Numerical Mathematics

Patrick Erik Bradley

Lecture Notes for the course
in the master programmes
Geodesy and Geoinformatics
and
Remote Sensing and Geoinformatics

Karlsruhe, 2019

Preface

These lecture notes are designed as an accompanying text for the course *Numerical Mathematics* in the master programs *Geodesy and Geoinformatics* and *Remote Sensing and Geoinformatics*. The topics treated here are common in a usual course on numerical mathematics, and are supplemented with example applications from the world of geodesy and (geo-)informatics. Here, the internet plays a prominent role by including google's PageRank, and cryptography. Furthermore, inspired by the book *The Simpsons and their Mathematical Secrets* by Simon Singh, related mathematical topics from the animated sitcoms *The Simpsons* and *Futurama* are treated, too. I would like to express my particular gratitude to Father Nil from the Skete of Saint Spiridon in Geilnau for his invaluable help with the spiritual examples.

Hints on typographical and other errors are welcome. This can lead to an improved further edition.

Karlsruhe, February 5, 2019 P.E. Bradley

Contents

Chapter 1

Floating Point Arithmetic

1.1 Floating Point Numbers

A floating point number

$$a = m \cdot \beta^e$$

consists of a *mantissa* m, a *base* β and an *exponent* e. It is *normalised*, if

$$\beta^{-1} \leq m < 1$$

i.e. if

$$m = 0.\, x_1 x_2 \ldots$$

with $x_1 \neq 0$.

Remark 1.1.1. *Also other normalisations are possible, e.g.*

$$2.597\,\mathrm{E} - 03 = 2.597 \cdot 10^{-3}$$

instead of $0.2597 \cdot 10^{-2}$.

IEEE Standard, Double Precision

A *double precision* number in the im *IEEE standard* is representable as a 64-bit word over the alphabet $\{0, 1\}$:

$$\underbrace{\sigma}_{\text{sign}} \underbrace{a_1 \ldots a_{52}}_{\text{mantissa}} \underbrace{e_0 \ldots e_{10}}_{\text{exponent}}$$

The value assigned to such a word is

$$x = (-1)^\sigma \left(1 + \sum_{i=1}^{51} 2^{-i} a_{52-i}\right) \cdot 2^{e-1023}$$

with

$$e = \sum_{i=0}^{10} e_i 2^i$$

1.2 Overflow and Underflow

Not all real numbers can be represented by a word of finite length. As the mantissa has finite length, rounding errors occur. *Overflow* and *Underflow* occur due to the finite length of the exponent.

Overflow

Overflow means that an arithmetic operation produces a number whose exponent is too big.

Underflow

Underflow means that an arithmetic operation produces a number whose exponent is too small.

Remark 1.2.1. *Overflow always leads to an error message, i.e. it is fatal.*
Underflow yields a number which is almost zero. This means that if the number is set to zero, then the calculation can be continued with this value.

Often, overflow can be elmiminated at the cost of introducing harmless underflows.

Example 1.2.2. *Let*

$$c = \sqrt{a^2 + b^2}$$

with $a = 10^{60}$ *and* $b = 1$ *in a decimal system with* 2*-digit exponent. Here,* a^2 *leads to an overflow. This can be eliminated as follows:*

$$c = s\sqrt{\left(\frac{a}{s}\right)^2 + \left(\frac{b}{s}\right)^2}, \quad s = \max\{|a|, |b|\} = 10^{60}$$

yields

$$c = 10^{60}\sqrt{1^2 + \left(\frac{1}{10^{60}}\right)^2}$$

with the underflow

$$\left(\frac{1}{10^{60}}\right)^2$$

Setting this to zero yields $c = 10^{60}$.

1.3 Rounding Errors

Not every real number can be represented exactly on a computer. E.g.

$$\sqrt{7} = 2.6457513\ldots$$

On a 5-digit decimal calculator, the last digits must be discarded. There are two possibilities:

1. Rounding:
$$\sqrt{7} \approx 2.6458$$
2. Truncation:
$$\sqrt{7} \approx 2.6457$$

Error Bounds

Round up to 5 digits. Then

$$a = X.XXXXY$$

is represented by the number

$$b = X.XXXZ$$

If $Y \geq 5$, then round to the next higher digit, and if $Y < 5$, then truncate. The error satisfies

$$|b - a| \leq 5 \cdot 10^{-5}$$

If the leading digit is $\neq 0$, i.e. $|a| \geq 1$, then

$$\frac{|b-a|}{|a|} \leq 5 \cdot 10^{-5} = \frac{1}{2} \cdot 10^{-4}$$

In general, we have:

1. For rounding to t decimal digits, the relative error is

$$\frac{|b-a|}{|a|} \leq \frac{1}{2} \cdot 10^{-t+1}$$

2. For truncating to t decimal digits, we have:

$$\frac{|b-a|}{|a|} \leq 10^{-t+1}$$

For binary numbers, it holds true that:

$$\frac{|b-a|}{|a|} \leq \begin{cases} 2^{-t} & \text{(Rounding)} \\ 2^{-t+1} & \text{(Truncation)} \end{cases}$$

1.4 Machine Epsilon

Let $b = \text{fl}(a)$ be the machine representation of a real number $a \in \mathbb{R}$. Let ϵ_M be the smallest upper bound for the relative error:

$$\epsilon = \frac{b-a}{a} \quad \text{und} \quad |\epsilon| \leq \epsilon_M$$

In other words:

$$\text{fl}(a) = a(1+\epsilon), \quad |\epsilon| \leq \epsilon_M$$

The number ϵ_M is called *machine epsilon* and is a characteristic of the floating point arithmetic on a given machine.

Remark 1.4.1. *ϵ_M is a bit larger than the largest number x, for which*

$$\text{fl}(1+x) = 1$$

Example 1.4.2. *In a 6-digit binary arithmetic, the number $x = 2^{-7}$ yields:*

$$\text{fl}(1+x) = 1$$

Consequence. *An approximation of ϵ_M is given through the following algorithm:*
Start. $x_0 = 1$
Step n. *If* $\mathrm{fl}(1 + x_{n-1}) \neq 1$, *then set* $x_n := \frac{x_{n-1}}{2}$.

Or in pseudo code:

```
x=1;
while (1 + x != 1)
    x = x/2;
```

1.5 Arithmetic

The result of an arithmetical operation can be represented on a computer only by approximation: E.g. the product of two n-digit numbers can have up to $2n$ digits.

Ideally, the result of a floating point operation is the correct rounding of the expected result. I.e. for the operation $\square$ it should hold true that:

$$\mathrm{fl}(a \square b) = (a \square b)(1 + \epsilon), \quad |\epsilon| \leq \epsilon_M$$

In the IEEE standard this is realised, as long as there is no overflow or underflow.

When calculating differences, there can occur large relative errors.

Example 1.5.1. *Consider the calculation of the difference* $1 - 0.999999$ *in a* 6*-digit decimal arithmetic. If* 7 *digits were possible, then this would yield the correct result* $0.100000 \cdot 10^{-6}$. *However, with only* 6 *digits, we have:*

$$\begin{array}{r} 1.00000 \\ -0.99999 \\ \hline 0.00001 \end{array} = 0.10000 \cdot 10^{-5}$$

The relative error is

$$\frac{\left|0.1 \cdot 10^{-5} - 0.1 \cdot 10^{-6}\right|}{0.1 \cdot 10^{-6}} = 9.9$$

and this is quite large. By using an internal extra seventh digit, this error could have been avoided.

Example 1.5.2. *Consider a* 4*-digit artihmetic. We would like to add* 10.90 *and* 0.009. *The result in this arithmetic is* 10.90, *because the larger number uses two digits before the decimal point which means that the smaller number is truncated after the second digit after the decimal point during the addition.*

1.5.1 Calculation of Sums

Let us generalise the equation

$$\mathrm{fl}(a + b) = (a + b)(1 + \epsilon), \quad |\epsilon| \leq \epsilon_M$$

to sums of the form

$$S_n = \mathrm{fl}(x_1 + x_2 + \cdots + x_n)$$

Here, the result depends on the ordering of summands. We define:

$$\mathrm{fl}(x_1 + x_2 + \cdots + x_n) := \mathrm{fl}(\ldots(\mathrm{fl}(\mathrm{fl}(x_1 + x_2) + x_3)\ldots) + x_n)$$

It holds true that:

$$
\begin{aligned}
S_2 &= \mathrm{fl}(x_1 + x_2) = (x_1 + x_2)(1 + \epsilon_1) = x_1(1 + \epsilon_1) + x_2(1 + \epsilon_1), \quad |\epsilon_1| \leq \epsilon_M \\
S_3 &= \mathrm{fl}(S_2 + x_3) = (S_2 + x_3)(1 + \epsilon_2) \\
&= x_1(1 + \epsilon_1)(1 + \epsilon_2) \\
&+ x_2(1 + \epsilon_1)(1 + \epsilon_2) \\
&+ x_3(1 + \epsilon_2) \\
S_n &= \mathrm{fl}(S_{n-1} + x_n) = (S_{n-1} + x_n)(1 + \epsilon_{n-1}) \\
&= x_1(1 + \epsilon_1)(1 + \epsilon_2) \cdots (1 + \epsilon_{n-1}) \\
&+ x_2(1 + \epsilon_1)(1 + \epsilon_2) \cdots (1 + \epsilon_{n-1}) \\
&+ x_3(1 + \epsilon_2) \cdots (1 + \epsilon_{n-1}) \\
&\dots \\
&+ x_{n-1}(1 + \epsilon_{n-2})(1 + \epsilon_{n-1}) \\
&+ x_n(1 + \epsilon_{n-1}), \qquad\qquad |\epsilon_i| \leq \epsilon_M, \quad i = 1, \dots, n-1
\end{aligned}
$$

Define η_i via

$$
\begin{aligned}
1 + \eta_1 &= (1 + \epsilon_1)(1 + \epsilon_2) \cdots (1 + \epsilon_{n-1}) \\
1 + \eta_2 &= (1 + \epsilon_1)(1 + \epsilon_2) \cdots (1 + \epsilon_{n-1}) \\
1 + \eta_3 &= (1 + \epsilon_2) \cdots (1 + \epsilon_{n-1}) \\
&\dots \\
1 + \eta_{n-1} &= (1 + \epsilon_{n-2})(1 + \epsilon_{n-1}) \\
1 + \eta_n &= 1 + \epsilon_{n-1}
\end{aligned}
$$

Then it holds true that:

$$
S_n = \sum_{i=1}^{n} x_i(1 + \eta_i)
$$

Examination of $1 + \eta_i$

$$
\begin{aligned}
1 + \eta_{n-1} &= (1 + \epsilon_{n-2})(1 + \epsilon_{n-1}) = 1 + \epsilon_{n-2} + \epsilon_{n-1} + \epsilon_{n-2}\epsilon_{n-1} \\
&\simeq 1 + \epsilon_{n-2} + \epsilon_{n-1} \qquad\qquad \text{(in 1. order)}
\end{aligned}
$$

Indeed, because of

$$
|\epsilon_{n-2}\epsilon_{n-1}| \leq \epsilon_M^2
$$

higher-order terms in ϵ_i can be discarded. It follows that:

$$
\eta_{n-1} \simeq \epsilon_{n-2} + \epsilon_{n-1} \qquad \text{(in 1. order)}
$$

resp.

$$
|\eta_{n-1}| \lesssim |\epsilon_{n-2}| + |\epsilon_{n-1}| \leq 2\epsilon_M \qquad \text{(in 1. order)}
$$

In general, we have:

$$
\begin{aligned}
|\eta_1| &\lesssim (n-1)\epsilon_M \\
|\eta_i| &\lesssim (n-i+1)\epsilon_M, \qquad\qquad i = 2, \dots, n
\end{aligned}
$$

The following holds true:

Theorem 1.5.3. *If $n\epsilon_M \leq 0.1$ and $|\epsilon_i| \leq \epsilon_M$ for $i = 1, \ldots, n$, then*

$$(1+\epsilon_1)(1+\epsilon_2)\cdots(1+\epsilon_n) = 1+\eta$$

with

$$\eta \leq 1.06 \cdot n\epsilon_M$$

Thus set

$$\epsilon'_M := 1.06 \cdot \epsilon_M$$

then the approximate bounds become rigourous:

$$|\eta_1| \leq (n-1)\epsilon'_M$$
$$|\eta_i| \leq (n-i+1)\epsilon'_M, i = 2, \ldots, n$$

Example 1.5.4. *The condition $n\epsilon_m \leq 0.1$ for $\epsilon_M = 10^{-15}$ means $n \leq 10^{14}$. A computer which takes for each addition $1\mu s = 10^{-6}$ seconds, needs for the addition of 10^{14} numbers*

$$10^8 s = 3.2 \, years$$

This implies that Theorem 1.5.3 is applicable for practical purposes.

1.6 Fermat's Last Theorem

The Greek mathematician Diophantos of Alexandria wrote in the third century A.D. a mathematical treatise named *Arithmetica*. It is a collection of 300 algebraic equations together with methods for finding solutions. One equation which occurs in this treatise is

$$x^2 + y^2 = z^2$$

where x, y, z are supposed to be positive natural numbers. The solutions are called *Pythagorean triples*. There are infinitely many of those.

Fermat writes in this place in his copy of this treatise on the margin:[1]

> "It is, however, not possible to decompose a cube into 2 cubes, or a bi-square into 2 bi-squares, or generally any power higher than the second, into 2 powers with the same exponent: I have discovered for this a truly wonderful proof, but this margin here is too small to comprise it."

In other words:

Theorem 1.6.1 (Fermat, ca. 1637)**.** *The equation*

$$x^n + y^n = z^n$$

has for $n \geq 3$ and $x, y, z > 0$ no integer solutions.

Strictly speaking, Theorem 1.6.1 is not a theorem, as Fermat does not provide a proof. It is the last in the collection of Fermat's results, without himself publishing a proof, for which a proof was found. Hence, the name *Fermat's Last Theorem*. A first valid proof was published by Andrew Wiles (1995). Up to then, special cases of this theorem, as well as the fact that it suffices to prove the theorem for $n = 4$ and for n an even prime number.

[1] In the original, this note is written in Latin.

Wiles' method can be sketched as follows: since 1990 it was known that for a counter example a, b, c mit $a^n + b^n = c^n$ it follows that the *elliptic curve*[2]

$$y^2 = x(x - a^n)(x + b^n)$$

(a so-called *Frey curve*) cannot be *modular*[3]. On the other hand, there is the *Taniyama-Shimura conjecture*, which says that all elliptic curves defined over $\mathbb{Q}$ are modular. Andrew Wiles and Richard Taylor proved this conjecture for a large class of elliptic curves containing the Frey curve. This yields a contradiction. Hence, there cannot exist a counter example to Fermat's Last Theorem.

Shortly afterwards, the following example appeared in the animated tv-sitcom *The Simpsons*:

Example 1.6.2 (Homer Simpson, 1995)**.**

$$1782^{12} + 1841^{12} = 1922^{12}$$

This is one of the apparitions which Homer Simpson has in the bachground of the scene *Homer*^{3} (Homer to the three) in the episode *Treehouse of Horror VI* when he gets transported to the third dimension. You can verify this "counter example" to Fermat's Last Theorem on a pocket calculator: the 12th root of the left hand side yields indeed 1922, an integer. The reason is the 10-digit floating point arithmetic on the pocket calculator. If you have more digits at your disposal, then you can see that this 12th root is a tiny bit bigger than 1922. Scriptwriter Cohen produced this near miss to Fermat's equation with the help of a computer program. Even without a calculator, one can see that this example must be wrong: the left hand side is the sum of an even with an odd number, hence odd. But the right hand side is even. This is a contradiction.

Three years later, in *The Wizard of Evergreen Terrace*, Homer writes onto his blackboard a prediction of the mass of the Higgs boson (14 years before its discovery), the density of the universe, a strange topological transformation of a donut into a sphere, and also another near miss of a counter example to Fermat's Last Theorem:

Beispiel 1.6.3 (Homer Simpson, 1998)**.**

$$3987^{12} + 4365^{12} = 4472^{12}$$

This can be "verified" again on a pocket calculator. Again, it can be seen that this example must be wrong: the right hand side is the sum of two numbers which are divisible by 3 (check the sum of the digits!), while the right hand side is not. Namely:

$$4472^{12} \equiv 2^{12} \equiv 4^6 \equiv 1 \not\equiv 0 \mod 3$$

[2] cf. Chapter 7.2

[3] This margin is too small to comprise an explanation of this expression...

Chapter 2

Non-linear Equations

2.1 The basic problem

Here, we will treat the basic problem of solving an equation

$$f(x) = 0$$

in one unknown x. It is left open, if all solutions are looked for, or if there are restrictions to the possible solution space.

Example 2.1.1. *Consider the equation $f = 0$ with*

$$f = x^2 - 9$$

This equation has a solution in the integers $\mathbb{Z}$. One method for solving this equation is to use the prime decomposition of integers:

$$9 = 3^2 \quad \Rightarrow \quad x = \pm 3$$

In this example the following was used:

Theorem 2.1.2 (Fundamental Theorem of Arithmetic)**.** *Every positive natural number has a decomposition as a product of prime numbers, which is unique up to the order of the factors.*

This is a pure existence theorem. It is to date an unsolved problem if the prime factor decomposition can be obtained in *polynomial time*.

Example 2.1.3. *Let*

$$f = x^2 - 2$$

Here, $f(x) = 0$ has no solutions in the rational numbers $\mathbb{Q}$. The reason is that $\sqrt{2}$ is irrational. But we can approximate $\sqrt{2}$ with rational numbers. It holds true that

$$f(1) = -1 < 0 \quad \textit{and} \quad f(2) = 2 > 0$$

Hence, by the intermediate value theorem, f has a zero $x \in [1, 2]$. Further, we have

$$f(3/2) = \frac{1}{4} > 0$$

Hence, f has a zero $x \in [1, 3/2]$. This can be continued for ever, where in every step the interval containing x is subdivided in half.

Here, the following was used:

Theorem 2.1.4 (Intermediate Value Theorem)**.** *Let $f\colon [a,b] \to \mathbb{R}$ be a continuous function. Then for every u between $f(a)$ and $f(b)$ there exists some $c \in [a,b]$ with $f(c) = u$.*

Example 2.1.5. *Let*

$$f = x^2 + 1$$

Here, $f(x) = 0$ has no solutions in the real numbers $\mathbb{R}$. However, the following extension of the domain $\mathbb{Q}$ leads to a solution: Let $i := \sqrt{-1}$ (symbolically). Then let

$$\mathbb{Q}(i) := \mathbb{Q} \oplus \mathbb{Q}i$$

be the set of numbers of the form

$$z = x + yi \quad \textit{(symbolically)}$$

The usual rules for addition and multiplication are used under the observance that $i^2 = -1$. The equation $f(z) = 0$ is exactly solvable in $\mathbb{Q}(i)$:

$$z = \pm i$$

Remark 2.1.6. *The method from 2.1.5 also works with $\sqrt{2}$:*

$$\mathbb{Q}(\sqrt{2}) := \mathbb{Q} \oplus \mathbb{Q}\sqrt{2}$$

Then $z^2 - 2 = 0$ is exactly solvable with $z = \pm\sqrt{2}$.

