Otmar Scherzer

CO-MAT1

Vorlesungsskriptum WS 2014/15

Computational Science Center
Universität Wien
Oskar-Morgenstern-Platz 1
1090 Wien

Contents

The reference list to numerical analysis is huge. References, which have been used to prepare this manuscript are [11, 13, 12, 8, 3, 1, 5, 6] . We have used in particular [6, 9, 2].

Chapter 1

Basic Concepts of Numerical Analysis

1.1 Vector- and Matrix Norms

We denote by $\mathbb{R}^n$ the space of n-dimensional real vectors, with elements

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad x_i \in \mathbb{R} .$$

$\mathbb{R}^{m \times n}$ denotes the space of $m \times n$ matrices, with elements

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \quad a_{ij} \in \mathbb{R} .$$

Each vector can be written as a matrix

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n \times 1} .$$

For $x \in \mathbb{R}^n$, we denote the *transpose vector* by

$$x^T = [x_1, x_2, ..., x_n] \in \mathbb{R}^{1 \times n} .$$

For matrices A, the definition of the transpose matrix A^T is accordingly.

The set $\{e_1, ..., e_n\} \subset \mathbb{R}^n$ of vectors with components $e_i = [\delta_{ij}]_{j=1}^n$ is called the *cartesian* basis. Thereby

$$\delta_{ij} = \begin{cases} 1, & i = j\,, \\ 0, & i \neq j\,, \end{cases}$$

denotes the *Kronecker* symbol.

There exist several different norms for matrices:

Definition 1.1. *A function $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ is called norm, if*

1. $\|x\| \geq 0$, $\forall x \in \mathbb{R}^{m\times n}$ *and* $\|x\| = 0$ *if and only if* $x = 0$.
2. $\|\lambda x\| = |\lambda| \, \|x\|$, $\forall x \in \mathbb{R}^{m\times n}$ *and* $\lambda \in \mathbb{R}$.
3. $\|x + y\| \leq \|x\| + \|y\|$, $\forall x, y \in \mathbb{R}^{m\times n}$.

Example 1.2. *Some common vector norms on $\mathbb{R}^n$ are*

1. $\|x\|_1 := \sum_{i=1}^n |x_i|$,
2. $\|x\|_2 := \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{x^T x}$,
3. $\|x\|_\infty := \max_{i=1,...,n} |x_i|$.

Some common matrix norms on $\mathbb{R}^{m\times n}$ are

1. $\|A\|_{1,\infty} := \max_{1\leq j\leq n} \sum_{i=1}^m |a_{i,j}|$.
2. $\|A\|_{\infty,1} := \max_{1\leq i\leq m} \sum_{j=1}^n |a_{i,j}|$.
3. Frobeniusnorm: $\|A\|_F := \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{i,j}^2}$.

The most important norm however is the spectral norm. For defining this one, we need *eigenvalues*:

Definition 1.3. *Let $C \in \mathbb{R}^{n\times n}$. Then λ is called an eigenvalue of C if there exists a vector $x \in \mathbb{R}^n \setminus \{0\}$ such that*

$$Cx = \lambda x\ .$$

The according x is called eigenvector*. Typically, an eigenvector x with norm* 1 *is called the* normalised eigenvector*.*

We denote by $\sigma(C)$ the set of all eigenvalues of the matrix C. That is,

$$\sigma(C) = \{\lambda : \exists\, x \in \mathbb{R}^n \setminus \{0\} \ \textit{such that}\ Cx = \lambda x\}\ .$$

Remark 1.4. *$\lambda \in \sigma(C)$ if and only if* $\det(\lambda I - C) = 0$. *This relation is obvious, when we know that* $\det(\lambda I - C) = 0$ *is equaivalent to the fact that $\lambda I - C$ is singular - or in other word has a non-trivial nullspace* $\{x : (\lambda I - C)x = 0\} \neq \{0\}$. *This nullspace contains all eigenvectors.*

Definition 1.5. *Then the* spectral radius *$\rho(C)$ is the largest absolute values of the eigenvalues, that is*

$$\boxed{\rho(C) := \max\{|\lambda| : \lambda \in \sigma(C)\} \ .} \tag{1.1}$$

Example 1.6. *The eigenvalues of matrices can be complex, although the coefficients of C are real. Let*

$$C = \begin{bmatrix} 2 & 5 \\ -1 & -2 \end{bmatrix}$$

Then the zeros of the characteristic polynomial are $\pm\mathrm{i}$. *Thus the spectral radius of C is* 1.

Definition 1.7. *For every matrix $A \in \mathbb{R}^{m\times n}$ we define* $\|A\|_2 := \sqrt{\rho(A^T A)}$.

Remark 1.8. *If C is a symmetric matrix, then all eigenvalues are real. In particular we have for $C = A^T A$ that all eigenvalues are real. Moreover, in the later case all eigenvalues are non-negative.*

Theorem 1.9. *Let $C \in \mathbb{R}^{m\times n}$, then there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ such that*

$$\boxed{U^T C V = S,}$$

where $S \in \mathbb{R}^{m\times n}$ is a rectangular diagonal matrix with non-negative diagonal elements $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min\{m,n\}} \geq 0$. In particular, if $m > n$, then

$$S = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n \\ 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix}$$

else if $m < n$,

$$S = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & & \vdots \\ 0 & \cdots & 0 & \sigma_m & 0 & \cdots & 0 \end{bmatrix}.$$

A matrix is called orthogonal if $U^T = U^{-1}$.

Remark 1.10. *If there exists a singular value decomposition of* $C = A^T A$, *then*

$$Av_i = \sigma_i u_i\,, \qquad A^T u_i = \sigma_i v_i\,, \qquad \forall i = 1 : \min\{m, n\}\,.$$

Here u_i *and* v_i *are the column vectors of* U *and* V, *respectively. Moreover,* v_i *are the eigenvectors of* $A^T A$ *with eigenvalues* $\lambda_i = \sigma_i^2$, *and* u_i *are the eigenvectors of* AA^T *with eigenvalues* $\lambda_i = \sigma_i^2$, *respectively. That is,*

$$A^T A v_i = \sigma_i^2 v_i \text{ and } AA^T u_i = \sigma_i^2 u_i\,, \quad \forall i = 1, \ldots, \min\{m, n\}\ .$$

Theorem 1.11. *Let* $C \in \mathbb{R}^{m\times m}$ *be symmetric, then there exist an orthogonal matrix* $U \in \mathbb{R}^{m\times m}$ *such that*

$$U^T C U = Diag(\lambda_1, \ldots, \lambda_m)\,,$$

That is, the singular values equal the eigenvalues.

Example 1.12. *The spectral norm should not be confused with the Frobenius norm: It is a standard result from linear algebra that*

$$\|A\|_F^2 = \sum_{\lambda\in\sigma(A^T A)} \lambda = \left\|\sigma(A^T A)\right\|_1\ .$$

On the other hand

$$\|A\|_2^2 = \sup_{\lambda\in\sigma(A^T A)} \lambda = \left\|\sigma(A^T A)\right\|_\infty\ .$$

Remark 1.13. *We review the confusing terminology:*
Let $C \in \mathbb{R}^{n\times n}$ *be a matrix.*

- *An* eigenvalue λ *is a solution of the equation* $Cx = \lambda x$ *for some* $x \neq 0$.
- *A* spectral value σ *is a solution of the equation* $C^T Cx = \sigma^2 x$ *for some* $x \neq 0$.
- *If* C *is symmetric, then* $C = A^T A$ *and* $\lambda \in \sigma(C)$ *if and only* $\sqrt{\lambda} \in \sigma(A)$.

1.2 Well- and Ill–Conditioned Systems of Linear Equations

We consider the difference between ill-conditioned and well-conditioned system of equations. A system of equations is considered to be well-conditioned if a small change in the coefficient matrix or a small change in the right hand side results in a small change in the solution vector. A system of equations is considered to be ill-conditioned if a small change in the coefficient matrix or a small change in the right hand side results in a large change in the solution vector.

Example 1.14. *We consider*

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}.$$

The solution of this equation is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

Now we consider the system with a perturbed right hand side

$$\begin{bmatrix} 1 & 2 \\ 2 & 3.999 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.998 \end{bmatrix}.$$

Now, the solution is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -3.999 \\ 4.000 \end{bmatrix}.$$

If we make a small change of the coefficients of A we observe that

$$\begin{bmatrix} 1.001 & 2.001 \\ 2.001 & 3.998 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}.$$

Here the solution is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3.994 \\ 0.001388 \end{bmatrix}.$$

We say that the system is ill-conditioned because small variations in the coefficient matrix or the right hand side result in a large change of the solution.

Example 1.15. *We consider*

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix} .$$

The solution of this equation is again

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} .$$

Now we consider the system with a perturbed right hand side

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.001 \end{bmatrix} .$$

Now, the solution is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1.999 \\ 1.001 \end{bmatrix} .$$

If we make a small change of the coefficients of A we observe that

$$\begin{bmatrix} 1.001 & 2.001 \\ 2.001 & 3.001 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix} .$$

Here the solution is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2.003 \\ 0.997 \end{bmatrix} .$$

We say that the system is well-conditioned because small variations in the coefficient matrix or the right hand side result in a small change in the solution.

To differ between ill– and well–posed equations one uses the condition number:

$$\boxed{\mathrm{Cond}(A) = \left\|A^{-1}\right\|_2 \|A\|_2 \ .}$$

The reason can be the following derivation

$$\begin{aligned}\frac{\|\Delta x\|_2}{\|x\|_2} &= \frac{\|A^{-1}\Delta b\|_2}{\|b\|_2}\frac{\|b\|_2}{\|x\|_2}\\ &\leq \frac{\|A^{-1}\|_2\,\|\Delta b\|_2}{\|b\|_2}\frac{\|Ax\|_2}{\|x\|_2}\\ &\leq \frac{\|A^{-1}\|_2\,\|\Delta b\|_2}{\|b\|_2}\frac{\|A\|_2\,\|x\|_2}{\|x\|_2}\\ &\leq \left\|A^{-1}\right\|_2\|A\|_2\frac{\|\Delta b\|_2}{\|b\|_2}\\ &= \mathrm{Cond}(A)\frac{\|\Delta b\|_2}{\|b\|_2}\ .\end{aligned}$$

That is, the condition number gives a qualitative figure of the error magnification of the relative error. The absolute error can be estimated as follows:

$$\|\Delta x\|_2 = \left\|A^{-1}\Delta b\right\|_2 \leq \left\|A^{-1}\right\|_2\|\Delta b\|_2\ ,$$

Therefore the error enhancement can be estimated to be of the order $\|A^{-1}\|_2$.

Example 1.16. • *We consider the matrix A from Example 1.14. The square-roots of the eigenvalues of A^TA are*

$$\begin{array}{c}0.000200032003451\\ 4.999200032003841\end{array}$$

The square roots of the eigenvalues of $(A^TA)^{-1}$ are

$$\begin{array}{c}1/0.000200032003451\\ 1/4.999200032003841\end{array}$$

and thus

- $\|A^{-1}\|_2 = 1/0.000200032003451 \sim 4999$. *$\|A^{-1}\|_2$ predicts therefore an error magnification of* 4999. *The actual error in the right hand side is* $0.0014 = \sqrt{2\cdot 0.001^2}$ *while the error in the solution is* $6.7073 = \sqrt{5.999^2+3^2}$ *Thus the absolute error is magnified by a factor* 4743. *Thus the actual error it overestimates the error by a factor* 1.05.

  - *The condition number is* 24992. *The relative error in the right hand side is* $0.00016 \sim 0.0014/\sqrt{4^2 + 7.999^2}$ *The relative error of the solution is* $2.9996 = 6.7073/\sqrt{2^2 + 1^2}$. *Thus the relative error magnification is* $18748 = 2.9996/0.00016$. *Here the condition number overestimates the error only by a factor* $1.3 \sim 24992/18748$.

- *For the matrix* A *from Example 1.15 we have the roots of the eigenvalues of* $A^T A$ *are*

$$\begin{array}{c} 0.236067977499791 \\ 4.236067977499790 \end{array}$$

  and the square roots of the eigenvalues of $(A^T A)^{-1}$ *are*

$$\begin{array}{c} 1/0.236067977499791 \\ 1/4.236067977499790 \end{array} .$$

  and thus

  - $\|A^{-1}\|_2 = 1/0.236067977499791 \sim 4.24$. $\|A^{-1}\|_2$ *predicts therefore an error magnification of* 4.24. *The errors of the right hand side and of the solution are both* $0.0014 = \sqrt{2 \cdot 0.001^2}$. *Thus the absolute error is not magnified at all.*
  - *Thus the condition number is* 17.94. *The relative error in the right hand side is* $0.00017365 \sim 0.0014/\sqrt{4^2 + 7^2}$. *The relative error of the solution is* $0.00063 = 0.0014/\sqrt{2^2 + 1^2}$. *Thus the relative error magnification is* $3.6280 = 0.00063/0.00017365$. *In this case the relative error is overestimated by a factor* 4.96.

- *The condition number of the first matrix is of a factor* 1388 *larger than the condition of the second matrix. The first matrix could be called ill-conditioned, while the second one is well–conditioned.*

1.3 Nonlinear Problems

For nonlinear problems the error analysis is much more involved. We shortly sketch it here. In principal the problem is again reduced to linear problems.

Let $F : \mathbb{R}^n \to \mathbb{R}$ by a function.

Problem 1.17. *We want to evaluate F at a vector $x \in \mathbb{R}^n$. We assume that some measurement, accumulated, and data errors force use to evaluate F at $x+\Delta x$. How can the error in the evaluation, that is $\Delta y := F(x+\Delta x)-F(x)$, be estimated?*

From the mean value theorem it follows

$$\begin{aligned}\Delta y &= F(x_1 + \Delta x_1, x_2 + \Delta x_2, ..., x_n + \Delta x_n) - F(x_1, x_2, ..., x_n) \\ &= \sum_{i=1}^{n} \frac{\partial F}{\partial x_i}(\zeta)\Delta x_i \ ,\end{aligned}$$

where ζ is a point on the line in between x and $x + \Delta x$.

Since Δx is considered small (errors should always stay small), it seems plausible to use the approximation:

$$\Delta y \approx \sum_{i=1}^{n} \frac{\partial F}{\partial x_i}(x)\Delta x_i \ .$$

The number

$$\mathcal{K}_{\text{abs}} := \left\| \left[\frac{\partial F}{\partial x_i}(x) \right]_{i=1,...,n} \right\|_2 = \|\nabla F(x)\|_2$$

is called *absolute condition number* of F at x and is considered a measure for the enhancement of the absolute error.

If we have a vector valued function $F : \mathbb{R}^n \to \mathbb{R}^n$, then we generalize this concept to

$$\mathcal{K}_{\text{abs}} := \left\| \left[\frac{\partial F_j}{\partial x_i}(x) \right]_{i,j=1,...,n} \right\|_2 = \|\nabla F(x)\|_2 \ .$$

Example 1.18. *If F is a linear function, that is $F(x) = Ax$ with some matrix $[a_{ij}]_{i,j=1,...,n} = A \in \mathbb{R}^{n\times n}$, then*

$$\mathcal{K}_{abs} = \|A\|_2 \ .$$

For a function $F : \mathbb{R}^n \to \mathbb{R}^n$ the relative error is estimated as follows: Let $A = \left[\frac{\partial F_j}{\partial x_i}(x) \right]_{i,j=1,...,n}$, then the relative condition number of F at a point x is $\text{Cond}(A)$.

1.4 Order

Let $\{a_\varepsilon : \varepsilon \neq 0\}$ and $\{b_\varepsilon : \varepsilon \neq 0\}$ be parametrized families of numbers, which are all different from zero, that is $a_\varepsilon, b_\varepsilon \neq 0$.

- We use the notation $a_\varepsilon = \mathcal{O}(b_\varepsilon)$, if there exist a constant $\varepsilon_0 > 0$ and $C > 0$ such that

$$|a_\varepsilon| \leq C\,|b_\varepsilon| \ , \qquad \forall \varepsilon \in (0, \varepsilon_0) \ .$$

- We say $a_\varepsilon = o(b_\varepsilon)$, if

$$\lim_{\varepsilon \to 0} \frac{|a_\varepsilon|}{|b_\varepsilon|} = 0 \ .$$

Example 1.19. *Take* $\{b_x = x : x \neq 0\}$ *and* $\{a_x = \sin(x) : x \neq 0\}$, *then* $\lim_{x\to 0} \frac{\sin x}{x} = 1$. *Thus*

$$\sin(x) = \mathcal{O}(x) \ .$$

Take $\left\{b_x = \sqrt{|x|} : x \neq 0\right\}$ *and* $\{a_x = \sin(x) : x \neq 0\}$, *then* $\lim_{x\to 0} \frac{\sin x}{x} \frac{x}{\sqrt{x}} = \lim_{x\to 0} \frac{\sin x}{x} \lim_{x\to 0} \sqrt{x} = 0$. *Thus*

$$\sin(x) = o(\sqrt{|x|}) \ .$$

Chapter 2

Elimination Algorithms

2.1 Gauss-Elimination

We describe this algorithm first exemplary and then we do it formally.

We consider the linear equation

$$\begin{aligned} 4x_1 + 1x_2 + 0x_3 + 0x_4 &= 3 \\ 1x_1 + 4x_2 + 1x_3 + 0x_4 &= 3 \\ 0x_1 + 1x_2 + 4x_3 + 1x_4 &= 0 \\ 0x_1 + 0x_2 + 1x_3 + 4x_4 &= 2 \end{aligned} \tag{2.1}$$

In the first step we eliminate the entry 1 in front of x_1 in the second line by multiplication of the first line with $-1/4$ and adding it to the second line. This gives the matrix equation

$$\begin{aligned} 4x_1 + x_2 + 0x_3 + 0x_4 &= 3 \\ 0x_1 + \frac{15}{4}x_2 + 1x_3 + 0x_4 &= \frac{9}{4} \\ 0x_1 + 1x_2 + 4x_3 + 1x_4 &= 0 \\ 0x_1 + 0x_2 + 1x_3 + 4x_4 &= 2 \end{aligned} \tag{2.2}$$

To eliminate the factor 1 in front of x_2 in the third line we multiply the

second line by $-\frac{4}{15}$ and add it to the third line. We get

$$\begin{aligned} 4x_1 + 1x_2 + 0x_3 + 0x_4 &= 3 \\ 0x_1 + \frac{15}{4}x_2 + 1x_3 + 0x_4 &= \frac{9}{4} \\ 0x_1 + 0x_2 + \frac{56}{15}x_3 + x_4 &= -\frac{9}{15} \\ 0x_1 + 0x_2 + x_3 + 4x_4 &= 2 \end{aligned}$$

To eliminate the factor 1 in front of x_3 in the fourth line we multiply the third line by $-\frac{15}{56}$ and add it to the fourth line. We get

$$\begin{aligned} 4x_1 + x_2 + 0x_3 + 0x_4 &= 3 \\ 0x_1 + \frac{15}{4}x_2 + x_3 + 0x_4 &= \frac{9}{4} \\ 0x_1 + 0x_2 + \frac{56}{15}x_3 + x_4 &= -\frac{9}{15} \\ 0x_1 + 0x_2 + 0x_3 + \frac{209}{56}x_4 &= \frac{121}{56} \end{aligned}$$

From this we get a solution by first resolving the last line for $x_4 = \frac{121}{209}$. Then we get $x_3 = -(\frac{9}{15} + \frac{121}{209})\frac{15}{56}$ and so on.

