1^n - \lambda_1^{n-1}\lambda_0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \lambda_n - \lambda_0 & \lambda_n^2 - \lambda_n\lambda_0 & \dots & \lambda_n^n - \lambda_n^{n-1}\lambda_0 \end{vmatrix}. \tag{D.5}$$

Applying the rule for determinants of block matrices to (D.5) yields

$$\det(V) = 1 \cdot \begin{vmatrix} \lambda_1 - \lambda_0 & \lambda_1^2 - \lambda_1\lambda_0 & \dots & \lambda_1^n - \lambda_1^{n-1}\lambda_0 \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_n - \lambda_0 & \lambda_n^2 - \lambda_n\lambda_0 & \dots & \lambda_n^n - \lambda_n^{n-1}\lambda_0 \end{vmatrix}. \tag{D.6}$$

As we also know from Linear Algebra that determinants are linear in each row, for each k, we can factor out $(\lambda_k - \lambda_0)$ from the kth row of (D.6), arriving at

$$\det(V) = \prod_{k=1}^{n} (\lambda_k - \lambda_0) \begin{vmatrix} 1 & \lambda_1 & \dots & \lambda_1^{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \lambda_n & \dots & \lambda_n^{n-1} \end{vmatrix}. \tag{D.7}$$

However, the determinant in (D.7) is precisely the Vandermonde determinant of the $n-1$ numbers $\lambda_1, \dots, \lambda_n$, which is given according to the induction hypothesis, implying

$$\det(V) = \prod_{k=1}^{n} (\lambda_k - \lambda_0) \prod_{\substack{k,l=1\\k>l}}^{n} (\lambda_k - \lambda_l) = \prod_{\substack{k,l=0\\k>l}}^{n} (\lambda_k - \lambda_l), \tag{D.8}$$

completing the induction proof of (D.2). ■

E Parameter-Dependent Integrals

Theorem E.1. *Let (X,d) be a metric space and let $(\Omega,\mathcal{A},\mu)$ be a measure space. Let $n \in \mathbb{N}$ and $f : X \times \Omega \longrightarrow \mathbb{R}^n$. Fix $x_0 \in X$ and assume f has the following properties:*

(a) *For each $x \in X$, the function $\omega \mapsto f(x,\omega)$ is integrable (i.e. all its components are integrable).*

(b) *For each $\omega \in \Omega$, the function $x \mapsto f(x,\omega)$ is continuous in x_0.*

(c) *f is uniformly dominated by an integrable function h, i.e. there exists an integrable $h : \Omega \longrightarrow \mathbb{R}^n$ such that*

$$\forall_{i=1,\dots,n} \quad \forall_{(x,\omega)\in X\times\Omega} \quad |f_i(x,\omega)| \leq h_i(\omega). \tag{E.1}$$

Then the function

$$\phi : X \longrightarrow \mathbb{R}^n, \quad \phi(x) := \int_\Omega f(x,\omega)\,\mathrm{d}\omega\,, \tag{E.2}$$

is continuous in x_0.

Proof. Let $(x_m)_{m\in\mathbb{N}}$ be a sequence in X such that $\lim_{m\to\infty} = x_0$. According to (b), this gives rise to a sequence $(g_m)_{m\in\mathbb{N}}$ of integrable functions

$$g_m : \Omega \longrightarrow \mathbb{R}^n, \quad g_m(\omega) := f(x_m,\omega), \tag{E.3}$$

converging pointwise on Ω to the function

$$g : \Omega \longrightarrow \mathbb{R}^n, \quad g(\omega) := f(x_0,\omega). \tag{E.4}$$

Since, by (c), all g_m are dominated by the integrable function h (i.e. $|g_{mi}| \leq h_i$ for each $m \in \mathbb{N}$, $i \in \{1,\dots,n\}$), we can apply the dominated convergence theorem (DCT) to each component to obtain

$$\begin{aligned}\lim_{m\to\infty}\phi(x_m) &= \lim_{m\to\infty}\int_\Omega f(x_m,\omega)\,\mathrm{d}\omega = \lim_{m\to\infty}\int_\Omega g_m(\omega)\,\mathrm{d}\omega \overset{\text{DCT}}{=} \int_\Omega g(\omega)\,\mathrm{d}\omega \\ &= \int_\Omega f(x_0,\omega)\,\mathrm{d}\omega = \phi(x_0),\end{aligned} \tag{E.5}$$

proving ϕ is continuous at x_0. ■

F Inner Products and Orthogonality

Definition and Remark F.1. Let X be a real vector space. A function $\langle\cdot,\cdot\rangle : X \times X \longrightarrow \mathbb{R}$ is called an *inner product* or a *scalar product* on X if, and only if, the following three conditions are satisfied:

(i) $\langle x, x\rangle > 0$ for each $0 \neq x \in X$.

(ii) $\langle \lambda x + \mu y, z\rangle = \lambda\langle x, z\rangle + \mu\langle y, z\rangle$ for each $x, y, z \in X$ and each $\lambda, \mu \in \mathbb{R}$.

(iii) $\langle x, y\rangle = \langle y, x\rangle$ for each $x, y \in X$.

If $\langle\cdot,\cdot\rangle$ is an inner product on X, then $\big(X, \langle\cdot,\cdot\rangle\big)$ is called an *inner product space* and the map

$$\|\cdot\| : X \longrightarrow \mathbb{R}_0^+, \quad \|x\| := \sqrt{\langle x, x\rangle}, \tag{F.1}$$

defines a norm on X, called the norm *induced* by the inner product.

Definition F.2. Let $\big(X, \langle\cdot,\cdot\rangle\big)$ be an inner product space.