Often used as possible solution spaces are:

1. $\mathbb{R}$ (real numbers)
2. $\mathbb{C} = \mathbb{R} \oplus \mathbb{R}i$ (complex numbers)

The main reason for choosing $\mathbb{C}$ is

Theorem 2.1.7 (Fundamental Theorem of Algebra)**.** *Every polynomial*

$$f(X) = a_0 + a_1 X + \cdots + a_n X^n$$

with coefficients $a_0, \ldots, a_n \in \mathbb{C}$ has a zero in $\mathbb{C}$.

In general, there are further restrictions to the possible solution space. In the case of $\mathbb{R}$, the solution must e.g. be taken in some given interval $[a,b]$. In the case $\mathbb{C}$, the solution space could be e.g. in some domain or disc.

When using a calculator, further restrictions are made. Thus, in a real solution space, the solution is often to be approximated by finite decimal numbers. In the case of complex solutions, one oten approximates the real and imaginary parts with finite decimal numbers.

2.2 Bisection Method

Here, the problem is to solve the equation

$$f(x) = 0$$

for a continuous real valued function f on an interval $[a, b]$.

Let it be assumed that $f(a) \cdot f(b) < 0$. Then a *binary search* can be done. For this, set

$$x_1 := \frac{a+b}{2}, \quad I_0 := [a, b]$$

There are 3 possibilities:

1. If $f(a)f(x_1) < 0$, then set $I_1 := [a, x_1]$.
2. If $f(x_1)f(b) < 0$, then set $I_1 := [x_1, b]$.
3. If $f(x_1) = 0$, then we are done: a solution is $x = x_1$.

By continuing, we obtain a sequence x_n as the midpoint of the interval I_{n-1}.

Theorem 2.2.1. *The sequence x_n converges to*

$$x := \lim_{n\to\infty} x_n$$

and it holds true that

$$f(x) = 0$$

The approximation error for x_n is

$$\epsilon_n := |x_n - x| \leq \frac{b-a}{2^n}$$

and the convergence is linear.

We need a definition. Let $x_n \to x$ be a convergent sequence.

Definition 2.2.2. *x_n converges* with order q, *if there is a $\rho > 0$ such that for all n we have:*

$$|x_{n+1} - x| \leq \rho |x_n - x|^q$$

ρ is called the rate of convergence.

Definition 2.2.3. *x_n is called* R-linearly convergent, *if there is a sequence $\alpha_n > 0$ gibt, which converges with order* 1 *and rate of convergence $\rho \in (0, 1)$ to* 0, *such that for all n it holds true that:*

$$|x_n - x| \leq \alpha_n$$

Proof of Theorem 2.2.1. By the intermediate value theorem (Theorem 2.1.4), each of the intervals I_n contains a zero of f. As the intersection of all these intervals is a point ξ, it follows that ξ is at the same time a zero of f and the limit of the sequence x_n.

The convergence is R-linear because

$$\epsilon_n \leq \kappa_n := \frac{b-a}{2^n} \quad \Rightarrow \quad \lim_{n\to\infty} \kappa_n = 0 \quad \text{and} \quad \kappa_{n+1} \leq \frac{1}{2}\kappa_n$$

□

2.3 Fixed Point Methods

We want to solve the equation

$$f(x) = 0$$

where $f = x^3 - x - 1$, iteratively with

$$x_n = \phi(x_{n-1})$$

Here, let

1. $\phi(x) := (x+1)^{\frac{1}{3}}$
2. $\phi(x) := x^3 - 1$

In each case, the solution is given as a *fixed point*

$$\phi(x) = x$$

By inspectiong Figure 2.1 we choose as starting point $x_0 = 1.5$. The graphs of the iterators ϕ are given in Figure 2.2.

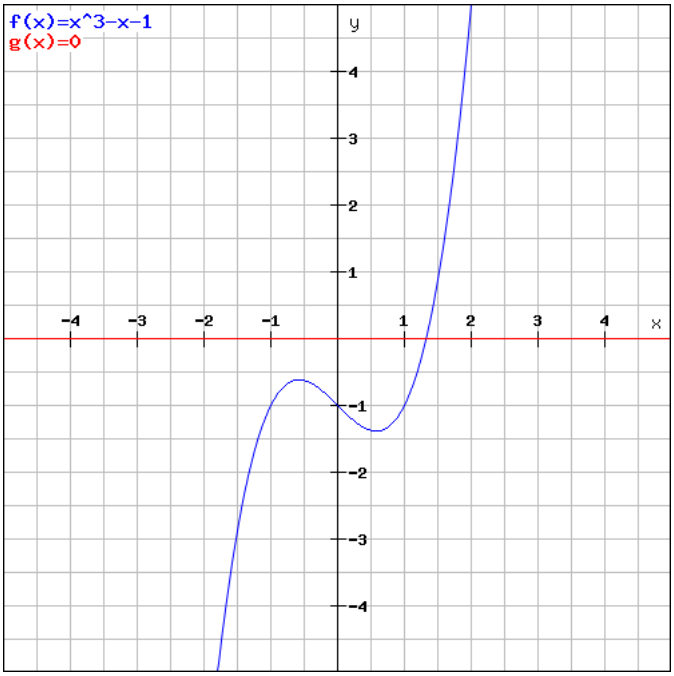


Figure 2.1: The graph of $x^3 - x - 1$.

Let us calculate the first few iterations:

1. $\phi(x) = (x+1)^{\frac{1}{3}}$.

$$x_1 = 1.3572,\ x_2 = 1.3309,\ x_3 = 1.3259,\ x_4 = 1.3249$$

2. $\phi(x) = x^3 - 1$.

$$x_1 = 2.375,\ x_2 = 12.396,\ x_3 = 1904.003,\ x_4 = 6.902 \cdot 10^9$$

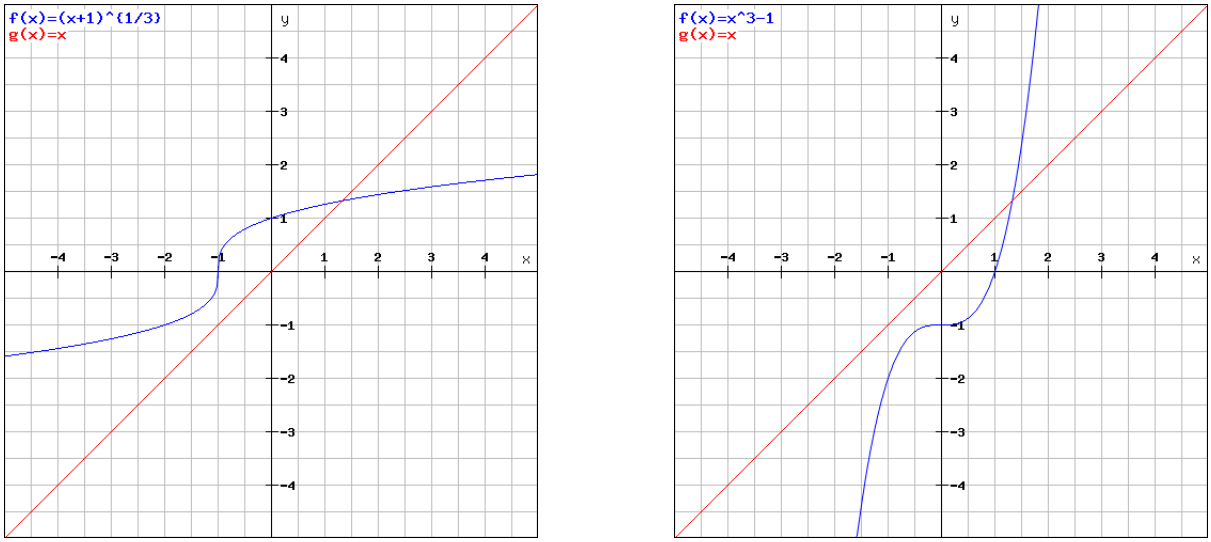


Figure 2.2: The graphs of the iterators ϕ.

The sequence in 1. could possibly converge, whereas the sequence in 2. probably diverges. A method for deciding this, is

Theorem 2.3.1 (Banach Fixed Point Theorem)**.** *Let* (X, d) *be a non-empty* metric space*. If* (X, d) *is* complete *and*

$$T\colon X \to X$$

a contraction*, then* T *has precisely one fixed point in* X*.*

We will now define the emphasized terms.

Definition 2.3.2. *A mapping*

$$d\colon X \times X \to \mathbb{R}$$

is a metric *on* X*, if* $d(x, y) \geq 0$ *for all* $x, y \in X$*, and if for all* $x, y, z \in X$ *it holds true that:*

1. $d(x, y) = 0$ *if and only if* $x = y$. (definiteness)
2. $d(x, y) = d(y, x)$ (symmetry)
3. $d(x, y) \leq d(x, z) + d(z, y)$ (triangle inequality)

Definition 2.3.3. *A metric space is called* complete *if every* Cauchy sequence *converges in* X*.*

Definition 2.3.4. *A sequence* (x_n) *is called a* Cauchy sequence*, if for every* $\epsilon > 0$ *there exists an* $N \in \mathbb{N}$ *such that for all* $m, n > N$ *we have:*

$$d(x_m, x_n) < \epsilon$$

Definition 2.3.5. *A mapping* $T\colon X \to X$ *is called a* contraction*, if there exists a non-negative real number* $L < 1$ *with*

$$d(T(x), T(y)) \leq L \cdot d(x, y)$$

The Banach Fixed Point Theoream (Theorem 2.3.1) can be formulated in a constructive manner:

Theorem 2.3.6. *Let $T\colon X \to X$ be a contraction in a complete metric space. Then the sequence*

$$x_{n+1} = T(x_n)$$

converges for every starting point $x_0 \in X$ to the (unique) fixed point of T.

The limit $x := \lim x_n$ for a contraction T is indeed a fixed point:

Proof.

$$\lim x_n = \lim T(x_{n-1}) \overset{(*)}{=} T(\lim x_{n-1})$$

$(*)$ holds true, because contractions are continuous. Hence, $x = T(x)$. □

The number L for contraction T is called *Lipschitz constant*.

For continuously differentiable real-valued functions on an interval, there is a criterion for being a contraction.

Theorem 2.3.7. *If $I \subset \mathbb{R}$ is a closed interval and $\phi\colon I \to \mathbb{R}$ is continuously differentiable with $\phi(I) \subseteq I$ and*

$$\left|\phi'(x)\right| \leq L < 1$$

for all $x \in I$, then ϕ is a contraction.

Proof. By the mean value theorem (Theorem 2.3.8), for all $x < y$ in I there exists some $\xi \in (x, y)$ with

$$\frac{|\phi(y) - \phi(x)|}{|y - x|} = \left|\phi'(\xi)\right| \leq L < 1$$

Hence, ϕ is a contraction. □

The following was used:

Theorem 2.3.8 (Mean Value Theorem)**.** *Let $f\colon [a, b] \to \mathbb{R}$ be a continuous function which is differentiable in the open interval (a, b). Then there is a $\xi \in (a, b)$ with*

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

Example 2.3.9. *Take $\phi(x) = (x + 1)^{\frac{1}{3}}$. Then*

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

For $I = [1, 2]$ it holds true that:

$$\left|\phi'(x)\right| < 0.21 =: L < 1$$

for $x \in I$ (cf. Figure 2.3).

Example 2.3.10. *For $\phi(x) = x^3 - 1$ we have $\phi'(x) = 3x^2$. With $x_0 = 1.5$ it follows that*

$$\phi'(x_0) = 6.75 > 1$$

Indeed, ϕ is on no interval containing x_0 a contraction.

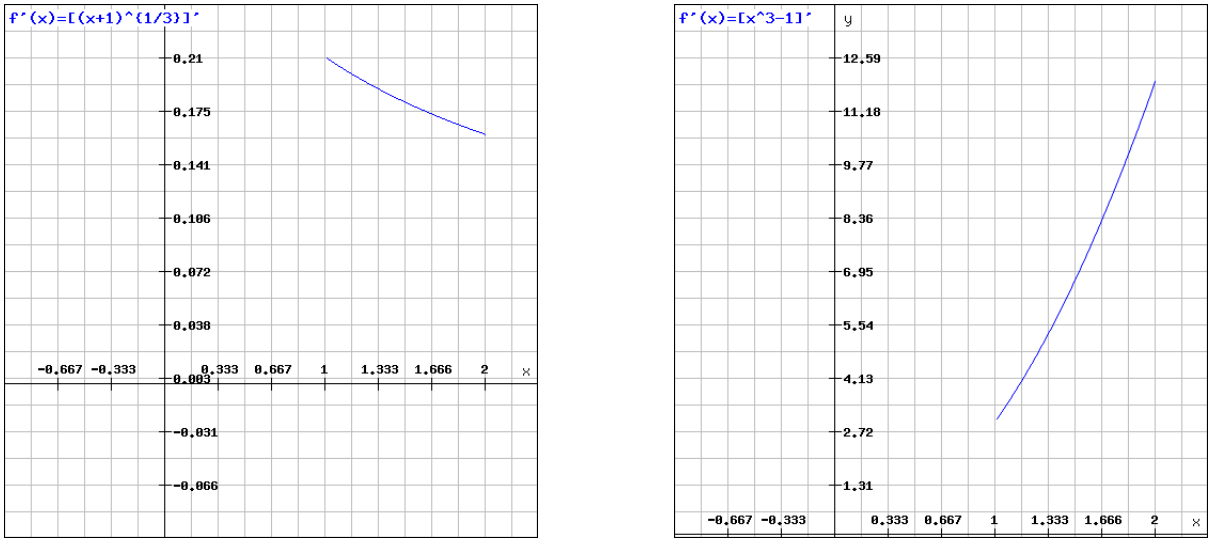


Figure 2.3: The derivatives of the iterators on the interval $[1, 2]$.

2.3.1 Error bounds for Contractions

A priori

Let $T\colon X \to X$ be a contraction with fixed point $x \in X$. For $x_n = T(x_{n-1})$ the error is

$$\epsilon_n := d(x_n, x)$$

We have:

$$d(x_k, x_{k-1}) \leq L \cdot d(x_{k-1}, x_{k-2}) \leq \cdots \leq L^{k-1} \cdot d(x_1, x_0)$$

where L is the Lipschitz constant of T. Further, we have:

$$\begin{aligned} d(x_{m+n}, x_n) &\leq d(x_{m+n}, x_{m+n-1}) + \cdots + d(x_{n+1}, x_n) \\ &\leq (L^{m+n-1} + \cdots + L^n) \cdot d(x_1, x_0) \end{aligned}$$

With $m \to \infty$ we obtain the *a priori error bound*:

$$\epsilon_n \leq \frac{L^n}{1-L} \cdot d(x_1, x_0)$$

Here, the *geometric series*

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}, \quad \text{falls} \quad |x| < 1$$

was used.

From the a priori error bound, the number of iterations can be bounded if the the error may be at most $\epsilon > 0$:

$$\epsilon_n \leq \frac{L^n}{1-L} \cdot d(x_1, x_0) \leq \epsilon \quad \Rightarrow \quad n \geq \frac{\log \frac{\epsilon \cdot (1-L)}{d(x_1, x_0)}}{\log L}$$

A posteriori

For a contraction T with $x_n = T(x_{n-1})$ we have:

$$d(x_{n+k}, x_{n+k-1}) \leq L^k \cdot d(x_n, x_{n-1})$$

This implies:

$$\begin{aligned} d(x_{n+m}, x_n) &\leq d(x_{n+m}, x_{n+m-1}) + \cdots + d(x_{n+1}, x_n) \\ &\leq (L^m + \cdots + L) \cdot d(x_n, x_{n-1}) \end{aligned}$$

With $m \to \infty$ this yields the *a posteriori error bound*:

$$\epsilon_n \leq \frac{L}{1-L} \cdot d(x_n, x_{n-1})$$

A priori vs. a posteriori

The a priori error bound

$$\tilde{\epsilon}_n := \frac{L^n}{1-L} \cdot d(x_1, x_0)$$

can be calculated after the first iteration, whereas the a posteriori error bound

$$\hat{\epsilon}_n := \frac{L}{1-L} \cdot d(x_m, x_{n-1})$$

is only known, once x_n is known.

Theorem 2.3.11. *For contractions, the a posteriori error bound is smaller than the a priori bound.*

Proof. As T is a contraction, we have:

$$\epsilon_n \leq \hat{\epsilon}_n = \frac{L}{1-L} \cdot d(x_n, x_{n-1}) \leq \frac{L}{1-L} \cdot L^{n-1} d(x_1, x_0) = \tilde{\epsilon}_n$$

□

2.4 Newton's Method

Here, we assume that $f\colon I \to \mathbb{R}$ is a twice continuously differentiable function, and that $\xi \in I$ be a *simple* zero of f. This means:

$$f(\xi) = 0 \quad \text{and} \quad f'(\xi) \neq 0$$

Expanding f into a Taylor series in $x_0 \in I$ yields:

$$f(\xi) = f(x_0) + f'(x_0)(\xi - x_0) + R(\xi, x_0)$$

with

$$R(\xi, x_0) = f''(\alpha)\frac{(\xi - x_0)^2}{2}$$

where α is strictly between ξ and x_0. For $|\xi - x_0| \to 0$, we have:

$$R(\xi, x_0) \to 0$$

Hence,

$$0 \approx f(x_0) + f'(x_0)(\xi - x_0)$$

i.e.

$$\xi \approx x_0 - \frac{f(x_0)}{f'(x_0)}$$

where $f'(x_0) \neq 0$ for $|\xi - x_0|$ sufficiently small. This gives us the iterator

$$\phi(x) = x - \frac{f(x)}{f'(x)}$$

with starting point x_0 near the zero ξ.

Lemma 2.4.1. *There exists in I a neighbourhood $U(\xi) = [\xi - h, \xi + h]$ of ξ, in which f' has no zero and in which*

$$\left|\phi'(x)\right| < 1$$

holds true.