Now, we do it in matrix notation. The equation (2.1) can be written in matrix notation:

$$\begin{bmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 0 \\ 2 \end{bmatrix} .$$

Each of the elimination steps can be implemented by multiplying a matrix from the left. In the first step we use the matrix

$$L^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Then the equation

$$\begin{bmatrix} 4 & 1 & 0 & 0 \\ 0 & \frac{15}{4} & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix} x = L^{(1)}Ax = L^{(1)}b = \begin{bmatrix} 3 \\ \frac{9}{4} \\ 0 \\ 2 \end{bmatrix}$$

is equivalent to (2.2). The second step in matrix form requires

$$L^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -\frac{4}{15} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Then we get the matrix equation

$$\begin{bmatrix} 4 & 1 & 0 & 0 \\ 0 & \frac{15}{4} & 1 & 0 \\ 0 & 0 & \frac{56}{15} & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix} x = L^{(2)}L^{(1)}Ax = L^{(2)}L^{(1)}b = \begin{bmatrix} 3 \\ \frac{9}{4} \\ -\frac{9}{15} \\ 2 \end{bmatrix}$$

In the last step we use the matrix

$$L^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{15}{56} & 1 \end{bmatrix}$$

and get the matrix equation

$$\begin{bmatrix} 4 & 1 & 0 & 0 \\ 0 & \frac{15}{4} & 1 & 0 \\ 0 & 0 & \frac{56}{15} & 1 \\ 0 & 0 & 0 & \frac{209}{56} \end{bmatrix} x = L^{(3)}L^{(2)}L^{(1)}Ax = L^{(3)}L^{(2)}L^{(1)}b = \begin{bmatrix} 3 \\ \frac{9}{4} \\ -\frac{9}{15} \\ \frac{121}{56} \end{bmatrix}$$

2.2 Roadmap

The Gauss-eliminiation described above, a–priori, is only suitable for solving linear equations. Now, we show that all information is already available to derive a decomposition of the form

$$A = LR\,,$$

where L is a lower triangular matrix, and R is an upper triangular matrix, without additional complexity. For this purpose we formulate the Gauss-eliminiation steps in matrix notation. We show that each step of Gauss-eliminiation can be modeled by a matrix, which can be easily inverted. These inverse matrices are used to derive the decomposition.

2.3 Gauss-Eliminiation in Matrix Notation

In the following we prescribe the elimination steps in a formal way, independent of the dimension n:

Let $x = [x_1, ..., x_n]^T \in \mathbb{R}^n$ such that $x_k \neq 0$. Then we define the matrix

$$L^{(k)} = I - l_k e_k^T \,, \tag{2.3}$$

where $l_k = [0, ..., 0, l_{k+1,k}, ..., l_{n,k}]^T \in \mathbb{R}^n$ with $l_{jk} = x_j/x_k$, $j = k+1, ..., n$. Then we have

$$L^{(k)}x = \begin{bmatrix} 1 & 0 & & & \cdots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ & \ddots & 1 & 0 & & \\ & & -l_{k+1,k} & 1 & \ddots & \\ \vdots & & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & -l_{n,k} & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ x_{k+1} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ 0 \\ \vdots \\ 0 \end{bmatrix} . \tag{2.4}$$

Denote $A = A^{(1)} = [a_{ij}]_{ij}$ the $n \times n$–Matrix and $x = [a_{i1}]_i \in \mathbb{R}^n$ the first column of $A^{(1)}$. If $a_{11} \neq 0$, then the matrix $L^{(1)} = I - l_1 e_1^T$ results in

$$\begin{aligned} L^{(1)}A^{(1)} &= \begin{bmatrix} 1 & 0 & \cdots & & 0 \\ -l_{21} & 1 & 0 & \cdots & 0 \\ -l_{31} & 0 & 1 & & 0 \\ \vdots & & \ddots & \ddots & 0 \\ -l_{n1} & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & a_{32}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix} =: A^{(2)} \,. \end{aligned} \tag{2.5}$$

This is the first step of the Gauss-elimination. If $a_{22}^{(2)} \neq 0$, then we select in the second step ($k = 2$) for x the second column of $A^{(2)}$, that is $x = [a_{12}, a_{22}^{(2)}, ..., a_{n2}^{(2)}]^T$. The according matrix $L^{(2)} = I - l_2 e_2^T$ then satisfies $A^{(3)} =$

$L^{(2)}A^{(2)}$, where,

$$L^{(2)} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & -l_{32} & \ddots & \cdots & \cdots & 0 \\ 0 & -l_{42} & 0 & 1 & 0 & \cdots \\ \vdots & \vdots & & \ddots & \ddots & 0 \\ 0 & -l_{n2} & & \cdots & \cdots & 1 \end{bmatrix} \quad A^{(3)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3n}^{(3)} \\ 0 & 0 & a_{43}^{(3)} & \cdots & a_{4n}^{(3)} \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \cdots & a_{nn}^{(3)} \end{bmatrix} .$$

Inductive, if all *pivot elements* $a_{ii}^{(i)}$ are different from 0, then, after $(n-1)$ steps, we get an upper tridiagonal matrix

$$R = L^{(n-1)}A^{(n-1)} = L^{(n-1)}L^{(n-2)} \cdots L^{(1)}A \; . \tag{2.6}$$

In the following we derive a decomposition $A = LR$, where L is a lower triangular matrix with diagonal entries 1 and R is an upper diagonal matrix. The structure of the Gauss-elimination explained above only provides

$$LA = R \; .$$

Essentially we have to prove that inverse of L is again a lower diagonal matrix.

2.4 The LR-Decomposition

From (2.6) it follows that

$$A = LR \text{ with } L = L^{(1)-1}L^{(2)-1} \cdots L^{(n-1)-1} \; . \tag{2.7}$$

The matrices $L^{(i)-1}$ can be explicitly calculated: First of all we remark that

$$e_i^T l_j = [0, \cdots, 0, 1, 0, \cdots, 0] \begin{bmatrix} 0 \\ \vdots \\ 0 \\ l_{j+1,j} \\ \vdots \\ l_{n,j} \end{bmatrix} = \begin{cases} 0 & \text{for} \quad i \leq j \\ l_{i,j} & \text{for} \quad i \geq j+1 \end{cases} \; . \tag{2.8}$$

From this it follows that:

$$\begin{aligned}(I - l_i e_i^T)(I + l_i e_i^T) &= I - l_i e_i^T + l_i e_i^T - l_i e_i^T l_i e_i^T \\ &= (e_i^T l_i = 0 (2.8)) I \ .\end{aligned}$$

The form of L is calculated inductive: Thereby we assume that for $1 \leq k < n$:

$$L^{(1)-1} \cdots L^{(k)-1} = I + l_1 e_1^T + \cdots + l_k e_k^T \ .$$

For $k = 1$ this assertion follows from the first part. From $L^{(k+1)-1} = I + l_{k+1} e_{k+1}^T$ it follows that

$$L^{(1)-1} \cdots L^{(k+1)-1} = (I + l_1 e_1^T + \cdots + l_k e_k^T)(I + l_{k+1} e_{k+1}^T) \ ,$$

and from (2.8) it follows that

$$\begin{aligned}L^{(1)-1} \cdots L^{(k+1)-1} =& I + l_1 e_1^T + \cdots + l_k e_k^T + l_{k+1} e_{k+1}^T + \sum_{i=1}^{k} l_i e_i^T l_{k+1} e_{k+1}^T \\ =& (e_i^T l_{k+1} = 0 \text{ for } i = 1, \cdots, k) \\ & I + l_1 e_1^T + \cdots + l_k e_k^T + l_{k+1} e_{k+1}^T \ .\end{aligned}$$

Thus we have seen that $L^{(i)-1} = I + l_i e_i^T$ (note $L^{(i)} = I - l_i e_i^T$) and $L = I + l_1 e_1^T + \cdots + l_{n-1} e_{n-1}^T$.

If during the Gauss-elimination a pivot element $a_{ii}^{(i)}$ becomes 0, then the Gauss-algorithm fails. Otherwise an LR-decomposition can be determined.

After a matrix A has been LR-decomposed the resulting equation $Ax = b$ can be easily solved with forward and backward substitution:

Solve $Ax = LRx = b$ in two steps:

- Solve $Ly = b$ by forward substitution
- Solve $Rx = y$ by backward substitution

Remark 2.1. *The total amount of products and division required for LR-decomposition is $\frac{1}{3}n^3 + \mathcal{O}(n^2)$. The number of computations for forward and backward substitutions is comparable small: It requires*

$$\frac{(n-1)^2}{2}(\textit{forward}) + \frac{(n-1)^2}{2} + n(\textit{backward substitution}) = n^2 - n + 1$$

Multiplications and divisions.

2.5 Pivoting

Example 2.2. *The matrix*

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

is a simple example where the elimination fails. However, if one interchanges the first and second row, the solution can be calculated by elimination. This is the topic of pivoting.

In practice the Gauss-elimination works well with partial pivoting, where in the i-th step the element $a_{ji}^{(i)}$ with

$$j = \underset{i \le k \le n}{\operatorname{argmax}} \frac{\left|a_{ki}^{(i)}\right|}{\sum_{l=i}^{n} \left|a_{kl}^{(i)}\right|}$$

is selected and the i-th and j-th rows are interchanged.

Example 2.3. *Solve $Mx = b$, for*

$$M = \begin{bmatrix} 2 & 6 & 10 \\ 1 & 3 & 3 \\ 3 & 14 & 28 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ -8 \end{bmatrix}.$$

To determine the first pivot element, we denote

$$M^{(1)} = M$$

and we compute the maximal factor

$$j = \underset{1 \le k \le 3}{argmax} \frac{\left|m_{k1}^{(1)}\right|}{\sum_{l=1}^{n} \left|m_{kl}^{(1)}\right|} = argmax\left\{c_1 = \frac{2}{18}, c_2 = \frac{1}{7}, c_3 = \frac{3}{45}\right\} = 2\,.$$

Then the system of linear equations is rewritten by interchanging the first and second rows of the original system:

$$\tilde{M}^{(1)} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} := \begin{bmatrix} 1 & 3 & 3 \\ 2 & 6 & 10 \\ 3 & 14 & 28 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \\ -8 \end{bmatrix}.$$

The first elimination step consists in multiplication of the second row by $-1/2$ *and to sum it with the first line, to give the second line. The second step consists in multiplication of the third row by* $-1/3$ *and sum it with the first line:*

$$M^{(2)} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} := \begin{bmatrix} 1 & 3 & 3 \\ 0 & 0 & -2 \\ 0 & -\frac{5}{3} & -\frac{19}{3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ \frac{14}{3} \end{bmatrix} .$$

The next pivot element is

$$j = \underset{2 \le k \le 3}{argmax} \frac{\left|m_{k2}^{(2)}\right|}{\sum_{l=1}^{n} \left|m_{kl}^{(2)}\right|} = argmax \left\{ c_2 = \frac{0}{2}, c_3 = \frac{5}{24} \right\} = 3 .$$

After pivoting we get the equation

$$\tilde{M}^{(2)} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} := \begin{bmatrix} 1 & 3 & 3 \\ 0 & -\frac{5}{3} & -\frac{19}{3} \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ \frac{14}{3} \\ 2 \end{bmatrix} .$$

No further elimination step is necessary. Now we can solve the equation system by backward substitution and get:

$$x_3 = -1, x_2 = 1, x_1 = 2 .$$

In some case even partial pivoting may fail, and then we have to use total pivoting. Thereby we select the index (j,k), with $i \le j,k \le n$, for pivoting, such that $\left|a_{j,k}^{(i)}\right|$ is maximal - in absolute terms - note that partial pivoting is formulated in relative terms. We then interchange the j-the and i-th row, and also the k-th and i-th column.

We summarize the numerical effort of Gauss-elimination in the following table:

Method	Computation cost
without pivoting	$\frac{1}{3}n^3 + \mathcal{O}(n^2)$
with partial pivoting	$\frac{1}{3}n^3 + \mathcal{O}(n^2)$
with total pivoting	$\frac{1}{3}n^3 + \mathcal{O}(\sum_{i=1}^{n} i^2)$

The numerical effort for total pivoting is high, and thus, typically, it is **not** implemented.

2.6 Cholesky-Decomposition

First, we recall that a matrix A is called *positive definite* if

$$x^T Ax > 0\,, \quad \forall x \neq 0\,.$$

With LR-decomposition a symmetric and positive definite matrix A is decomposed into the product of L (lower diagonal triangular matrix, which has entries 1 in the diagonal) and an upper diagonal triangular matrix. The Cholesky decomposition provides a decomposition $A = LL^T$ for symmetric and positive definite matrices. Here L does not need to have entries 1 in the diagonal necessarily.

Example 2.4. *Let*

$$A = \begin{bmatrix} 4 & 1 & 3 \\ 1 & 4 & 0 \\ 3 & 0 & 4 \end{bmatrix}\,.$$

The Cholesky-decomposition can be determined by the comparison of the coefficients. We do it first in this particular example and then in a general form.

$$\begin{bmatrix} 4 & 1 & 3 \\ 1 & 4 & 0 \\ 3 & 0 & 4 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}\,.$$

From this we immediately get:

$$l_{11}^2 = 4\,, \quad l_{11}l_{21} = 1\,, \quad l_{11}l_{31} = 3\,,$$

thus

$$l_{11} = 2\,, \quad l_{21} = \frac{1}{2}\,, \quad l_{31} = \frac{3}{2}\,.$$

Moreover, we have

$$l_{21}^2 + l_{22}^2 = 4\,, \quad l_{21}l_{31} + l_{22}l_{32} = 0\,,$$

which gives

$$l_{22} = \frac{\sqrt{15}}{2}, \quad l_{32} = -\frac{\sqrt{15}}{10}.$$

Finally $l_{31}^2 + l_{32}^2 + l_{33}^2 = 4$, which implies

$$l_{33} = \frac{2\sqrt{10}}{5}.$$

Thus

$$\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ \frac{1}{2} & \frac{\sqrt{15}}{2} & 0 \\ \frac{3}{2} & -\frac{\sqrt{15}}{10} & \frac{2\sqrt{10}}{5} \end{bmatrix} .$$

This comparison of coefficients can be generalized to an abstract setting:

$$\begin{bmatrix} a_{11} & a_{12} & \vdots & a_{1n} \\ a_{21} & a_{22} & \vdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \vdots & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & \cdots & l_{n1} \\ 0 & l_{22} & & l_{n2} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & l_{nn} \end{bmatrix}$$

There we obtain successively:

$$\begin{array}{cc} a_{11} = l_{11}^2 & l_{11} = \sqrt{a_{11}} \\ \\ a_{21} = l_{21} l_{11} & l_{21} = a_{21}/l_{11} \\ a_{22} = l_{21}^2 + l_{22}^2 & l_{22} = \sqrt{a_{22} - l_{21}^2} \\ & \Rightarrow \\ a_{31} = l_{31} l_{11} & l_{31} = a_{31}/l_{11} \\ a_{32} = l_{31} l_{21} + l_{32} l_{22} & l_{32} = (a_{32} - l_{31} l_{21})/l_{22} \\ a_{33} = l_{31}^2 + l_{32}^2 + l_{33}^2 & l_{33} = \sqrt{a_{33} - l_{31}^2 - l_{32}^2} \\ \vdots & \vdots \end{array}$$

From the algorithm above we see that the complexity for calculation of l_{ij} (with $i \geq j$) is at most j $*, /, \sqrt{}$ operations. Thus the total complexity of $*, /, \sqrt{}$ operations is

$$\sum_{j=1}^{n} (n+1-j)j = \frac{n(n+1)^2}{2} - \frac{n(n+1)(2n+1)}{6} = \frac{1}{6} n^3 + \mathcal{O}(n^2) .$$

2.7 QR-Decomposition

Definition 2.5. *Let $v \in \mathbb{R}^r \backslash \{0\}$: The matrix*

$$\boxed{H = I - \frac{2}{v^T v} v v^T \in \mathbb{R}^{r \times r}}$$

is called Householder-*transformation.*

Example 2.6. *Let $v = [1,1]^T$, then*

$$H = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \frac{2}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}[1,1] = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} .$$

Let $w = [1,-1]^T$ be a vector that is orthogonal to v, then we have

$$Hv = H\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix} = -v\,,$$
$$Hw = H\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} = w.$$

For an arbitrary vector ζ we write it with respect to the orthonormal coordinate system $\left\{\frac{v}{\sqrt{2}}, \frac{w}{\sqrt{2}}\right\}$:

$$\begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} = \frac{\lambda}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \frac{\mu}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix} .$$

This gives the local coordinate parameters

$$\lambda = \frac{\zeta_1 + \zeta_2}{\sqrt{2}} \text{ and } \mu = \frac{\zeta_1 - \zeta_2}{\sqrt{2}} .$$

Thus

$$H\zeta = \frac{\lambda}{\sqrt{2}}Hv + \frac{\mu}{\sqrt{2}}Hw = -\frac{\lambda}{\sqrt{2}}v + \frac{\mu}{\sqrt{2}}w .$$

Thus matrix H serves as a mirror transformation for the v component along the mirror w.

This observation holds true in every dimension:

Lemma 2.7. *The Householder transformation H is a symmetric matrix, which satisfies:*

$$Hv = -v \text{ and } Hw = w\,, \qquad \forall w \in [v]^{\perp} = \{\lambda v : \lambda \in \mathbb{R}\}^{\perp} .$$

To introduce the QR decomposition we construct an Householder matrix H, which transforms a given, arbitrary, vector $x \in \mathbb{R}^r \backslash \{0\}$ into $e_1 \in \mathbb{R}^r$. That is, we are looking to a vector $v \in \mathbb{R}^r \backslash \{0\}$ such that

$$Hx = x - \frac{2}{v^T v}v(v^T x) = \zeta e_1 . \tag{2.9}$$

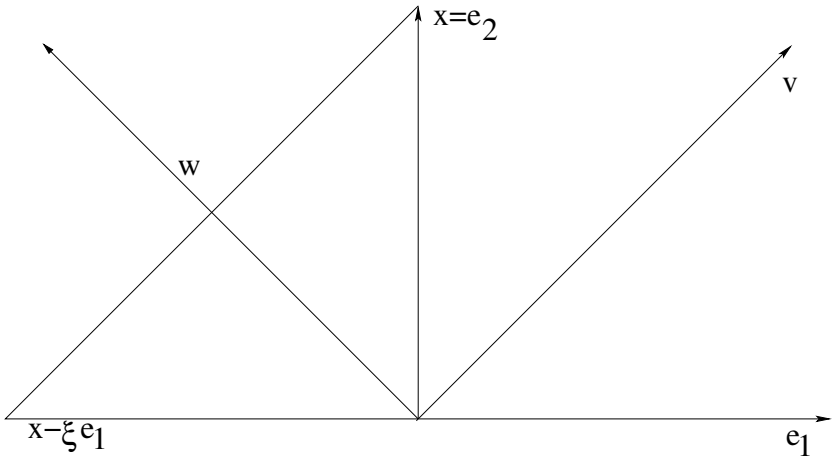

Figure 2.1: Here $x = e_2$

Geometrical speaking v and w are the complementary vectors between x and e_1.

We derive v from some necessary conditions: From (2.9) it follows that

$$x - \zeta e_1 = \frac{2v^T x}{\|v\|_2^2} v \ . \tag{2.10}$$

This means that v is proportional to $x - \zeta e_1$. In fact every $v = \lambda(x - \zeta e_1)$, $\lambda \neq 0$ satisfies (2.10). Now we determine the necessary relation between x and ζ. Let $v = \lambda(x - \zeta e_1)$, then from (2.10) it follows that

$$x - \zeta e_1 = \frac{2\|x\|_2^2 - 2\zeta x_1}{\|x\|_2^2 - 2\zeta x_1 + \zeta^2}(x - \zeta e_1)\,,$$

which implies that

$$|\zeta| = \|x\|_2 \ . \tag{2.11}$$

We use the following choice for (ζ, λ) (every other choice is possible as well)

$$\zeta = \begin{cases} -\|x\|_2 & \text{for } x_1 = 0 \\ -\frac{x_1}{|x_1|}\|x\|_2 & \text{for } x_1 \neq 0 \end{cases} \quad \text{and } \lambda = \frac{1}{\|x\|_2} \ . \tag{2.12}$$

With this choice

$$v = \begin{cases} x/\|x\|_2 + e_1 & \text{for } x_1 = 0\,, \\ \frac{1}{\|x\|_2}\left[x + \frac{x_1\|x\|_2}{|x_1|}e_1\right] & \text{for } x_1 \neq 0 \end{cases} \tag{2.13}$$

and $Hx = \zeta e_1$. Thus (2.9) is satisfied.

Example 2.8. *Let* $x = [1, 3, 4]^T$. *Then* $\|x\|_2 = \sqrt{26}$

$$v = \frac{1}{\sqrt{26}} \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} .$$

Then the Householder matrix is

$$H = \begin{bmatrix} -0.1961 & -0.5883 & -0.7845 \\ -0.5883 & 0.7106 & -0.3859 \\ -0.7845 & -0.3859 & 0.4855 \end{bmatrix} .$$

We check and see that with $w = [0.7845, 0, -1.1961]^T$, *which is orthogonal to* v,

$$Hx = \begin{bmatrix} -5.0990 \\ 0 \\ 0 \end{bmatrix} , \quad Hw = w , \quad Hv = -v .$$

The Housholder-algorithm for decomposing a matrix

$$A = [a_{ij}]_{\substack{1 \le i \le m \\ 1 \le j \le n}} = [a_1, \cdots, a_n]$$

is realized as follows (note that the matrix A does not have to be quadratic): We assume that $m \ge n$.