(a) Vectors $x, y \in X$ are called *orthogonal* or *perpendicular* (denoted $x \perp y$) if, and only if, $\langle x, y\rangle = 0$. An *orthogonal system* is a family $(x_\alpha)_{\alpha\in I}$, $x_\alpha \in X$, I being some index set, such that $\langle x_\alpha, x_\beta\rangle = 0$ for each $\alpha, \beta \in I$ with $\alpha \neq \beta$. An orthogonal system is called an *orthonormal system* if, and only if, it consists entirely of unit vectors (with respect to the induced norm).

(b) An orthonormal system $(x_\alpha)_{\alpha\in I}$ is called an *orthonormal basis* if, and only if, $x = \sum_{\alpha\in I}\langle x, x_\alpha\rangle\, x_\alpha$ for each $x \in X$.

(c) If V is a linear subspace of X, then we define

$$V^\perp := \{x \in X : x \perp v \text{ for each } v \in V\}. \tag{F.2}$$

Lemma F.3. *Let $\big(X, \langle\cdot,\cdot\rangle\big)$ be an inner product space and let V be a linear subspace of X.*

(a) *$V^\perp$ is a linear subspace of X.*

(b) *$V \cap V^\perp = \{0\}$.*

(c) *If I is some index set and $(x_\alpha)_{\alpha\in I}$, $x_\alpha \in X$, is an orthogonal system such that $x_\alpha \neq 0$ for each $\alpha \in I$, then the x_α are all linearly independent.*

(d) *Let $0 < \dim X = n < \infty$. Then an orthonormal system $(x_\alpha)_{\alpha\in I}$ is an orthonormal basis if, and only if, it is a basis in the usual sense of linear algebra.*

Proof. (a): Let $x, y \in V^\perp$ and $\lambda \in \mathbb{R}$. Then, for each $v \in V$, it holds that $\langle x + y, v\rangle = \langle x, v\rangle + \langle y, v\rangle = 0$ and $\langle \lambda x, v\rangle = \lambda\langle x, v\rangle = 0$, showing that $x + y \in V^\perp$ as well as $\lambda x \in V^\perp$, implying that $V^\perp$ is a linear subspace.

(b): If $v \in V$ and $v \in V^\perp$, then $\langle v, v\rangle = 0$ by the definition of $V^\perp$. However, $\langle v, v\rangle = 0$ implies $v = 0$ according to F.1(i).

(c): Suppose $n \in \mathbb{N}$ and $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ together with $x_{\alpha_1}, \dots, x_{\alpha_n}$ are such that $\sum_{i=1}^n \lambda_i x_{\alpha_i} = 0$. Then, for each $j \in \{1, \dots, n\}$, the relations $\langle x_{\alpha_i}, x_{\alpha_j}\rangle = 0$ for each $i \neq j$ imply $0 = \langle 0, x_{\alpha_j}\rangle = \langle\sum_{i=1}^n \lambda_i x_{\alpha_i}, x_{\alpha_j}\rangle = \sum_{i=1}^n \lambda_i\langle x_{\alpha_i}, x_{\alpha_j}\rangle = \lambda_j\langle x_{\alpha_j}, x_{\alpha_j}\rangle$, which yields $\lambda_j = 0$ by F.1(i). Thus, we have shown that $\lambda_j = 0$ for each $j \in \{1, \dots, n\}$, which establishes that the x_α are all linearly independent.

(d): If $(x_\alpha)_{\alpha \in I}$ is an orthonormal basis, then the x_α are linearly independent by (c) (since $\|x_\alpha\| = 1$ for each $\alpha \in I$). In particular, $\dim X = n$ implies $\#I \leq n$. As Def. F.2(b) implies $\operatorname{span}\{x_\alpha : \alpha \in I\} = X$ (here we use $n < \infty$, otherwise the sums from Def. F.2(b) could have infinitely many nonzero terms), we see that $(x_\alpha)_{\alpha \in I}$ is a basis in the usual sense of linear algebra. Conversely, if $(x_\alpha)_{\alpha \in I}$ is a basis in the usual sense of linear algebra, then, for each $x \in X$, there are $\lambda_\alpha \in \mathbb{R}$, $\alpha \in I$, such that $x = \sum_{\alpha \in I} \lambda_\alpha \, x_\alpha$. Thus,

$$\begin{aligned} \sum_{\alpha \in I} \langle x, x_\alpha \rangle \, x_\alpha &= \sum_{\alpha \in I} \left\langle \sum_{\beta \in I} \lambda_\beta \, x_\beta, \, x_\alpha \right\rangle x_\alpha = \sum_{\alpha \in I} \sum_{\beta \in I} \lambda_\beta \langle x_\beta, x_\alpha \rangle \, x_\alpha \\ &= \sum_{\alpha \in I} \sum_{\beta \in I} \lambda_\beta \, \delta_{\beta,\alpha} \, x_\alpha = \sum_{\alpha \in I} \lambda_\alpha \, x_\alpha = x, \end{aligned} \tag{F.3}$$

showing that $(x_\alpha)_{\alpha \in I}$ is an orthonormal basis. ■

G Baire Category and the Uniform Boundedness Principle

Theorem G.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*

$$X = \bigcup_{k=1}^{\infty} A_k, \tag{G.1}$$

then there exists $k_0 \in \mathbb{N}$ such that A_{k_0} has nonempty interior: $\operatorname{int} A_{k_0} \neq \emptyset$.

Proof. Seeking a contradiction, we assume (G.1) holds with closed sets A_k such that $\operatorname{int} A_k = \emptyset$ 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

$$\overline{B}_\epsilon(x) \subseteq O \setminus A_k. \tag{G.2}$$

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

$$\overline{B}_{\epsilon_k}(x_k) \subseteq B_{\epsilon_{k-1}}(x_{k-1}) \setminus A_k, \quad \epsilon_k \leq \frac{1}{k} \quad \text{for each } k \in \mathbb{N}. \tag{G.3}$$

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 (G.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 (G.1). ■

Remark G.2. The purpose of this remark is to explain the name *Baire Category Theorem* in Th. G.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. G.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. G.1.