Proof. As f' is continuous, there is a neighbourhood $V(\xi)$ of ξ not containing a zero of f'. Now,

$$\phi'(x) = \frac{f(x)f''(x)}{f'(x)^2}$$

is continuous as the product of continuous functions, and

$$\phi'(\xi) = 0$$

Hence, there exists in I a neighbourhood $W(\xi)$ of ξ with $|\phi'| < 1$. Any neighbourhood of ξ inside $V(\xi) \cap W(\xi)$ now satisfies our needs. □

Consequence. *The iterator $\phi\colon U(\xi) \to \mathbb{R}$ is a contraction with Lipschitz constant*

$$L = \max\left\{\left|\phi'(x)\right| \mid x \in U(\xi)\right\} < 1$$

Proof. We need to show that $\phi(U(\xi)) \subseteq U(\xi)$. This follows from:

$$|\phi(x) - \xi| = |\phi(x) - \phi(\xi)| \leq L|x - \xi| < |x - \xi| \leq h$$

where $U(\xi) = \{x \in \mathbb{R} \mid |x - \xi| \leq h\}$. □

We can also say something about the convergence of the Newton iterator:

Theorem 2.4.2. *Let $f\colon [a, b] \to \mathbb{R}$ be twice continuously differentiable with simple zero $\xi \in [a, b]$. Then there exists in $[a, b]$ a neighbourhood $U(\xi)$, such that the iterator ϕ converges quadratically (i.e. with order 2) for every starting point $x_0 \in U(\xi)$.*

Proof. The Taylor series in $x_n \in U(\xi)$ is

$$0 = f(\xi) = f(x_n) + f'(x_n)(\xi - x_n) + \frac{1}{2}f''(\alpha_n)(\xi - x_n)^2$$

As

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

it follows that:

$$\xi - x_{n+1} = \frac{f(x_n)}{f'(x_n)} + (\xi - x_n) = \frac{-f''(\alpha_n)}{2f'(x_n)}(\xi - x_n)^2$$

Hence, for the error $\epsilon_n = |\xi - x_n|$ we have:

$$\epsilon_{n+1} = \underbrace{\left|\frac{-f''(\alpha_n)}{2f'(x_n)}\right|}_{=:C_n} \epsilon_n^2$$

with

$$\lim_{n\to\infty} C_n = \left|\frac{f''(\xi)}{2f'(\xi)}\right|$$

Hence, C_n is bounded. With $\rho = \min\{C_n\}$, we have

$$\epsilon_{n+1} \leq \rho\epsilon_n^2$$

and this means that the order of convergence is 2. □

Let us summarize that Newton's method converges quadratically, but only *locally*, i.e. in a neighbourhood of a zero.

In order to find a suitable starting point, one can e.g. use the bisection method.

2.4.1 Two Applications

Optimization

Let $f\colon I \to \mathbb{R}$ be three times continuously differentiable. The problem here is to find maxima and minima of f. These are zeros of the derivative f'. This yields the Newton iterator

$$\phi(x) = x - \frac{f'(x)}{f''(x)}$$

Newton-Raphson Division

Hier, the problem is to calculate $\frac{1}{D}$ for $D \neq 0$ numerically. For this task, solve the equation

$$f(x) = 0$$

with $f = \frac{1}{x} - D$ and $x \neq 0$. The iterator is

$$\phi(x) = x - \frac{f(x)}{f'(x)} = x(2 - Dx)$$

and uses only multiplication and subtraction.

2.5 Secant Method

Newton's method

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

uses in every step the derivative of the function f at the point x_n. If this is not known or too difficult to compute, the differntial quotient f' can be approximated with the difference quotient:

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

This yields the iteration

$$x_{n+1} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}$$

An advantage is that only function values $f(x)$ are used. A disadvantage is that convergence is slower. The error satisfies:

$$\epsilon_{n+1} \approx C \cdot \epsilon_n^{\alpha}$$

with

$$\alpha = \frac{1+\sqrt{5}}{2} \approx 1.618 < 2$$

2.6 Newton Fractal

Now, we describe the Newton iterator over the complex numbers. For this, let p be a non-constant *meromorphic* function on $\mathbb{C}$, i.e.

$$p = \frac{f}{g}$$

with f, g *holomorphic*. The latter means that those functions can be expanded in every point of $\mathbb{C}$ locally (d.h. in a neighbourhood) into a Taylor series. Then the complex Newton iterator for the equation $p = 0$ is:

$$\phi(z) = z - \frac{p(z)}{p'(z)}$$

The results for the real Newton iterator carry over to $\phi(z)$. Convergence depends on the starting point $z_0 \in \mathbb{C}$.

For the following considerations, we use the abbreviation:

$$\phi^n := \underbrace{\phi \circ \cdots \circ \phi}_{n\text{-mal}}$$

I.e.

$$\phi^n(z) = \underbrace{\phi(\phi(\ldots(\phi}_{n \text{ times}}(z))\ldots)$$

For $z \in \mathbb{C}$ we consider the sequence

$$F_z(w)\colon |z-w|, |\phi(z) - \phi(w)|, \left|\phi^2(z) - \phi^2(w)\right|, \ldots$$

For $F_z(w)_n = |\phi^n(z) - \phi^n(w)|$ there are two possibilities:

1. There is a neighbourhood $U(z)$ of z, such that for all $w \in U(z)$:

$$\lim_{n\to\infty} F_z(w)_n = 0$$

2. In every neighbourhood $U(z)$ there is some w, such that

$$F_z(w)_n \not\to 0 \qquad (n \to \infty)$$

Definition 2.6.1. *The set*

$$\mathcal{F}(\phi) := \{z \in \mathbb{C} \mid 1. \text{ holds true for } z\}$$

is called the Fatou set *of* ϕ*. The set*

$$\mathcal{J}(\phi) := \{z \in \mathbb{C} \mid 2. \text{ holds true for } z\}$$

is called the Julia set *or also the* Newton fractal *of* ϕ*.*

Theorem 2.6.2. *Let* ϕ *be a complex Newton iterator. If the starting point* z_0 *is in* $\mathcal{F}(\phi)$*, then* $\phi^n(z_0)$ *converges to a periodic cycle of finite length. If* $z_0 \in \mathbb{J}(\phi)$*, then* $\phi^n(z_0)$ *does not converge.*

If in the first case, the length of the cycle is 1, then the Newton iterator converges to the fixed point for this particular starting point z_0.

Example 2.6.3. *Let* $p = z^3 - 1$*. This yields the Newton iterator*

$$\phi = \frac{2z^3 + 1}{3z^2}$$

Notice that the solutions of $p = 0$ *are the third roots of unity:*

$$\zeta = e^{\frac{2\pi i}{3}},\ \zeta^2 = e^{\frac{4\pi i}{3}},\ \zeta^3 = 1$$

with $i = \sqrt{-1} \in \mathbb{C}$*.*

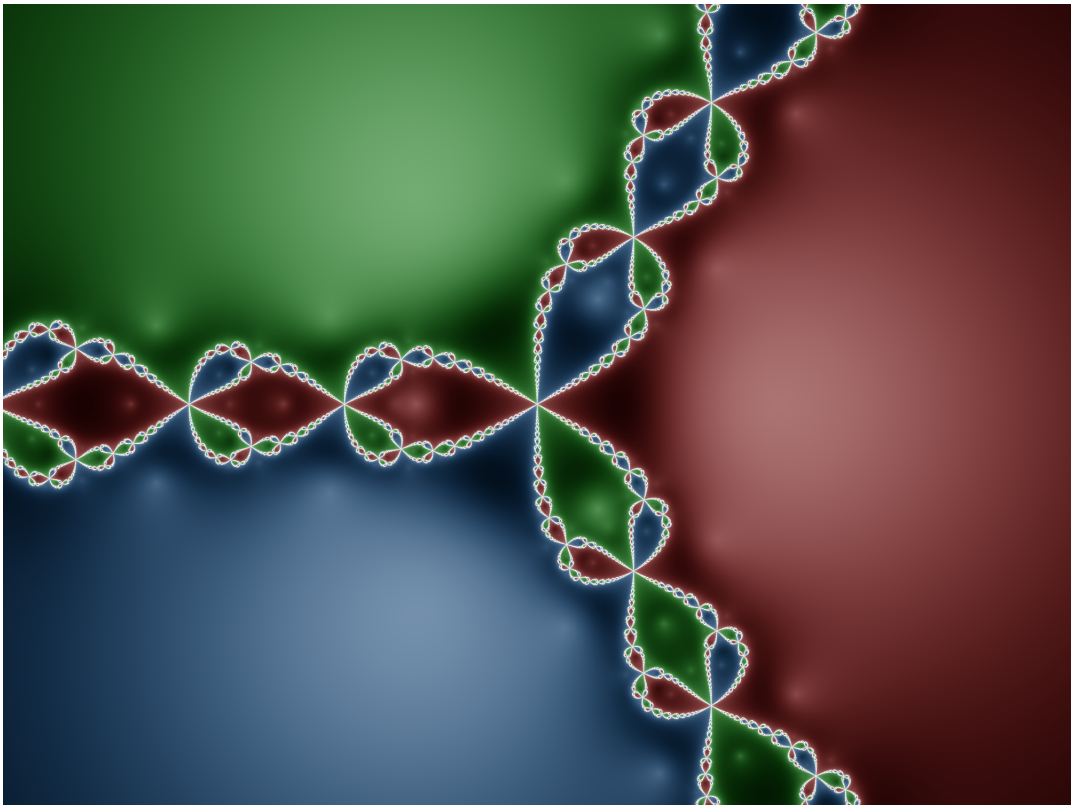


Figure 2.4: Newton fractal and Fatou set for $z^3 - 1 = 0$ (Source: Wikipedia, Author: Georg-Johann Lay).

In Figure 2.4, the Julia set is white. Red means convergence to 1*, green convergence to* ζ*, and blue convergence to* ζ^2*. The brightness of the coloured points represents the convergence rate: bright means fast, and dark slow convergence.*

For the next example, we need

Lemma 2.6.4. *Every polynomial f with real coefficients and odd degree has a real zero.*

Proof. View f as a complex polynomial. Then, by the fundamental theorem of algebra (Theorem 2.1.7), f has a complex zero. Later, in Section 3.2, we will see that f can be written as:

$$f(X) = \alpha \cdot \prod_{\mu}(X - \alpha_\mu), \quad \alpha, \alpha_\nu \in \mathbb{C} \tag{2.1}$$

with $\alpha_\mu \in \mathbb{C}$. As f is real, we have that if $\xi \in \mathbb{C}$ is a zero, then also the complex conjugate $\bar{\xi}$ is a zero of f, because:

$$f(\bar{\xi}) = \sum_{\nu} a_\nu \bar{\xi}^\nu = \sum_{\nu} \bar{a}_\nu \overline{\xi^\nu} = \overline{\sum_{\nu} a_\nu \xi^\nu} = \bar{0} = 0$$

if $f(X) = \sum_{\nu} a_\nu X^\nu$. It follows, because of (2.1), that in the case of odd degree at least one zero ξ has to satisfy:

$$\bar{\xi} = \xi$$

i.e. ξ is real. □

Example 2.6.5. *Let $p = z^3 - 2z + 2$. Then the Newton iterator*

$$\phi = \frac{2z^3 - 2}{3z^2 - 2}$$

for p has at least one and at most three real zeroes. The critical values are given as:

$$p'(x) = 0 \quad \Leftrightarrow \quad x = \pm\sqrt{\frac{2}{3}}$$

As

$$0 < p\left(\sqrt{\frac{2}{3}}\right) = 2 - \frac{4}{3}\sqrt{\frac{2}{3}} < p\left(-\sqrt{\frac{2}{3}}\right) = 2 + \frac{4}{3}\sqrt{\frac{2}{3}}$$

it follows that $-\sqrt{\frac{2}{3}}$ is a local maximum, $\sqrt{\frac{2}{3}}$ a local minimum, and between those two points, there is no real zero. Hence, we have the decomposition

$$p(X) = (Z - \alpha)(X - \beta)(X - \bar{\beta})$$

with $\alpha < 0$ and $\beta \in \mathbb{C} \setminus \mathbb{R}$. A real plot of the polynomial p is given in Figure 2.5.

The next idea is that outside of the Julia set, there is not always convergence. Namely,

$$\phi(0) = 1, \quad \phi(1) = 0$$

i.e. here we have a cycle of length 2. There are starting points which converge to this cycle. In Figure 2.6, the Julia set is white; red means convergence to the cycle $\{0, 1\}$, beige means convergence to the zero α, green convergence to the zero β, and blue convergence to the zero $\bar{\beta}$.

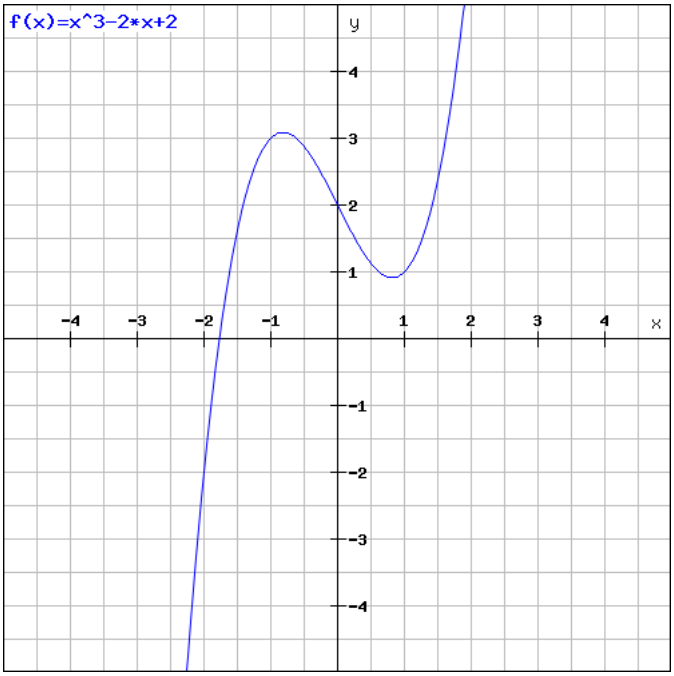


Figure 2.5: Plot of $x^3 - 2x + 2$

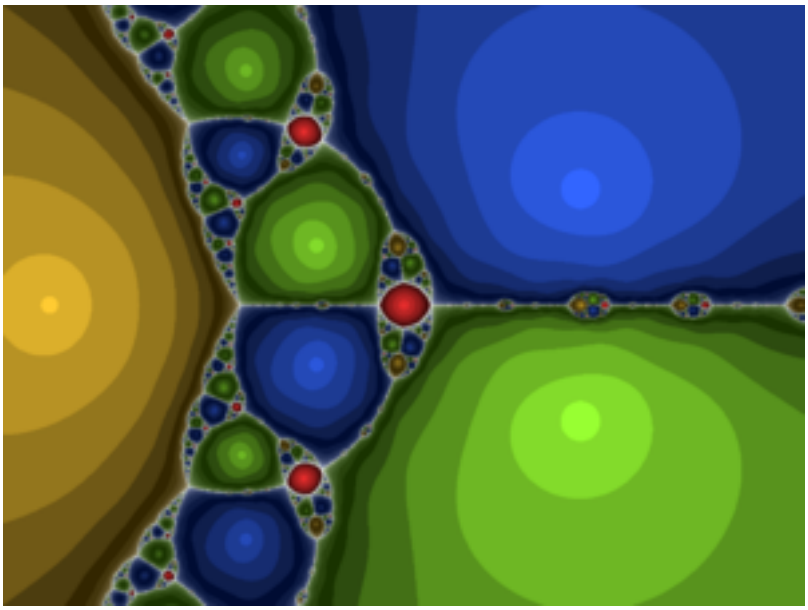


Figure 2.6: Newton fractal and Fatou set for $z^3 - 2z + 2 = 0$ (Source: Wikipedia, Author: Georg-Johann Lay).

Chapter 3

Polynomials

Let $K[X]$ denote the set of all polynomials with coefficients in K, where $K = \mathbb{Q}$, $\mathbb{R}$ or $\mathbb{C}$ is the field of rational, real or complex numbers.

Let $f = \sum_{\nu \in \mathbb{N}} a_\nu X^\nu \in K[X]$ be a polynomial. For $f \neq 0$ there exists the number

$$\deg(f) := \max\{\nu \mid a_\nu \neq 0\}$$

This number is the *degree* of f. Further, we define

$$\deg(0) := -\infty$$

3.1 Euclidian Algorithm

The Euclidean Algorithm relies on *division with remainder*:

Division with Remainder

Let $f, g \in K[X]$ with $g \neq 0$. Then there exist $q, r \in K[X]$ with $\deg(r) < \deg(g)$, such that

$$f(X) = q(X) \cdot g(X) + r(X)$$

In case $r = 0$, we write

$$g \mid f$$

("g divides f"). The *largest common divisor* $\mathrm{lcd}(f, g)$ is a polynomial $d \in K[X]$ satisfying:

1. $d \mid f$ and $d \mid g$ (i.e. d is a common divisor)
2. $e \mid f$ and $e \mid g \;\Rightarrow\; e \mid d$ (i.e. d is a maximal common divisor)

Notice that the largest commond divisor is not uniquely determined. But we have:

Lemma 3.1.1. *Let $f, g \neq 0$. If d_1 and d_2 are largest common divisors of f and g, then:*

$$d_1 = c_2 \cdot d_2$$

with $\deg(c_2) = 0$.

Proof. As d_1, d_2 are both largest common divisors, we have:

$$d_1 \mid d_2 \quad \text{and} \quad d_2 \mid d_1$$

This means:

$$d_2 = c_1 \cdot d_1 \quad \text{and} \quad d_1 = c_2 \cdot d_2$$

As $d_1 \neq 0$ and $d_2 \neq 0$, it follows that:

$$\begin{aligned} \deg(d_2) = \deg(c_1) + \deg(d_1) \quad &\text{and} \quad \deg(d_1) = \deg(c_2) + \deg(d_2) \\ \Rightarrow \quad \deg(d_2) &= \deg(c_1) + \deg(c_2) + \deg(d_2) \\ \Rightarrow \quad 0 &= \deg(c_1) + \deg(c_2) \end{aligned}$$

As $\deg(c_1)$ and $\deg(c_2)$ are natural numbers, it follows that:

$$\deg(c_1) = \deg(c_2) = 0$$

□

The Euclidean Algorithm

"[The Euclidean Algorithm] is the granddaddy of all algorithms, because it is the oldest non-trivial algorithm that has survived to the present day."

(Donald Knuth, *The Art of Computer Programming*, Vol. 2: *Seminumerical Algorithms*, 2nd edition (1981), p. 318.)

Algorithm 3.1.2 (Euclid)**.** Input: *Polynomials* $a, b \in K[X] \setminus \{0\}$.

Repeat until for some remainder $r_{N+1} = 0$ *holds true:*

$$\begin{aligned} a &= q_0 \cdot b + r_0, && \deg(r_0) < \deg(b) \\ b &= q_1 \cdot r_0 + r_1, && \deg(r_1) < \deg(r_0) \\ r_0 &= q_2 \cdot r_1 + r_2, && \deg(r_2) < \deg(r_1) \\ &\vdots \\ r_{N-1} &= q_{N+1} \cdot r_N && (r_{N+1} = 0) \end{aligned}$$

Output: r_N.