- We initialize $A^{(1)} = A$ and $x = a_1$ (the first column of the matrix A). With this vector x we can determine the Householder matrix $H^{(1)} \in \mathbb{R}^{m \times m}$ from (2.10) and get

  $$H^{(1)} a_1 = r_{11} e_1 , \quad r_{11} = \pm \|a_1\|_2 \neq 0 .$$

  Here the notation $\pm$ just means that we do not specify the sign in front of $\|a_1\|_2$ - it is, however, completely determined by (2.12). Consequently

  $$H^{(1)} A = \left[\begin{array}{c|ccc} r_{11} & * & * & * \\ \hline 0 & & A^{(2)} & \end{array}\right] \text{ where } A^{(2)} \in \mathbb{R}^{(m-1)\times(n-1)} .$$

- Now, we proceed inductively, and assume that after i steps we have constructed Householder matrices $H^{(1)}, \cdots, H^{(i)}$ such that

$$H^{(i)} \cdots H^{(1)} A = \left[\begin{array}{ccc|c} r_{11} & \cdots & r_{1i} & \\ & \ddots & \vdots & R^{(i)\prime} \\ 0 & & r_{ii} & \\ \hline 0 & \cdots & 0 & A^{(i+1)} \end{array}\right] \tag{2.14}$$

  where $R^{(i)\prime} \in \mathbb{R}^{i \times (n-i)}$ and $A^{(i+1)} \in \mathbb{R}^{(m-i)\times(n-i)}$.

  In the inductive step we select for $x \in \mathbb{R}^{m-i}$ the first column a_{i+1} of $A^{(i+1)}$ and construct the Householder matrix $H^{(i+1)\prime} \in \mathbb{R}^{(m-i)\times(m-i)}$ and the according vector $v' \in \mathbb{R}^{m-i}$ from (2.10). Application of the Householder transform gives

$$H^{(i+1)\prime} A^{(i+1)} = \left[\begin{array}{c|ccc} r_{i+1,i+1} & * & * & * \\ \hline 0 & \multicolumn{3}{c}{A^{(i+2)}} \end{array}\right]$$

  where $r_{i+1,i+1} = \pm \left\|a_{i+1}\right\|_2 \neq 0$. Again $\pm$ means that we have to choose a particular sign. Using the notation, we get

$$H^{(i+1)} := \left[\begin{array}{c|c} I & 0 \\ \hline 0 & H^{(i+1)\prime} \end{array}\right]$$

$$H^{(i+1)} H^{(i)} \cdots H^{(1)} A = \left[\begin{array}{ccc|ccc} r_{11} & \cdots & r_{1i} & \multicolumn{3}{c}{\multirow{2}{*}{$R^{(i)\prime}$}} \\ & \ddots & \vdots & & & \\ 0 & & r_{ii} & & & \\ \hline 0 & \cdots & 0 & r_{i+1,i+1} & * & * \; * \\ \hline 0 & \cdots & 0 & 0 & \multicolumn{2}{c}{A^{(i+2)}} \end{array}\right].$$

  We observe here, that in each step the first i rows are *not* changed.

Each Householder matrix H is unitary, that is $H^T = H^{-1}$, and the product of unitary matrices is again unitary. This shows that the Householder decomposition of matrix A with rank n can be decomposed as follows:

$$A = QR = Q \left[\begin{array}{ccc} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \\ \hline 0 & \cdots & 0 \end{array}\right].$$

where $Q \in \mathbb{R}^{m\times m}$ is unitary and $R \in \mathbb{R}^{m\times n}$ is triangular matrix with $r_{ii} \neq 0$.

Example 2.9. *We consider QR-decomposition of the matrix*

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 4 \\ 1 & 6 & 7 \end{bmatrix} .$$

- *In the first step we use* (2.12) *and* (2.13) *to calculate*

$$\zeta = -\sqrt{3} \text{ and } v = \frac{1}{\sqrt{3}} \begin{bmatrix} 1+\sqrt{3} \\ 1 \\ 1 \end{bmatrix} .$$

Thus the first Householder matrix is given by

$$H^{(1)} = \begin{bmatrix} -0.5774 & -0.5774 & -0.5774 \\ -0.5774 & 0.7887 & -0.2113 \\ -0.5774 & -0.2113 & 0.7887 \end{bmatrix} .$$

This shows then that

$$H^{(1)}A = \begin{bmatrix} -1.7321 & -5.7735 & -6.9282 \\ 0 & 0.5207 & 1.0981 \\ 0 & 3.5207 & 4.0981 \end{bmatrix} .$$

We select the submatrix

$$A^{(2)} = \begin{bmatrix} 0.5207 & 1.0981 \\ 3.5207 & 4.0981 \end{bmatrix} .$$

and calculate the second Householder matrix according to $x = [0.5207, 3.5207]^T$. *Using* (2.12) *and* (2.13) *we calculate*

$$\zeta = -3.5590 \text{ and } v = \begin{bmatrix} 1.1463 \\ 0.9892 \end{bmatrix} .$$

The according Householder transformation is

$$H^{(2)\prime} = \begin{bmatrix} -0.1463 & -0.9892 \\ -0.9892 & 0.1463 \end{bmatrix} .$$

Moreover,

$$H^{(2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.1463 & -0.9892 \\ 0 & -0.9892 & 0.1463 \end{bmatrix} .$$

Because,

$$H^{(2)\prime}A^{(2)} = \begin{bmatrix} -3.5590 & -4.2147 \\ 0 & -0.4867 \end{bmatrix},$$

we have found now the decomposition

$$R = H^{(2)}H^{(1)}A = \begin{bmatrix} -1.7321 & -5.7735 & -6.9282 \\ 0 & -3.5590 & -4.2147 \\ 0 & 0 & -0.4867 \end{bmatrix}.$$

Finally the Q-matrix of the QR decomposition is given by

$$Q = H^{(1)T}H^{(2)T} = H^{(1)}H^{(2)} = \begin{bmatrix} -0.5774 & 0.6556 & 0.4867 \\ -0.5774 & 0.0937 & -0.8111 \\ -0.5774 & -0.7493 & 0.3244 \end{bmatrix}.$$

Remark 2.10. *The QR-algorithm is often used for solving least-squares problems: Minimization of $\|Ax - b\|_2^2$ is equivalent to solving $A^T(Ax - b)$ (which we will see later). If we have a QR decomposition, then it is equivalent to*

$$R^T Rx = R^T b$$

So the solution can be found by backward and forward substitution.

2.8 Conclusion

At the end we compare the complexity of all elimination algorithms for quadratic matrices $A \in \mathbb{R}^{n\times n}$:

QR	$\frac{2}{3}n^3 + \mathcal{O}(n^2)$
LR	$\frac{1}{3}n^3 + \mathcal{O}(n^2)$
Cholesky	$\frac{1}{6}n^3 + \mathcal{O}(n^2)$

Chapter 3

Interpolation

We study the problem of interpolation of function samples $y_0 = y(x_0), \ldots, y_l = y(x_l)$ from a function $y : [a, b] \to \mathbb{R}$ on the grid

$$\Delta = \{a = x_0 < x_1 < \ldots < x_l = b\} \ . \tag{3.1}$$

The *grid size* is defined by

$$h := \max_{i=1,\ldots,l} h_i \, , \quad h_i = x_i - x_{i-1} \ .$$

Notation: m is the degree of the polynomial, $l + 1$ is the number of interpolation points.

3.1 Lagrange Interpolation

Historically, the first interpolation methods are based on polynomials:

Definition 3.1. Π_m *denotes the space of polynomials of degree* $\leq m$.

Polynomial interpolation consists in determining a polynomial $p \in \Pi_m$ such that

$$p(x_i) = y_i \, , \quad i = 0, \ldots, m \ . \tag{3.2}$$

Definition 3.2. *We denote by*

$$w(x) := \prod_{i=0}^{m} (x - x_i) \in \Pi_{m+1}$$

the nodal *polynomial at* Δ*. The polynomial*

$$\boxed{l_i(x) := \frac{w(x)}{(x - x_i)w'(x_i)} = \prod_{j=0, j\neq i}^{m} \frac{x - x_j}{x_i - x_j} \in \Pi_m, \quad x \neq x_i} \tag{3.3}$$

is called Lagrange-polynomial.

The Lagrange-polynomial satisfies

$$l_i(x_j) = \delta_{ij} \ . \tag{3.4}$$

The polynomial

$$p(x) = \sum_{i=0}^{m} y_i l_i(x) \ .$$

satisfies $p(x_j) = \sum_{i=0}^{m} y_i l_i(x_j) = y_j$, that is the interpolation exercise. The polynomial is unique in the space of functions in Π_m.

Example 3.3. *Every function* $f(x)$ *is interpolated at nodal points* a *and* b *by the linear polynomial*

$$\boxed{p(x) = f(a) - \frac{f(b) - f(a)}{b - a}(x - a) \ .}$$

3.2 Trigonometric Interpolation

Here we consider the grid

$$\Delta = \left\{ t_0 = 0 < t_1 = \frac{2\pi}{l} < \ldots < t_{l-1} = (l-1)\frac{2\pi}{l} \right\} \ , \tag{3.5}$$

which equally subdivides the interval $[0, 2\pi)$ into l subintervals.

The goal is to interpolate sample values $\{y_0, y_1, \ldots, y_{l-1}\}$ at Δ with a function of the form

$$y(t) = \frac{a_0}{2} + \sum_{j=1}^{m} a_j \cos(jt) + \sum_{j=1}^{m} b_j \sin(jt) \ .$$

Such functions are called *trigonometric polynomial* of degree m.

We restrict attention to the case that $l = 2m + 1$, in which case we can expect that we can solve the trigonometric interpolation exercise uniquely ($2m + 1$ nodal values and $2m + 1$ interpolation values): The interpolation exercise reads as follows:

Given $\{y_0, y_1, \ldots, y_{2m}\}$ determine $a_0, \{a_1, \ldots, a_m\}, \{b_1, \ldots, b_m\}$ such that

$$\boxed{y_k = \frac{a_0}{2} + \sum_{j=1}^{m} a_j \cos(jt_k) + \sum_{j=1}^{m} b_j \sin(jt_k)\,, \quad \forall k = 0, 1, \ldots, 2m\,.} \tag{3.6}$$

The coefficients $\{a_i, b_i\}$ can be determined analytically: For this purpose we use the following expression of the sums of cos and sin:

$$\begin{aligned}
&\sum_{k=0}^{l-1} \cos(\hat{j}t_k)\cos(jt_k) = \begin{cases} 0 & \text{for} \quad j \neq \hat{j} \text{ and } j, \hat{j} \in \left\{0, 1, \ldots, \frac{l-1}{2}\right\}, \\ \frac{l}{2} & \text{for} \quad j = \hat{j} \in \left\{1, \ldots, \frac{l-1}{2}\right\}, \\ l & \text{for} \quad j = \hat{j} = 0\,, \end{cases} \\
&\sum_{k=0}^{l-1} \cos(\hat{j}t_k)\sin(jt_k) = 0 \text{ for } j, \hat{j} \in \left\{0, 1, \ldots, \frac{l-1}{2}\right\}, \\
&\sum_{k=0}^{l-1} \sin(\hat{j}t_k)\sin(jt_k) = \begin{cases} 0 & \text{for} \quad j \neq \hat{j} \text{ and } j, \hat{j} \in \left\{0, 1, \ldots, \frac{l-1}{2}\right\}, \\ \frac{l}{2} & \text{for} \quad j = \hat{j} \in \left\{1, \ldots, \frac{l-1}{2}\right\}, \\ 0 & \text{for} \quad j = \hat{j} = 0\,. \end{cases}
\end{aligned} \tag{3.7}$$

These equalities are determined from the summation formulas

$$\begin{aligned}
\cos(jt_k)\cos(\hat{j}t_k) &= \frac{1}{2}\Big(\cos((j-\hat{j})t_k) + \cos((j+\hat{j})t_k)\Big) \\
&= \frac{1}{2}\mathrm{Re}\left(e^{\mathrm{i}(j-\hat{j})t_k} + e^{\mathrm{i}(j+\hat{j})t_k}\right), \\
\cos(jt_k)\sin(\hat{j}t_k) &= \frac{1}{2}\Big(\sin((j+\hat{j})t_k) - \sin((j-\hat{j})t_k)\Big) \\
&= \frac{1}{2}\mathrm{Im}\left(e^{\mathrm{i}(j+\hat{j})t_k} - e^{\mathrm{i}(j-\hat{j})t_k}\right), \\
\sin(jt_k)\sin(\hat{j}t_k) &= \frac{1}{2}\Big(\cos((j-\hat{j})t_k) - \cos((j+\hat{j})t_k)\Big) \\
&= \frac{1}{2}\mathrm{Re}\left(e^{\mathrm{i}(j-\hat{j})t_k} - e^{\mathrm{i}(j+\hat{j})t_k}\right).
\end{aligned}$$

We only show the first identity of (3.7), the others are left as exercises:

Theorem 3.4.

$$\sum_{k=0}^{l-1} \cos(\hat{j}t_k)\cos(jt_k) = \begin{cases} 0 & \textit{for} \quad j \neq \hat{j} \ \textit{and} \ j, \hat{j} \in \left\{0, 1, \ldots, \frac{l-1}{2}\right\} , \\ \frac{l}{2} & \textit{for} \quad j = \hat{j} \in \left\{1, \ldots, \frac{l-1}{2}\right\} , \\ l & \textit{for} \quad j = \hat{j} = 0 , \end{cases}$$

Proof: Denoting

$$q_+ = e^{\mathrm{i}(j+\hat{j})\frac{2\pi}{l}} \text{ and } q_- = e^{\mathrm{i}(j-\hat{j})\frac{2\pi}{l}} ,$$

it follows that

$$\begin{aligned} \sum_{k=0}^{l-1} \cos(\hat{j}t_k)\cos(jt_k) &= \frac{1}{2}\mathrm{Re}\sum_{k=0}^{l-1}\left(e^{\mathrm{i}(j-\hat{j})t_k} + e^{\mathrm{i}(j+\hat{j})t_k}\right) \\ &= \frac{1}{2}\mathrm{Re}\left(\sum_{k=0}^{l-1} q_-^k + \sum_{k=0}^{l-1} q_+^k\right) . \end{aligned}$$

Let us denote by

$$\sum := \sum_{k=0}^{l-1} q_-^k + \sum_{k=0}^{l-1} q_+^k .$$

- If $j = \hat{j} = 0$, then $q_+ = q_- = 1$. therefore

  $$\sum := 2l .$$

  Thus in turn

  $$\sum_{k=0}^{l-1} \cos(\hat{j}t_k)\cos(jt_k) = l .$$

- If $j = \hat{j} \in \left\{1, \ldots, \frac{l-1}{2}\right\}$, then $q_- = 1$, and $\sum_{k=0}^{l-1} q_-^k = \sum_{k=0}^{l-1} 1 = l$.
  Because $j + \hat{j} = 2j \in \{2, \ldots, l-1\}$ we see that $q_+ \neq 1$.
  The second term of $\sum$ is a geometric sum, that is

  $$\sum_{k=0}^{l-1} q_+^k = \frac{1 - q_+^l}{1 - q_+} = \frac{1 - e^{\mathrm{i}(j+\hat{j})2\pi}}{1 - q_+} = 0 .$$

  Therefore

  $$\sum_{k=0}^{l-1} \cos(\hat{j}t_k)\cos(jt_k) = \frac{1}{2}\mathrm{Re}(0 + l) = \frac{l}{2} .$$

- if $j \neq \hat{j}$, then $0 \neq j - \hat{j}$ and $j + \hat{j} \in \{1, \ldots, l-2\}$. Therefore, both $l(j \pm \hat{j})\frac{2\pi}{l}$ are multipliers of 2π, and thus $q_{\pm}^{l} = 1$, which means that $\sum = 0$. Therefore

$$\sum_{k=0}^{l-1} \cos(\hat{j}t_k)\cos(jt_k) = 0 \ .$$

□

Now, we continue with determining the coefficient $\{a_j, b_j\}$. Thereby we use three types of equalities (3.7) above :

- From (3.6) it follows that by taking into account that $\cos(0t_k) = \cos(0) = 1$,

$$\sum_{k=0}^{l-1} y_k = l\frac{a_0}{2} + \sum_{j=1}^{m} a_j \underbrace{\sum_{k=0}^{l-1} \cos(0t_k)\cos(jt_k)}_{(1.\text{in }(3.7)\text{ with }\hat{j}=0)=0} + \sum_{j=1}^{m} b_k \underbrace{\sum_{k=0}^{l-1} \cos(0t_k)\sin(jt_k)}_{(2.\text{in }(3.7)\text{ with }\hat{j}=0)=0} \ .$$

  That is

$$a_0 = \frac{2}{l}\sum_{k=0}^{l-1} y_k \ .$$

- Let $\hat{j} \in \left\{1, \ldots, \frac{l-1}{2}\right\}$. Then, by multiplication of (3.6) with cosine

functions and summation gives

$$\begin{aligned}
&\sum_{k=0}^{l-1} y_k \cos(\hat{j}t_k) \\
=&\frac{a_0}{2} \underbrace{\sum_{k=0}^{l-1} \cos(\hat{j}t_k)}_{(1.\text{in (3.7) with } j=0)=0} \\
&+ \sum_{j=1}^{m} a_j \underbrace{\sum_{k=0}^{l-1} \cos(jt_k)\cos(\hat{j}t_k)}_{(1.\text{in (3.7)})=\frac{l}{2}\delta_{j,\hat{j}}} \\
&+ \sum_{j=1}^{m} b_j \underbrace{\sum_{k=0}^{l-1} \sin(jt_k)\cos(\hat{j}t_k)}_{(2.\text{in (3.7)})=0}, \\
=&\frac{l}{2} a_{\hat{j}} . \qquad \forall 1 \le \hat{j} < \frac{l}{2} .
\end{aligned}$$

That is

$$a_{\hat{j}} = \frac{2}{l} \sum_{k=0}^{l-1} y_k \cos(\hat{j}t_k) . \tag{3.8}$$

- Let $\hat{j} \in \left\{1, \ldots, \frac{l-1}{2}\right\}$. Multiplication of (3.6) with sine functions and summation gives

$$\begin{aligned}
&\sum_{k=0}^{l-1} y_k \sin(\hat{j}t_k) \\
=&\sum_{k=0}^{l-1} \left(\frac{a_0}{2} + \sum_{j=1}^{m} a_j \cos(jt_k) + \sum_{j=1}^{m} b_j \sin(jt_k) \right) \sin(\hat{j}t_k) , \\
=&\frac{l}{2} b_{\hat{j}} .
\end{aligned}$$

Or in other words:

$$b_{\hat{j}} = \frac{2}{l} \sum_{k=0}^{l-1} y_k \sin(\hat{j}t_k) . \tag{3.9}$$

It is common to change to a complex number notation:

$$\boxed{c_{\hat{j}} = a_{\hat{j}} + \mathrm{i}b_{\hat{j}} = \frac{2}{l}\sum_{k=0}^{l-1} y_k(\cos(\hat{j}t_k) + \mathrm{i}\sin(\hat{j}t_k)) = \frac{2}{l}\sum_{k=0}^{l-1} y_k \exp(\mathrm{i}\hat{j}t_k)\ .} \quad (3.10)$$

Definition 3.5. *The discrete Fourier transform (DFT) of a set of n complex data values* $\{y_k : k = 0, \ldots, l-1\}$*, which are evenly spaced in* $[0, 2\pi)$ *is the set*

$$\left\{ c_{\hat{j}} = \sum_{k=0}^{l-1} y_k \exp(\mathrm{i}\hat{j}t_k) : \hat{j} = 0, 1, \ldots, l-1 \right\}\ .$$

Note, that in comparison with (3.10) *the prefactor* $\frac{2}{l}$ *is left out.*

3.3 Fast Fourier Transform (FFT)

Is an algorithm for fast evaluation of the DFT.