Theorem G.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*

$$M_x := \sup\big\{\|f(x)\| : f \in \mathcal{F}\big\} < \infty \quad \textit{for each } x \in X, \tag{G.4}$$

then there exists $x_0 \in X$ and $\epsilon_0 > 0$ such that

$$\sup\big\{M_x : x \in \overline{B}_{\epsilon_0}(x_0)\big\} < \infty. \tag{G.5}$$

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. Let $f \in \mathcal{F}$, $k \in \mathbb{N}$. As both f and the norm are continuous, the set

$$A_{k,f} := \big\{x \in X : \|f(x)\| \leq k\big\} \tag{G.6}$$

is an continuous inverse image of the closed set $[0,k]$ and, hence, closed. Since arbitrary intersections of closed sets are closed, so is

$$A_k := \bigcap_{k=1}^{\infty} A_{k,f} = \big\{x \in X : \|f(x)\| \leq k \text{ for each } f \in \mathcal{F}\big\}. \tag{G.7}$$

Now (G.4) implies

$$X = \bigcup_{k=1}^{\infty} A_k \tag{G.8}$$

and the Baire Category Th. G.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,

$$\sup\left\{M_x : x \in \overline{B}_{\epsilon_0}(x_0)\right\} \leq k_0, \tag{G.9}$$

proving (G.5). ■

We now specialize the uniform boundedness principle to bounded linear maps:

Theorem G.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*

$$\sup\left\{\|T(x)\| : T \in \mathcal{T}\right\} < \infty \quad \textit{for each } x \in X, \tag{G.10}$$

then $\mathcal{T}$ is a bounded subset of $\mathcal{L}(X,Y)$, i.e.

$$\sup\left\{\|T\| : T \in \mathcal{T}\right\} < \infty. \tag{G.11}$$

Proof. To apply Th. G.3, define, for each $T \in \mathcal{T}$:

$$f_T : X \longrightarrow \mathbb{R}, \quad f_T(x) := \|Tx\|. \tag{G.12}$$

Then $\mathcal{F} := \{f_T : T \in \mathcal{T}\} \subseteq C(X,\mathbb{R})$ and (G.10) implies that $\mathcal{F}$ satisfies (G.4). Thus, if M_x is as in (G.4), we obtain $x_0 \in X$ and $\epsilon_0 > 0$ such that

$$\sup\left\{M_x : x \in \overline{B}_{\epsilon_0}(x_0)\right\} < \infty, \tag{G.13}$$

i.e. there exists $C \in \mathbb{R}^+$ satisfying

$$\|Tx\| \leq C \quad \text{for each } T \in \mathcal{T} \text{ and each } x \in X \text{ with } \|x - x_0\| \leq \epsilon_0. \tag{G.14}$$

In consequence, for each $x \in X \setminus \{0\}$,

$$\|Tx\| = \frac{\|x\|}{\epsilon_0} \cdot \left\| T\left(x_0 + \epsilon_0 \frac{x}{\|x\|}\right) - Tx_0 \right\| \leq \frac{\|x\|}{\epsilon_0} \cdot 2\,C, \tag{G.15}$$

showing $\|T\| \leq \frac{2C}{\epsilon_0}$ for each $T \in \mathcal{T}$, thereby establishing the case. ■

H Linear Algebra

H.1 Gaussian Elimination Algorithm

Definition H.1. Let A be an $n \times m$ matrix, $m, n \in \mathbb{N}$. For each row, i.e. for each $i \in \{1, \dots, n\}$, let $\nu(i) \in \{1, \dots, m\}$ be the smallest index k such that $a_{ik} \neq 0$ and $\nu(i) := m + 1$ if the ith row consists entirely of zeros. Then A is said to be in *echelon form* if, and only if, for each $i \in \{2, \dots, n\}$, one has $\nu(i) > \nu(i-1)$ or $\nu(i) = m + 1$. Thus, A is in echelon form if, and only if, it looks as follows:

$$A = \begin{pmatrix} 0 & \dots & 0 & \square & * & * & * & * & \dots & * & * & * & * \\ 0 & \dots & 0 & 0 & \dots & 0 & \square & * & * & * & \dots & * & * \\ 0 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 & \square & * & \dots & * \\ \vdots & & & & & & & & & & & & \vdots \end{pmatrix}. \tag{H.1}$$

The first nonzero elements in each row are called *pivot elements* (in (H.1), the positions of the pivot elements are marked by squares $\square$). The columns containing a pivot element are called *pivot columns*, the corresponding variables are called *pivot variables.* All remaining variables are called *free variables.*

Example H.2. The following matrix is in echelon form:

$$\begin{pmatrix} 0 & 0 & 3 & 3 & 0 & -1 & 0 & 3 \\ 0 & 0 & 0 & 4 & 0 & 2 & -3 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

It remains in echelon form if one adds zero rows at the bottom. Here x_3, x_4, x_7 are pivot variables, whereas x_1, x_2, x_5, x_6, x_8 are free variables.

—

The set of solutions to a linear system are easily determined if its augmented matrix is in echelon form (see Rem H.10 below). Thus, it is desirable to transform the augmented matrix of a linear system into echelon form without changing the set of solutions. This can be achieved by the well-known Gaussian elimination algorithm.

Definition H.3. The following three operations, which transform an $n \times m$ matrix into another $n \times m$ matrix, are known as *elementary row operations*:

(a) *Row Switching:* Switching two rows.

(b) *Row Multiplication:* Replacing a row by some nonzero multiple of that row.

(c) *Row Addition:* Replacing a row by the sum of that row and a multiple of another row.