Theorem 3.1.3 (Euklid)**.** *The Euclidean Algorithm terminates. The last remainder* $r_N \neq 0$ *is the largest common divisor of* a *and* b*.*

Proof. The sequence $\deg(r_n) \in \mathbb{N} \cup \{-\infty\}$ is strictly decreasing. Hence, there exists a smallest N, such that $r_{N+1} = 0$. Then $d := r_N \neq 0$ is the last non-zero remainder. Now we prove that r_N satisfies the properties of an lcd.

- $r_N \mid a$ and $r_N \mid b$.

  $$\begin{aligned} r_{N-1} &= q_{N+1} \cdot r_N && \Rightarrow r_N \mid r_{N-1} \\ r_{N-2} &= q_N \cdot r_{N-1} + r_N && \Rightarrow r_N \mid r_{N-2} \end{aligned}$$

  From each preceding equation in Algorithmus 3.1.2 it follows that $r_N \mid r_n$ for all n. In particular, it holds true that:

  $$r_N \mid a \quad \text{and} \quad r_N \mid b$$

- $e \mid a$ and $e \mid b \;\Rightarrow\; e \mid r_N$. From

$$e \mid a = q_0 \cdot b + r_0 \quad \text{and} \quad e \mid b$$

it follows that $e \mid r_0$. From the next equation of Algorithmus 3.1.2 it then follows that $e \mid r_1$ etc. until finally it follows that $e \mid r_N$.

This proves the assertion. □

In general the Euclidean Algorithm works for so-called *Euclidean Rings*, in which there exists a division with remainder.

Beispiel 3.1.4. *The ring $\mathbb{Z}$ of integers is a Euclidean Ring. The role of the function* deg *is played by the absolute value $|\cdot|$:*

$$a = q \cdot b + r, \quad |r| < |b|$$

is here the division with remainder.

Importance for Euclid

Euclid himself uses his algorithm in order to prove the Fundamental Theorem of Arithmetic (Theorem 2.1.2). Over the integers, it can be formulated as follows: Every number $n \in \mathbb{Z} \setminus \{0\}$ has a representation

$$n = \pm p_1 \cdots p_r$$

with uniquely determined prime numbers p_i up to the ordering of factors.

3.2 Sturm Chain

If the zeros of a polynomial are to be determined, then one can, in the case of small degree, express the zeros in terms of radicals. This yields an explicit representation of the zeros. An example is the well-known formula for the solutions of a quadratic equation. There are also such formula for the zeros of polynomials of degree three and four (the *Cardano formulae*). If the degree is five or higher, then we have:

Theorem 3.2.1 (Abel-Ruffini)**.** *The general polynomial equation of degree five or higher has no solution in radicals.*

This means that the zeros of a general polynomial cannot be expressed in terms of radicals. Consequently, for higher degree polynomials, numerical methods for calculating their zeros are required.

In this section, we solve the problem of finding the number of zeros of a real polynomial $f \in \mathbb{R}[X]$ in a given interval $[a, b]$.

First some general statements for $K = \mathbb{Q}, \mathbb{R}$ or $\mathbb{C}$:

Lemma 3.2.2. *Let $f \in K[X] \setminus \{0\}$. Then $\xi \in K$ is a zero of f if and only if*

$$(X - \xi) \mid f$$

holds true.

Proof. $\Rightarrow$: Let $\xi \in K$ be a zero of f. Division with remainder yields:

$$f(X) = q(X)(X - \xi) + r(X)$$

with $\deg(r) < 1$. Now,

$$0 = f(\xi) = r(\xi)$$

As $\deg(r) < 1$, it follows that $r = 0$. Hence, $(X - \xi) \mid f$.

$\Leftarrow$: Assume $(X - \xi) \mid f$, this implies

$$f(X) = q(X) \cdot (X - \xi)$$

It follows that $f(\xi) = 0$. □

A consequence is that a polynomial $f \in K[X] \setminus \{0\}$ has at most $\deg(f)$ zeros in K.

Proof. Let $\xi \in K$ be a zero of f. From

$$f = q \cdot (X - \xi)$$

it follows that

$$\deg(q) = \deg(f) - 1$$

and after at most $\deg(f)$ zeros, this process stops. □

For $K = \mathbb{C}$ it follows that, a non-constant polynomial $f \in K[X]$ has a representation

$$f(X) = \alpha \cdot \prod_{\mu} (X - \alpha_\mu), \quad \alpha, \alpha_\nu \in \mathbb{C}$$

hat.

Definition 3.2.3. *A polynomial $f \in K[X]$ is* sqare-free*, if for no non-constant polynomial $g \in K[X]$ it holds true that $g^2 \mid f$.*

Lemma 3.2.4. *Square-free polynomials have only simple zeros.*

Proof. Let $\xi \in K$ be a zero of $f \in K[X]$. Then:

$$f = q \cdot (X - \xi)$$

with $q(\xi) \neq 0$, as f is square-free. We need to show that $f'(\xi) \neq 0$. For this, we have:

$$f' = q' \cdot (X - \xi) + q \quad \Rightarrow \quad f'(\xi) = q(\xi) \neq 0$$

□

Now, we can define the Sturm chain.

Definition 3.2.5. *Let $f \in \mathbb{R}[X]$ be a polynomial. Then the sequence $p_0 := f$, $p_1 := f'$, $p_2, \ldots, p_N$ with*

$$\begin{aligned}
f &= q_1 \cdot f' - p_2, && \deg(p_2) < \deg(f') \\
f' &= q_2 \cdot p_2 - p_3, && \deg(p_3) < \deg(p_2) \\
p_2 &= q_3 \cdot p_3 - p_4, && \deg(p_4) < \deg(p_3) \\
&\vdots \\
p_{N-1} &= q_N \cdot p_N && (p_{N+1} = 0)
\end{aligned}$$

is called a Sturm chain *for f.*

The elements of a Sturm chain are, up to a sign, the remainders occurring in the Euclidean Algorithm for calculating $\operatorname{lcd}(f, f')$.

Theorem 3.2.6 (Sturm)**.** *Let $f \in \mathbb{R}[X]$ be square-free, and $p_0, \ldots, p_N$ a Sturm chain for f. Then for $a < b$ the number of zeros in the interval $(a, b]$ eqals*

$$\sigma(a) - \sigma(b)$$

where $\sigma(\xi)$ is the number of sign changes in the sequence

$$p_0(\xi), \ldots, p_N(\xi)$$

Remark 3.2.7. *Sturm's Theorem also holds true if a Sturm chain $p_0, \ldots, p_N$ is replaced with*

$$\alpha_0 \cdot p_0, \ldots, \alpha_N \cdot p_N \tag{3.1}$$

where $\alpha_0, \ldots, \alpha_N > 0$. The sequence (3.1) is also called a Sturm chain *for f. Using the α_i, it is possible to remove denominators in fractions occurring in a Sturm chain.*

A Sturm chain tells us if f is square-free or not.

Lemma 3.2.8. *$f \in K[X]$ is square-free if and only if*

$$\operatorname{lcd}(f, f') = 1$$

Proof. $\Rightarrow$: If $d(X) = \operatorname{lcd}(f, f')$ is not constant, then $d(X)$ has a zero $\xi \in \mathbb{C}$, by the Fundamental Theorem of Algebra (Theorem 2.1.7). As d divides f and f', it follows that

$$f(\xi) = f'(\xi) = 0$$

Hence, the zero ξ is not simple. This implies, by Lemma 3.2.4, that f is not square-free.

$\Leftarrow$: Let $\operatorname{lcd}(f, f') = 1$. If $f = g^2 \cdot q$, then:

$$f' = 2gg'q + g^2q' = g \cdot (2g'q + gq')$$

Hence, $g \mid \operatorname{lcd}(f, f') = 1$, i.e. g is constant. Hence f is square-free. □

What can be done if f is not square-free?

Theorem 3.2.9. *Let $f \in K[X] \setminus \{0\}$. Then*

$$g := \frac{f}{\operatorname{lcd}(f, f')}$$

is square-free and has the same zeros as f.

Proof. Let

$$f = (X - \xi)^k \cdot h$$

with $h(\xi) \neq 0$. Then

$$f' = k \cdot (X - \xi)^{k-1} \cdot h + (X - \xi)^k \cdot h'$$

Hence, $(X - \xi)^{k-1} \mid f'$ and $(X - \xi)^k \nmid f'$, as otherwise

$$(X - \xi)^k \mid f' - (X - \xi)^k \cdot h' = k \cdot (X - \xi)^{k-1}$$

which is impossible. It follows that

$$(X - \xi)^{k-1} \mid \operatorname{lcd}(f, f') \quad \text{and} \quad (X - \xi)^k \nmid \operatorname{lcd}(f, f')$$

I.e. the multiplicity of the zero ξ in $\operatorname{lcd}(f, f')$ is One less than its multiplicity in f. Hence, g is square-free and has the same zeros as f. □

Example 3.2.10. *Let $f(X) = X^3 - 2X^2$. We would like to count the number of real zeros of f. Applying Sturm's method yields the Sturm chain*

$$f, f' = 3X^2 - 4X, X$$

and we see that $\operatorname{lcd}(f, f') = X$*, i.e. f is not square-free. So, we take*

$$g := \frac{f}{\operatorname{lcd}(f, f')} = X^2 - 2X$$

and obtain the Sturm chain

$$g, g' = 2X - 2, p_2 = 1$$

The following table of signs

	$a >> 0$	$-a << 0$
g	$+$	$+$
g'	$+$	$-$
p_2	$+$	$+$
σ	0	2

yields

$$\sigma(-a) - \sigma(a) = 2$$

as the number of real zeros of f (without multiplicities).

3.3 Prime, perfect and narcissistic numbers

In the episode *Marge and Homer Turn a Couple Play*, just as Tabitha wants to give her husband a declaration of love, a call appears on the screen in the baseball stadium to estimate the number of spectators:

a) 8191

b) 8128

c) 8208

d) No way to tell

The first number is a prime number. It is a so-called *Mersenne prime number*. Such are prime numbers of the form $2^p - 1$, where p itself is a prime number. In fact:

$$8191 = 2^{13} - 1$$

Mersenne prime numbers are record holders: the ten largest of these are the largest prime numbers which have been discovered. The largest known prime number is $2^{74207281} - 1$, a Mersenne prime number discovered in the year 2016.

The second number is a *perfect number*, i.e. a number which equals the sum of its divisors other than itself. The smallest perfect number is

$$6 = 1 + 2 + 3$$

The next one is

$$28 = 1 + 2 + 4 + 7 + 14$$

This one is followed by 496 and by 8128. It is not known whether there are infinitely many perfect numbers or not. Also unknown to this date is if all perfect numbers are even.

The third number is a *narcissistic number*. These numbers have the property that the sum of their digits, each taken to the power of the number of digits, equals the number itself. Indeed,

$$8208 = 8^4 + 2^4 + 0^4 + 8^4$$

It has been proven that there are only finitely many narcissistic numbers. They are 88 in number, and the largest one is

$$115\,132\,219\,018\,763\,992\,565\,095\,597\,973\,971\,522\,401$$

Chapter 4

Interpolation

Let $K = \mathbb{Q}, \mathbb{R}$ or $\mathbb{C}$. Let $n+1$ pairs $(x_i, f_i) \in K^2$ and a function $\Phi(x, a_0, \ldots, a_n)$ be given. The problem now is to choose the parameters $a_0, \ldots, a_n$ in such a way that

$$\Phi(x_i, a_0, \ldots, a_n) = f_i$$

We consider the *linear interpolation problem*:

Linear Interpolation Problem

Here,

$$\Phi(x, a_0, \ldots, a_n) = a_0 \cdot \Phi_0(x) + \cdots + a_n \cdot \Phi(x)$$

with linearly independent functions $\Phi_0, \ldots, \Phi_n$.

4.1 Polynomial Interpolation

Here the problem is to find the polynomial P of degree $\leq n$ which takes the values $f(\alpha_i)$ in $n+1$ distinct places α_i.

The space in which the solution is sought, is

$$K[X]_n := \{\text{Polynomials of degree} \leq n\}$$

This is a so-called linear interpolation problem: Solve the system of linear equations

$$P(\alpha_i) = f(\alpha_i)$$

depending on a basis $b_i(X)$ of the vector space $K[X]_n$. It holds true that

$$P(X) = \sum_{i=0}^{n} \rho_i b_i(X) =: \Phi(X, \rho_0, \ldots, \rho_n)$$

4.1.1 Standard Basis

The standard basis for $K[X]_n$ is

$$b_i(X) = X^i, \quad i = 0, \ldots, n$$

With respect to this basis, the polynomial has the usual representation

$$P(X) = \sum_{i=0}^{n} \rho_i X^i$$

The interpolation problem leads to the linear system of equations

$$\underbrace{\begin{pmatrix} 1 & \alpha_0 & \dots & \alpha_0^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_n & \dots & \alpha_n^n \end{pmatrix}}_{=V(\alpha_0,\dots,\alpha_n)} \begin{pmatrix} \rho_0 \\ \vdots \\ \rho_n \end{pmatrix} = \begin{pmatrix} f(\alpha_0) \\ \vdots \\ f(\alpha_n) \end{pmatrix}$$

Definition 4.1.1. *The matrix $V(\alpha_0, \dots, \alpha_n)$ is called* Vandermonde matrix.

Lemma 4.1.2. *The Vandermonde matrix is regular if and only if the α_i are pairwise distinct.*

Sketch of proof. We have

$$\det V(\alpha_0, \dots, \alpha_n) = \prod_{0 \le j < k \le n} (\alpha_j - \alpha_k)$$

This determinant is non-zero if and only if the α_i are pairwise distinct. □

A consequence is:

Theorem 4.1.3. *The polynomial interpolation problem has a unique solution.*

Proof. According to Lemma 4.1.2, the polynomial interpolation problem has a unique solution for the standard basis of $K[X]_n$. Hence, it also has a unique solution for an arbitrary basis $b_0(X), \dots, b_n(X)$. □

Remark 4.1.4. *Lemma 4.1.2 is the reason why for $n+1$ distinct points it is required that the polynomial degree be $\le n$ for the polynomial interpolation problem.*

Example 4.1.5. *Two distinct points in the Euclidean plane determine a unique straight line. But there are infinitely many parabolas containing these two points.*

The method which offers itself for solving the polynomial interpolation problem for the standard basis is Gauss' algorithm. However, its complexity of $O(n^3)$ is high.

4.1.2 Lagrange Polynomials

The *Lagrange polynomials*

$$\ell_i(X) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{X - \alpha_j}{\alpha_i - \alpha_j}$$

satisfy the property

$$\ell_i(\alpha_j) = \delta_{ij} := \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

δ_{ij} is the *Kronecker delta*. The Lagrange polynomials are a basis of $K[X]_n$.

Proof. Let

$$B(X) := \sum_{\nu=0}^{n} \beta_\nu \ell_\nu(X) = 0$$

Then

$$0 = B(\alpha_\mu) = \sum_{\nu=0}^{n} \beta_\nu \delta_{\nu\mu} = \beta_\mu$$

Hence, the $\ell_0, \ldots, \ell_n$ are linearly independent. □

The solution of the interpolation problem is given as:

$$P(X) = \sum_{i=0}^{n} f(\alpha_i)\ell_i(X)$$

Proof.

$$P(\alpha_\mu) = \sum_{\nu=0}^{n} f(\alpha_\nu)\ell_\nu(\alpha_\mu) = \sum_{\nu=0}^{n} f(\alpha_\nu)\delta_{\nu\mu} = f(\alpha_\mu)$$

□

The coefficients ρ_i are simply the values:

$$\rho_i = f(\alpha_i)$$

A disadvantage of the Lagrange polynomials is that in case a further place α_{n+1} is taken, then all $\ell_i(X)$ become different.

Example 4.1.6. *We interpolate using the Lagrange polynomials:*

$$f(0) = 3,\ f(1) = 2,\ f(3) = 1,\ f(4) = 0 \tag{4.1}$$

and estimate the Value $f(2.5)$. We have

$$\begin{aligned}
\ell_0(X) &= -\frac{1}{12}(X-1)(X-3)(X-4) \\
\ell_1(X) &= \frac{1}{6}X(X-3)(X-4) \\
\ell_2(X) &= -\frac{1}{6}X(X-1)(X-4)
\end{aligned}$$

The interpolation polynomial is

$$P(X) = 3 \cdot \ell_0 + 2 \cdot \ell_1 + 1 \cdot \ell_2 + 0 \cdot \ell_3 = \frac{1}{12}\left(-X^3 + 6X^2 - 17X + 36\right)$$

and $f(2.5) = P(2.5) = 1.28125$. The polynomial and the values at the places are given in Figure 4.1.

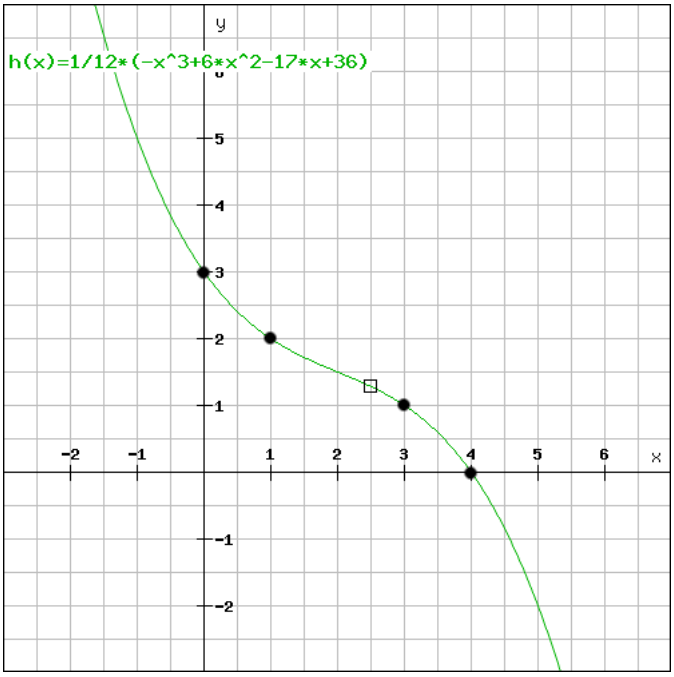


Figure 4.1: The interpolation polynomial for the values in (4.1).