Let

$$\omega = \omega_l = \exp\left(\mathrm{i}\frac{2\pi}{l}\right)\ .$$

With this notation the DFT becomes

$$\left\{ c_{\hat{j}} = \sum_{k=0}^{l-1} y_k \omega^{\hat{j}k} : \hat{j} = 0, 1, \ldots, l-1 \right\}\ .$$

We explain the FFT for a 4×4 system, that is for $l = 4$. In this case $\omega = \exp\left(\mathrm{i}\frac{2\pi}{4}\right) = \mathrm{i}$. The linear relation of the DFT is as follows:

$$\begin{aligned}
\omega^0 y_0 + \omega^0 y_1 + \omega^0 y_2 + \omega^0 y_3 &= c_0\ , \\
\omega^0 y_0 + \omega^1 y_1 + \omega^2 y_2 + \omega^3 y_3 &= c_1\ , \\
\omega^0 y_0 + \omega^2 y_1 + \omega^4 y_2 + \omega^6 y_3 &= c_2\ , \\
\omega^0 y_0 + \omega^3 y_1 + \omega^6 y_2 + \omega^9 y_3 &= c_3\ .
\end{aligned}$$

Let

$$F_4 := \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega & \omega^2 & \omega^3 \\ 1 & \omega^2 & \omega^4 & \omega^6 \\ 1 & \omega^3 & \omega^6 & \omega^9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \omega & \omega^2 & \omega^3 \\ 1 & \omega^2 & 1 & \omega^2 \\ 1 & \omega^3 & \omega^2 & \omega \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \mathrm{i} & -1 & -\mathrm{i} \\ 1 & -1 & 1 & -1 \\ 1 & -\mathrm{i} & -1 & \mathrm{i} \end{bmatrix}\ .$$

be the *Fourier*-matrix.

Thus the system in matrix vector notation reads as follows:

$$F_4 \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} c_0 \\ c_2 \\ c_1 \\ c_3 \end{bmatrix} .$$

The matrix F_4 can be factorized as follows:

$$F_4 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & \mathrm{i} \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -\mathrm{i} \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} .$$

The last matrix puts the odd indices in front of the even ones. The middle matrix consists of two Fourier matrices of half size. In general we have

$$F_{2n} = \begin{bmatrix} I_n & D_n \\ I_n & -D_n \end{bmatrix} \begin{bmatrix} F_n & 0 \\ 0 & F_n \end{bmatrix} \begin{bmatrix} 1 & 0 & & & & & & \\ & 0 & 1 & 0 & & & & \\ & & & 0 & 1 & 0 & & \\ & & & & & & & \\ 0 & 1 & 0 & & & & & \\ & & 0 & 1 & 0 & & & \\ & & & & 0 & 1 & 0 & \\ & & & & & & & \end{bmatrix} ,$$

where I_n is the n-dimensional unitary matrix and $D_n = \mathrm{diag}(1, \omega, \ldots, \omega^{n-1})$ with $\omega = \omega_n = \exp(2\pi \mathrm{i}/n)$. The last matrix puts the odd lines on the top of the matrix and shuffles the even to the end. See Strang [14].

We calculate matrix vector multiplications after the factorization:

- We have to perform and index renumbering. Since there are no multiplications needed it does not count to the complexity.
- we need to perform two times the Fourier matrix multiplication of size $n/2$.
- We need n multiplications, when multiplying with the diagonal matrices $D_n F_n$.

Thus we have the recursive complexity

$$e(2n) = 2e(n) + \mathcal{O}(n) .$$

$\mathcal{O}(n)$ here refers to the fact that at most order n operations, such as multiplications with ω are performed. Let $n = 2^p$ (in our example $n = 4$ and $p = 2$). From the master theorem it follows that $e(n) = \mathcal{O}(n \log_2(n))$.

3.4 Spline Interpolation

Let $\Delta = \{a = x_0 < x_1 < \ldots < x_l = b\}$ be a grid on the interval $[a, b]$. A *step function* is a function, which satisfies

$$s(x) = s_i , \quad x_{i-1} \leq x < x_i , \quad i = 1, \ldots, l .$$

The set of all step functions is denoted by $S_{0,\Delta}$. It is a vector space of dimension l. As basis functions we use the characteristic functions $\chi_i = \chi_{[x_{i-1},x_i)}$, $i = 1, \ldots, l$. Thus

$$s(x) = \sum_{i=1}^{l} s_i \chi_i(x) .$$

Remark 3.6. *Let $f : [a, b] \to \mathbb{R}$. The step function $s(x) = \sum_{i=1}^{l} s_i \chi_i(x)$ with*

$$s_i = \frac{1}{h_i} \int_{x_{i-1}}^{x_i} f(x)\, dx , \quad i = 1, \ldots, l \tag{3.11}$$

is the best approximating step function with respect to the norm $\rho \to \sqrt{\int_a^b \rho^2(x)\, dx}$. That is the functional

$$\rho \in S_{0,\Delta} \to \int_a^b (f(x) - \rho(x))^2\, dx$$

is minimal for s.

3.5 Linear Splines

Definition 3.7. *A* spline of degree *n is a function s, which is $(n-1)$–times differentiable in (a, b) and on every interval $[x_{i-1}, x_i)$ a polynomial of degree n. The space of splines of order n is denoted by $S_{n,\Delta}$.*

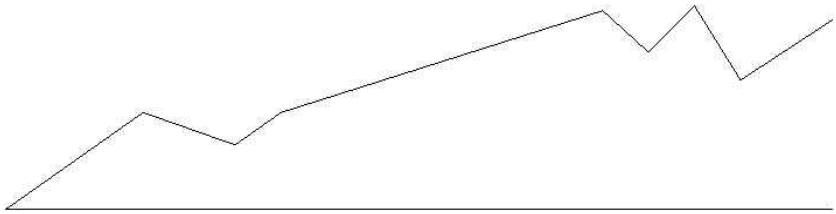

Figure 3.1: A linear spline

Of particular importance are the linear splines ($n = 1$) and cubic splines ($n = 3$).

Remark 3.8.

- *$S_{n,\Delta}$ is an $(n + l)$-dimensional space. Thus, in order to determines a spline of degree n, we have to specify the values at l nodal points and n additional conditions.*
- *A basis for $S_{1,\Delta}$ are formed by the* hat functions *Λ_i, $i = 0, 1, \ldots, l$, which are continuous, piecewise linear, and satisfy*

$$\Lambda_i(x_j) = \delta_{ij}\,, \quad i, j = 0, \ldots, l\,. \tag{3.12}$$

- (3.12) *allows for an easy computation of the interpolating spline: Let $y_0, \ldots, y_l$ sample values. Then, $s = \sum_{i=0}^{l} y_i \Lambda_i \in S_{1,\Lambda}$ is the unique spline, which satisfies*

$$s(x_i) = y_i\,, \quad i = 0, \ldots, l\,.$$

3.6 Cubic Splines

Cubic splines, that are the elements of $S_{3,\Delta}$, are used frequently in computer graphics.

We summarize some basic facts:

1. a cubic spline is two times differentiable.
2. A cubic spline is determined from $(l+3)$ measurements and conditions.
3. If $s \in S_{3,\Delta}$, then $s'' \in S_{1,\Delta}$. Thus

$$s'' = \sum_{i=0}^{l} \gamma_i \Lambda_i\,, \tag{3.13}$$

where $\gamma_i = s''(x_i)\,, i = 0, \ldots, l\,,$. γ_i are called *moments* of s.

In the following we derive the conditions for determining a cubic spline. First, we need some auxiliary result: For an arbitrary function ρ, which is twice differentiable in $[x_{i-1}, x_i]$, we have:

$$\begin{aligned} &\rho(x) - \rho(x_i) \\ &= \int_{x_i}^{x} \rho'(t) \cdot 1 \, dt \\ \underbrace{=}_{\text{Integration by parts}} \quad &\rho'(t)(t-x)\big|_{t=x_i}^{x} - \int_{x_i}^{x} \rho''(t)(t-x)\, dt \\ &= -\rho'(x_i)(x_i - x) - \int_{x_i}^{x} \rho''(t)(t-x)\, dt\,. \end{aligned} \tag{3.14}$$

Moreover, for an arbitrary $t \in [x_{i-1}, x_i]$ we have

$$\begin{aligned} s''(t) &= \gamma_{i-1}\Lambda_{i-1}(t) + \gamma_i \Lambda_i(t) \\ &= \gamma_{i-1}\frac{x_i - t}{x_i - x_{i-1}} + \gamma_i \frac{t - x_{i-1}}{x_i - x_{i-1}} \\ &= -\frac{\gamma_{i-1}}{h_i}(t - x_i) + \frac{\gamma_i}{h_i}(t - x_{i-1}) \\ &= \frac{\gamma_i - \gamma_{i-1}}{h_i}(t - x_i) + \gamma_i\,, \end{aligned} \tag{3.15}$$

which implies that for every $x \in [x_{i-1}, x_i)$

$$\begin{aligned}
& s(x) - s(x_i) + s'(x_i)(x_i - x) \\
\underbrace{=}_{(3.14)} & - \int_{x_i}^{x} s''(t)(t-x)\,dt \\
\underbrace{=}_{(3.15)} & - \frac{\gamma_i - \gamma_{i-1}}{h_i} \int_{x_i}^{x} (t-x_i)(t-x)\,dt - \gamma_i \int_{x_i}^{x} t - x\,dt \\
\underbrace{=}_{\text{Integration by parts}} & \frac{\gamma_i - \gamma_{i-1}}{2h_i} \int_{x_i}^{x} (t-x_i)^2\,dt \\
& + \frac{\gamma_i - \gamma_{i-1}}{2h_i}(t-x_i)^2(t-x)\Big|_{t=x_i}^{x} \\
& - \frac{\gamma_i}{2}(t-x)^2\Big|_{t=x_i}^{x} \\
= & \frac{\gamma_i - \gamma_{i-1}}{h_i}\frac{(x-x_i)^3}{6} + \gamma_i \frac{(x-x_i)^2}{2}\,.
\end{aligned} \tag{3.16}$$

Using the abbreviations we get

$$s_i = s(x_i) \text{ and } s'_i = s'(x_i) \text{ for } i = 0, \ldots, l\,.$$

Thus from (3.16) it follows that for every $x \in [x_{i-1}, x_i)$ and $i = 1, \ldots, l$

$$s(x) = s_i + s'_i(x - x_i) + \gamma_i \frac{(x-x_i)^2}{2} + \frac{\gamma_i - \gamma_{i-1}}{h_i}\frac{(x-x_i)^3}{6}\,. \tag{3.17}$$

In particular for $i = 1, \ldots, l$ we have

$$\begin{aligned}
s_{i-1} &= s_i - s'_i h_i + \frac{\gamma_i h_i^2}{2} - \frac{(\gamma_i - \gamma_{i-1})h_i^2}{6} \\
&= s_i - s'_i h_i + \frac{h_i^2}{6}(2\gamma_i + \gamma_{i-1})\,, \\
s'_{i-1} &= s'_i - \frac{h_i}{2}(\gamma_{i-1} + \gamma_i)\,.
\end{aligned} \tag{3.18}$$

Combinations of these equations shows

$$\begin{aligned}
&\frac{s_{i+1}-s_i}{h_{i+1}} - \frac{s_i - s_{i-1}}{h_i} \\
=&s'_{i+1} - \gamma_i \frac{h_{i+1}}{6} - \gamma_{i+1}\frac{h_{i+1}}{3} - s'_i + \gamma_{i-1}\frac{h_i}{6} + \gamma_i \frac{h_i}{3} \\
=&(\gamma_i + \gamma_{i+1})\frac{h_{i+1}}{2} - \gamma_i \frac{h_{i+1}}{6} - \gamma_{i+1}\frac{h_{i+1}}{3} + \gamma_{i-1}\frac{h_i}{6} + \gamma_i \frac{h_i}{3} \\
=&\frac{1}{6}(h_{i+1}\gamma_{i+1} + 2\gamma_i(h_i + h_{i+1}) + \gamma_{i-1}h_i) \; .
\end{aligned}$$

Writing this system in matrix notation we get

$$\begin{aligned}
&\frac{1}{6}\underbrace{\begin{bmatrix} h_1 & 2(h_1+h_2) & h_2 & & & 0 \\ & h_2 & 2(h_2+h_3) & \ddots & & \\ & & \ddots & \ddots & h_{l-1} & \\ & & & h_{l-1} & 2(h_{l-1}+h_l) & h_l \end{bmatrix}}_{\in \mathbb{R}^{(l-1)\times(l+1)}} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{l-1} \\ \gamma_l \end{bmatrix} \\
=&-\underbrace{\begin{bmatrix} -h_1^{-1} & h_1^{-1}+h_2^{-1} & -h_2^{-1} & & & 0 \\ & -h_2^{-1} & h_2^{-1}+h_3^{-1} & \ddots & & \\ & & \ddots & \ddots & -h_{l-1}^{-1} & \\ & & & -h_{l-1}^{-1} & h_{l-1}^{-1}+h_l^{-1} & -h_l^{-1} \end{bmatrix}}_{\in \mathbb{R}^{(l-1)\times(l+1)}} \begin{bmatrix} s_0 \\ s_1 \\ \vdots \\ s_{l-1} \\ s_l \end{bmatrix} .
\end{aligned} \tag{3.19}$$

The matrices in (3.19) have dimension $(l-1)\times(l+1)$, and thus are underdetermined. Thus, additional conditions are required: For a *natural cubic spline* we request in addition that

$$s''(a) = s''(b) = 0 \; . \tag{3.20}$$

However, (3.13) then shows, that

$$\gamma_0 = s''(a) = 0 \text{ and } \gamma_l = s''(b) = 0 \; . \tag{3.21}$$

Thus the system (3.19) simplifies to

$$\begin{aligned} &\frac{1}{6}\underbrace{\begin{bmatrix} 2(h_1+h_2) & h_2 & & 0 \\ h_2 & 2(h_2+h_3) & \ddots & \\ & \ddots & \ddots & h_{l-1} \\ & & h_{l-1} & 2(h_{l-1}+h_l) \end{bmatrix}}_{:=\mathcal{G}\in\mathbb{R}^{(l-1)\times(l-1)}} \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{l-1} \end{bmatrix} \\ = &\begin{bmatrix} d_1 \\ \vdots \\ d_{l-1} \end{bmatrix}, \end{aligned} \tag{3.22}$$

where

$$\begin{aligned} d_i = \frac{s_{i+1}-s_i}{h_{i+1}} - \frac{s_i - s_{i-1}}{h_i} = \frac{s_{i+1}}{h_{i+1}} - s_i\left(\frac{1}{h_i}+\frac{1}{h_{i+1}}\right) + \frac{s_{i-1}}{h_{i-1}}, \\ i = 1,\dots,l-1\,. \end{aligned} \tag{3.23}$$

Example 3.9. *We consider an equidistant grid with step size h. For given $j = 1,\dots,l-1$ we determine the natural cubic spline s which satisfies*

$$s(x_i) = s_i = \delta_{ij}\,, \quad i = 0,\dots,l\,. \tag{3.24}$$

The system (3.22), (3.23) *reads as follows:*

$$
\begin{aligned}
&\frac{1}{6}\begin{bmatrix} 4 & 1 & & 0 \\ 1 & 4 & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 4 \end{bmatrix}\begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{l-1} \end{bmatrix} \\
=&-\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix}\begin{bmatrix} s_1 \\ \vdots \\ s_{l-1} \end{bmatrix} \\
=&-\frac{1}{h^2}\begin{bmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix}\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (3.25) \\
=&\frac{1}{h^2}\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ -2 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.
\end{aligned}
$$

Chapter 4

Numerical Quadrature

In this chapter we investigate the *numerical approximation* of integrals

$$I[f] = \int_a^b f(x)\,dx\,, \tag{4.1}$$

without determining analytically the anti-derivative of f.

Example 4.1. *The simplest approximation formulas are the* mid point

$$\int_a^b f(x)\,dx \;\approx (b-a) f\left(\frac{a+b}{2}\right) \tag{4.2}$$

and the trapezoidal formula

$$\int_a^b f(x)\,dx \;\approx \frac{b-a}{2} f(a) + \frac{b-a}{2} f(b)\,, \tag{4.3}$$

respectively. Obviously, in general, in (4.2) and (4.3) there is no equality.

In general, we are considering quadrature formulas of the form:

$$\boxed{Q[f] = \sum_{i=0}^{m} \omega_i f(x_i) \approx I[f]\,,}$$

with *nodal points* $\{x_i : i = 0, \ldots, m\}$ and *weights* $\{\omega_i : i = 0, \ldots, m\}$.

In particular we have

Mid point rule	$m = 0$	$x_0 = \frac{a+b}{2}$	$\omega_0 = b - a$
Trapezoidal rule	$m = 1$	$x_0 = a, x_1 = b$	$\omega_0 = \omega_1 = \frac{b-a}{2}$

4.1 Compound Formulas

In practice an interval $[\hat{a}, \hat{b}]$ is subdivided into n subintervals and (4.2) and (4.3) are applied to each subinterval.

We assume for the sake of simplicity of presentation a uniform subdivision:

$$\begin{aligned} &\hat{a} = \hat{x}_0 < \hat{x}_1 < \hat{x}_2 < \cdots < \hat{x}_n = \hat{b}\,, \\ &\hat{x}_j = \hat{a} + jh\,, \quad h = \frac{\hat{b} - \hat{a}}{n}\,. \end{aligned} \tag{4.4}$$

We call $\hat{x}_i$ as the sampling points. The compound formula is defined as follows:

$$\boxed{Q_n[f] := \sum_{j=1}^{n} Q[f_{[x_{j-1},x_j]}] \approx \int_{\hat{a}}^{\hat{b}} f(x)\,dx\,.}$$

Example 4.2. *The trapezoidal rule results in the compound formula*

$$\begin{aligned} Q_n[f] &:= \sum_{j=1}^{n} \frac{\hat{x}_j - \hat{x}_{j-1}}{2}(f(\hat{x}_j) + f(\hat{x}_{j-1})) \\ &= \frac{h}{2}f(\hat{a}) + h\sum_{j=1}^{n-1} f(\hat{x}_j) + \frac{h}{2}f(\hat{b})\,. \end{aligned} \tag{4.5}$$

We denote the $\hat{x}_j$ the sampling points.

4.2 Order of Quadrature Formulas

Qualitative properties of a quadrature formula are the *degree of exactness* and *order of convergence*

Definition 4.3. *Let Π_m be the vector space of polynomials of degree $\leq m$.*

1. *A quadrature formula $Q[f]$ has* degree q *if*

$$Q[p] = I[p]\,, \quad \forall p \in \Pi_q\,.$$

2. *A compound quadrature formula converges against $I[f]$ with order s if*

$$|Q_n[f] - I[f]| = O(n^{-s})\,, \quad n \to \infty\,.$$

Note the different terminology:

$\hat{x}_i$	sampling points	compound
x_i	nodal points	small interval
$\hat{a}, \hat{b}$	boundaries	compound
a, b	boundaries	small interval
n	# intervals	compound
m	# nodal intervals	small interval

4.3 Newton-Cotes-Formulas

Using polynomial interpolation we can construct quadrature formulas for $I[f]$ for arbitrary degrees of freedom q.

Proposition 4.4. *Let $x_0 < x_1 < \cdots < x_m$ nodal points in $[a, b]$ and let*

$$\omega_i := \int_a^b l_i(x)\,dx\ . \tag{4.6}$$

Then the Newton-Cotes *quadrature formula*

$$Q[f] = \sum_{i=0}^{m} \omega_i f(x_i)$$

has exactness degree of at least $q = m$.

Example 4.5. *We constrain ourselves to equidistant nodal points $a = x_0 < x_1 < \cdots < x_m = b$.*

For $m = 1$ we have the trapezoidal rule (4.3) and for $m = 2$ we have the Simpson formula

$$\int_a^b f(x)\,dx \approx \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)\ . \tag{4.7}$$

According to Proposition 4.4 the Simpson-formula has exactness degree 2. In fact it has even exactness degree 3. The trapezoidal rule has exactness degree 1.