Theorem H.4. *Applying elementary row operations to the augmented matrix* $(A|b)$ *of the linear system* (5.1) *does not change the system's set of solutions.* ■

Definition H.5. Given a real $n \times m$ matrix A, $m, n \in \mathbb{N}$, and $b \in \mathbb{R}^n$, the *Gaussian elimination algorithm* is the following procedure that successively applies elementary row operations to the augmented matrix $(A|b)$ of the linear system (5.1):

Let $(A^{(1)}|b^{(1)}) := (A|b)$, $r(1) := 1$. For each $k \geq 1$, as long as $r(k) < n$ and $k \leq m+1$, the Gaussian elimination algorithm transforms $(A^{(k)}|b^{(k)})$ into $(A^{(k+1)}|b^{(k+1)})$ and $r(k)$ into $r(k+1)$ by performing precisely one of the following actions:

(a) If $a^{(k)}_{r(k),k} \neq 0$, then, for each $i \in \{r(k)+1, \dots, n\}$, replace the ith row by the ith row plus $-a^{(k)}_{ik}/a^{(k)}_{r(k),k}$ times the $r(k)$th row. Set $r(k+1) := r(k)+1$.

(b) If $a^{(k)}_{r(k),k} = 0$, and there exists $i \in \{r(k)+1, \dots, n\}$ such that $a^{(k)}_{ik} \neq 0$, then one chooses such an $i \in \{r(k)+1, \dots, n\}$ and switches the ith with the $r(k)$th row. One then proceeds as in (a).

(c) If $a^{(k)}_{ik} = 0$ for each $i \in \{r(k), \dots, n\}$, then set $A^{(k+1)} := A^{(k)}$, $b^{(k+1)} := b^{(k)}$, and $r(k+1) := r(k)$.

—

Note that the Gaussian elimination algorithm stops after at most $m+1$ steps. Moreover, in its kth step, it can only manipulate elements that have row number at least $r(k)$ *and* column number at least k.

Theorem H.6. *Given a real $n \times m$ matrix A, $m, n \in \mathbb{N}$, and $b \in \mathbb{R}^n$, applying the Gaussian elimination algorithm to the augmented matrix $(A|b)$ of the linear system* (5.1) *yields a matrix $(\tilde{A}|\tilde{b})$ in echelon form, where $\tilde{A}$ is a real $n \times m$ matrix and $\tilde{b} \in \mathbb{R}^n$ such that the linear system $\tilde{A}x = \tilde{b}$ has precisely the same set of solutions as the original linear system $Ax = b$.* ■

Remark H.7. To avoid mistakes, especially when performing the Gaussian elimination algorithm manually, it is advisable to check the row sums after each step. It obviously suffices to consider how row sums are changed if (a) of Def. H.5 is carried out. Moreover, let $i \in \{r(k)+1, \dots, n\}$, as only rows with row numbers $i \in \{r(k)+1, \dots, n\}$ can be changed by (a) in the kth step. If $s^{(k)}_i = b^{(k)}_i + \sum_{j=1}^m a^{(k)}_{ij}$ is the sum of the ith row before (a) has been carried out in the kth step, then

$$\begin{aligned} s^{(k+1)}_i &= b^{(k+1)}_i + \sum_{j=1}^m a^{(k+1)}_{ij} = b^{(k)}_i - \frac{a^{(k)}_{ik}}{a^{(k)}_{r(k),k}}\, b^{(k)}_{r(k)} + \sum_{j=1}^m \left(a^{(k)}_{ij} - \frac{a^{(k)}_{ik}}{a^{(k)}_{r(k),k}}\, a^{(k)}_{r(k),j} \right) \\ &= s^{(k)}_i - \frac{a^{(k)}_{ik}}{a^{(k)}_{r(k),k}}\, s^{(k)}_{r(k)} \end{aligned} \tag{H.2}$$

must be the sum of the ith row after (a) has been carried out in the kth step.

—

We recall the notion of rank as well as the following general result on linear systems from Linear Algebra:

Definition and Remark H.8. The rank of a matrix A, denoted by $\mathrm{rk}(A)$, is the dimension of the vector space spanned by the row vectors of A. This number is always the same as the dimension of the vector space spanned by the column vectors of A.

Theorem H.9. *Given a real $n \times m$ matrix A, $m, n \in \mathbb{N}$, and $b \in \mathbb{R}^n$, let $(A|b)$ be the augmented matrix of the linear system $Ax = b$, and let $(\tilde{A}|\tilde{b})$ be the augmented matrix in echelon form that results from applying the Gaussian elimination algorithm to $(A|b)$ (cf. Def.* H.5 *and Th.* H.6*). Then the following assertions are equivalent:*

(i) *The linear system $Ax = b$ has at least one solution $x \in \mathbb{R}^m$.*

(ii) $\mathrm{rk}(A) = \mathrm{rk}(A|b)$.

(iii) *The last column $\tilde{b}$ in the augmented matrix in echelon form $(\tilde{A}|\tilde{b})$ has no pivot elements (i.e. if there is a zero row in $\tilde{A}$, then the corresponding entry of $\tilde{b}$ also vanishes).*

If $Ax = b$ has at least one solution, then the set of all solutions is an affine space $\mathcal{A}$, which can be represented in the form $\mathcal{A} = v_0 + \ker A$, where v_0 is an arbitrary solution of $Ax = b$ (in other words, one obtains all solutions to the inhomogeneous system $Ax = b$ by adding a particular solution of the inhomogeneous system to the set of all solutions to the homogeneous system $Ax = 0$). The dimension of the solution space is

$$\dim \mathcal{A} = \dim(\ker A) = m - \mathrm{rk}(A). \tag{H.3}$$

$\dim \mathcal{A}$ *also equals the number of free variables occurring in the echelon form $(\tilde{A}|\tilde{b})$ (cf. Def.* H.1*). In particular, if $Ax = b$ has at least one solution, then the solution is unique if, and only if,* $\mathrm{rk}(A) = m$. ■

Remark H.10. If the augmented matrix $(A|b)$ of the linear system (5.1) is in echelon form and the set $\mathcal{A}$ of solutions is nonempty, then one obtains a parameterized representation of $\mathcal{A}$ by the following procedure known as *back substitution*: Starting at the bottom with the first nonzero row, one solves each row for the corresponding pivot variables, in each step substituting the expressions for pivot variables that were obtained in previous steps. The free variables are treated as parameters. A particular element of $\mathcal{A}$ can be obtained by setting all free variables to 0. This is illustrated in the following example.