4.1.3 Newton Polynomials

Let $\alpha_0, \ldots, \alpha_n \in K$ be pairwise distinct. The *Newton Polynomials* are

$$N_i(X) = \prod_{k=0}^{i-1}(X - \alpha_k), \quad i = 0, \ldots, n$$

For $i > 0$ we have:

$$N_i(\alpha_j) = \begin{cases} \prod_{k=0}^{i-1}(\alpha_j - \alpha_i), & j \geq i \\ 0, & j < i \end{cases}$$

Hence, the approach

$$P(X) = \sum_{i=0}^{n} \rho_i N_i(X)$$

leads to the system of linear equations

$$\begin{pmatrix} 1 & & & 0 \\ 1 & (\alpha_1 - \alpha_0) & & \\ 1 & (\alpha_2 - \alpha_0) & (\alpha_2 - \alpha_0)(\alpha_2 - \alpha_1) & \\ \vdots & \vdots & \ddots & \\ 1 & (\alpha_n - \alpha_0) & \cdots & \prod_{i=0}^{n-1}(\alpha_n - \alpha_i) \end{pmatrix} \begin{pmatrix} \rho_0 \\ \vdots \\ \rho_n \end{pmatrix} = \begin{pmatrix} f(\alpha_0) \\ \vdots \\ f(\alpha_n) \end{pmatrix} \tag{4.2}$$

since

$$P(\alpha_j) = \sum_{i=0}^{j} \rho_i N_i(\alpha_j)$$

As the coefficient matrix in (4.2) is invertible, it follows that the Newton polynomials $N_0(X)$ to $N_n(X)$ are a basis of $K[X]_n$.

As the coefficient matrix is an upper triangular matrix, the solution can be found by *forward substitution*, starting from the top equation

$$\rho_0 = f(\alpha_0)$$

and then substituting this into the second equation:

$$f(\alpha_0) + (\alpha_1 - \alpha_0)\rho_1 = f(\alpha_1)$$

which is one linear equation with one unkown, etc.

4.1.4 Interpolation error

A typical problem is to approximate a continuous function with an interpolation polynomial. A basis for this is:

Theorem 4.1.7 (Weierstraß)**.** *Let $f\colon [a,b] \to \mathbb{R}$ be continuos. Then for every $\epsilon > 0$ there exists a polynomial $P(X) \in \mathbb{R}[X]$ such that*

$$\|f - P\|_\infty := \max\{|f(t) - P(t)| \mid t \in [a,b]\} < \epsilon$$

This gives the interpolation error. Here, the following holds true:

Theorem 4.1.8. *Let $f\colon [a,b] \to \mathbb{R}$ be $(n+1)$-fold continuously differentiable. Then for every $t \in [a,b]$ there exists a $\xi \in I_t$ such that*

$$f(t) - P(t) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \cdot N_{n+1}(t)$$

Here, $P(X)$ is the interpolation polynomial for $n+1$ places $\alpha_0, \ldots, \alpha_n$.

In the theorem, $N_{n+1}(X)$ is the $(n+1)$-th Newton polynomial and I_t the smallest interval which contains the places $\alpha_0, \ldots, \alpha_n$ and $t \in \mathbb{R}$.

Consequence. *The following error estimation holds true:*

$$|f(t) - P(t)| \leq \frac{\left\|f^{(n+1)}\right\|_\infty}{(n+1)!} \prod_{\nu=0}^{n} |t - \alpha_\nu|$$

where

$$\|\cdot\|_\infty\colon C[a,b] \to \mathbb{R}, \quad g \mapsto \max\{|g(t)| \mid t \in [a,b]\}$$

is the maximum norm *on the vector space $C[a,b]$ of continuous functions $[a,b] \to \mathbb{R}$.*

Example 4.1.9. *Let $f(X) = \frac{1}{12}(-X^3 + 6X^2 - 17X + 36)$. We want to interpolate this function in $(0,3), (1,2), (3,1)$. The interpolation polynomial is*

$$P_2(X) = 3 - \frac{7}{6}X + \frac{1}{6}X^2$$

The interpolation error is shown in Figure 4.2. It equals the error estimation, as f is a polynomial.

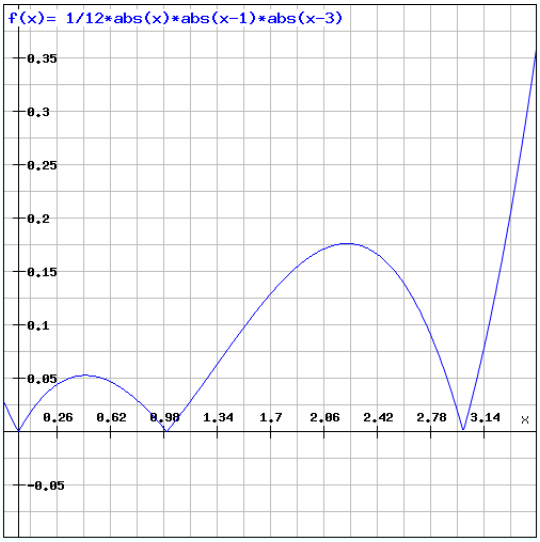


Figure 4.2: The interpolation error equals the estimate in this example.

4.1.5 Runge's Phenomenon

As for polynomials $P \in \mathbb{R}[X]$ it holds true that:

$$\lim_{t \to \pm\infty} P(t) = \pm\infty$$

it is appropriate to interpolate only values of functions with the same limiting behavious. Otherwise, there will be strong oscillations near the boundary, in particular in the case of equidistant places.

Runge's Example

Runge considers

$$f(x) = \frac{1}{1+x^2}$$

on the intervall $[-5, 5]$. Cf. Figure 4.3.

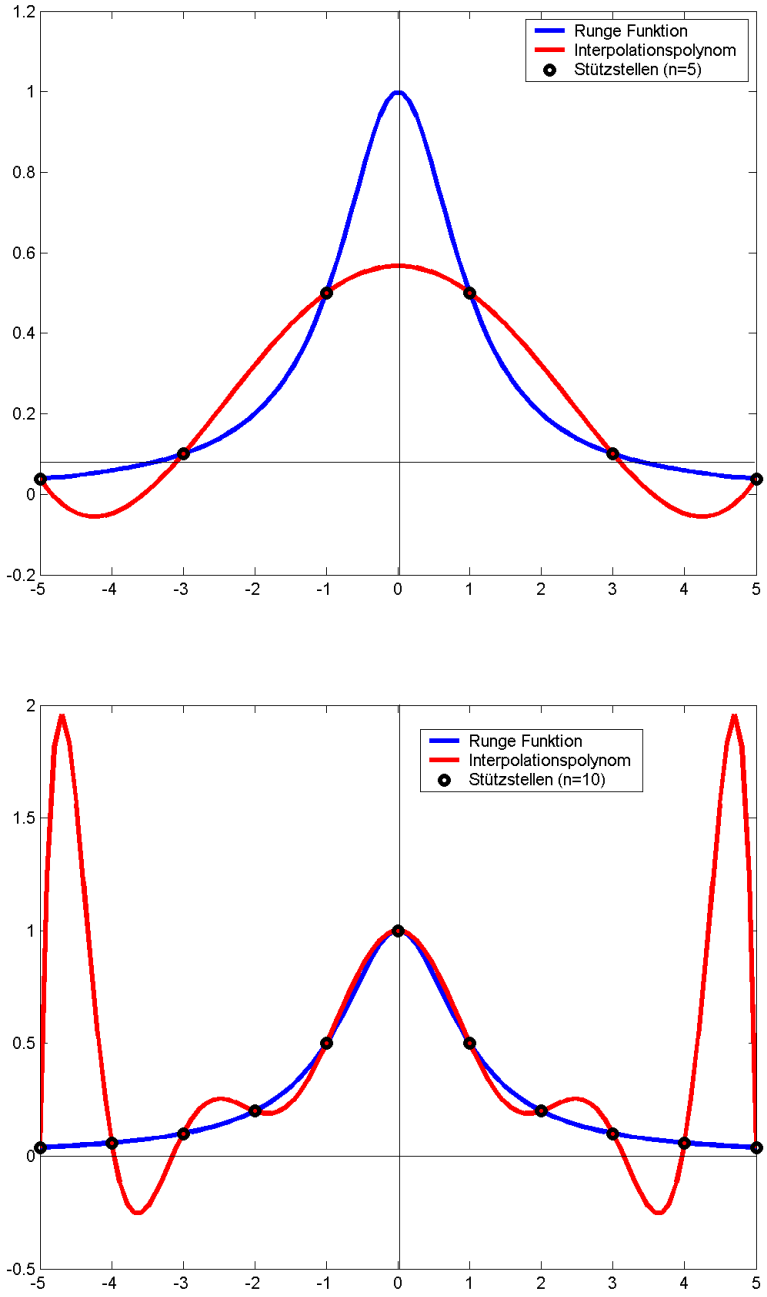


Figure 4.3: Interpolation of f with 5 resp. 10 equidistant places (Source: Wikipedia, author: Mártin Pieper).

4.2 Spline-Interpolation

A *spline* is an elastic rod (cf. Figure 4.4). It is used in naval architecture for lines without sudden change of the radius of curvature.

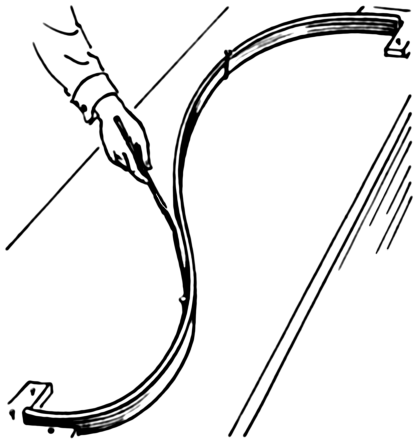


Figure 4.4: A spline (source: Wikipedia, author: Pearson Scott Foresman).

4.2.1 Polygonal chain

For real places $x_0 < \cdots < x_n$ and values $f(x_0), \ldots, f(x_n)$ there is the *knot basis* $\phi_i(x)$, used for piece-wise linear interpolation as in Figure 4.5, such that

$$\phi_i(x_j) = \delta_{ij}$$

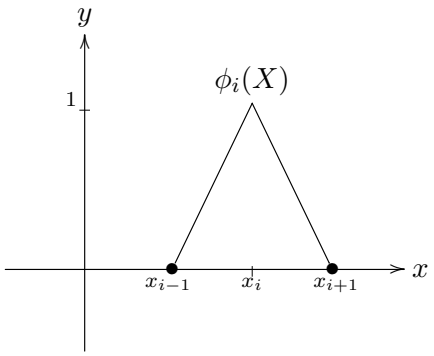


Figure 4.5: A knot basis function.

The interpolating function is

$$P(x) = \sum_{i=0}^{n} f(x_i)\phi_i(x)$$

4.2.2 Spline Spaces

Let the interval $[a, b]$ be given with places

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$$

This leads to a partition of $[a, b]$

$$\mathcal{Z} := \{I_i = [x_{i-1}, x_i] \mid i = 1, \ldots, n\}$$

The *fineness* of the partition $\mathcal{Z}$ is

$$h_{\mathcal{Z}} = \max\{x_i - x_{i-1} \mid i = 1, \ldots, n\}$$

This leads to the *spline space* for $\mathbb{Z}$:

$$S_{\mathcal{Z}}^{(k,r)}[a,b] := \{P \in C^r[a,b] \colon P|_{I_i} \in \mathbb{R}[X]_k,\ i = 1, \ldots, n\}$$

where

$$C^r[a,b] = \{f \colon [a,b] \to \mathbb{R} \mid f \text{ is } r\text{-fold continuously differentiable}\}$$

and $P|_{I_i}$ is the restriction of P to I_i:

$$P|_{I_i} \colon I_i \to \mathbb{R}, \quad t \mapsto P(t)$$

Remark 4.2.1. *The spline space $S_{\mathcal{Z}}^{(k,r)}$ is a vector space.*

Interpolation Error

Let q_i be an interpolating polynomial on I_i. From Section 4.1.4 we recall that

$$|f(t) - q_i(t)| \leq \frac{\left\|f^{(k+1)}\right\|_\infty}{(k+1)!} \prod_{\nu=0}^{k} |t - \alpha_i|$$

where $\alpha_0 = x_{i-1}, \ldots, \alpha_k = x_i \in I_i = [x_{i-1}, x_i]$ are further places. As

$$|t - \alpha_i| \leq |x_i - x_{i-1}| \leq h_{\mathcal{Z}}$$

it follows that:

$$|f(t) - P(t)| \leq \frac{\left\|f^{(k+1)}\right\|_\infty}{(k+1)!} \cdot h_{\mathcal{Z}}^{k+1}$$

Hence, we have:

Theorem 4.2.2. *The interpolation error for interpolating f with $P \in S_{\mathcal{Z}}^{(k,r)}$ is given as*

$$|f(t) - P(t)| \leq \frac{\left\|f^{(k+1)}\right\|_\infty}{(k+1)!} \cdot h_{\mathcal{Z}}^{k+1}$$

Example 4.2.3. *For piece-wise linear interpolation we have*

$$P \in S_{\mathcal{Z}}^{(1,0)}$$

Hence, for the interpolation error we have

$$|f(t) - P(t)| \leq \left\|f'\right\|_\infty \cdot h_{\mathcal{Z}}^2$$

4.2.3 Cubic Splines

A spline $P \in S_{\mathcal{Z}}^{(3,2)}$ is called a *cubic spline*.

Theorem 4.2.4. *The interpolating cubic spline P exists and is uniquely determined by the additional specification of $P''(a)$ and $P''(b)$.*

Proof. Existence. Each polynomial $q_i(X) = P|_{I_i}$ has 4 coefficients. This means there are $4n$ parameters. For these there are

- $2n$ linear equations $q_i(x_i) = f(x_i)$, $q_{i+1}(x_i) = f(x_i)$
- $n-1$ linear equations for P' continuous
- $n-1$ linear equations for P'' continuous
- 2 linear extra equations through the specification of $P''(a), P''(b)$

Hence, there are $4n$ linear equations in $4n$ unknowns. I.e. in the case of uniqueness, this system also has a solution.

Uniqueness. Let $P, Q \in S_{\mathbb{Z}}^{(3,2)}[a,b]$ with the same extra specifications. Then

$$s := P - Q \in \left\{ w \in C^2[a,b] \mid w(x_i) = 0,\ i = 0, \dots, n \right\} =: N \tag{4.3}$$

More precisely, $s \in N \cap S_{\mathbb{Z}}^{(3,2)}[a,b]$. For $w \in N$ arbitrary, it holds true that:

$$\int_a^b s''(x)w''(x)\,dx = 0 \tag{4.4}$$

since:

$$\begin{aligned}
\int_a^b s''(x)w''(x)\,dx &= \sum_{i=0}^{n} \int_{x_i}^{x_{i+1}} s''(x)w''(x)\,dx = \sum_{i=0}^{n} s''(x)w'(x)\Big|_{x_i}^{x_{i+1}} - \int_{x_i}^{x_{i+1}} s'''(x)w'(x)\,dx \\
&= \sum_{i=0}^{n} s''(x)w'(x)\Big|_{x_i}^{x_{i+1}} - \underbrace{s'''(x)w(x)\Big|_{x_i}^{x_{i+1}}}_{=0} + \int_{x_i}^{x_{i+1}} \underbrace{s''''(x)}_{=0} w(x)\,dx \\
&= \underbrace{s''(b)}_{=0} w'(b) - \underbrace{s''(a)}_{=0} w'(a) = 0
\end{aligned}$$

With $w = s$ it follows for the curvature

$$\int_a^b \left| s''(x) \right|^2 dx = 0$$

Hence, s is linear. But as $s \in N$, it follows that: $s = 0$. Hence, $P = Q$. □

Definition 4.2.5. *The cubic spline P with $P''(a) = P''(b) = 0$ is called* natural.

Remark 4.2.6. *The natural cubic spline minimises the total curvature*

$$\int_a^b \left| f''(x) \right|^2 dx$$

under all interpolating functions $f \in C^2[a,b]$.

Proof. Let $P \in S_{\tilde{z}}^{(3,2)}$. Then $w := f - P \in N$ (defined as in (4.3)), and it holds true that:

$$\begin{aligned}\int_a^b \left|f''(x)\right|^2 dx &= \int_a^b \left|P''(x) + w''(x)\right|^2 dx \\ &= \int_a^b \left|P''(x)\right|^2 dx + 2\underbrace{\int_a^b P''(x)w''(x)\,dx}_{=0} + \underbrace{\int_a^b \left|w''(x)\right|^2 dx}_{\geq 0}\end{aligned}$$

(The middle integral vanishes according to (4.4)). Hence, $\int_a^b |P''(x)|^2\,dx$ is minimal. □

Calculating the natural cubic splines

We set up the $4n$ linear equations. Let

$$q_i(X) = a_0^{(i)} + a_1^{(i)}(X - x_i) + a_2^{(i)}(X - x_i)^2 + a_3^{(i)}(X - x_i)^3, \quad i = 1, \ldots, n$$

The condition $q_i(x_i) = f(x_i)$ leads to:

$$a_0^{(i)} = f(x_i), \quad i = 1, \ldots, n \tag{4.5}$$

With $h_i := x_{i-1} - x_i$ it follows that

$$q_i(x_{i-1}) = a_0^{(i)} + a_1^{(i)}h_i + a_2^{(i)}h_i^2 + a_3^{(i)}h_i^3$$

Then the condition $q_i(x_{i-1}) = f(x_{i-1})$ leads to:

$$f(x_{i-1}) - f(x_i) = a_1^{(i)}h_i + a_2^{(i)}h_i^2 + a_3^{(i)}h_i^3, \quad i = 1, \ldots, n \tag{4.6}$$

The condition $q_1''(x_0) = q_n''(x_n) = 0$ leads to:

$$a_2^{(1)} + 3a_3^{(1)}h_1 = 0, \quad a_2^{(n)} = 0 \tag{4.7}$$

The condition $q_i'(x_i) = q_{i+1}'(x_i)$ leads to:

$$a_1^{(i)} = a_1^{(i+1)} + 2a_2^{(i+1)}h_{i+1} + 3a_3^{(i+1)}h_{i+1}^2, \quad i = 1, \ldots, n-1 \tag{4.8}$$

Finally, the condition $q_i''(x_i) = q_{i+1}''(x_i)$ leads to:

$$a_2^{(i)} = a_2^{(i+1)} + 3a_3^{(i+1)}h_{i+1}, \quad i = 1, \ldots, n-1 \tag{4.9}$$

(4.5) to (4.9) simplify to

$$h_i a_2^{(i-1)} + 2(h_i + h_{i+1})a_2^{(i)} + h_{i+1}a_2^{(i+1)} = 3\left(\frac{f(x_{i+1}) - f(x_i)}{h_{i+1}} - \frac{f(x_i) - f(x_{i-1})}{h_i}\right),$$

with $i = 1, \ldots, n-1$. This is a system of linear equations whose coefficient matrix has the following form:

$$A = \begin{pmatrix} 2(h_1+h_2) & h_2 & & & & & 0 \\ h_2 & 2(h_2+h_3) & h_3 & & & & \\ & h_3 & \ddots & \ddots & & & \\ & & \ddots & & & & \\ & & & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} & \\ 0 & & & & h_{n-1} & 2(h_{n-1}+h_n) & \end{pmatrix}$$

This matrix is regular, as we know from uniqueness.