Remark 4.6. *We emphasize that for all quadrature formulas*

$$b-a=\int_a^b 1\,dx=I[1]\underbrace{=}_{1\in\Pi_0}Q[1]=\sum_{i=0}^{m}\omega_i\ . \tag{4.8}$$

For $m=2$ we get the compound Simpson-rule: With $\hat{x}_i=\hat{a}+ih$, $i=0,\dots,2m$ and $h=\frac{\hat{b}-\hat{a}}{2n}$ we have

$$\int_{\hat{a}}^{\hat{b}} f(x)\,dx\sim$$
$$\frac{h}{3}\left\{f(\hat{a})+4f(\hat{x}_1)+2f(\hat{x}_2)+4f(\hat{x}_3)+\cdots+2f(\hat{x}_{2n-2})+4f(\hat{x}_{2n-1})+f(\hat{b})\right\}$$

Remark 4.7. *The compound trapezoidal rule has degree 2: On an arbitrary interval $[a,b]$ we have by Taylor's theorem:*

$$\begin{aligned}
f(x)&=f(a)+f'(a)(x-a)+\frac{f''(\zeta_a)}{2}(x-a)^2\\
f(x)&=f(a)+\frac{f(b)-f(a)}{b-a}(x-a)\\
&\qquad+\left(f'(a)-\frac{f(b)-f(a)}{b-a}\right)(x-a)+\frac{f''(\zeta_a)}{2}(x-a)^2\\
&=f(a)+\frac{f(b)-f(a)}{b-a}(x-a)\\
&\qquad-\frac{f''(\zeta_b)}{2}(b-a)(x-a)+\frac{f''(\zeta_a)}{2}(x-a)^2
\end{aligned}$$

Denoting by $C:=\sup_{x\in[a,b]}|f''(\xi)|$ and integration over the interval $[a,b]$ we get

$$\begin{aligned}
&\left|\int_a^b f(x)\,dx-(b-a)\frac{f(a)+f(b)}{2}\right|\\
&\left|\int_a^b f(x)\,dx-\int_a^b f(a)+\frac{f(b)-f(a)}{b-a}(x-a)\,dx\right|\\
\le&\frac{C}{2}\int_a^b(x-a)(b-x)\,dx=\frac{C}{6}(b-a)^3\ .
\end{aligned}$$

This shows that the compound trapezoidal formula on a uniform grid satisfies:

$$\left|\int_{\hat{a}}^{\hat{b}} f(x)\,dx-Q_n[f]\right|\le n\frac{C}{6}h^3=\frac{C(b-a)}{6}h^2\ .$$

where $C := \sup_{x\in[\hat{a},\hat{b}]} |f''(\xi)|$.

Thus the compound trapezoidal formula is a second order method.

4.4 Gaussian-Quadrature

Let $x_0 < x_1 < \cdots < x_m$ denote nodal points and denote the according weights $\omega_0, \omega_1, \cdots, \omega_m$. Wie are interested in how large the maximal degree of exactness of the following quadrature formula can be:

$$Q[f] = \sum_{i=0}^{m} \omega_i f(x_i) \approx I[f] = \int_a^b f(x)\,dx\ . \tag{4.9}$$

Proposition 4.8. *Then the degree of exactness of a quadrature formula* $Q[\cdot]$ *from (4.9) can be at most* $q = 2m + 1$.

With Gaussian-quadrature formulas it is possible to achieve the degree $2m+1$. The weights ω_i and the nodes are tabelized. The higher accuracy is obtained by optimizing the nodal points.

Gauß-Legendre-Integration on the Interval $[-1, 1]$

n	x_i	ω_i
1	0	2
2	$-\sqrt{\frac{1}{3}}, \sqrt{\frac{1}{3}}$	$1, 1$
3	$-\sqrt{\frac{3}{5}}, 0, \sqrt{\frac{3}{5}}$	$\frac{5}{9}, \frac{8}{9}, \frac{5}{9}$
4	$-\sqrt{\frac{3}{7}+\frac{2}{7}\sqrt{\frac{6}{5}}}, -\sqrt{\frac{3}{7}-\frac{2}{7}\sqrt{\frac{6}{5}}},$ $\sqrt{\frac{3}{7}-\frac{2}{7}\sqrt{\frac{6}{5}}}, \sqrt{\frac{3}{7}+\frac{2}{7}\sqrt{\frac{6}{5}}}$	$\frac{18-\sqrt{30}}{36}, \frac{18+\sqrt{30}}{36},$ $\frac{18+\sqrt{30}}{36}, \frac{18-\sqrt{30}}{36}$
5	$-\frac{1}{3}\sqrt{5+2\sqrt{\frac{10}{7}}}, -\frac{1}{3}\sqrt{5-2\sqrt{\frac{10}{7}}}, 0,$ $\frac{1}{3}\sqrt{5-2\sqrt{\frac{10}{7}}}, \frac{1}{3}\sqrt{5+2\sqrt{\frac{10}{7}}}$	$\frac{322-13\sqrt{70}}{900}, \frac{322+13\sqrt{70}}{900}, \frac{128}{225},$ $\frac{322+13\sqrt{70}}{900}, \frac{322-13\sqrt{70}}{900}$

Chapter 5

Ordinary Differential Equations

In this chapter we study the numerical solution of ordinary differential equations

$$y' = f(t, y)\ , t \in [0, T] \text{ with the initial condition } y(0) = y_0\ . \tag{5.1}$$

Here y is a vector valued function. We call this equation a *system of first order*.

We use the following convention:

y	$\in \mathbb{R}^{\nu \pm 1}$
x	$\in \mathbb{R}^{\nu+1}$
$\hat{i}, \hat{j}$	Index for x, y
i, j	Index of iterations of numerical method
y_i, t_i	iterate y_i approximating $y(t_i)$

Example 5.1. *(Exponential Grow) A population has infinite resources. At time t the population is $P(t)$. We assume that the rate of change in the population is constant: That is, there exists a constant α such that*

$$\frac{P'(t)}{P(t)} = \alpha \tag{5.2}$$

and the population doubles every year. From the later we determine α. Because $(\log P)'(t) = \alpha$ it follows that

$$\log P(t) = \alpha t + \beta\ .$$

If we choose the time units in years and assume that the population doubles every year (that is $\alpha = 2$) it follows that:

$$2 = \frac{P(t+1)}{P(t)} = e^{\alpha} \text{ and } P(0) = e^{\beta} .$$

Several further easy examples can be found in the book of Heuser [7].

The second example provides a relation between ordinary and partial differential equations, and how ordinary differential equation solvers can be used for the solution of partial differential equations.

Example 5.2. *Let $u(x,t)$, $-1 \le x \le 1$, be the temperature distribution at time t in a slab of length $l = 2$. Assuming constant conductivity $\sigma = 1$, u satisfies the* heat conduction *equation:*

$$u_t = \sigma u_{xx} = u_{xx} , \quad -1 < x < 1 , 0 < t < T . \tag{5.3}$$

This is now a partial differential equation *because it depends on derivatives of* **at least two** *variables x, t. By discretization of the x variable we can transform the partial differential equation in a system of ordinary differential equations.*

Let $v : [-1,1] \to \mathbb{R}$ be an arbitrary function satisfying $v(-1) = v(1) = 0$, then we get by integration by parts

$$\int_{-1}^{1} u_t(t,x)v(x)\,dx = \int_{-1}^{1} u_{xx}(t,x)v(x)\,dx = -\int_{-1}^{1} u_x(t,x)v_x(x)\,dx . \tag{5.4}$$

Assume that the temperatures $u(-1,t) := u_0(t)$ and $u(1,t) := u_1(t)$ are measured, then, for every $t > 0$, $u(t,x)$ can be approximated by a linear spline in space over the grid $\Delta = \{-1 = x_0 < x_1 < ... < x_\nu = 1\}$, that is

$$u(t,x) = \sum_{\hat{i}=0}^{\nu} y_{\hat{i}}(t)\Lambda_{\hat{i}}(x) , \tag{5.5}$$

where $\Lambda_{\hat{i}}$ is a linear hat function with peak at $x_{\hat{i}}$. Taking into account the boundary conditions we see that $y_0 = u_0(t)$ and $y_\nu = u_1(t)$. All other functions $y_{\hat{i}}$ are unknown.

Inserting (5.5) in (5.4) we get a system of differential equations for $y_1, ..., y_{\nu-1}$*:*

$$\sum_{\hat{i}=0}^{\nu} y_{\hat{i}}'(t) \int_{-1}^{1} \Lambda_{\hat{i}}(x) v(x)\, dx = -\sum_{\hat{i}=0}^{\nu} y_{\hat{i}}(t) \int_{-1}^{1} \Lambda_{\hat{i}}'(x) v_x(x)\, dx \,,$$

where we choose $v(x) \in \left\{ \Lambda_{\hat{j}}(x) : \hat{j} = 1, ..., \nu - 1 \right\}$ *- this means that* v *is a hat function, which satisfies homogenous boundary conditions.*

Denote by

$$G := [\langle \Lambda_{\hat{i}}, \Lambda_{\hat{j}} \rangle]_{1 \le \hat{i}, \hat{j} \le \nu - 1} = \frac{h}{6} \begin{bmatrix} 4 & 1 & 0 & \cdots & \cdots & 0 \\ 1 & 4 & 1 & 0 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & 1 & 4 & 1 \\ \vdots & . & . & 0 & 1 & 4 \end{bmatrix}$$

and

$$A := [\langle \Lambda_{\hat{i}}', \Lambda_{\hat{j}}' \rangle]_{1 \le \hat{i}, \hat{j} \le \nu - 1} = h \begin{bmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & -1 & 2 & -1 \\ \vdots & \ddots & \ddots & 0 & -1 & 2 \end{bmatrix}$$

we get a compact description of the system

$$Gy'(t) + Ay(t) = b(t)\,, \tag{5.6}$$

where b *is an appropriate vector, which depends on* u_0 *and* u_1*.*

To completely specify the system (5.6) *we need initial values for* $y_1, ..., y_{\nu-1}$*, which are typically determined from interpolation of the initial temperature* $u(0, x)$*. Note however, that this is a system of equations (the solution* y *is vector valued).*

5.1 The Euler Method

The *(explicit) Euler-method* approximates y on a uniform grid

$$\Delta = \{0 = t_0 < t_1 < t_2 < ... < t_n\} \subseteq I$$

with the recursive formula

$$\boxed{y_{i+1} = y_i + (t_{i+1} - t_i) f(t_i, y_i)\ .}$$

The *implicit Euler-method* approximates the solution by

$$\boxed{y_{i+1} = y_i + (t_{i+1} - t_i) f(t_{i+1}, y_{i+1})\ .} \tag{5.7}$$

Thereby in each step an equation has to be solved. This method is stable in a sense which has to be specified afterward. However, the method is rather slow.

5.2 Runge-Kutta Method

The disadvantage of both Euler-methods is the slow convergence (in dependence of the time discretization). Faster convergence can be obtained with an ansatz

$$\boxed{y_{i+1} = y_i + h \sum_{j=1}^{s} b_j f(t_i + c_j h, \eta_j)\,, \quad \sum_{j=1}^{s} b_j = 1\,,} \tag{5.8}$$

where η_j is an approximation for $y(t_i + c_j h)$.

Such methods are called *Runge-Kutta-methods* of *degree* s. In particular:

- The explicit Euler method is with $s = 1$ and $c_1 = 0$, $\eta_1 = y_i$.
- For the implicit Euler method we have $s = 1$, $c_1 = 1$, $\eta_1 = y_{i+1}$.

Because in (5.8) one calculates and approximation $y_{i+1} \approx y(t_{i+1})$ starting from $y_i \approx y(t_i)$ the method is called *single step* method. If other previous approximations $y_{i-1}, ...$ are used to determine y_{i+1}, then the method is called *multi-step* method.

5.3 Single Step Runge-Kutta Methods

We assume that $y_i = y(t_i)$, then by the fundamental theorem of differential calculus

$$\begin{aligned} y(t_{i+1}) - y_{i+1} &\underbrace{=}_{y(t_i)=y_i} y(t_{i+1}) - y(t_i) - h\sum_{j=1}^{s} b_j f(t_i + c_j h, n_j) \\ &= \int_{t_i}^{t_{i+1}} y'(t)\,dt - h\sum_{j=1}^{s} b_j f(t_i + c_j h, n_j) \\ &\underbrace{=}_{ODE} \int_{t_i}^{t_{i+1}} f(t, y(t))\,dt \; - h\sum_{j=1}^{s} b_j f(t_i + c_j h, n_j)\,. \end{aligned}$$

We see that the *local error* gets small if

$$h\sum_{j=1}^{s} b_j f(t_i + c_j h, n_j) \approx \int_{t_i}^{t_{i+1}} f(t, y(t))\,dt\,.$$

This suggest to take quadrature formulas for choosing $\{b_j\}$, $\{c_j\}$ and $\{\eta_j\}$.

Example 5.3.

- *Using the mid-point rule we get*

$$y_{i+1} = y_i + hf(t_i + h/2, \eta_1)\,, \tag{5.9}$$

where ideally $\eta_1 = y(t_i + h/2)$. Because this value of the solution y is not known we are looking for an approximation: The method of Runge *(1895) used the approximation*

$$\eta_1 = y(t_i) + \frac{h}{2}y'(t_i) \approx y_i + \frac{h}{2}f(t_i, y_i)\,.$$

- *With the trapezoidal rule we find*

$$y_{i+1} = y_i + \frac{h}{2}f(t_i, y_i) + \frac{h}{2}f(t_{i+1}, \tilde{\eta}_1)\,,$$

where $\tilde{\eta}_i \approx y(t_i + h)$. If we proceed as in the Runge method and if we use the approximation

$$\tilde{\eta}_1 = y_i + hy'(t_i)\,,$$

then we get the method of Heun.

The *Runge-Kutta methods* rely on the following choice of coefficients:

$$\boxed{\eta_j \approx y(t_i + c_j h) = y(t_i) + \int_{t_i}^{t_i + c_j h} y'(t)\, dt \; = y(t_i) + \int_{t_i}^{t_i + c_j h} f(t, y(t))\, dt \; .} \tag{5.10}$$

For the approximate evaluation there are used again quadrature formulas which, for the evaluation of $f(t, y)$, use the same nodal values $f(t_i + c_j h, \eta_j)$, $j = 1, ..., s$, as they are used for calculating y_{i+1}. Thus we make the following ansatz:

$$\eta_j = y_i + h \sum_{k=1}^{s} a_{jk} f(t_i + c_k h, \eta_k) \, , \quad \sum_{k=1}^{s} a_{jk} = c_j \; . \tag{5.11}$$

In practice the coefficients $\{a_{jk}, b_j, c_j\}$ are summarized in a quadratic tableau (*Runge-Kutta Abc* or *Butcher-scheme*):

$$\begin{array}{c|c} \mathrm{c} & \mathrm{A} \\ \hline & \mathrm{b}^T \end{array} \quad = \quad \begin{array}{c|cccccc} c_1 & a_{1,1} & ... & ... & ... & a_{1,s} \\ c_2 & a_{2,1} & a_{2,2} & ... & ... & ... \\ c_3 & a_{3,1} & a_{3,2} & ... & ... & ... \\ ... & ... & ... & ... & ... & ... \\ c_s & a_{s,1} & ... & ... & a_{s,s-1} & a_{s,s} \\ \hline & b_1 & b_2 & ... & b_{s-1} & b_s \end{array}$$

where $A = [a_{j,k}] \in \mathbb{R}^{s \times s}$, $b = [b_1, ..., b_s]^T \in \mathbb{R}^s$ and $c = [c_1, ..., c_s]^T \in \mathbb{R}^s$.

Example 5.4. *For the explicit and implicit Euler method we have the following tableau, respectively:*

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

The method of Runge requires to add a trivial equation to transform it into the general scheme:

$$\begin{array}{rcl} \eta_0 & = & y_i + h \sum_{k=0}^{1} 0 \cdot f(t_i + c_k h, \eta_k) \\ \eta_1 & = & y_i + \frac{h}{2} f(t_i + h/2, \eta_0) \\ y_{i+1} & = & y_i + h f(t_i + h/2, \eta_1) \; . \end{array}$$

The tableau then reads as follows:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}$$

5.4 Stiff ODE's

Stiffness is a phenomenon rather than a definition in a rigorous mathematical setting. The terminology *stiff* probably originates from chemical reaction problems which exhibit tight coupling of various reactions of different scales.

Since there is no rigorous mathematical definition of stiffness, we can only describe it phenomenologically.

Example 5.5. *Consider the differential equation*

$$y'(t) = -15y(t)\,, \quad t \geq 0\,, \qquad y(0) = 1\,. \tag{5.12}$$

The exact solution is

$$y(t) = e^{-15t}\,,$$

which satisfies $y(t) \to 0$ *for* $t \to \infty$.

Numerically we see a completely different behavior for various methods.

1. *The Euler method with a step size of* $h = 1/4$ *oscillates and the solution blows up very rapidly. The iterates* $y_i, i = 0, \ldots, 10$,

$$[1, -3, 8, -21, 57, -157, 433, -1189, 3271, -8995, 24736]^T\,.$$

While the exact solution is

$$[1, 0.0235, 0.0006, 0, 0, 0, 0, 0, 0, 0, 0]^T\,.$$

2. *The iterates with Euler's method with step size* $h = 1/8$ *are bounded:*

$$\begin{aligned}[1, &-0.8750, 0.7656, -0.6699, 0.5862, -0.5129,\\ &0.4488, -0.3927, 0.3436, -0.3007, 0.2631, -0.2302,\\ &0.2014, -0.1762, 0.1542, -0.1349, 0.1181, -0.1033,\\ &0.0904, -0.0791, 0.0692]^T\,.\end{aligned}$$

The exact solution is

$$[1, 0.1534, 0.0235, 0.0036, 0.0006, 0.0001, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]\,.$$

3. The trapezoidal method*, defined by ,*

$$\begin{aligned} y_{i+1} &= y_i + \frac{h}{2}(f(t_i, y_i) + f(t_{i+1}, y_{i+1})) \\ &= \frac{2 - 15h}{2 + 15h} y_i \, . \end{aligned}$$

With step size $h = 1/8$ we get

$$[1, 0.0323, 0.0010, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],$$

which decreases monotonically to zero.

Example 5.6. *One of the most prominent examples of a stiff ODEs is a system that describes the chemical reaction of Robertson:*

$$\begin{aligned} y_1' &= -4.10^{-2} y_1 + 10^4 y_2 y_3 \, , \\ y_2' &= 4.10^{-2} y_1 - 10^4 y_2 y_3 - 3.10^7 y_2^2 \, , \\ y_3' &= 3.10^7 y_2^2 \, . \end{aligned} \tag{5.13}$$

On a short time interval the numerical solution of the system does not make problems, however for large t (let us say 10^{11}) it does.

5.4.1 Stiffness Ratio

Consider the linear inhomogeneous system

$$y'(t) = \mathcal{A} y(t) + f(t) \, , \tag{5.14}$$

where $y = y(t), f = f(t) \in \mathbb{R}^\nu$ and $\mathcal{A} \in \mathbb{R}^{\nu \times \nu}$ is symmetric with eigenvalues $\lambda_{\hat{i}} \in \mathbb{C}$ and eigenvectors $y_{\hat{i}}$, $\hat{i} = 1, \ldots, \nu$. We assume that the matrix $\mathcal{A}$ can be diagonalized: That is, there exists a matrix Y, consisting of the columns of $y_{\hat{i}}$, such that

$$\mathcal{A} = Y \Lambda Y^{-1} \, ,$$

where Λ is the diagonal-matrix consisting of eigenvalues of $\mathcal{A}$, and $y_{\hat{i}}$, $\hat{i} = 1, \ldots, \nu$ forms an orthonormal basis of $\mathbb{R}^\nu$ [4]. So, if $\mathcal{A}$ can be diagonalized, then with

$$\boxed{z(t) = Y^{-1} y(t) \text{ and } g(t) = Y^{-1} f(t) \, .} \tag{5.15}$$

$$z'(t) = Y^{-1} y'(t) = \Lambda Y^{-1} y(t) + Y^{-1} f(t) = \Lambda z(t) + g(t) \, , \tag{5.16}$$

or in other words

$$z_{\hat{i}}'(t) = \lambda_{\hat{i}} z_{\hat{i}}(t) + g_{\hat{i}}(t) \; . \tag{5.17}$$

The solution of this system is determined by the method of *variations of constants*: This procedure makes use of the ansatz

$$\boxed{z_{\hat{i}}(t) = c_{\hat{i}}(t) e^{\lambda_{\hat{i}} t} \; .} \tag{5.18}$$

Then

$$z_{\hat{i}}'(t) = c_{\hat{i}}'(t) e^{\lambda_{\hat{i}} t} + c_{\hat{i}}(t) \lambda_{\hat{i}} e^{\lambda_{\hat{i}} t} \; .$$

To satisfy the differential equation (5.17) we have to satisfy

$$c_{\hat{i}}'(t) e^{\lambda_{\hat{i}} t} + \underline{c_{\hat{i}}(t) \lambda_{\hat{i}} e^{\lambda_{\hat{i}} t}} = \underline{c_{\hat{i}}(t) \lambda_{\hat{i}} e^{\lambda_{\hat{i}} t}} + g_{\hat{i}}(t) \, ,$$

or in other words

$$c_{\hat{i}}'(t) = e^{-\lambda_{\hat{i}} t} g_{\hat{i}}(t) \; .$$

Thus we get

$$c_{\hat{i}}(t) - c_{\hat{i}}^{(0)} = \int_0^t c_{\hat{i}}'(\tau) \, d\tau = \int_0^t g_{\hat{i}}(\tau) e^{-\lambda_{\hat{i}} \tau} \, d\tau \; .$$

And thus

$$\begin{aligned} z_{\hat{i}}(t) &= c_{\hat{i}}(t) e^{\lambda_{\hat{i}} t} \\ &= \left(\int_0^t g_{\hat{i}}(\tau) e^{-\lambda_{\hat{i}} \tau} \, d\tau + c_{\hat{i}}^{(0)} \right) e^{\lambda_{\hat{i}} t} \\ &= \int_0^t g_{\hat{i}}(\tau) e^{\lambda_{\hat{i}} (t-\tau)} \, d\tau + c^{(0)} e^{\lambda_{\hat{i}} t} \, , \end{aligned}$$

or in compact vector notation

$$z(t) = \int_0^t g(\tau) e^{\Lambda (t-\tau)} \, d\tau + c_{\hat{i}}^{(0)} e^{\Lambda t} \, ,$$

where here $e^{\Lambda(t-\tau)} = [e^{\lambda_{\hat{i}}(t-\tau)}]_{1 \le \hat{i} \le \nu}$.