Example H.11. Consider the linear system $Ax = b$, where

$$A := \begin{pmatrix} 1 & 2 & -1 & 3 & 0 \\ 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad b := \begin{pmatrix} 1 \\ 2 \\ 3 \\ 0 \end{pmatrix}. \tag{H.4}$$

Since $(A|b)$ is in echelon form and b does not have any pivot elements, the set of solutions $\mathcal{A}$ is nonempty, and it can be obtained using back substitution as described in Rem. H.10:

$$x_5 = 3, \tag{H.5a}$$
$$x_3 = 2 + x_5 - x_4 = 5 - x_4, \tag{H.5b}$$
$$x_1 = 1 - 3x_4 + x_3 - 2x_2 = 6 - 4x_4 - 2x_2. \tag{H.5c}$$

Thus,

$$\begin{aligned}\mathcal{A} = \begin{pmatrix}6\\0\\5\\0\\3\end{pmatrix} + \ker A &= \left\{\begin{pmatrix}6-2x_2-4x_4\\x_2\\5-x_4\\x_4\\3\end{pmatrix} : x_2, x_4 \in \mathbb{R}\right\}\\ &= \left\{\begin{pmatrix}6\\0\\5\\0\\3\end{pmatrix} + x_2\begin{pmatrix}-2\\1\\0\\0\\0\end{pmatrix} + x_4\begin{pmatrix}-4\\0\\-1\\1\\0\end{pmatrix} : x_2, x_4 \in \mathbb{R}\right\}.\end{aligned} \tag{H.6}$$

H.2 Orthogonal Matrices

Definition H.12. Let $n \in \mathbb{N}$. A real $n \times n$ matrix Q is called *orthogonal* if, and only if, Q is invertible with $Q^{-1} = Q^{\mathrm{t}}$. The set of all orthogonal $n \times n$ matrices is denoted by $O(n)$.

Theorem H.13. *Let $n \in \mathbb{N}$. For a real $n \times n$ matrix Q, the following statements are equivalent:*

(i) *Q is orthogonal.*

(ii) *Q^{t} is orthogonal.*

(iii) *The columns of Q form an orthonormal basis of $\mathbb{R}^n$.*

(iv) *The rows of Q form an orthonormal basis of $\mathbb{R}^n$.*

(v) *$\langle Qx, Qy\rangle = \langle x, y\rangle$ holds for each $x, y \in \mathbb{R}^n$.*

(vi) *Q is isometric with respect to the Euclidean norm $\|\cdot\|_2$, i.e. $\|Qx\|_2 = \|x\|_2$ for each $x \in \mathbb{R}^n$.*

Proof. "(i)⇔(ii)": $Q^{-1} = Q^{\mathrm{t}}$ is equivalent to $E = QQ^{\mathrm{t}}$, which is equivalent to $(Q^{\mathrm{t}})^{-1} = Q = (Q^{\mathrm{t}})^{\mathrm{t}}$.

"(i)⇔(iii)": $Q^{-1} = Q^{\mathrm{t}}$ implies

$$\left\langle \begin{pmatrix} q_{1i} \\ \vdots \\ q_{ni} \end{pmatrix}, \begin{pmatrix} q_{1j} \\ \vdots \\ q_{nj} \end{pmatrix} \right\rangle = \sum_{k=1}^{n} q_{ki} q_{kj} = \sum_{k=1}^{n} q^{\mathrm{t}}_{ik} q_{kj} = \begin{cases} 0 & \text{for } i \neq j, \\ 1 & \text{for } i = j, \end{cases} \tag{H.7}$$

showing that the columns of Q form an orthonormal system, i.e. n linearly independent vectors according to Lem. F.3(c), i.e. an orthonormal basis of $\mathbb{R}^n$ according to Lem. F.3(d). Conversely, if the columns of Q form an orthonormal basis of $\mathbb{R}^n$, then they satisfy (H.7), which implies $Q^{\mathrm{t}} Q = \mathrm{Id}$.

"(i)⇔(iv)": Since the rows of Q are the columns of Q^{t}, the equivalence of (i) and (iv) is immediate from (iii) and (ii).

"(i)⇒(v)": For each $x, y \in \mathbb{R}^n$: $\langle Qx, Qy \rangle = \langle Q^{\mathrm{t}} Q x, y \rangle = \langle \mathrm{Id}\, x, y \rangle = \langle x, y \rangle$.

"(v)⇒(vi)": For each $x \in \mathbb{R}^n$: $\|Qx\|_2 = \langle Qx, Qx \rangle = \langle x, x \rangle = \|x\|_2$.

"(vi)⇒(i)": The equality $\langle Qx, Qy \rangle = \langle Q^{\mathrm{t}} Q x, y \rangle = \langle \mathrm{Id}\, x, y \rangle = \langle x, y \rangle$ holding for each $x, y \in \mathbb{R}^n$ implies $Q^{\mathrm{t}} Q = \mathrm{Id}$, i.e. Q is orthogonal. ■

Proposition H.14. *Let $n \in \mathbb{N}$.*

- *If $A, B \in O(n)$, then $AB \in O(n)$.*
- *$O(n)$ constitutes a group.*

Proof. (H.14): If $A, B \in O(n)$, then

$$(AB)^{\mathrm{t}} AB = B^{\mathrm{t}} A^{\mathrm{t}} AB = B^{\mathrm{t}}\, \mathrm{Id}\, B = B^{\mathrm{t}} B = \mathrm{Id}, \tag{H.8}$$

showing $AB \in O(n)$.