Chapter 5

Numerical Linear Algebra

5.1 The Power Method for Determining Eigen Vectors in the Example of PageRank

In google's PageRank, we learn about important properties of stochastic matrices, and also about the power method for calculating eigen vectors.

The idea behind PageRank is that the importance of a web page depends on the number of pages pointing to that page and their importance.

Assume that page P_j has ℓ_j links to other pages. If there is a link to page P_i (we denote this here as $P_j \to P_i$), then P_i obtains from P_j the fraction of $\frac{1}{\ell_j}$ of its importance. The *importance rank* $I(P_i)$ of P_i is the sum of all contributions of pages with a link to P_i:

$$I(P_i) = \sum_{P_j \to P_i} \frac{I(P_j)}{\ell_j} \tag{5.1}$$

In order to understand (5.1) better, we consider the *hyperlink matrix* $H = (H_{ij})$ with

$$H_{ij} = \begin{cases} \frac{1}{\ell_j}, & P_j \to P_i \\ 0 & \text{otherwise} \end{cases}$$

This has the properties:

- All entries are non-negative.
- The column sum is always either 1 or 0.

Definition 5.1.1. *A* stochastic matrix *is a square matrix with non-negative real entries whose column sums are all equal to* 1*.*

With the vector $I = (I(P_i))$, (5.1) can be written as:

$$I = H \cdot I$$

i.e. I is an *eigen vector* of H for the *eigen value* 1. Such a vector is called a *stationary vector* of H.

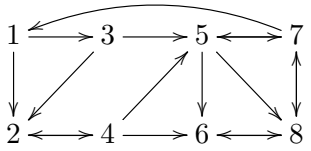


Figure 5.1: A miniature internet.

Example 5.1.2. *The graph in Figure 5.1 has the hyperlink matrix*

$$H = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 1 & \frac{1}{3} & 0 \end{pmatrix}$$

A stationary vector is

$$I = \begin{pmatrix} 0.0600 \\ 0.0675 \\ 0.0300 \\ 0.0675 \\ 0.0975 \\ 0.2025 \\ 0.1800 \\ 0.2950 \end{pmatrix}$$

Hence, the most important node is 8*.*

5.1.1 The power method

For calculating the stationary vector I of the hyperlink matrix H let it be said that:

- H has about 25 thousand million rows and columns.
- Most entries of H are zero.
- On average there are about 10 entries per column.

For these reasons, a fast as possible method for computing I is of interest. This is the *power method*:

- start vector $I^0 \neq 0$.
- $I^{k+1} := H \cdot I^k$

The principle is $I^k \to I$ for $k \to \infty$.

In Example 5.1.2, we already have $I^{60} = I$.

The following questions arise:

- Does I^k always converge?

- Is the limit vector independent of the start vector?
- Do the importance ranks contain the desired information?

The answers are three times: *No.*

The way out is a modification of the hyperlink matrix.

Example 5.1.3. *Let the graph* $1 \to 2$ *be given. Its hyperlink matrix is*

$$H = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

For the start vector $I^0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$*, we have* $I^1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ *and* $I^2 = I = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$*. The deeper reason for this is that node* 2 *does not have any outgoing links. Such a node is called a* dangling node*.*

If we interpret the PageRank $I(P_i)$ as the fraction of time during which a random surfer stays on a page, then we can ask for the column sum of I to equal 1. If the surfer is on a dangling node, then he or she should simply jump to any page.

Example 5.1.4. *Let again the graph* $1 \to 2$ *be given. Then the random surfer jumps according to the matrix*

$$S = \begin{pmatrix} 0 & \frac{1}{2} \\ 1 & \frac{1}{2} \end{pmatrix}$$

A stationary vector of this matrix is $I = \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \end{pmatrix}$*. Now, node* 2 *is twice as important as node* 1*, which agrees with intuition.*

We now replace the hyperlinkmatrix H with

$$S = H + A$$

where A has for every dangling node the column $\begin{pmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{pmatrix}$ and zeros otherwise.

An *eigen value* of a square matrix S is a scalar λ for which there exists a vector $v \neq 0$ such that

$$S \cdot v = \lambda \cdot v \tag{5.2}$$

The vector v is called an *eigen vector* of S for the eigen value λ. The set of all vectors satisfying (5.2) forms a vector space and is called the *eigen space* of S for eigen value λ.

For eine stochastc matrix, 1 is the eigenvalue with the largest absolute value.

We now assume that for the eigen values λ_i of an $n \times n$-matrix S holds true that:

$$1 = \lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$$

Further, we assume that there is a basis $v_1, \ldots, v_n$ of $\mathbb{R}^n$ consisting of eigen values of S. Then

$$\begin{aligned} I^0 &= c_1 v_1 + c_2 v_2 + \cdots + c_n v_n \\ I^1 = SI^0 &= c_1 v_1 + c_2 \lambda_2 v_2 + \cdots + c_n \lambda_n v_n \\ I^2 = SI^1 &= c_1 v_1 + c_2 \lambda_2^2 v_2 + \cdots + c_n \lambda_n^2 v_n \\ &\vdots \\ I^k = SI^{k-1} &= c_1 v_1 + c_2 \lambda_2^k v_2 + \cdots + c_n \lambda_n^k v_n \end{aligned}$$

As $\lambda_j^k \to 0$ for $k \to \infty$ and $j \geq 2$, it follows that:

$$I^k \to I = c_1 v_1, \quad \text{ein stationärer Vektor} \quad (k \to \infty) \tag{5.3}$$

If $|\lambda_2|$ is very small, then the convergence (5.3) is very fast.

However, not always does $1 = \lambda_1 > |\lambda_2|$ hold true.

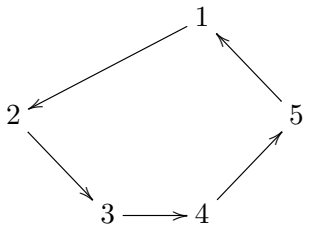


Figure 5.2: A cyclic internet.

Example 5.1.5. *In the graph of Figure 5.2, we have*

$$S = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Here, with $I^0 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$, *we have* $I^5 = I^0$, *i.e.* I^k *does not converge. The reason is that* $|\lambda_2| = 1$.

Definition 5.1.6. *The matrix S is called* primitive, *if for some m S^m has only positive entries.*

The meaning of S being primitive is that for any two pages A, B there is a sequence of links $A \to \cdots \to B$.

Definition 5.1.7. *A graph in which for any two nodes A, B there is a directed path $A \to \cdots \to B$, is called* strongly connected.

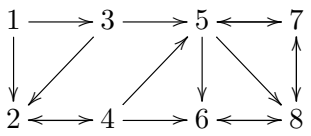


Figure 5.3: A subnet in the internet: it is possible to enter $\{5, 6, 7, 8\}$ by following links, but not to get out of this subnet.

Example 5.1.8. *A stationary vector for Figure 5.3 is*

$$I = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0.12 \\ 0.24 \\ 0.24 \\ 0.4 \end{pmatrix}$$

Here, the zero entries are problematic. The reason is that there is a subnet: it is possible to enter $\{5, 6, 7, 8\}$ *via links, but not possible to escape from this subnet. The corresponding matrix is*

$$S = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{3} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 1 & \frac{1}{2} & 0 \end{pmatrix}$$

This matrix is reducible.

Definition 5.1.9. *A square matrix S is called* reducible *if, after suitable permutations of rows and columns, it has the form*

$$S = \begin{pmatrix} * & 0 \\ * & * \end{pmatrix}$$

Otherwise, S is called irreducible.

Lemma 5.1.10. *If S is irreducible, then there exists a stationary vector whose entries are all positive.*

If one replaces in S all non-zero entries by 1, then one obtains the *adjacency matrix* of the network.

Lemma 5.1.11. *A graph is strongly connected, if and only if its adjacency matrix is irreducible.*

Lemma 5.1.12. *Any primitive matrix is irreducible.*

The theorem of Perron-Frobenius now delivers what we need:

Theorem 5.1.13 (Perron-Frobenius)**.** *Let S be a primitive stochastic matrix. Then:*

1. *1 is an eigen value of S with multiplicity 1 (i.e. the eigen space for eigen value 1 is one-dimensional).*

2. *1 is the eigen value of S with the largest absolute value, all other eigen values have a smaller absolute value.*

3. *The eigen vectors for eigen value 1 have either only positive or only negative entries. In particular, there exists an eigen vector for eigen value 1 whose sum of entries equals one.*

Final modifikation

The final modification is done with a parameter $\alpha \in (0,1)$. With probability α a random surfer follows matrix S, and with probability $1-\alpha$ he or she jumps to any page.

Let

$$\mathbb{1} := \begin{pmatrix} 1 & 1 & \dots \\ 1 & 1 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

Then the *google matrix* is

$$G = \alpha S + (1-\alpha)\frac{1}{n}\mathbb{1}$$

Sometimes, α is called the *damping constant*.

Remark 5.1.14. *The google matrix G is stochastic and primitiv.*

Consequence. *G has a stationary vector whose entries are all positive and which can be found by the power method.*

$\alpha = 1$ yields the original network, and for $\alpha = 0$ we have the complete graph: every node is equally probable. For using the network structure it is preferred to have α near 1.

Remark 5.1.15. *For the second eigen value λ_2 of G it holds true that:*

$$|\lambda_2| = \alpha$$

This means that α should not be too close to 1. Sergey Brin und Larry Page choose $\alpha = 0.85$.

Application to the internet

The google matrix has the structure

$$G = \alpha H + \alpha A + \frac{1-\alpha}{n}\mathbb{1}$$

With the power method we have

$$GI^k = \alpha H I^k + \alpha A I^k + \frac{1-\alpha}{n}\mathbb{1} I^k$$

Notice that H contains mostly zeros, and that in A and $\mathbb{1}$ all rows are identical. For computing the last two summands, one needs to add the importance ranks of the dangling nodes or all nodes, respectively. This has to be done only once.

It is claimed that after 50-100 iterations, I is sufficiently well approximated. This needs a few days. It is rumored that the PageRank I is updated about every month.

5.1.2 Some Topology

"Mmm... donuts." H. Simpson

Inspired by the internet graph of the previous section, we note that graph theory can be viewed as a subdomain of topology. Topology deals with properties of spaces which remain unchanged under continuous deformations as stretching and bending, but not e.g. tearing apart. The aim of this section is to understand Homer's last line on the blackboard in *The Wizard of Evergreen Terrace* (cf. Section 1.6). This line is reproduced in Figure 5.4.

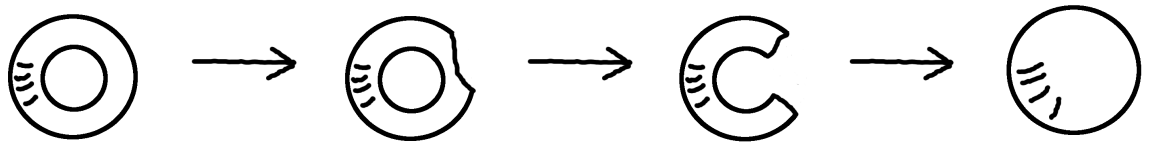


Figure 5.4: Nibbling (followed by a topological deformation) is a topological transformation allowed by Homer Simpson, in order to transform a donut into a sphere.

First back to graphs. A graph consists of nodes (also called vertices), some of which are connected with others by edges. Between a pair of nodes there is at most one edge. Let b_0 be the number of connected components and b_1 the number of "holes", then there is a relation to the number V of vertices and the number E of edges:

Lemma 5.1.16. *For finite graphs, it holds true that:*

$$V - E = b_0 - b_1$$

Proof. Neither side of the equation changes if an edge is contracted: by this, an edge with two distinct vertices is replaced by a vertex. By performing edge contractions until only edges are left which have a single vertex each, then every connected component is a graph with precisely one vertex. Edges look like "petals". Every edge corresponds to precisely one hole and vice versa. Hence, $V = b_0$ and $E = b_1$, and the equation holds true. □

The right hand side of the equation in in Lemma 5.1.16 is called *Euler charakteristic*, und b_i is called *i-th Betti number*.

Betti numbers also exist in higher dimension: b_2 is the number of cavities of a closed orientable surface t die Anzahl der Hohlkörper einer geschlossenen orientierbaren Fläche (or also in higher dimension). *Orientiable* means that the surface has two sides: an *interior* and an *exterior*. The Betti numbers are topological invariants in the sense that they do not change under continuous deformations. One speaks of a *homeomorphism* if there is a continuous deformation between the two objects. For example, a square is homeomorphich to a circle, or a coffee cup is homeomorphic to a torus (for Homer Simpson: a donut).

For a sphere, the first Betti number is zero, as there is no "hole" or "tunnel". For a donut, the first Betti number is two, as there is the exterior hole in the middle, and the tunnel in the interior. For a sphere with g handles, we have $b_1 = 2g$, as each handle has a contribution of 2: an exterior hole and an interior tunnel.

For closed orientable surfaces there is the following classification:

Theorem 5.1.17. *For every closed orientable surface there is a $g \geq 0$, such that it is homeomorphic to a sphere with g handles.*

The number g is called the *genus* of the surface. As the genus is a topological invariant, it follows that:

Consequence. *A donut and a sphere are not homeomorphic.*

Proof. A donut has genus 1, whereas a sphere has genus 0. As the genus does not change under homeomorphismds, it follows that a donut and a sphere cannot be homeomorphic. □

As Homer Simpson loves donuts, he allows not only continuos deformations, but also a transformation named *nibble*. In this way, he can transform a donut into a sphere. The expression "homeromorphism" offers itself.

Theorem 5.1.18 (Homer Simpson, 1998)**.** *Donut and sphere are "homeromorphic".*

Proof. Transform the donut by nibbling until it is homeomorphic to the third object from the left in Figure 5.4. Afterwards, perform a continuous deformation. □

There exist also non-orientable surfaces. In the episode *Möbius Dick* of the animated sitcom *Futurama*[1] the space ship *Planet Express* flies through the *Bermuda Tetrahedron*. Similar to the Bermuda Triangle on Earth, there are heaps of lost space ships. There, a four-dimensional whale appears. The title of this episode is an allusion to this whale. Further, "Möbius" alludes to the *Möbius strip*, a non-orientable surface which can be constructed as follows:

Take a rectangular strip and identify the two short sides after a half twist. The result can be seen in Figure 5.5.

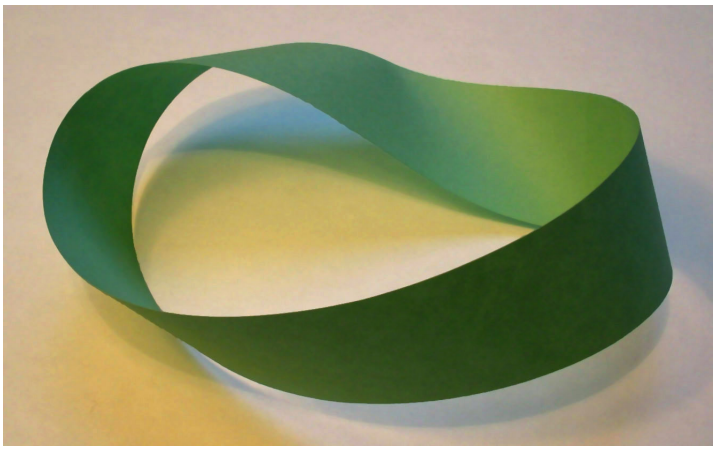


Figure 5.5: A Möbius strip made of paper (Source: Wikipedia, author: David Benbennick).

If one starts in the interior of the surface from some point and walks parallel to the boundary, then after one revolution, one reaches the starting point, but on the other "side". After another revolution, one is again at the starting point on the original "side". This shows that the surface has only one side. Hence, it is not orientable.

What kind of surface does one get if one slits the Möbius strip parallel to the boundary?

[1] created by Matt Groening, the inventor of *The Simpsons*

5.1.3 Alexandrov topologies

Definition 5.1.19. *Let $R \subset X \times X$ be a relation. R is called* reflexive*, if for all $x \in X$ we have*

$$(x, x) \in R$$

R is called transitive*, if for all $x, y, z \in X$ we have*

$$(x, y) \in R \text{ and } (y, z) \in R \;\Rightarrow\; (x, y) \in R$$

We often write

$$xRy$$

instead of $(x, y) \in R$.

Pavel Alexandrov (1896–1982) discovered that a reflexive and transitive relation R on a set X defines a topology, for which every point $x \in X$ has a *minimal neighbourhood*

$$U(x) = \{y \in X \mid (x, y) \in R\}$$

Interesting is the case of a *partial order* $\leq$ on X. Such is a reflexive and transitive relation which is furthermore *anti-symmetric*:

$$x \leq y \text{ and } y \leq x \;\Rightarrow\; x = y$$

for all $x, y \in X$. The corresponding topological space is called a T_0-space.

Definition 5.1.20. *Let $R \subset X \times X$ be a relation. Then*

$$\begin{aligned} R^0 &= \{(x, x) \in X \times X\} \\ R^2 = R \circ R &= \{(x, z) \in X \times X \mid \textit{there exists } y \in X \textit{ such that } (x, y) \in R \textit{ and } (y, z) \in R\} \\ R^{n+1} &= R^n \circ R \end{aligned}$$

The relation

$$R^* = \bigcup_{n \in \mathbb{N}} R^n$$

is called the reflexive and transitive closure *of R.*

Example 5.1.21. *In geoinformatics, there is the relation* bounded-by*, for which e.g. a room in a building is bounded by a wall, and this is bounded by an edge, and that by a vertex. This so-called* incidence topology *is the reflexive and transitive closure of* bounded-by*. This defines a T_0-space, whose points are the volumes, areas, lines and points, viewed as building components.*

Example 5.1.22. *Another example is the relation*

$$\preceq = \textit{is-requirement-for-acquisition-of}$$

on the set

$$M = \{\textit{prayer, illumination of the mind, attentiveness, self-observation, knowing of oneself, repentance, humility, grace of God}\}$$

By grace *we mean an (unmerited) gift. We have:*

$$\begin{aligned} \textit{prayer} \preceq \textit{illumination of the mind} \preceq \textit{attentiveness} \preceq \textit{self-observation} \\ \preceq \textit{knowing of oneself} \preceq \textit{repentance} \preceq \textit{humility} \preceq \textit{grace of God} \end{aligned}$$

Furthermore, the relation $\preceq$ is a partial order. This special case is called total order. *This implies that if one element of M is missing in a person, then the grace of God does not rest on that person:*

> *For whosoever shall keep the whole law, and yet offend in one point, he is guilty of all*
> *James 2:10*

In particular, without humility there is no grace of God! However, $\leq$ is not the only possible topology on M. This is to be understood more as an instruction for receiving the grace of God. In any case, we have

$$\textit{humility} \preceq \textit{grace of God}$$
$$\textit{repentance} \preceq \textit{grace of God}$$

The other elements of M are needed for obtaining repentance and humility. Examples are the parabola of the tollkeeper and the pharisee, as well as the two robbers crucified with Jesus. Saint John Climacus writes in The Ladder of Divine Ascent*:*

> *Some of the faithful, and even of the unfaithful, have been deserted by the passions, all except one [i.e. pride]; and that one has been left as a paramount evil which fully takes the place of all the others, for it is so harmful that it can even cast down from heaven.*
> *(Step 26, 62)*

For only when we humble ourselves, we become like God:

> *Take my yoke upon you, and learn of me; for I am meek and lowly in heart: and ye shall find rest unto your souls.*
> *(Matthew 11:29)*

5.2 One equation in one unknown

The problem in this section is: solve the equation

$$a \cdot X = b$$

with given a, b.