Thus, in total we get:

$$\boxed{y(t) = Y z(t) = \int_0^t e^{\Lambda (t-\tau)} f(\tau) \, d\tau + e^{\Lambda t} Y c^{(0)} \; .} \tag{5.19}$$

Let us assume that

$$\Re(\lambda_{\hat{i}}) < 0\,, \quad \forall \hat{i} = 1, 2, \ldots, \nu\,. \tag{5.20}$$

Let $\overline{\lambda}, \underline{\lambda} \in \{\lambda_{\hat{i}}, i = 1, 2, \ldots, n\}$ be the maximal absolute eigenvalues:

$$-\Re\overline{\lambda} = \left|\Re(\overline{\lambda})\right| \geq |\Re(\lambda_{\hat{i}})| \geq |\Re(\underline{\lambda})| = -\Re(\underline{\lambda}), \qquad i = 1, 2, \ldots, n\,.$$

We now define the *stiffness ratio* as

$$\frac{\Re(\overline{\lambda})}{\Re(\underline{\lambda})}.$$

The crux with the siffness ratio is that it is severely affected by the smallest negative eigenvalue (equivalently the one with highest absolute value). This one however, is the best behaving analytically. Surprisingly, it affects the numerics most striking.

Remark 5.7. *The solution of the homogenous equation according to* (5.17) *(that is with $g_{\hat{i}} \equiv 0$) is given by*

$$z_{\hat{i}}(t) = ce^{\lambda_{\hat{i}} t}\,.$$

The name variations of constant for the ansatz (5.18) *is due to the fact that the constant c of the solution of the homogenuous system is replaced by the function $c_{\hat{i}}(t)$. That means that the constant is replaced by a function, that is it is varied now.*

Example 5.8. *Also the Example 5.2 results into a system of stiff ODEs' if n is large.*

5.4.2 A-Stability

The behavior of numerical methods on stiff problems can be analyzed by applying these methods to the test equation

$$y'(t) = \lambda y(t) \text{ with } y(0) = 1 \tag{5.21}$$

for some $\lambda \in \mathbb{C}$. The solution of this equation is $y(t) = e^{\lambda t}$. This solution is monotonically decreasing and approaches zero for $t \to \infty$ when $\Re(\lambda) < 0$.

Definition 5.9. *If the numerical method also exhibits the monotonicity behavior, then the method is said to be A-stable.*

Now, we return to Runge-Kutta methods and define the *stability function*:

Definition 5.10.

$$R : \mathbb{C}\backslash \left\{ \frac{1}{\sigma} : 0 \neq \sigma \in \sigma(A) \right\} \to \mathbb{C}\,,$$

$$\zeta \to 1 + \zeta b^T (\mathbb{1} - \zeta A)^{-1} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

The stability domain of a Runge-Kutta method is defined by

$$\mathcal{S} := \{\zeta : |R(\zeta)| \leq 1\}\ .$$

For applications one should choose the step-size h in such a way that $h\lambda_i \in \mathcal{S}$ for all eigenvalues.

Note also, that the stability function and stability domain only depend on A and b, but **not** *on c.*

Theorem 5.11. *A Runge-Kutta method is A-stable if for given*

$$\zeta \in \mathbb{C}^- := \left\{ \zeta = \zeta^r + \mathrm{i}\zeta^i \in \mathbb{C} : \zeta^r \leq 0 \right\}$$

we have $|R(\zeta)| \leq 1$.

Example 5.12. *For the explicit Euler method we have*

$$R(\zeta) = 1 + \zeta\,,$$

and thus it is not A-stable. Because the set of ζ where

$$|R(\zeta)|^2 = |1 + \zeta|^2 = (1 + \zeta^r)^2 + \zeta^{i2} \leq 1$$

is a circle with center $(-1, 0)$ and radius 1 in the complex plane, the explicit Euler-method is only stable if the step-size h is chosen such that

$$\boxed{h\lambda \in B_{(-1,0)}(1) \text{ (ball of radius } 1 \text{ and center } (-1,0)).}$$

For (5.12) *we can guarantee stability for* $h \leq 1/15$*, which supports the numerical results of Example 5.5.*

If we are considering again a system of ODEs (5.14)*, where the matrix* $\mathcal{A}$ *has eigenvalues* $\{\lambda_i = \lambda_i^r + \mathrm{i}\lambda_i^i : i = 1, \ldots, n\}$*, then for A-stability it is required that*

$$(h\lambda_i^r + 1)^2 + (h\lambda_i^i)^2 \leq 1 \; .$$

Example 5.13. *The implicit Euler method is A-stable, because here* $R(\zeta) = (1-\zeta)^{-1}$ *and*

$$|1-\zeta|^2 = (1-\zeta^r)^2 + \zeta^{i2} = 1 - 2\zeta^r + |\zeta|^2 \geq 1 \text{ for } \zeta^r \leq 0 \; .$$

The step size is **not** *essential for A-stability.*

Example 5.14. *The simplest Gauß-Quadrature formula* ($s = 1$) *is the* implicit mid-point rule*. The according tableau is*

$$\begin{array}{c|c} c = 1/2 & A = 1/2 \\ \hline & b^T = 1 \end{array}$$

and the Runge-Kutta method has the form

$$y_{i+1} = y_i + hf(t_i + h/2, \eta_1) \, , \quad \eta_1 = y_i + \frac{h}{2} f(t_i + h/2, \eta_1) \; . \tag{5.22}$$

By combination of the two equations we get:

$$y_{i+1} = y_i + hf(t_i + h/2, (y_i + y_{i+1})/2) \; .$$

According to the definition of the stability function we have:

$$\begin{aligned} R(\zeta) &= 1 + \zeta b(1 - \zeta A)^{-1} \\ &= 1 + \zeta \left(1 - \frac{\zeta}{2}\right)^{-1} \\ &= \frac{1 + \frac{\zeta}{2}}{1 - \frac{\zeta}{2}} \\ &= 1 + \zeta + \zeta^2/2 + \zeta^3/4 + \ldots \, , \end{aligned}$$

which is a Möbius-function. This function satisfies

$$|R(\zeta)|^2 = \frac{\left(1 + \frac{\zeta^r}{2}\right)^2 + \left(\frac{\zeta^i}{2}\right)^2}{\left(1 - \frac{\zeta^r}{2}\right)^2 + \left(\frac{\zeta^i}{2}\right)^2} \leq 1 \, , \qquad \forall \zeta = \zeta^r + \mathrm{i}\zeta^i \in \mathbb{C}^- \; .$$

That is, the implicit mid-point rule is A-stable. Every choice of the step size is feasible.

5.5 Ill-Conditioned ODE

There exist ODEs for which error and noise significantly influence the solution. Such problems are called *ill–conditioned*, and cannot be cured by a numerical approach. As an illustration we consider the system

$$u_1' = 2u_2 \text{ and } u_2' = 2u_1$$

for which the general solution is

$$u_1 = ae^{2x} + be^{-2x} \text{ and } u_2 = ae^{2x} - be^{-2x} \, .$$

Taking the initial conditions

$$u_1(0) = 3 \text{ and } u_2(0) = -3$$

we have

$$a + b = 3 \text{ and } a - b = -3 \, ,$$

and therefore $a = 0$ and $b = 3$, and the solution of the system is

$$u_1 = 3e^{-2x} \text{ and } u_2 = -3e^{-2x} \, .$$

However, if we put

$$u_1(0) = 3 + \varepsilon \text{ and } u_2(0) = -3 \, ,$$

(assume that ε is some noise), then we have

$$a + b = 3 + \varepsilon \text{ and } a - b = -3 \, ,$$

which gives $a = \frac{\varepsilon}{2}$ and $b = 3 + \frac{\varepsilon}{2}$, and therefore the solution is

$$u_1 = \frac{\varepsilon}{2} e^{2x} + (3 - \varepsilon) e^{-2x} \text{ and } u_2 = \frac{\varepsilon}{2} e^{2x} - \left(3 - \frac{\varepsilon}{2}\right) e^{-2x} \, .$$

For fixed $\varepsilon > 0$ the term $\frac{\varepsilon}{2} e^{2x}$ gets dominant for large x.

Ill-conditioning can also occur for a single first-order ODE: Consider for example

$$y' = 3y - t^2$$

for which the general solution is

$$y = Ce^{3t} + \frac{t^2}{3} + \frac{2t}{9} + \frac{2}{27} \, .$$

If we take as initial condition $y(0) = \frac{2}{27} + \varepsilon$, then $C = \varepsilon$. Again, the term Ce^{3t} gets dominant for large t. Thus for every small error ε the error term will dominate the exact solution.

5.6 Multi-Step Methods

The basic idea consists in approximating the integrand on the right hand side of

$$y(t_{i+l}) = y(t_{i-k}) + \int_{t_{i-k}}^{t_{i+l}} y'(\tau)\, d\tau = y(t_{i-k}) + \int_{t_{i-k}}^{t_{i+l}} f(\tau, y(\tau))\, d\tau$$

over an intervall $[t_{i-k}, t_{i+l}]$. Given some $s \in \mathbb{N}$ let

$$(t_j, f_j) := (t_j, f(t_j, y_j)) \text{ for } j = i-s, i-s+1, \ldots, i\,,$$

where $y_j \approx y(t_j)$, which we assume to be calculated already. The polynomial of degree s interpolating these values is given by

$$P_s(\tau) = \sum_{j=i-s}^{i} f_j L_j(\tau) \text{ with } L_j(\tau) = \prod_{j \neq \hat{j}=i-s}^{i} \frac{\tau - t_{\hat{j}}}{t_j - t_{\hat{j}}}\,.$$

The functions L_j are the basic Lagrange polynomials.

The s-th order multi-step method is defined by

$$y_{i+l} = y_{i-k} + \int_{t_{i-k}}^{t_{i+l}} P_s(\tau)\, d\tau\,.$$

The different methods depend on the choice of s, k and l.

- The s-th order *Adams-Bashford* methods is explicit and $l = 1$ and $k = 0$.
- The s-th order *Adams-Moulton* method is implicit and $l = 0$ and $k = 1$.

We are only studying the first five members of the Adams-Bashford for constant step-size:

Order s	Interpolant	Interpolation points
0	constant	(t_i, f_i)
1	linear	$(t_i, f_i), (t_{i-1}, f_{i-1})$
2	quadratic	$(t_i, f_i), (t_{i-1}, f_{i-1}), (t_{i-2}, f_{i-2})$
2	cubic	$(t_i, f_i), (t_{i-1}, f_{i-1}), (t_{i-2}, f_{i-2}), (t_{i-3}, f_{i-3})$
4	quartic	$(t_i, f_i), (t_{i-1}, f_{i-1}), (t_{i-2}, f_{i-2}), (t_{i-3}, f_{i-3}), (t_{i-4}, f_{i-4})$

- If $s = 0$, $l = 1$ and $k = 0$ then the Adams-Bashfort method satisfies $P_0 = f(t_i, y_i)$ and thus

$$y_{i+1} = y_i + h_i f(t_i, y_i) \text{ with } h_i = t_i - t_{i-1}$$

  is exactly the Euler method.

- If $s = 1$, $l = 1$ and $k = 0$ we have

$$P_1(\tau) = f_{i-1} + \frac{f_i - f_{i-1}}{h_{i-1}}(\tau - t_{i-1}) \; .$$

  Thus we obtain

$$\begin{aligned} \int_{t_i}^{t_{i+1}} f(\tau, y(\tau))\, d\tau &\approx \int_{t_i}^{t_{i+1}} \left(f_{i-1} + \frac{f_i - f_{i-1}}{h_{i-1}}(\tau - t_{i-1}) \right) d\tau \\ &= \frac{h_i}{2} \left(\frac{h_i + 2h_{i-1}}{h_{i-1}} f_i - \frac{h_i}{h_{i-1}} f_{i-1} \right) \; . \end{aligned}$$

  In particular if $h = h_i = h_{i-1}$, then

$$y_{i+1} = y_i + \frac{h}{2}(3f_i - f_{i-1}) \; .$$

The derivation of higher order methods is analogous. Here only the results for constant step-size are summarized:

Order s	Adam-Bashfort
0	$y_{i+1} = y_i + h f_i$
1	$y_{i+1} = y_i + \frac{h}{2}(3f_i - f_{i-1})$
2	$y_{i+1} = y_i + \frac{h}{12}(23f_i - 16f_{i-1} + 5f_{i-2})$
3	$y_{i+1} = y_i + \frac{h}{24}(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3})$
4	$y_{i+1} = y_i + \frac{h}{720}(1901f_i - 2774f_{i-1} + 2616f_{i-2} - 1274f_{i-3} + 251f_{i-4})$

Chapter 6

Statistical Testing

6.1 Probability Space

Definition 6.1. *An experiment is said to be random if it has more than one possible outcome, and deterministic if it has only one. The set of possible outcomes describes the sample space.*

Example 6.2. *Examples of random experiments are:*

Random experiment	*Sample space* Ω
Throwing a coin	$\{H(ead), T(ails)\}$
Lifetime of a lamp	$\mathbb{R}^+$
Throwing a dice three times	$\{(a_1, a_2, a_3) : a_i \in \{1, 2, \ldots, 6\}\}$

Below we describe a random experiment in a formal manner:

Definition 6.3. *A* probability space $(\Omega, \mathcal{A}, \mathbb{P})$ *consists of a sample space* Ω*, a* σ*-algebra* $\mathcal{A}$ *and a probability measure* $\mathbb{P}$*. Subsets of* Ω *are called* events*.*

A σ*-algebra is a set of subsets of* Ω *which satisfies:*

1. $\Omega \in \mathcal{A}$*.*
2. *If* $A \in \mathcal{A}$*, then also the complement, denoted by* A^c*, is in* $\mathcal{A}$*.*
3. *Let* $A_1, A_2, \ldots \in \mathcal{A}$*, then* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$*.*

The mathematical concept of σ*-algebra is rather technical. It is necessary for a rigorous definition, but in practice one can think of the* σ*-algebra as*

all reasonable *subsets of* Ω. *For instance if* Ω *is finite, then the* σ*-algebra consists of all subsets.*

The probability measure *is a function*

$$\mathbb{P} : \mathcal{A} \to [0, 1] ,$$

which satisfies

1. $\mathbb{P}(\Omega) = 1$ *and* $\mathbb{P}(\{\}) = 0$.
2. *Let* $A_1, A_2, \ldots \in \mathcal{A}$ *be pairwise disjoint, then*

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) .$$

In the following we review some important examples of probability measures:

Remark 6.4. *If* Ω *is finite, then* $\mathcal{A} = 2^{\Omega}$, *which denotes all subsets of* Ω.

6.2 Laplace-Experiments

Definition 6.5. *A Laplace-experiment is a probability experiment, with finitely many events, where each event has equal probability. In this case we define*

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} , \qquad \forall A \text{ is event. .}$$

Example 6.6. *Examples of Laplace-experiments are throwing of dices and coins, respectively. As an example, for throwing of dices, we consider the event A of even spots: Then*

$$A = \{2, 4, 6\} \text{ and } \Omega = \{1, 2, \ldots, 6\} .$$

Therefore

$$\mathbb{P}(A) = \frac{1}{2} .$$

In the following we consider general counting principles:

Definition 6.7. *1. Let $k \in \mathbb{N}$ and $n \in N$ satisfying $n \geq k$.* Permutations without repetitions of n elements *are given by the set*

$$\boxed{A_n^k := \{(x_1, x_2, \ldots, x_k) : x_i \in \{1, 2, \ldots, n\} - \{x_1, \ldots, x_{i-1}\}\} \ .}$$

This definition only makes sense for $n \geq k$, because otherwise some x_k would have to be chosen from the empty set.

2. Let $k \in \mathbb{N}$ and $n \in N$. Permutations with repetitions of n elements *are given by the set:*

$$\boxed{B_n^k := \{(x_1, x_2, \ldots, x_k) : x_i \in \{1, 2, \ldots, n\}\} \ .}$$

3. Let $k \in \mathbb{N}_0$ and $n \in N$ satisfying $n \geq k$. Combinations without repetitions of n elements *are given by the set:*

$$\boxed{C_n^k := \left\{(x_1, x_2, \ldots, x_n) : x_i \in \{0, 1\} \ , \sum_{i=1}^{n} x_i = k\right\} \ .}$$

Note, an element of C_n^k is an indicator, like for instance,

$$(1, 1, 0, 1, ...1),$$

which components of a combination are used.

4. $k \in \mathbb{N}_0$ and $n \in N$. Combinations with repetitions of n elements *are given by the set:*

$$\boxed{D_n^k := \left\{(x_1, x_2, \ldots, x_n) : x_i \in \{0, 1, \ldots, k\} \ , \sum_{i=1}^{n} x_i = k\right\} \ .}$$

Theorem 6.8.

$$\left|A_n^k\right| = \frac{n!}{(n-k)!} \ , \left|B_n^k\right| = n^k \ , \left|C_n^k\right| = \binom{n}{k}, \ \textit{and} \ \left|D_n^k\right| = \binom{n+k-1}{k} \ .$$

Proof. While the assertions are relatively easy to prove for $\left|A_n^k\right|, \left|B_n^k\right|, \left|C_n^k\right|$, the proof of the assertion for $\left|D_n^k\right|$ requires an interesting trick: We define the mapping

$$\begin{aligned} \mathcal{L} : D_n^k &\to C_{n+k-1}^k \,. \\ (x_1, \ldots, x_n) &\to (\underbrace{1, \ldots, 1}_{x_1 \times}, 0, \ldots, 0, \underbrace{1, \ldots, 1}_{x_k \times}) \end{aligned}$$

If some $x_i = 0$, it is repeated 0-times, which means it is just left out; nevertheless the separator 0 appears. We illustrate this by an example: Let $n = 4$ and $k = 3$, then

x_1, x_2, x_3	$\mathcal{L}(x_1, x_2, x_3)$
$(1, 1, 1)$	$(1, 0, 1, 0, 1)$
$(1, 0, 2)$	$(1, 0, 0, 1, 1)$
$(0, 3, 0)$	$(0, 1, 1, 1, 0)$

This mapping is one-to-one. And thus $\left|D_n^k\right| = \left|C_{n+k-1}^k\right| = \binom{n+k-1}{k}$. □

Example 6.9. *We consider four combinations:*

1. k **distinguishable** *particles are distributed randomly to n energy states. In this case we assign each particle x_i a state in $\{1, \ldots, n\}$.*

2. k **un-distinguishable** *particles are distributed randomly to n energy states. In this case we assign each state the number of particles in the box.*

   (a) *in each box there can only be* **one** *particle.*

   (b) *in each box there can be* **infinitely** *many particles.*

- *In the case 1a, we find iteratively that*

  $$x_1 \in \{1, \ldots, n\}, x_2 \in \{1, \ldots, n\} \setminus \{x_1\}, \ldots$$

  Therefore we have $n(n-1) \cdot \ldots \cdot (n-k+1)$ possibilities.

- *In the case 1b, we find that $x_i \in \{1, \ldots, n\}$, for all $i = 1, \ldots, k$. Therefore we have n^k possibilities. If all of this possibilities have the same probability, then it is called* **Maxwell-Boltzmann-model**.

- *In the case 2a we have* $x_i \in \{0,1\}$ *and the total elements of particles is* k*, which is equivalent to* $\sum_{i=1}^{n} x_i = k$*. Therefore we have* $\binom{n}{k}$ *possibilities. If all of this possibilities have the same probability, then it is called* **Fermi-Dirac-model**.