(H.14): Due to (H.14) and Th. H.13(ii), $O(n)$ is a subgroup of the general linear group $GL(n)$ of all invertible real $n \times n$ matrices. ■

H.3 Positive Definite Matrices

Notation H.15. If A is an $n \times n$ matrix, $n \in \mathbb{N}$, then, for $1 \leq k \leq l \leq n$, let

$$A^{kl} := \begin{pmatrix} a_{kk} & \dots & a_{kl} \\ \vdots & \ddots & \vdots \\ a_{lk} & \dots & a_{ll} \end{pmatrix} \tag{H.9}$$

denote the $(1 + l - k) \times (1 + l - k)$ principal submatrix of A, i.e.

$$\underset{\substack{k,l \in \{1,\dots,n\}, \\ 1 \leq k \leq l \leq n}}{\forall} \quad \underset{i,j \in \{1,\dots,1+l-k\}}{\forall} \quad a^{kl}_{ij} := a_{i+k-1, j+k-1}. \tag{H.10}$$

Proposition H.16. *Let A be a real $n \times n$ matrix, $n \in \mathbb{N}$. Then A is positive definite (cf. Def.* 2.37(c)*) if, and only if, every principal submatrix A^{kl} of A, $1 \leq k \leq l \leq n$, is positive definite.*

Proof. If all principal submatrices of A are positive definite, then, as $A = A^{nn}$, A is positive definite. Now assume A to be positive definite and fix $k, l \in \{1, \dots, n\}$ with $1 \leq k \leq l \leq n$. Let $x = (x_k, \dots, x_l) \in \mathbb{R}^{1+l-k} \setminus \{0\}$ and extend x to $\mathbb{R}^n$ by 0, calling the extended vector y:

$$y = (y_1, \dots, y_n) \in \mathbb{R}^n \setminus \{0\}, \quad y_i = \begin{cases} x_i & \text{for } k \leq i \leq l, \\ 0 & \text{otherwise.} \end{cases} \tag{H.11}$$

We now consider x and y as column vectors and compute

$$x^{\mathrm{t}} A^{kl} x = \sum_{i,j=k}^{l} a_{ij} x_i x_j = \sum_{i,j=1}^{n} a_{ij} y_i y_j > 0, \tag{H.12}$$

showing A^{kl} to be positive definite. ■

Proposition H.17. *Let A be a symmetric positive definite real $n \times n$ matrix, $n \in \mathbb{N}$.*

(a) *All eigenvalues of A are positive real numbers.*

(b) $\det A > 0$.

Proof. (a): Let λ be an eigenvalue of A. According to Th. 2.39, $\lambda \in \mathbb{R}_0^+$ and there is a real eigenvector $x \in \mathbb{R}^n \setminus \{0\}$ for λ. Thus,

$$0 < x^{\mathrm{t}} A x = x^{\mathrm{t}} \lambda x = \lambda \|x\|_2^2. \tag{H.13}$$

As $\|x\|_2^2$ is positive, λ must be positive as well.

(b) follows from (a), as $\det A$ is the product of the eigenvalues of A. ■

H.4 Cholesky Decomposition for Positive Semidefinite Matrices

In this section, we show that the results of Sec. 5.3 (except uniqueness of the decomposition) extend to symmetric positive *semi*definite matrices.

Theorem H.18. *Let Σ be a symmetric positive semidefinite real $n \times n$ matrix, $n \in \mathbb{N}$. Then there exists a lower triangular real $n \times n$ matrix A with nonnegative diagonal entries (i.e. with $a_{jj} \geq 0$ for each $j \in \{1, \dots, n\}$) and*

$$\Sigma = AA^{\mathrm{t}}. \tag{H.14}$$

In extension of the term for the positive definite case, we still call this decomposition a Cholesky decomposition *of Σ. It is recalled from Th. 5.19 that, if Σ 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. 5.19. The general (positive semidefinite) case can be deduced from the positive definite case as follows: If Σ is symmetric positive semidefinite, then each $\Sigma_k := \Sigma + \frac{1}{k}\,\mathrm{Id}$, $k \in \mathbb{N}$, is positive definite:

$$x^{\mathrm{t}}\Sigma_k x = x^{\mathrm{t}}\Sigma x + \frac{x^{\mathrm{t}}x}{k} > 0 \quad \text{for each } 0 \neq x \in \mathbb{R}^n. \tag{H.15}$$

Thus, for each $k \in \mathbb{N}$, there exists a lower triangular matrix A_k with positive diagonal entries such that $\Sigma_k = A_k A_k^{\mathrm{t}}$.

On the other hand, Σ_k converges to Σ with respect to each norm on $\mathbb{R}^{n^2}$ (since all norms on $\mathbb{R}^{n^2}$ are equivalent). In particular, $\lim_{k\to\infty} \|\Sigma_k - \Sigma\|_2 = 0$. Thus,

$$\|A_k\|_2^2 = r(\Sigma_k) = \|\Sigma_k\|_2, \tag{H.16}$$

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 $\mathbb{R}^{n^2}$ 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 $A \in \mathbb{R}^{n^2}$. As this convergence is with respect to the norm topology on $\mathbb{R}^{n^2}$, each entry of the A_{k_l} must converge (in $\mathbb{R}$) to the respective entry of A. In particular, A is lower triangular with all nonnegative diagonal entries. It only remains to show $AA^{\mathrm{t}} = \Sigma$. However,

$$\Sigma = \lim_{l\to\infty} \Sigma_{k_l} = \lim_{l\to\infty} A_{k_l} A_{k_l}^{\mathrm{t}} = AA^{\mathrm{t}}, \tag{H.17}$$

which establishes the case. ■

Example H.19. The decomposition

$$\begin{pmatrix} 0 & 0 \\ \sin x & \cos x \end{pmatrix} \begin{pmatrix} 0 & \sin x \\ 0 & \cos x \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{for each } x \in \mathbb{R} \tag{H.18}$$

shows a symmetric positive semidefinite matrix (which is not positive definite) can have uncountably many different Cholesky decompositions.