Example 5.2.1. *Over $K = \mathbb{Q}$ or $\mathbb{R}$ the inverse a^{-1} can be computed e.g. with Newton-Raphson division (cf. Section 2.4.1). Then $X = a^{-1} \cdot b$.*

Example 5.2.2. *Solve*

$$2 \cdot X \equiv 1 \mod 3$$

Solution: $X \equiv 2 \mod 3$, since:

$$2 \cdot 2 \equiv 4 \equiv 1 \mod 3$$

and for $X \equiv 0$ or $1 \mod 3$ it holds true that:

$$2 \cdot X \not\equiv 1 \mod 3$$

We found the solution with brute force. Is it possible to do this more efficiently?

Answer. *Calculate*

$$1 = \operatorname{lcd}(2,3) = x \cdot 2 + y \cdot 3$$

Then

$$x \cdot 2 \equiv 1 \mod 3$$

E.g. $x = 2$, $y = -1$ *satisfy this condition.*

Theorem 5.2.3. *The largest common divisor* $d = \operatorname{lcd}(a,b)$ *of* $a, b \in \mathbb{Z}$ *has a linear representation*

$$d = x \cdot a + y \cdot b$$

with $x, y \in \mathbb{Z}$.

Consequence. *If* a *and* n *are coprime, then the congruence*

$$a \cdot X \equiv b \mod n$$

has a unique solution.

Proof. The euclidean algorithm yields

$$\begin{aligned} a &= b \cdot q + r, \quad |r| < |b| \\ b &= r \cdot q_1 + r_1, \quad |r_1| < |r| \\ &\vdots \\ r_{n-2} &= r_{n-1} \cdot q_{n-2} + r_n, \quad |r_n| < |r_{n-1}| \end{aligned}$$

and $r_n = d$. This yields:

$$\begin{aligned} r &= a - b \cdot q \\ r_1 &= b - r \cdot q_1 = b - (a - b \cdot q) \cdot q_1 = b \cdot (1 + qq_1) - a \cdot q_1 \\ &\vdots \\ r_{n-1} &= r_{n-3} - r_{n-2} \cdot q_{n-3} \\ d &= r_{n-2} - r_{n-1} \cdot q_{n-2} \\ &= \underbrace{r_{n-2} - (r_{n-3} - r_{n-2} \cdot q_{n-3}) \cdot q_{n-2}}_{= r_{n-2} \cdot (\dots) + r_{n-3} \cdot q_{n-2}} = \cdots = a \cdot x + b \cdot y \end{aligned}$$

□

The method in the proof of Theorem 5.2.3 is called *extended euclidean algorithm.*

Consequence. *If* p *is a prime number, then any equation*

$$a \cdot X \equiv b \mod p$$

with $a \not\equiv 0 \mod p$ *has a unique solution.*

In particular, every $a \not\equiv 0 \mod p$ has a multiplicative inverse modulo p. This means that

$$\mathbb{F}_p := \{0, \dots, p-1\}$$

is a *field*, if addition and multiplication are taken modulo p. In this field, we have the *euclidean division.*

5.3 Gauß algorithm

The $n \times n$-matrices

$$K^{n\times n} := \{A = (a_{ij}) \mid a_{ij} \in K\}$$

with entries in a field K form a *unitary ring* under addition and matrix multiplication. For $n \geq 2$, this ring is non-commutative. This means that the usual calculation laws for $+$ and $\cdot$ are valid[2] with the exceptions that not every non-zero $n \times n$-matrix is invertible, and in general the rule $A \cdot B \neq B \cdot A$ holds true. E.g. for $n = 2$:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \neq \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Elementary matrices

Let

$$E_{ij} = (\epsilon_{pq}^{ij})$$

with

$$\epsilon_{pq}^{ij} = \begin{cases} 1, & (p,q) = (i,j) \\ 0 & \text{sonst} \end{cases}$$

This matrix is called *elementary matrix* and has a 1 in the i-th row and j-th column, otherwise 0. It holds true that:

$$E_{ij} \cdot E_{k\ell} = \begin{cases} E_{i\ell}, & j = k \\ 0 & \text{sonst} \end{cases}$$

Proof. $E_{ij} \cdot E_{k\ell} = (\gamma_{rs})$ with

$$\gamma_{rs} = \sum_t \epsilon_{rt}^{ij} \epsilon_{ts}^{k\ell} = \epsilon_{rj}^{ij} \epsilon_{js}^{k\ell} = \begin{cases} 1, & (r,s) = (i,\ell) \quad \text{und} \quad j = k \\ 0 & \text{otherwise} \end{cases}$$

□

By linear combination, other matrices can be built from elementary matrices, e.g.:

$$I := \sum_{i=1}^{n} E_{ii} \qquad \text{(unity matrix)}$$

$$\text{Diag}(\alpha_1, \ldots, \alpha_n) := \sum_{i=1}^{n} \alpha_i E_{ii} \qquad \text{(diagonal matrix)}$$

$$M_i(\alpha) := I + (\alpha - 1) \cdot E_{ii} \qquad \text{(multiplication matrix)}$$

$$A_{ij}(\alpha) := I + \alpha E_{ij} \qquad (i \neq j) \qquad \text{(addition matrix)}$$

$$V_{ij} := I - E_{ii} - E_{jj} + E_{ij} + E_{ji} \qquad \text{(transposition matrix)}$$

The meaning of the last three matrices becomes clear through the following examples in $K^{4\times 4}$.

Example 5.3.1.

$$M_1(\alpha) = \begin{pmatrix} \alpha & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix} = I + (\alpha - 1)E_{11}$$

[2]The role of the zero is played by the zero matrix, the role of the one by the unity matrix.

Multiplication from the left and from the right to any 4×4-matrix yields:

$$M_1(\alpha) \cdot \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix} = \begin{pmatrix} \alpha \cdot \alpha_{11} & \alpha \cdot \alpha_{12} & \alpha \cdot \alpha_{13} & \alpha \cdot \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix}$$

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix} \cdot M_1(\alpha) = \begin{pmatrix} \alpha \cdot \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha \cdot \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha \cdot \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha \cdot \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix}$$

Example 5.3.2.

$$A_{23}(\beta) = \begin{pmatrix} 1 & & & \\ & 1 & \beta & \\ & & 1 & \\ & & & 1 \end{pmatrix} = I + \beta \cdot E_{23}$$

$$A_{23} \cdot \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix}$$

$$= \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} + \beta \cdot \alpha_{31} & \alpha_{22} + \beta \cdot \alpha_{32} & \alpha_{23} + \beta \cdot \alpha_{33} & \alpha_{24} + \beta \cdot \alpha_{34} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix}$$

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix} \cdot A_{23}(\beta) = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} + \beta \cdot \alpha_{12} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} + \beta \cdot \alpha_{22} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} + \beta \cdot \alpha_{32} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} + \beta \cdot \alpha_{42} & \alpha_{44} \end{pmatrix}$$

Example 5.3.3.

$$V_{24} = \begin{pmatrix} 1 & & & \\ & & & 1 \\ & & 1 & \\ & 1 & & \end{pmatrix} = I - E_{22} - E_{44} + E_{24} + E_{42}$$

$$V_{24} \cdot \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \end{pmatrix}$$

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{pmatrix} \cdot V_{24} = \begin{pmatrix} \alpha_{11} & \alpha_{14} & \alpha_{13} & \alpha_{12} \\ \alpha_{21} & \alpha_{24} & \alpha_{23} & \alpha_{22} \\ \alpha_{31} & \alpha_{34} & \alpha_{33} & \alpha_{32} \\ \alpha_{41} & \alpha_{44} & \alpha_{43} & \alpha_{42} \end{pmatrix}$$

Lemma 5.3.4. *The matrix $M_i(\alpha)$ is invertible for $\alpha \neq 0$. $A_{ij}(\alpha)$ is invertible for all $\alpha \in K$. V_{ij} is invertible. These matrices perform the following operations on a given matrix A with compatible size:*

- *$M_i(\alpha)$ from the left: multiplies the i-th row of A with α.*
- *$A_{ij}(\alpha)$ from the left: adds α times the j-th row of A onto the i-th row of A.*
- *V_{ij} from the left: swaps i-th and j-th row of A.*
- *$M_i(\alpha)$ from the right: multiplies the i-th column of A with α.*
- *$A_{ij}(\alpha)$ from the right: adds α times the i-th column onto the j-th column of A.*
- *V_{ij} from the right: swaps i-th and j-th column of A.*

The inverses are respectively:

$$M_i(\alpha)^{-1} = M_i(\alpha^{-1})$$
$$A_{ij}(\alpha)^{-1} = A_{ij}(-\alpha)$$
$$V_{ij}^{-1} = V_{ij}$$

Consequently, for solving a system of linear equations

$$A \cdot x = b$$

with $A \in K^{m\times n}$ and $b \in K^m$: we have the Gauß algorithm (with only row operations): Multiplication from the left with an invertible matrix $B \in K^{m\times m}$ whose result is the stair-case normal form

$$T = \begin{pmatrix} 1 & * & & & * & * & \\ & & 1 & & * & * & \\ & & & 1 & * & * & \\ & & & & & & 1 \end{pmatrix}$$

The stair-case normal form yields a basis of the solution space of the homogeneous system

$$Ax = 0$$

as follows: insert rows with precisely one -1-entry below every non-step in order to obtain the extended stair-case

$$\tilde{T} = \begin{pmatrix} 1 & * & & & * & * & \\ & -1 & & & & & \\ & & 1 & & * & * & \\ & & & 1 & * & * & \\ & & & & -1 & & \\ & & & & & -1 & \\ & & & & & & 1 \end{pmatrix}$$

The columns with the new -1's are a basis of the homogeneous system of equations:

$$\mathbb{L} = \left\langle \begin{pmatrix} * \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} * \\ 0 \\ * \\ * \\ -1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} * \\ 0 \\ * \\ * \\ 0 \\ -1 \\ 0 \end{pmatrix} \right\rangle$$

This leads to our first decomposition of A:

$$A = C \cdot T$$

with C invertible and T in stair-case normal form.

5.4 LU decomposition

Let $A \in K^{n\times n}$.

Definition 5.4.1. *An LU decomposition of A is a factorisation of the form*

$$A = L \cdot U,$$

where L is a lower triangular matrix whose diagonal entries are all equal to one, and U is an upper triangular matrix.

Doolittle algorithm

The *doolittle algorithm* can under certain conditions obtain an LU decpomposition of A.

Let $A_0 := A$. For $\nu = 1, \dots, n$ let $A_\nu := L_\nu \cdot A_{\nu-1} = (\alpha_{ij}^{(\nu)})$ with

$$L_\nu = A_{n,\nu}(\ell_{n,\nu}) \cdot A_{n-1,\nu}(\ell_{n-1,\nu}) \cdots A_{\nu+1,\nu}(\ell_{\nu+1,\nu}) = \begin{pmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & \ell_{\nu+1,\nu} & \ddots & \\ & & \vdots & & \ddots \\ & & \ell_{n,\nu} & & 1 \end{pmatrix}$$

and

$$\ell_{i,\nu} := -\frac{\alpha_{i\nu}^{(\nu-1)}}{\alpha_{\nu\nu}^{(\nu-1)}}, \quad (i = \nu+1, \dots, n)$$

Observe that L_ν eliminates in the ν-th row of $A_{\nu-1}$ everything below the diagonal. In particular, we have

$$U := A_{n-1} = L_{n-1} \cdots L_1 \cdot A$$

is an upper triangular matrix. Then

$$L := L_1^{-1} \cdots L_{n-1}^{-1} = \begin{pmatrix} 1 & & & & & 0 \\ -\ell_{2,1} & \ddots & & & & \\ & & 1 & & & \\ \vdots & & -\ell_{\nu+1,\nu} & \ddots & & \\ & & \vdots & & 1 & \\ -\ell_{n,1} & & -\ell_{n,\nu} & & -\ell_{n,n-1} & 1 \end{pmatrix}$$

is a lower triangular matrix whose diagonal entries are all one, and

$$A = L \cdot U$$

However, in each step it is used that $\alpha_{\nu\nu}^{(\nu-1)} \neq 0$. If this is not the case, then swap the ν-th row with a row below. In the end this yields a *PLU decomposition*

$$A = P \cdot L \cdot U$$

with a *permutation matrix* $P = (\pi_{ij})$, where

$$\pi_{ij} = \delta_{i,\sigma(j)}$$

for a permutation σ of the numbers $1, \dots, n$.

Applications of the LU decomposition

1. Solution of $Ax = b$ with $A \in K^{n\times n}$, $b \in K^n$. The decomposition $A = L \cdot U$ leads to

$$L\underbrace{(Ux)}_{=y} = b \quad \rightsquigarrow \quad Ux = y$$

   These are two systems of linear equations. The first one is solved by *forward substitution*, and the second one by *backward substitution* (cf. the following paragraph).

2. The inverse of $A \in K^{n\times n}$. Apply 1. simultaneously on every column of I as the right hand side b.

3. Determinant. It holds true that:

$$\det(A) = \det(L) \cdot \det(U) = \det(U)$$

Forward substitution

Let a triangular system of linear equations be given:

$$\begin{aligned}
\ell_{11}y_1 &= b_1 \\
\ell_{21}y_1 + \ell_{22}y_2 &= b_2 \\
&\vdots \\
\ell_{n1}y_1 + \cdots + \ell_{nn}y_n &= b_n
\end{aligned}$$

Then

$$\begin{aligned}
y_1 &= \frac{b_1}{\ell_{11}} \\
y_i &= \frac{1}{\ell_{ii}}\left(b_i - \sum_{k=1}^{i-1} \ell_{ik}y_k\right), \qquad i = 2, \ldots, n
\end{aligned}$$

is the solution if all $\ell_{ii} \neq 0$.

Backward substitution

Let the triangular system of linear equations be given:

$$\begin{aligned}
u_{nn}x_n &= y_n \\
u_{n-1,n-1}x_{n-1} + u_{n-1,n}x_n &= y_{n-1} \\
&\vdots \\
u_{11}x_1 + \cdots + u_{1n}x_n &= y_1
\end{aligned}$$

Then

$$\begin{aligned}
x_n &= \frac{y_n}{u_{nn}} \\
x_i &= \frac{1}{u_{ii}}\left(y_i - \sum_{k=i+1}^{n} u_{ik}x_k\right), \qquad i = n-1, \ldots, 1
\end{aligned}$$

is the solution, if all $u_{ii} \neq 0$.

5.4.1 Sorting pancakes

Permutations as in the previous section occurr in *The Simpsons* in a hidden manner, namely in the form of the "Municipal House of Pancakes" in Springfield, Homer Simpson's home town. This appeared first in the episode *The Twisted World of Marge Simpson* (1997). Assume that one of its waiters serves n pancakes in random order. He can reverse pancakes on the serving plate by taking a few from the top with a fish slice (also called "spatula") and then flip that stack. The question now is, how often does he have to flip stacks of pancakes in the worst case until they are all sorted by size. The number of flips is then called *pancake number* and is denoted as P_n. Wanted is a formula which describes P_n.

Computer scientists like sorting data, and there are parallels with pancakes. The number P_n is only known up to $n = 19$. This is why the pancake sorting problem is of interest.

One can calculate the pancake number for the first few values of n by going through all combinations of different-sized pancakes and determining the number of flips.

$P_1 = 0$, as the only pancake is already in the correct order.

$P_2 = 1$, as in the worst case, the big pancake lies on the small pancake, and then there is one flip.

To determine P_3 is already a bit more difficult. Threre are the six possibilities $(1, 2, 3)$, $(1, 3, 2)$, $(2, 3, 1)$, $(2, 1, 3)$, $(3, 1, 2)$, $(3, 2, 1)$, where the number corresponds to the size and the position from the left corresponds to the position of the pancake from top to bottom. The number of flips is given in the following table:

Permutation	$(1,2,3)$	$(1,3,2)$	$(2,3,1)$	$(2,1,3)$	$(3,1,2)$	$(3,2,1)$
Number of flips	0	3	3	2	2	1

Hence, $P_3 = 3$.

In the year 1979, an upper bound for P_n was found. Namely:

Theorem 5.4.2 (William H. Gates & Christo H. Papadimitriou, 1979)**.** *It holds true that*

$$P_n \leq \frac{5n + 5}{3}$$

William H. Gates is better known as Bill Gates and is a co-founder of the company Microsoft.

David S. Cohen, one of the authors of *The Simpsons*, published in the year 1995 an article about the *burnt pancake problem*, where the objective is to sort pancakes which are burnt on one side in such a way that they are sorted by size and the burnt side is always facing down. Let the number of flips in this problem be denoted as V_n. We have:

Theorem 5.4.3 (David S. Cohen, 1995)**.** *The following holds true:*

$$\frac{3n}{2} \leq V_n \leq 2n - 2$$

5.5 The Spectral Theorem

5.5.1 Eigen spaces

Let $A \in K^{n \times n}$.

Definition 5.5.1. *The* eigen space $E_\lambda(A)$ *for the eigen value* λ *is defined as the solution space of*

$$A - \lambda \cdot I$$

if it is $\neq 0$.

Example 5.5.2. *Let*

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

Then $E_1(A) = K \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ *and* $E_{-1} = K \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

Lemma 5.5.3. $\lambda \in K$ *is an eigen value of* A*, if and only if* λ *is a zero of*

$$f_A(X) = \det(X \cdot I - A)$$

The latter is the charakteristic polynomial *of* A*.*

Proof. $A - \lambda \cdot I$ is not invertible, if and only if the corresponding solution space is non-trivial. □

Definition 5.5.4. *An* eigen vector *of* A *is a non-trivial element of an eigen space of* A*.*

5.5.2 Base change

Definition 5.5.5. *A matrix* $A \in K^{n \times n}$ *is called* diagonalisable*, if there exists an invertible matrix* S *such that*

$$S^{-1}AS$$

is a diagonal matrix.