- *That is* $x_i \in \{0,1,\ldots,k\}$ *and the total elements of particles is* k*. That is we have* $\left|D_n^k\right| = \binom{n+k-1}{k}$ *possibilities. If all of this possibilities have the same probability, then it is called* **Bose-Einstein-model**.

Example 6.10. *In front of a theater there are queuing up* $2n$ *people. The entrance fee is* 50 *Euro. Each person has either a* 50 *or* 100 *Euro bill available. There are exactly* n *persons with a* 50 *Euro bill and* 50 *with a* 100 *Euro bill. There is no money in the theater counter in the beginning. What is the probability that the entrance is smoothly without requeuing.*

The possible events are identified with the elements of the sets

$$\Omega_n := \left\{ (x_1, x_2, \ldots, x_{2n}) : x_i = \pm 1 \,, \sum_{i=1}^{2n} x_i = 0 \right\} .$$

$x_i = 1$ *means that the person* i *pays with a* 50 *Euro bill and* $x_i = -1$ *means that the person pays with a* 100 *Euro bill.* Ω_n *is isomorphic to the set*

$$\left\{ (x_1, x_2, \ldots, x_{2n}) : x_i \in \{0,1\} \,, \sum_{i=1}^{2n} x_i = n \right\} = C_{2n}^n \,.$$

Thus $|\Omega_n| = \binom{2n}{n}$.

The events where **no** *requeuing is necessary are characterized as follows:*

$$A_n := \left\{ (x_1, x_2, \ldots, x_{2n}) : x_i = \pm 1 \,, \sum_{i=1}^{2n} x_i = 0 \,, \sum_{i=1}^{k} x_i \geq 0 \quad \forall k \leq 2n \right\} .$$

We interpret each x_i *as the slope. For an event that requires requeuing the path meets the height* -1 *at a point* ρ *for the first time (see Figure 6.1). From* ρ *to* $2n$ *we mirror the path at the* -1 *axis. The mirrored path then arrives at* $2n$ *at height* -2*. All paths starting at* 0 *and arriving at* -2 *are*

$$\bar{A}_n := \left\{ (x_1, x_2, \ldots, x_{2n}) : x_i = \pm 1 \,, \sum_{i=1}^{2n} x_i = -2 \right\} .$$

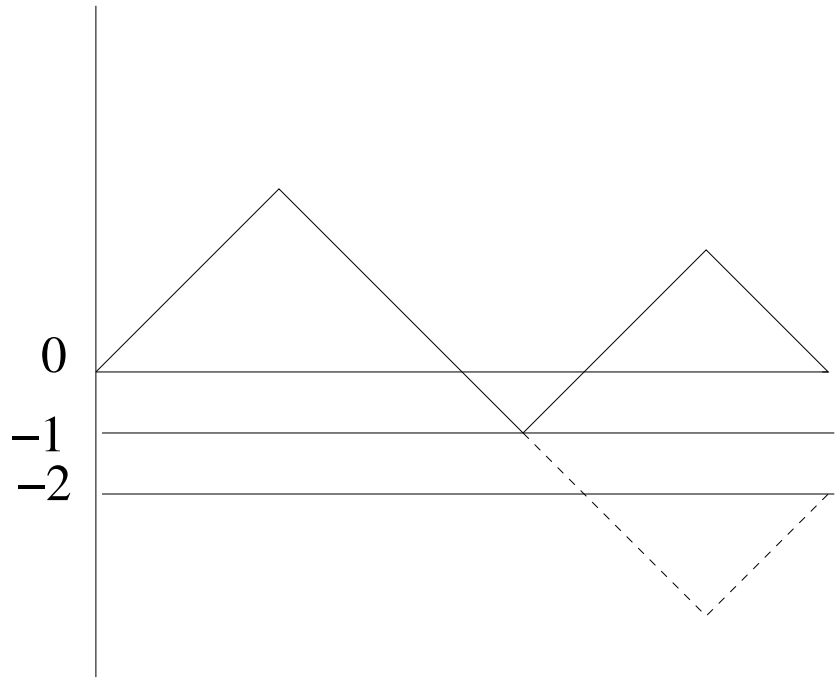

Figure 6.1: As soon as the -1 axis is intersected the path is mirrored around the -1 axis, and it ends up at -2 at $2n$.

To get to the endpoint -2 *we have to have* $n-1$ *values of* $+1$ *and* $n+1$ *values of* -1.

Again, we transform $-1 \to 0$, *and then we have* $n-1$ *values of* 1 *and* $n+1$ *values of* 0. *Thus*

$$\bar{A}_n \cong \left\{ (x_1, x_2, \ldots, x_{2n}) : x_i = \{0,1\} \ , \sum_{i=1}^{2n} x_i = n-1 \right\} \cong C_{2n}^{n-1} \ .$$

Because $A_n = \Omega_n \backslash \bar{A}_n$, *we have* $|A_n| = \binom{2n}{n} - \binom{2n}{n-1}$ *and thus*

$$\mathbb{P}(A_n) = \frac{\binom{2n}{n} - \binom{2n}{n-1}}{\binom{2n}{n}} = \frac{1}{n+1} \ .$$

6.3 Probability Distributions

Until now we have been considering only discrete events. Now we also considering also continuous (geometrical) events.

Example 6.11. *We are testing the life span of lamps. The life span is non-negative, so we use* $\Omega = \mathbb{R}^+$. *We have determined from experiments that the probability that a lamp stays intact in the time interval* $[0, t_e)$ *is*

$$\mathbb{P}([0, t_e)) = \int_0^{t_e} f(\lambda)\, d\lambda\,,$$

where

$$f(\lambda) = \frac{1}{1000} e^{-x/1000}\,.$$

Note that

$$\mathbb{P}([0, +\infty)) = \int_0^{+\infty} f(\lambda)\, d\lambda = 1\,,$$

which means that $\mathbb{P}$ *is a probability measure, in fact. The function* f *is called* probability density.

Let A *denote the event that the life time of the tested lamp is more than* 1000 *hours, that is that the event* A *is* $[1000, +\infty]$. *For this event the probability is given by*

$$\mathbb{P}(A) = \frac{1}{1000} \int_{1000}^{+\infty} e^{-x/1000}\, dx = -e^{-x/1000}\Big|_{1000}^{+\infty} = \frac{1}{e} \approx 0.37\,.$$

Definition 6.12.

- *A measurable function* $X : \Omega \to \mathbb{R}$ *from the sample space to* $\mathbb{R}$ *is called* random variable. *Measurable means that for every* reasonable *set* $A \subseteq \mathbb{R}$

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{A}\,.$$

  X *is called* discrete *is the range of* X, $\mathcal{R}(X)$, *is finite.*

  Note that the definition of measurability is very similar to the definition of continuity. A function $X : \Omega \to \mathbb{R}$ *is called continuous if for every* open *set* $A \subseteq \mathbb{R}$, $X^{-1}(A)$ *is* **open**.

- *A measurable function* $X : \Omega \to \mathbb{R}^n$ *is called* random *or* statistical vector.

- *Given a random variable X the* associated probability distribution *is defined as follows:*

$$\mathbb{P}_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A)) \,, \qquad \forall A \in \mathcal{A} \,.$$

- *If X is discrete, then $\mathbb{P}$ it is called* discrete probability distribution. *In this case we have*

$$\mathbb{P}(X \in A) = \sum_{a \in A} \mathbb{P}(X = a) \,.$$

The map

$$a \to \mathbb{P}(X = a) \tag{6.1}$$

is called discrete probability density *of X.*

- *A random variable X is called absolutely continuous, if there exists a density function $f : \mathbb{R}^n \to [0, \infty)$ such that*

$$\mathbb{P}(X \in A) = \int_A f(\lambda) \, d\lambda \,, \qquad \forall A \in \mathcal{A} \,. \tag{6.2}$$

The formal relation to the discrete setting is as follows: Let X be discrete with range $\{a_1, \ldots, a_n\}$. The, by setting

$$f(x) = \sum_{i=1}^{n} \alpha_i \delta(x - a_i) \,,$$

where δ is a delta-distribution, it follows from (6.2)*:*

$$\mathbb{P}(X \in A) = \sum_{a_i \in A} \alpha_i \,,$$

meaning that $\alpha_i = \mathbb{P}(X = a_i)$

Example 6.13. *We consider the probabilistic experiment of throwing a dice twice. The sample space is*

$$\Omega = \{(i, j) : i, j = 1, 2, \ldots, 6\} \,.$$

Let $\mathcal{A}$ be the union of all subsets of Ω and define

$$\mathbb{P}(A) := \frac{|A|}{|\Omega| = 36} \,, \qquad \forall A \in \mathcal{A} \,.$$

The triple $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space.

For instance, a random variable is

$$X : (i, j) \to i + j \ .$$

Then

$$\left|X^{-1}(\{2\})\right| = |\{(1,1)\} = 1| \ ,$$
$$\left|X^{-1}(\{3\})\right| = |\{(1,2),(2,1)\}| = 2 \, , \ldots .$$

Thus we have $\mathbb{P}(X = 2) = 1/36$, $\mathbb{P}(X = 3) = 1/18, \ldots$.

Example 6.14. *We bring some examples of discrete and continuous distributions:*

Discrete distributions:

- *A random variable X is said to have a* binomial *distribution with parameters $n = \{1, 2, \ldots\}$ and $\theta \in [0, 1]$ if $X(\Omega) = \{0, 1, \ldots, n\}$ and $k \in X(\Omega)$*

$$\mathbb{P}(X = k) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} \ . \tag{6.3}$$

  The formula can be understood as follows: we want k successes (the probability of each success is θ, and thus the probability of k success is θ^k) and $n - k$ failures (probability is $(1 - \theta)^{n-k}$). However, the k successes can occur anywhere among the n trials, and there are $\binom{n}{k}$ different ways of distributing k successes in n trials.

- *A random variable X is said to be* geometrically distributed *if*

$$X(\Omega) = \{1, \ldots, n\} \ \text{ and } \mathbb{P}(X = k) = \theta(1 - \theta)^{k-1} \ .$$

  It is the probability that the first occurrence of success require k number of independent trials, each with success probability θ.

- *A random variable X is said to be* Poisson distributed *if*

$$X(\Omega) = \{0, 1, 2, \ldots\} \ \text{ and } \mathbb{P}(X = k) = e^{-\alpha} \frac{\alpha^k}{k!}, \quad \alpha > 0.$$

Continuous distributions:

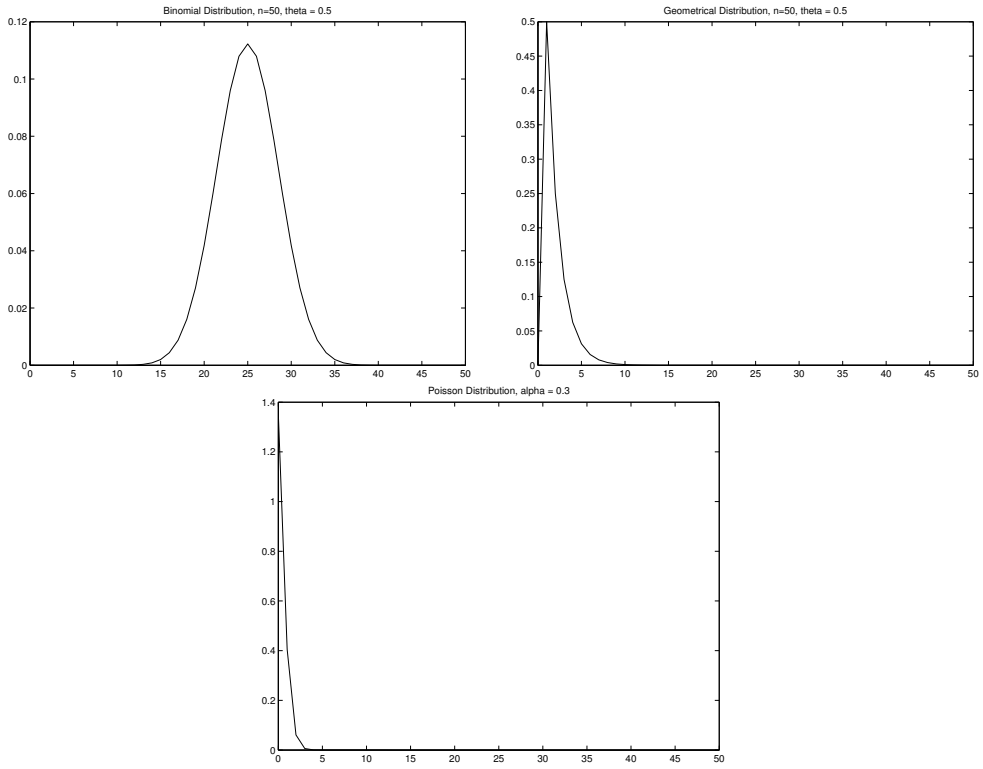

Figure 6.2: Binomial, geometrical, and Poisson distribution

- *The density of the* normal distribution *is given by*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} .$$

Therefore

$$\mathbb{P}(X \in A) = \int_A f(\lambda)\, d\lambda .$$

In this case we say that X is N_{μ,σ^2}-distributed.
Let $A = [0, z]$, $\sigma = 1$, $\mu = 0$, then

$$\mathbb{P}(A) = \frac{1}{\sqrt{2\pi}} \int_0^z \exp\left(-\frac{x^2}{2}\right) dx = \frac{1}{2} erf\left(\frac{z}{\sqrt{2}}\right) ,$$

where erf$(\cdot)$ denotes the error *function. Usually it is implemented in standard software, such as MATHEMATICA and MATLAB, but it is not to be calculated analytically.*
The normal distribution satisfies remarkable properties: Let X be $N(\mu, \sigma^2)$-distributed (that is, the distribution has density f). Then

- *$pX + q$ with $(p \neq 0)$ is $N(p\mu + q, p^2\sigma^2)$ distributed.*
- *$(X - \mu)/\sigma$ is $N(0, 1)$-distributed.*
- *Suppose that $X_1, \ldots, X_n$ are independent random variables, which are $N(\mu_i, \sigma_i^2)$ distributed, then*

$$\hat{X} := X_1 + \ldots + X_n$$

is $N(\mu_1 + \ldots + \mu_n, \sigma_1^2 + \ldots + \sigma_n^2)$ distributed.

- *The density of the* exponential distribution *is given by*

$$f(x) = \begin{cases} 0 & for \quad x \leq 0 , \\ \alpha e^{-\alpha x} & for \quad x > 0 . \end{cases}$$

- *The density of the* logarithmic normal distribution *is given by*

$$f(x) = \begin{cases} 0 & for \quad x \leq 0 , \\ \frac{1}{2\pi\sigma^2}\frac{1}{x} e^{-(\ln x - \mu)^2/2\sigma^2} & for \quad x > 0 . \end{cases}$$

If a random variable Y is log-normal distributed, then $X = \log(Y)$ is normal distributed.

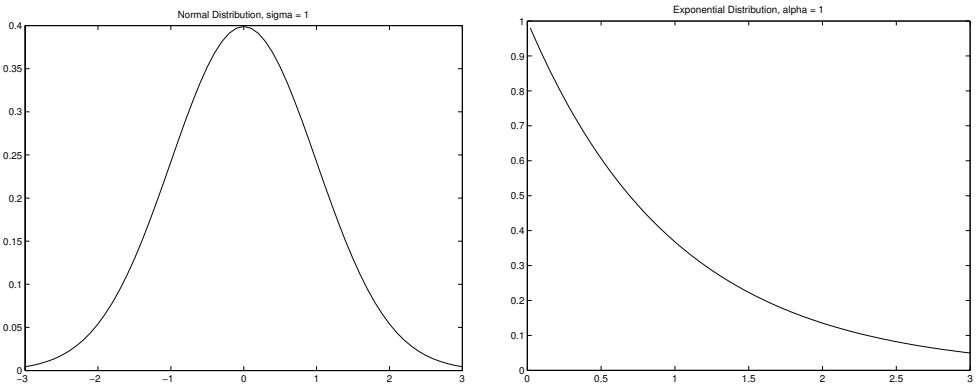

Figure 6.3: Normal and exponential distribution

- *The density of the* χ^2-distribution *is given by*

$$f(x) = \begin{cases} 0 & \text{for} \quad x \leq 0\,, \\ \frac{e^{-x/2}x^{k/2-1}}{2^{k/2}\Gamma(k/2)} & \text{for} \quad x > 0\,, \end{cases}$$

  where Γ *denotes the* Γ*-function. It is the distribution of a sum of the squares of* k *independent normal random variables.*

- *The density of the* uniform distribution *is given by*

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for} \quad x \in [a,b]\,, \\ 0 & \text{else.} \end{cases}$$

6.4 Expectation, Variance and Covariance

Definition 6.15. *Let* X *be an* n*-dimensional random vector with associated probability measure* $\mathbb{P}_X$ *and* $g : \mathbb{R}^n \to \mathbb{R}$ *a function.*

- *If* X *is discrete, then the* expectation *is defined as follows*

$$\boxed{\mathbb{E}[g(X)] := \sum_{a \in \mathcal{R}(X)} g(a)\mathbb{P}_X(a)\,.} \tag{6.4}$$

- *If X is continuous, then the* expectation *is defined as*

$$\boxed{\mathbb{E}[g(X)] := \int_\Omega g(x) d\mathbb{P}_X(x) \, .} \tag{6.5}$$

The integral is defined via a sequence of step functions

$$g^{(m)}(x) = \sum_{i=1}^{l_m} \alpha_i^{(m)} \chi_{A_i^{(m)}}(x) \, ,$$

which satisfy $g^{(m)} \leq g^{(m+1)}$ and which are convergent to g:

$$\int_\Omega g d\mathbb{P}_X = \lim_{m\to\infty} \sum_{i=1}^{l_m} \alpha_i^{(m)} \mathbb{P}_X(A_i^{(m)}) \, .$$

In particular if X is a random variable $(n = 1)$ and $g(x) = x$, then

$$\mathbb{E}[X] := \int_\Omega x d\mathbb{P}_X(x) \, .$$

If in addition the density of an absolutely continuous probability measure $\mathbb{P}_X$ is f, then

$$\mathbb{E}[X] := \int_\Omega x f(x) dx \, .$$

- *The variance is defined as follows: Let X be a random variable. Then*

$$\boxed{var(X) := \mathbb{E}[X^2] - \mathbb{E}[X]^2 \, .}$$

- *Let X and Y be random variables, then*

$$\boxed{Cov(X, Y) := \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]} \tag{6.6}$$

denotes the covariance.