Theorem H.20. *Let Σ be a symmetric positive semidefinite real $n \times n$ matrix, $n \in \mathbb{N}$. Define the index set*

$$I := \big\{(i,j) \in \{1,\dots,n\}^2 : j \leq i\big\}. \tag{H.19}$$

Then a matrix $A = (A_{ij})$ providing a Cholesky decomposition $\Sigma = AA^{\mathrm{t}}$ of $\Sigma = (\sigma_{ij})$ is obtained via the following algorithm, defined recursively over I, using the order $(1,1) < (2,1) < \dots < (n,1) < (2,2) < \dots < (n,2) < \dots < (n,n)$ (which corresponds to traversing the lower half of Σ by columns from left to right):

$$A_{11} := \sqrt{\sigma_{11}}. \tag{H.20a}$$

For $(i,j) \in I \setminus \{(1,1)\}$:

$$A_{ij} := \begin{cases} \left(\sigma_{ij} - \sum_{k=1}^{j-1} A_{ik}A_{jk}\right)/A_{jj} & \text{for } i > j \text{ and } A_{jj} \neq 0, \\ 0 & \text{for } i > j \text{ and } A_{jj} = 0, \\ \sqrt{\sigma_{ii} - \sum_{k=1}^{i-1} A_{ik}^2} & \text{for } i = j. \end{cases} \tag{H.20b}$$

Proof. A lower triangular $n \times n$ matrix A provides a Cholesky decomposition of Σ if, and only if,

$$AA^{\mathrm{t}} = \begin{pmatrix} A_{11} & & & \\ A_{21} & A_{22} & & \\ \vdots & \vdots & \ddots & \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{pmatrix} \begin{pmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ & A_{22} & \dots & A_{n2} \\ & & \ddots & \vdots \\ & & & A_{nn} \end{pmatrix} = \Sigma, \tag{H.21}$$

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:

$$\sum_{k=1}^{j} A_{ik}A_{jk} = \sigma_{ij}, \quad (i,j) \in I. \tag{H.22a}$$

Using the order on I introduced in the statement of the theorem, (H.22a) takes the form

$$\begin{aligned} A_{11}^2 &= \sigma_{11}, \\ A_{21}A_{11} &= \sigma_{21}, \\ &\ \vdots \\ A_{n1}A_{11} &= \sigma_{n1}, \\ A_{21}^2 + A_{22}^2 &= \sigma_{22}, \\ A_{31}A_{21} + A_{32}A_{22} &= \sigma_{32}, \\ &\ \vdots \\ A_{n1}^2 + \dots + A_{nn}^2 &= \sigma_{nn}. \end{aligned} \tag{H.22b}$$

From Th. H.18, we know (H.22) must have at least one solution with $A_{jj} \geq 0$ for each $j \in \{1,\dots,n\}$. In particular, $\sigma_{jj} \geq 0$ for each $j \in \{1,\dots,n\}$ (this is also immediate from Σ being positive semidefinite, since $\sigma_{jj} = e_j^{\mathrm{t}}\,\Sigma e_j$, where e_j denotes the jth standard unit vector of $\mathbb{R}^n$). We need to show that (H.20) yields a solution to (H.22). The proof is carried out by induction on n. For $n = 1$, we have $\sigma_{11} \geq 0$ and $A_{11} = \sqrt{\sigma_{11}}$, i.e. there is nothing to prove. Now let $n > 1$. If $\sigma_{11} > 0$, then

$$A_{11} = \sqrt{\sigma_{11}}, \quad A_{21} = \sigma_{21}/A_{11}, \quad \dots, \quad A_{n1} = \sigma_{n1}/A_{11} \tag{H.23a}$$

is the unique solution to the first n equations of (H.22b) satisfying $A_{11} > 0$, and this solution is provided by (H.20). If $\sigma_{11} = 0$, then $\sigma_{21} = \dots = \sigma_{n1} = 0$: Otherwise, let $s := (\sigma_{21}\ \dots\ \sigma_{n1})^{\mathrm{t}} \in \mathbb{R}^{n-1} \setminus \{0\}$, $\alpha \in \mathbb{R}$, and note

$$(\alpha,\, s^{\mathrm{t}}) \begin{pmatrix} 0 & s^{\mathrm{t}} \\ s & \Sigma_{n-1} \end{pmatrix} \begin{pmatrix} \alpha \\ s \end{pmatrix} = (\alpha,\, s^{\mathrm{t}}) \begin{pmatrix} s^{\mathrm{t}}s \\ \alpha s + \Sigma_{n-1}s \end{pmatrix} = 2\alpha\|s\|^2 + s^{\mathrm{t}}\Sigma_{n-1}s < 0$$

for $\alpha < -s^{\mathrm{t}}\Sigma_{n-1}s/(2\|s\|^2)$, in contradiction to Σ being positive semidefinite. Thus,

$$A_{11} = A_{21} = \dots = A_{n1} = 0 \tag{H.23b}$$

is a particular solution to the first n equations of (H.22b), and this solution is provided by (H.20). We will now denote the solution to (H.22) given by Th. H.18 by $B_{11}, \dots, B_{nn}$ to distinguish it from the A_{ij} constructed via (H.20).