Example 6.16. *We consider again Example 6.13. We have a random Variable $X = X_1 + X_2$, consisting of two random variables X_1 and X_2 for throwing a dice each. Thus we have*

i	$\alpha_i = \lvert X^{-1}(\{i\})\rvert$
$i = 2, \ldots, 7$	$i - 1$
$i = 7, \ldots, 12$	$13 - i$
else	0

Then, according to (6.4)*, the expectation of* X *is given by*

$$\mathbb{E}[X] = \frac{1}{36} \sum_{i=2}^{12} i\alpha_i = 7;$$

Example 6.17. *Let* X *be* N_{μ,σ^2} *be distributed. Then* $\mathbb{E}[X] = \mu$ *and* $var(X) = \sigma^2$*.*

6.5 Statistical Terminology

Definition 6.18. • *A probability measure* $\mathbb{P}$ *on* $\mathbb{R}^n$ *is said to be the* product *of* $\mathbb{P}_1, \ldots, \mathbb{P}_n$ *if*

$$\mathbb{P}(A_1 \times A_2 \times \ldots \times A_n) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2) \ldots \mathbb{P}_n(A_n)\,, \quad \forall A_i \subseteq \mathbb{R}\,.$$

Let $X_1, \ldots, X_n$ *be random variables on* Ω*, then* $\vec{X} = (X_1, \ldots, X_n)$ *is called an* n*-dimensional random variable on* Ω*. In this case*

$$\mathbb{P}_{\vec{X}}(A) := \mathbb{P}_{X_1,\ldots,X_n}(A) := \mathbb{P}(\vec{X}^{-1}(A))\,, \quad \forall A \in \mathbb{R}^n\,.$$

- *Random variables* $X_1, \ldots, X_n$ *are* statistically independent *if*

$$\mathbb{P}_{X_1,\ldots,X_n} = \mathbb{P}_{X_1} \cdots \mathbb{P}_{X_n}\,.$$

- *Suppose* $X_1, \ldots, X_n$ *are random variables with probability densities* $f_{X_1}, \ldots, f_{X_n}$*. The system is statistically independent if and only if the stochastic vector* $(X_1, \ldots, X_n)$ *has probability density* $f_{X_1} \cdots f_{X_n}$*.*

- *Suppose that* X *and* Y *are stochastically independent. Then* $Cov(X, Y) = 0$*.*

Definition 6.19. • *A* sample of size n *is a statistically independent sequence* $X_1, \ldots, X_n$ *of random variables, all of them identically distributed. The common probability distribution of the* X_i *is called* probability distribution of the sample*.*

Let $g : \mathbb{R}^n \to \mathbb{R}$ *be a function, then* $g(X_1, \ldots, X_n)$ *is called* statistics*.*

- *Concrete realizations (that are outcomes of experiments) are denoted by $x_1, \ldots, x_n$. Note that samples are written with capitals and realizations with small letters.*

Example 6.20. *A statistics of fundamental importance is the* sample mean*:*

$$\boxed{\overline{X} = \frac{X_1 + \ldots + X_n}{n} \, .} \tag{6.7}$$

Definition 6.21. *For every sample $X_1, \ldots, X_n$ of size $n \geq 2$ the statistics*

$$\boxed{S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2} \tag{6.8}$$

is called sample variance*.*

Proposition 6.22. *If $X_1, \ldots, X_n$ is a sample from a population with mean μ and variance σ^2, then*

$$\mathbb{E}[S^2] = \sigma^2 \, . \tag{6.9}$$

Proof. First, we note that

$$\begin{aligned} \mathbb{E}[\overline{X}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] &= \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu \, . \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}\left[(\overline{X} - \mu)^2\right] = \mathrm{var}[\overline{X}] &= \mathrm{var}\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2} \mathrm{var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{var}[X_i] = \frac{\sigma^2}{n} \, . \end{aligned} \tag{6.10}$$

Now, we have

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}((X_i-\overline{X})+(\overline{X}-\mu))^2\\ &= \sum_{i=1}^{n}(X_i-\overline{X})^2+2(\overline{X}-\mu)\underbrace{\sum_{i=1}^{n}(X_i-\overline{X})}_{=0}+\sum_{i=1}^{n}(\overline{X}-\mu)^2\\ &= \sum_{i=1}^{n}(X_i-\overline{X})^2+n(\overline{X}-\mu)^2\ .\end{aligned}$$

Thus, it follows from (6.10)

$$\begin{aligned}n\sigma^2 &= \sum_{i=1}^{n}\mathbb{E}\left[(X_i-\mu)^2\right]=\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]\\ &= \mathbb{E}\left[\sum_{i=1}^{n}(X_i-\overline{X})^2\right]+n\underbrace{\mathbb{E}\left[(\overline{X}-\mu)^2\right]}_{\sigma^2/n},\end{aligned}$$

which gives the assertion. □

Note that the expectation of S^2 is not dependent on n and therefore it is called *unbiased* estimator.

6.6 Maximum Likelihood Estimation

Definition 6.23. *Let $X_1,\dots,X_n$ be a sample from a population with density f. The* likelihood function *associated with this sample is the probability of the n-vector $(X_1,\dots,X_n)$. That is, the likelihood function is*

$$L(x_1,\dots,x_n) := f(x_1)\cdots f(x_n)\,,\qquad \forall (x_1,\dots,x_n)\in\mathbb{R}^n\ . \tag{6.11}$$

The according probability is defined as

$$\mathbb{P}((X_1,\dots,X_n)\in A)=\int_A L(x_1,\dots,x_n)\,dx_1\cdots dx_n\ .$$

We draw a sample $X_1, \ldots, X_n$ from a population with probability density $f(\cdot, \theta)$, where $\theta \in \Theta$ is a parameter (for instance (μ, σ^2) if the sample is normally distributed). The likelihood depends now on θ and therefore we write now L_θ instead of L.

Definition 6.24. *The experiment results in an outcome* $(x_1, \ldots, x_n)$ *of* $X_1, \ldots, X_n$. Maximum likelihood estimation *consists in choosing an element*

$$\boxed{\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n) \in \Theta\,,} \tag{6.12}$$

which maximizes the function

$$\boxed{\theta \to L_\theta(x_1, \ldots, x_n)\,.} \tag{6.13}$$

Example 6.25. • *Given an exponentially distributed sample with parameter* θ. *A sample* $(X_1, \ldots, X_n)$ *results in an output* $(x_1, \ldots, x_n)$. *We construct the maximum likelihood estimation. The probability density is given by*

$$f(x; \theta) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & \text{if } x \geq 0\,, \\ 0 & \text{else}\,. \end{cases}$$

Consequently, the likelihood function is given by

$$L_\theta(x_1, \ldots, x_n) = \frac{1}{\theta^n} e^{-(x_1 + x_2 + \ldots + x_n)/\theta}\,.$$

Let us assume that we know from tests that $x_i > 0$ *for all* $i = 1, 2, \ldots, n$. *By differentiation of* L_θ *with respect to* θ *we derive the optimality condition*

$$0 = \frac{\partial}{\partial \theta} L_\theta(x_1, \ldots, x_n) = \left(\frac{x_1 + \ldots + x_n}{\theta^{n+2}} - \frac{n}{\theta^{n+1}} \right) e^{-(x_1 + x_2 + \ldots + x_n)/\theta}\,.$$

The solution of this equation is

$$\boxed{\hat{\theta} = \frac{x_1 + \ldots + x_n}{n}\,.}$$

Therefore, the maximum likelihood estimator is given by

$$\boxed{\hat{\theta}(X_1, \ldots, X_n) = \overline{X}\,,}$$

where $\overline{X}$ *is the sample means* (6.7).

- *We draw a sample $X_1, \ldots, X_n$ of $N(\mu, \sigma^2)$-distributed population, where μ and σ are unknown. Given the outcome $(x_1, \ldots, x_n)$ of the sample $(X_1, \ldots, X_n)$ we wish to make the maximum likelihood estimation of the 2-vector (μ, σ^2). The probability density is*

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, \quad \forall \sigma > 0 .$$

Consequently, the likelihood function is given by

$$L_{\mu,\sigma^2}(x_1, \ldots, x_n) := \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\sum_{i=1}^n (x_i-\mu)^2/(2\sigma^2)} . \tag{6.14}$$

Maximimization of L_{μ,σ^2} is equivalent to maximizing

$$\begin{aligned} &\log L_{\mu,\sigma}(x_1, \ldots, x_n) \\ &= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2} . \end{aligned}$$

Differentiation with respect to μ and σ^2 gives

$$\begin{aligned} \frac{\partial}{\partial\mu} \log L_{\mu,\sigma^2}(x_1, \ldots, x_n) &= \frac{(x_1 + x_2 + \ldots + x_n) - n\mu}{\sigma^2} , \\ \frac{\partial}{\partial\sigma} \log L_{\mu,\sigma^2}(x_1, \ldots, x_n) &= -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (x_i-\mu)^2}{\sigma^3} . \end{aligned}$$

These derivatives vanish for

$$\begin{aligned} \mu = \overline{x} &= \frac{x_1 + x_2 + \ldots + x_n}{n} , \\ \sigma^2 &= \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2 . \end{aligned}$$

In summary, we have shown that the maximum likelihood estimator is given by

$$\boxed{\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n) = \left(\overline{x}, \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2\right) .}$$

Therefore, the maximum likelihood estimator is given by

$$\boxed{\hat{\theta}(X_1, \ldots, X_n) = \left(\overline{X}, \frac{n-1}{n} S^2\right) ,}$$

where $\overline{X}$ and S^2 are the sample mean (6.7) *and variance* (6.8)*, respectively.*

6.7 Decision Making with Tests

A *statistical hypothesis test* is a method of making decisions using data from a scientific study.

Example 6.26. *The first example concerns a clairvoyant test. A person is shown* 25*-times the backside of cards, which are randomly chosen. The person is asked about the colors, clubs (Kreuz), spade (Pik), hearts (Herz), square (Karo). The number of hits is denoted by the random variable* X*.*

- *We formulate the*

  null hypothesis H_0*, that the person is* not *a clairvoyant, and*

  the alternative hypothesis H_1 *that the person is a clairvoyant.*

  For each card the probability to guess the right color is $p = 1/4$*.*

- *For* $c = 0, 1, 2, \ldots, 25$ *the probability that the person guesses exactly* $k = 0, 1, \ldots, n = 25$ *times right is given by the bonomial distribution* (6.3)*:*

$$B(k \mid p, n) := \binom{n}{k} p^k (1-p)^{n-k}, \quad \forall k = 0, 1, \ldots, n \, .$$

  In particular we have for $B(k|1/4, 25)$

$k = 0$	1	2	3	4	5	6
0.0008	0.0063	0.0251	0.0641	0.1175	0.1645	0.1828
7	8	9	10	11	12	13
0.1654	0.1241	0.0781	0.0417	0.0189	0.0074	0.0025
14	15	16	17	18	19	20
0.0007	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
21	22	23	24	25		
0.0000	0.0000	0.0000	0.0000	0.0000		

  As a consequence the probability that the person guesses more than c *times right is*

$$\mathbb{P}(X \geq c) = \sum_{i=c}^{25} \binom{n}{i} \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{n-i} \, .$$

We summarize the probabilities $\mathbb{P}(X \geq c)$ *in the table:*

$c = 0$	1	2	3	4	5	6
1	0.99	0.99	0.97	0.90	0.79	0.62
7	8	9	10	11	12	$13 - 25$
0.44	0.27	0.15	0.07	0.03	0.00	0.00

- *Before the test we define the* probability of an error of the decision. *Typical values for decision making are between* 1% *and* 5%. *We accept the alternative references if the person guesses* k_0 *times right, where* k_0 *is chosen to satisfy* $\mathbb{P}(X \geq k_0) \leq 0.05 = 5\%$. *This means, that if the person guesses less than* 12 *right, he is not considered a clairvoyant.*

There are two typical errors in statistical tests:

Definition 6.27. *First order error: If* H_0 *is correct and the decision is made for* H_1. *This would occur in the above example if a person that is not a clairvoyant but guesses at least* 13 *times right.*

Second order error: If H_0 *is false and the decision is made for* H_0. *This would mean that the person is a clairvoyant, but not considered as such. For instance faking the guesses by giving systematically wrong answers.*

We are now specific and consider particular hypotheses:

Definition 6.28. *A* statistical hypothesis H *is a conjecture about the probability distribution.*

In many practical applications it is assumed that the data is normally distributed. Then the testing is for the values of the mean μ and/or the variance σ^2.

The following example formulates the hypothesis in a testing framework:

Example 6.29. *A typical question is whether a population is* $N(500, 50)$*-distributed or the population is* $N(490, 50)$*-distributed?*

We draw two samples (X_1, X_2) *and specify to accept that the population is* $N(500, 50)$*-distributed if either one of the realizations* x_1 *or* x_2 *of the samples* X_1, X_2, *respectively, is greater than* 496. *Otherwise we accept* H_1. *Setting*

$$G := \big\{(x_1, x_2) \in \mathbb{R}^2 : x_1, x_2 \leq 496\big\}$$

we can restate the decision rule as follows: if $(x_1, x_2) \in G$ *then* H_1 *is accepted and if* $(x_1, x_2) \notin G$ *then* H_0 *is accepted.*

For this example it is possible to calculate analytically first and second order errors:

$$\begin{aligned}
\alpha =& \mathbb{P}(\textit{acceptance of } H_1 \mid H_0 \textit{ is true}) \\
=& \mathbb{P}((X_1, X_2) \in G \mid \mu = 500, \sigma^2 = 50) \\
=& \frac{1}{100\pi} \left(\int_{-\infty}^{496} \exp\left(-\frac{(s-500)^2}{100} \right) \, ds \right)^2 \\
\underbrace{=}_{s-500=-\sqrt{50}t} & \frac{50}{100\pi} \left(\int_{4/\sqrt{50}}^{\infty} \exp\left(-\frac{t^2}{2} \right) \, dt \right)^2 \\
=& \left(\frac{1}{\sqrt{2\pi}} \int_{4/\sqrt{50}}^{\infty} \exp\left(-\frac{t^2}{2} \right) \, dt \right)^2 \\
=& \left(\frac{1}{2} - \frac{1}{2} \textit{erf}(0.4) \right)^2 \\
=& 0.817
\end{aligned}$$

The second order error is determined as follows

$$\begin{aligned}
\beta =& \mathbb{P}(\textit{acceptance of } H_0 \mid H_1 \textit{ is true}) \\
=& \mathbb{P}((X_1, X_2) \notin G \mid \mu = 490\,,\ \sigma^2 = 50) \\
=& 1 - \frac{1}{100\pi} \left(\int_{-\infty}^{496} \exp\left(-\frac{(s-490)^2}{100} \right) \, ds \right)^2 \\
=& 0.356\ .
\end{aligned}$$

We summarize the essential new terms in an abstract definition.

Definition 6.30. *A* hypothesis test *is an ordered sequence*

$$(X_1, \ldots, X_n; H_0, H_1, G)\,,$$

where $X_1, \ldots, X_n$ *is a sample,* H_0 *and* H_1 *are hypotheses, and* G *is a critical set. The* level of significance *of the hypothesis test is the number*

$$\boxed{\alpha = \mathbb{P}^{H_0}_{X_1,\ldots,X_n}(G) = \mathbb{P}((X_1, \ldots, X_n) \in G \mid H_0)\ .} \tag{6.15}$$

We shall say that the critical region is of size α[1]. *Note, that this is the probability of first order errors.*

Definition 6.31. Distribution Tests: *From now on let*

$$(f(\cdot,\theta))_{\theta\in\Theta}$$

be a family of probability densities. Moreover, we shall assume that H_0 *and* H_1 *are statements of the type:*

$$\boxed{H_0 : \theta \in \Theta_0 \text{ and } H_1 : \theta \in \Theta_1\,,}$$

where

$$\Theta = \Theta_1 \cup \Theta_0 \text{ and } \Theta_0 \cap \Theta_1 = \{\}\ .$$

Let

$$\begin{aligned}\alpha(\theta) &:= \mathbb{P}^{\theta}_{X_1,\ldots,X_n}(G) := \mathbb{P}((X_1,\ldots,X_n)\in G \mid \theta)\,, \qquad \forall\theta\in\Theta_0\,,\\ \beta(\theta) &:= \mathbb{P}^{\theta}_{X_1,\ldots,X_n}(G^c) := \mathbb{P}((X_1,\ldots,X_n)\in G^c \mid \theta)\,, \qquad \forall\theta\in\Theta_1\ .\end{aligned}$$

then $\theta \to 1-\beta(\theta)$ *denotes the* power function.[2]

6.8 Regression

Regression models involve the following variables:

1. The unknown parameters, denoted as $\beta \in \mathbb{R}^k$, which may represent a scalar or a vector.
2. The independent variables, $\vec{X} \in \mathbb{R}^n$.
3. The dependent variable, $\vec{Y} \in \mathbb{R}^n$.

A regression model relates Y to a function of X and β:

$$Y \approx f(X,\beta)\ . \tag{6.16}$$

To carry out regression analysis, the form of the function f must be specified. Sometimes the form of this function is based on knowledge about the

[1]The superscipt H_0 means that it is a conditional probability under the assumption that H_0 is true.

[2]Typical values are $\alpha = 0.05$ or $\beta < 0.5$.

relationship between Y and X that does not rely on the data. If no such knowledge is available, a flexible or convenient form for f is chosen.

In order to perform a regression analysis the user must provide information about the dependent variable Y:

- If n data points of the form (Y, X) are observed, where $n < k$, most classical approaches to regression analysis cannot be performed: since the system of equations defining the regression model is under-determined, there are not enough data to recover β.
- If exactly $N = k$ data points are observed, and the function f is linear, the equations $Y = f(X, \beta)$ can be solved exactly rather than approximately. This reduces to solving a set of n equations with n unknowns (the elements of β), which has a unique solution as long as the X are linearly independent. If f is nonlinear, a solution may not exist, or many solutions may exist.
- The most common situation is where $n > k$ data points are observed. In this case, there is enough information in the data to estimate a unique value for β that best fits the data in some sense, and the regression model when applied to the data can be viewed as an overdetermined system in β.

In the last case, the regression analysis provides the tools for finding a solution for unknown parameters β that will, for example, minimize the distance between the measured and predicted values of the dependent variable Y (also known as method of least squares).

Example 6.32. *A person is carrying out n measurements and summarizes the results in the following table:*

x	x_1	x_2	$\dots$	x_n
y	y_1	y_2	$\dots$	y_n

Theoretically, if there are no read off errors, there can be expected a linear relation between x and y:

$$y = ax \ .$$

We wish to find out an estimate $\hat{a}$ for the number a.

We denote the error in the estimation by

$$e_i := y_i - \hat{a}x_i =: y_i - \hat{y}_i \ .$$

We therefore seek for an estimation $\hat{a}$ which minimizes the expression $\sum_{i=1}^{n} |e_i|$, or something similar. The problem with this approach is that

$$f : a \to \sum_{i=1}^{n} |y_i - ax_i|$$

is not differentiable with respect to a. This method is actually called robust *regression, and is a very appealing method for estimation if there are noise components with high outliers.*

It is more common to minimize the quadratic functional

$$f : a \to \sum_{i=1}^{n} |y_i - ax_i|^2 \ .$$

Differentiation with respect to a shows that $\hat{a}$ satisfies

$$f'(\hat{a}) = -2 \sum_{i=1} (y_i - \hat{a}x_i)x_i \, ,$$

and thus

$$\hat{a} = \sum_{i=1}^{n} y_i x_i / \sum_{i=1}^{n} x_i^2 \, ,$$

which is called the least squares estimator.

The general linear regression model one assumes that

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \, , \qquad \forall i = 1, \ldots, n \, ,$$

where x_{ij} is the i-th observation on the j-th independent variable.

Writing this in matrix notation we get

$$\vec{y} = X\vec{\beta} \, .$$

The least squares estimator $\hat{\beta}$ minimizes the functional

$$f : \beta \to \left\| \vec{y} - X\vec{\beta} \right\|_2^2 \ .$$

$\hat{\beta}$ then satisfies the *normal equation*

$$X^t X \vec{\beta} = X^t \vec{y} \, . \tag{6.17}$$

Note that now $X^tX \in \mathbb{R}^{k\times k}$ is quadratic and symmetric. In this case we can explicitly write down the solution:

$$\beta = (X^tX)^{-1}X^tY \ .$$

It is important to confirm the goodness of the fit. Typical test include the R-squared, analysis of the residuals and hypotheses testing.

Remark 6.33. *Numerical the normal equation (6.17) can be solved with Cholesky factorization and QR-decomposition.*

For the QR decomposition we use that $\hat{\beta}$ is the solution of the equation

$$R\beta = Q^T\vec{y},$$

where $X = QR$ is the QR decomposition.

A good introduction to mathematical statistics is [10].

Bibliography

[1] P. Deuflhard and A. Hohmann. *Numerische Mathematik I. Eine algorithmisch orientierte Einführung*. De Gruyter, Berlin, 1993. 2., überarb. Aufl.

[2] L.V. Fausett. *Numerical Methods, Algorithms and Applications*. Pearson Education, Upper Saddle River, NJ, 2003.

[3] G. H. Golub and J. M. Ortega. *Wissenschaftliches Rechnen und Differentialgleichungen*. Berliner Studienreihe zur Mathematik. Heldermann Verlag, Berlin, 1995.

[4] G. H. Golub and Ch. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 1996. third edition.

[5] G. Haemmerlin and K.-H. Hoffmann. *Numerische Mathematik*. Springer Verlag, Berlin, Heidelberg, New York, fourth edition, 1994.

[6] M. Hanke. *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Teubner, Stuttgart, Leipzig, Wiesbaden, 2002.

[7] H. Heuser. *Gewöhnliche Differentialgleichungen*. Teubner, Stuttgart, second edition, 1991.

[8] N. J. Higham. *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.

[9] C.F. Loan. *Introduction to scientific computing*. Prentice Hall, Upper Saddle River, NJ, 2nd edition, 2000.

[10] W. R. Pestman. *Mathematical statistics*. de Gruyter Textbook. Walter de Gruyter & Co., Berlin, second edition, 2009.

[11] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*. Springer Verlag, Berlin, 2000.

[12] H. R. Schwarz. *Numerische Mathematik*. B. G. Teubner, Stuttgart, fourth edition, 1997. With a contribution by Jörg Waldvogel.

[13] J. Stoer. *Numerische Mathematik 1*. Springer Verlag, Berlin, 1999.

[14] G. Strang. Wavelet transforms versus Fourier transforms. *Bull. Amer. Math. Soc.*, 28(2):288–305, 1993.