In each case, $A_{11}, \dots, A_{n1}$ are given by (H.23), and, for $(i,j) \in I$ with $i, j \geq 2$, we define

$$\tau_{ij} := \sigma_{ij} - A_{i1}A_{j1} \quad \text{for each } (i,j) \in J := \big\{(i,j) \in I : \, i,j \geq 2\big\}. \tag{H.24}$$

To be able to proceed by induction, we show that the symmetric $(n-1)\times(n-1)$ matrix

$$T := \begin{pmatrix} \tau_{22} & \dots & \tau_{n2} \\ \vdots & \ddots & \vdots \\ \tau_{n2} & \dots & \tau_{nn} \end{pmatrix} \tag{H.25}$$

is positive semidefinite. If $\sigma_{11} = 0$, then (H.23b) implies $\tau_{ij} = \sigma_{ij}$ for each $(i,j) \in J$ and T is positive semidefinite, as Σ being positive semidefinite implies

$$\begin{pmatrix} x_2 & \dots & x_n \end{pmatrix} T \begin{pmatrix} x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 & x_2 & \dots & x_n \end{pmatrix} \Sigma \begin{pmatrix} 0 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \geq 0 \tag{H.26}$$

for each $(x_2, \dots, x_n) \in \mathbb{R}^{n-1}$. If $\sigma_{11} > 0$, then (H.23a) holds as well as $B_{11} = A_{11}, \dots,$ $B_{n1} = A_{n1}$. Thus, (H.24) implies

$$\tau_{ij} := \sigma_{ij} - A_{i1}A_{j1} = \sigma_{ij} - B_{i1}B_{j1} \quad \text{for each } (i,j) \in J. \tag{H.27}$$

Then (H.22) with A replaced by B together with (H.27) implies

$$\sum_{k=2}^{j} B_{ik}B_{jk} = \tau_{ij} \quad \text{for each } (i,j) \in J \tag{H.28}$$

or, written in matrix form,

$$BB^{\mathrm{t}} = \begin{pmatrix} B_{22} & & \\ \vdots & \ddots & \\ B_{n2} & \dots & B_{nn} \end{pmatrix} \begin{pmatrix} B_{22} & \dots & B_{n2} \\ & \ddots & \vdots \\ & & B_{nn} \end{pmatrix} = \begin{pmatrix} \tau_{22} & \dots & \tau_{n2} \\ \vdots & \ddots & \vdots \\ \tau_{n2} & \dots & \tau_{nn} \end{pmatrix} = T, \tag{H.29}$$

which, once again, establishes T to be positive semidefinite (since, for each $x \in \mathbb{R}^{n-1}$, $x^{\mathrm{t}}Tx = x^{\mathrm{t}}BB^{\mathrm{t}}x = (B^{\mathrm{t}}x)^{\mathrm{t}}(B^{\mathrm{t}}x) \geq 0$).

By induction, we now know the algorithm of (H.20) yields a (possibly different from (H.29) for $\sigma_{11} > 0$) decomposition of T:

$$CC^{\mathrm{t}} = \begin{pmatrix} C_{22} & & \\ \vdots & \ddots & \\ C_{n2} & \dots & C_{nn} \end{pmatrix} \begin{pmatrix} C_{22} & \dots & C_{n2} \\ & \ddots & \vdots \\ & & C_{nn} \end{pmatrix} = \begin{pmatrix} \tau_{22} & \dots & \tau_{n2} \\ \vdots & \ddots & \vdots \\ \tau_{n2} & \dots & \tau_{nn} \end{pmatrix} = T \tag{H.30}$$

or

$$\sum_{k=2}^{j} C_{ik}C_{jk} = \tau_{ij} = \sigma_{ij} - A_{i1}A_{j1} \quad \text{for each } (i,j) \in J, \tag{H.31}$$

where

$$C_{22} := \sqrt{\tau_{22}} = \begin{cases} \sqrt{\sigma_{22}} & \text{for } \sigma_{11} = 0, \\ B_{22} & \text{for } \sigma_{11} > 0, \end{cases} \tag{H.32a}$$

and, for $(i,j) \in J \setminus \{(2,2)\}$,

$$C_{ij} := \begin{cases} \left(\tau_{ij} - \sum_{k=2}^{j-1} C_{ik}C_{jk}\right)/C_{jj} & \text{for } i > j \text{ and } C_{jj} \neq 0, \\ 0 & \text{for } i > j \text{ and } C_{jj} = 0, \\ \sqrt{\tau_{ii} - \sum_{k=2}^{i-1} C_{ik}^2} & \text{for } i = j. \end{cases} \tag{H.32b}$$

Substituting $\tau_{ij} = \sigma_{ij} - A_{i1}A_{j1}$ from (H.24) into (H.32) and comparing with (H.20), 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 (H.20) (i.e. all occurring square roots exist as real numbers). It also follows that $\{A_{ij} : (i,j) \in I\}$ is a solution to (H.22): The first n equations are satisfied according to (H.23); the remaining $(n-1)\,n/2$ equations are satisfied according to (H.31) combined with $C_{ij} = A_{ij}$. This concludes the proof that (H.20) furnishes a solution to (H.22). ■

References

[Alt06] Hans Wilhelm Alt. *Lineare Funktionalanalysis*, 5th ed. Springer-Verlag, Berlin, 2006 (German).

[Phi13] P. Philip. *Calculus I for Computer Science and Statistics Students*. Lecture Notes, Ludwig-Maximilians-Universität, Germany, 2012/2013, available in PDF format at http://www.math.lmu.de/~philip/publications/lectureNotes/calc1_forInfAndStatStudents.pdf.

[Phi14] P. Philip. *Calculus II for Statistics Students*. Lecture Notes, Ludwig-Maximilians-Universität, Germany, 2014, available in PDF format at http://www.math.lmu.de/~philip/publications/lectureNotes/calc2_forStatStudents.pdf.

[RF10] Halsey Royden and Patrick Fitzpatrick. *Real Analysis*, 4th ed. Pearson Education, Boston, USA, 2010.

[Yos80] K. Yosida. *Functional Analysis*, 6th ed. Grundlehren der mathematischen Wissenschaften, Vol. 123, Springer-Verlag, Berlin, 1980.