Example 5.5.6. *Let*

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

A basis of $\mathbb{R}^2$ *consisting of eigen vectors of* A *is given by the following matrix*

$$S = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

It holds true that

$$S^{-1} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix}$$

After base change, we have

$$S^{-1}AS = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Definition 5.5.7. *The matrix* $S^{-1}AS$ *is called* similar *to* A*.*

Remark 5.5.8. *Similar matrices have the same eigen values.*

Theorem 5.5.9 (Spektral theorem)**.** *Let* $A \in K^{n \times n}$ *be a diagonalisable matrix. Then* K^n *has a basis* S *consisting of eigen vectors of* A*. The diagonal entries of the diagonal matrix* $S^{-1}AS$ *are the eigen values of* A*.*

Proof. Let

$$S^{-1}AS = \operatorname{Diag}(\lambda_1, \ldots, \lambda_n)$$

Then we have:

$$AS = S \operatorname{Diag}(\lambda_1, \ldots, \lambda_n)$$

Hence, for the i-th column of S, we have:

$$As_i = \lambda_i s_i$$

and the basis S of K^n consists of eigen vectors s_i of A for the eigen value λ_i. □

Definition 5.5.10. *The set*

$$\mathrm{Spec}(A) = \{\lambda \mid \lambda \textit{ ist Eigenwert von } A\}$$

is the spectrum *of* A.

5.5.3 Determinant and trace

Let $A = (a_{ij}) \in K^{n\times n}$. Then the *determinant* of A is defined as

$$\det(A) = \sum_{\pi\in S_n} \mathrm{sgn}(\pi) \prod_{i=1}^{n} a_{i,\pi(i)}$$

where S_n is the set of all permutations of the numbers $1,\dots,n$ and $\mathrm{sgn}(\pi)$ is the *sign* of the permutation π:

$$\mathrm{sgn}(\pi) = \begin{cases} 1, & \pi \text{ is an even permutation} \\ -1, & \pi \text{ is an odd permutation} \end{cases}$$

Example 5.5.11. *For* $n = 1$, $A = (a)$ *and* $\det(A) = a$.

For $n = 2$,

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

and

$$\det(A) = a_{11}a_{22} - a_{12}a_{21}$$

As the identical permutation (1) *is even and the transposition* (12) *is an odd permutation.*

For $n = 3$,

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

and

$$\det(A) = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{21} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}$$

as the identical permutation (1) *and the cycles* (123) *and* (132) *are even, and the transpositions* (12), (13), (23) *are odd.*

Definition 5.5.12. *The* trace $\mathrm{trace}(A)$ *is defined as the sum of the diagonal entries of* A.

Example 5.5.13. *Let* $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. *Then*

$$\mathrm{trace}\, A = a + d, \quad \det(A) = ad - bc$$

Consider the characteristic polynomial:

$$f_A(X) = \det(X \cdot I - A) = X^2 - (a+d)X + ad - bc = X^2 - \mathrm{trace}(A)X + \det(A)$$

If $f_A(X) = (X - \lambda_1)(X - \lambda_2)$ *with* $\lambda_1, \lambda_2 \in \mathbb{C}$, *then*

$$f_A(X) = X^2 - (\lambda_1 + \lambda_2)X + \lambda_1\lambda_2$$

Comparing the coefficients yields:

$$\mathrm{trace}(A) = \lambda_1 + \lambda_2$$
$$\det(A) = \lambda_1 \cdot \lambda_2$$

Lemma 5.5.14. *Let $A \in \mathbb{C}^{n\times n}$. Then:*

$$\operatorname{trace}(A) = \sum_{\lambda\in\operatorname{Spec}(A)} \lambda$$
$$\det(A) = \prod_{\lambda\in\operatorname{Spec}(A)} \lambda$$

Proof. We have

$$\det(X\cdot I - A) = \prod_{\lambda\in\operatorname{Spec}(A)} (X-\lambda)$$

On the right hand side, the constant term equals

$$(-1)^n \cdot \prod_{\lambda\in\operatorname{Spec}(A)} \lambda$$

On the left hands side, the constant term equals

$$\det(-A) = (-1)^n \cdot \det(A)$$

On the right hand side, the coefficient of X^{n-1} equals

$$-\sum_{\lambda\in\operatorname{Spec}(A)} \lambda$$

On the left hand side, the coefficient of X^{n-1} equals

$$-\operatorname{trace}(A)$$

Comparing the coefficients yields the assertion. □

5.5.4 The Futurama Theorem

First, some preliminary remarks on permutations. Each permutation of the numbers $1,\dots,n$ can be written as the product of disjoint cycles as follows: start with 1, the follow with $\sigma(1)$, then $\sigma^2(1) := \sigma(\sigma(1))$, etc. until for the first time again $\sigma^k(1) = 1$. This is the first cycle. If in this cycle, a number in $\{1,\dots,n\}$ does not occur, then continue as above with one such number. Etc. At some point, each of the numbers $1,\dots,n$ occurs in precisely one cycle. This yields a decomposition of the permutation into disjoint cycles.

In the episode *The Prisoner of Benda* of *Futurama*, there is a machine invented by Professor Farnsworth which swaps minds. If applied to two persons A and B, then afterwards, the mind of A is in the body of B, and vice versa. In a group of n persons[3] all have swapped minds with different partners, and after some time they all want to get back into their original bodies. The problem, however, is that the machine will work only once for a given pair of persons.

Ken Keeler, the main author of this episode, faced the problem of figuring out how all minds can get back into their original bodies. After some effort, he proved the *Futurama Theorem*:

Theorem 5.5.15 (Ken Keeler, 2010)**.** *It suffices to add two more persons, such that each mind gets back into its original body.*

The following proof is developped in the episode on a black board:

[3] in the episode, there are 8 persons

Proof. Let π be the permutation of $[n] := \{1, \dots, n\}$, which is created by the sequence of mind swaps.

Case 1. Assume that π is a cycle of length k. Without restriction, we may assume that

$$\pi = \begin{pmatrix} 1 & 2 & \dots & k & k+1 & \dots & n \\ 2 & 3 & \dots & 1 & k+1 & \dots & n \end{pmatrix}$$

Here, the pre-images of the permutation (minds) are on top and the images (bodies) on the bottom. Let

$$\pi^* := \begin{pmatrix} 1 & 2 & \dots & k & k+1 & \dots & n & x & y \\ 2 & 3 & \dots & 1 & k+1 & \dots & n & x & y \end{pmatrix} \quad \text{mit} \quad x, y \notin [n]$$

and let $(a\ b)$ be the transposition which swaps a and b. Let

$$\sigma := (x\ 1) \circ (y\ 2) \cdots (y\ k) \circ (x\ 2) \circ (y\ 1)$$

Then

$$\pi^* \circ \sigma = \begin{pmatrix} 1 & 2 & \dots & n & x & y \\ 1 & 2 & \dots & n & y & x \end{pmatrix} \qquad (*)$$

Hence, only x and y need to swap their minds, which is possible, as they have not yet been attached to the machine yet.

Case 2. Let π be any permutation of $[n]$. Decompose π into a product of disjoint cycles, and apply case 1 up to $(*)$ on each cycle. Afterwards, swap x with y, if necessary. □

5.5.5 Positive definite matrices

Definition 5.5.16. *A matrix $A \in \mathbb{C}^{n \times n}$ is called* hermitean*, if*

$$A^* := \bar{A}^\top = A$$

where A^* is the complex conjugate of the transpose of A.

A special case is given by real matrices: a real matrix A is hermitean, if and only if it is symmetric:

$$A^\top = A$$

Rules of calculation

The following holds true:

$$(A^*)^* = A$$

and

$$(AB)^* = B^* A^*$$
$$(A^*)^{-1} = \left(A^{-1}\right)^*$$

whenever the expressions are defined.

Proof. Let $A = (\alpha_{ij})$, $B = (\beta_{ij})$, $AB = (\gamma_{ij})$ and $A^* = (\alpha'_{ij})$, $B^* = (\beta'_{ij})$, $(AB)^* = (\gamma'_{ij})$. Further, let $(A^*)^* = (\gamma'_{ij})$. Then

$$\gamma'_{ij} = \overline{\alpha'}_{ji} = \bar{\bar{\alpha}}_{ij} = \alpha_{ij}$$

Hence: $(A^*)^* = A$. Further,

$$\gamma'_{ij} = \bar{\gamma}_{ji} = \sum_k \bar{\alpha}_{jk}\bar{\beta}_{ki} = \sum_k \bar{\beta_{ki}}\bar{\alpha}_{jk} = \sum_k \beta'_{ik}\alpha'_{kj}$$

Hence: $(AB)^* = B^*A^*$. Further,

$$\left(A^{-1}\right)^* A^* = \left(AA^{-1}\right)^* = I^* = I$$

Hence, $(A^*)^{-1} = \left(A^{-1}\right)^*$. □

Definition 5.5.17. *A hermitean matrix $A \in \mathbb{C}^{n\times n}$ is called* positive (semi-)definite*, if for all $z \in \mathbb{C}^n$:*

$$z^*Az > 0 \qquad (z^*Az \geq 0)$$

if $z \neq 0$.

Remark 5.5.18. *If A is hermitean, then z^*Az is real.*

Proof. It holds true that

$$\overline{z^*Az} = (z^*Az)^* = z^*A^*z^{**} = z^*Az$$

□

Remark 5.5.19. *A symmetric real matrix $A \in \mathbb{R}^{n\times n}$ is positive (semi-)definite, if and only if for all $x \in \mathbb{R}^n$:*

$$x^\top Ax > 0 \qquad (x^\top Ax \geq 0)$$

if $x \neq 0$ ist.

Example 5.5.20. *The unity matrix I is positive definite. The reason: with $z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} \in \mathbb{C}^n$ we have*

$$z^*Iz = z^*z = \sum_{i=1}^n \bar{z}_i z_i = \sum_{i=1}^n |z_i|^2 > 0$$

if $z \neq 0$.

Example 5.5.21. *The symmetric real matrix*

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

is positive definite. The reason: with $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ we have

$$\begin{aligned} x^\top Ax &= x^\top \begin{pmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 - x_3 \\ -x_2 + 2x_3 \end{pmatrix} = 2x_1^2 - 2x_1x_2 + 2x_2^2 - 2x_2x_3 + 2x_3^2 \\ &= x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + x_3^2 > 0 \end{aligned}$$

if $x \neq 0$.

Example 5.5.22. $A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ *is not positive definite. The reason: with* $z = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ *we have*

$$z^\top Az = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = -2 < 0$$

but $z \neq 0$.

Lemma 5.5.23. *The eigen values of hermitean matrices are real.*

Proof. Let e be an eigen vector of the hermitean matrix A for eigen value $\lambda \in \mathbb{C}$. Then:

$$\mathbb{R} \ni e^*Ae = e^*(\lambda e) = \lambda \cdot \underbrace{e^*e}_{=\gamma}$$

and $\gamma > 0$ is real. Then also λ is real. □

Theorem 5.5.24 (Spektral Theorem II)**.** *A hermitean matrix* $A \in \mathbb{C}^{n\times n}$ *is diagonalisable and has only real eigen values. Further,* $\mathbb{C}^n$ *has an orthonormal basis consisting of eigen vectors of* A*. If* A *is furthermore real, then* $\mathbb{R}^n$ *has an orthonormal basis consisting of eigen vectors of* A*.*

Positive (semi-)definiteness can be read off the eigen values.

Remark 5.5.25. *Let* $A \in \mathbb{C}^{n\times n}$ *be hermitean. Then: A is positive (semi-)definite, if and only if all eigen values of A are positive (non-negative).*

Proof. $\Rightarrow$. Let e be an eigen vector of A for eigen values $\lambda \in \mathbb{R}$. Then

$$e^*Ae = e^*\lambda e = \lambda \cdot e^*e = \lambda$$

The expression on the left is positive (non-negative).

$\Leftarrow$. Let $z \in \mathbb{C}^n \setminus \{0\}$ and let $\{e_i\}$ be an orthonormal basis of $\mathbb{C}^n$ consisting of eigen vectors of A, and let λ_i be the eigen value corresponding to e_i. Then $z = \sum_i \alpha_i e_i$ and

$$z^\top Az = \sum_i \bar{\alpha}_i e_i^\top \sum_j \alpha_j Ae_j = \sum_{i,j} \bar{\alpha}_i \alpha_j \lambda_j \underbrace{e_i^\top e_j}_{=\delta_{ij}} = \sum_i |\alpha_i|^2 \lambda_i$$

The expression on the right is positive (non-negative). □

One can construct a hermitean matrix from a not necessarily square matrix:

Lemma 5.5.26. *Let* $A \in \mathbb{C}^{m\times n}$*. Then* A^*A *is hermitean and positive semidefinite.*

Proof. We have

$$(A^*A)^* = A^*A^{**} = A^*A$$

Hence, A^*A is hermitean. Further, for $x \in \mathbb{C}^n$:

$$x^*A^*Ax = (Ax)^*Ax \geq 0$$

Hence, A^*A is positive semi-definite. □

5.6 Principal Component Analysis (PCA)

Assume that for n Persons p attributes are measured. This yields n points in $\mathbb{R}^p$ as a random vector $X = (X_1, \ldots, X_n)$. The aim of the principal component analysis is to project these data points to a q-dimensional subspace ($q < p$) in such a way that as little information as possible is lost, and that redundancy (i.e. correlation) is compressed.

The idea is to make a base change such that the new variables are decorrelated. Then the covariance matrix is diagonal. The gain is that in case of normally distributed data, the new variables are statistically independent.

The covariance matrix

$$\mathrm{Cov}(X) = \mathbb{E}\left((X-\mu)(X-\mu)^\top\right) = (\mathrm{Cov}(X_i, X_j)) \in \mathbb{R}^{n\times n} \qquad (\mu = \mathbb{E}(X))$$

is symmetric, hence diagonalisable, according to the Spectral Theorem 5.5.24. It is even positive semi-definite. Namely, as

$$\mathrm{Cov}(S^\top X) = S^\top \mathrm{Cov}(X) S = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$$

(let S be an orthonormal basis of $\mathbb{R}^n$ consisting of eigen vectors of $\mathrm{Cov}(X)$) the diagonalisation itself is a covarince matrix. Its diagonal entries are variances, i.e. non-negative. According to Remark 5.5.25, it follows that $\mathrm{Cov}(X)$ is positiv semi-definite.

The columns of

$$Y := S^\top X$$

are called the *principal componentes* of X. We have:

$$\mathrm{Var}(Y_i) = \lambda_i$$

The method is now as follows: Order S in such a way that the eigen values λ_i are sorted in ascending order. Choose q with $\lambda_1 \geq \ldots \lambda_q$, such that the quotient

$$\tau_q := \frac{\sum_{i=1}^{q} \mathrm{Var}(Y_i)}{\sum_{i=1}^{n} \mathrm{Var}(X_i)}$$

is large. This expression is between 0 and 1. Notice that

$$\sum_{i=1}^{n} \mathrm{Var}(X_i) = \mathrm{trace}(\mathrm{Cov}(X)) = \mathrm{trace}(\mathrm{Cov}(Y)) = \sum_{i=1}^{n} \mathrm{Var}(Y_i)$$

is the total variance of X. This is the sum of all eigenvalues. The dimension of q is determined via the larges eigenvalues.

The principal components also yield the best linear approximation to X: The *first component* is the straight line H_1 through the center $\mu = \mathbb{E}(X)$ with smallest error. The *second component* is the straight line H_2 through μ, and orthogonal to H_1, such that the plane spanned by H_1, H_2 has smallest error, etc. Finally, we obtain the principal components $H_1, \ldots, H_q$ for the eigenvalues $\lambda_1 \geq \ldots \lambda_q$.

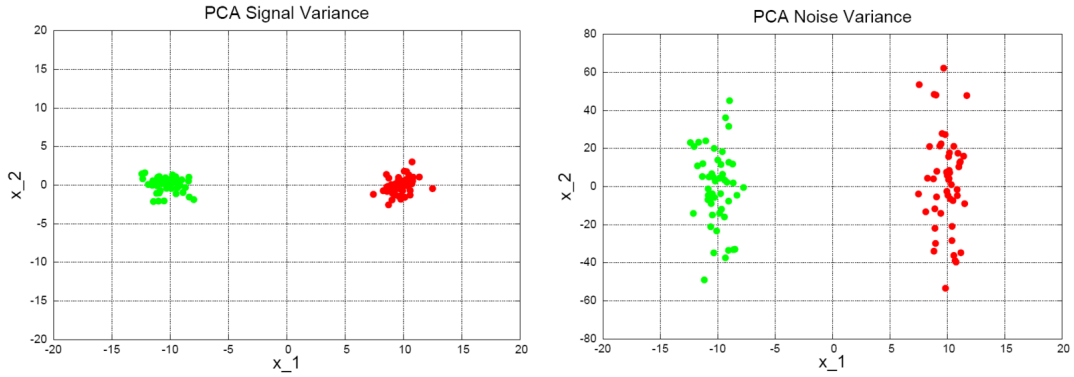


Figure 5.6: 2 Cluster mit Signal Variance (links) und Noise Variance (rechts). Source: Wikipedia, author: Rene Andrae.

Fundamental Assumption of PCA

In the principal component analysis it is assumed that the directions of largest variance contain most of the information.

This assumption, however, is not always satisfied. This shall be seen in the following example from cluster analysis.

In Figure 5.6 (left), the variance within the two clusters is low compared to the distance between the clusters. This is why the first component is the x_1-axis. This suffices to separate the clusters. The second component x_2 can be neglegted. The total variance is dominated by the signal, and we have two separated clusters.

In Figure 5.6 (right), the variance within the clusters has the main contribution to the total variance. It is assumed that variance is generated by noise. This is why this example is called "noise variance". The first component is x_2. It does not contain any information on the separability of the clusters.

Summarising, we can say that often, but not always, the dominating principal components contain most of the information *relevant for a given problem.*

5.7 Cholesky Decomposition

Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definit.

Definition 5.7.1. *A* Cholesky decomposition *of A is a factorisation*

$$A = G \cdot G^\top$$

where G is a lower triangular matrix with positive entries.

A Cholesky decomposition of A can be calculated as follows: With $A = (\alpha_{ij})$ and $G = (\gamma_{ij})$ we have

$$\alpha_{ij} = \sum_{k=1}^{j} \gamma_{ik}\gamma_{jk}, \qquad i \geq j$$

This yields:

$$\gamma_{ij} = \begin{cases} 0, & i < j \\ \sqrt{\alpha_{ii} - \sum_{k=1}^{i-1} \gamma_{ik}^2}, & i = j \\ \frac{1}{\gamma_{jj}} \left(\alpha_{ij} - \sum_{k=1}^{j-1} \gamma_{ik}\gamma_{jk} \right), & i > j \end{cases}$$

Example 5.7.2.

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix} = \begin{pmatrix} \gamma_{11}^2 & \gamma_{11}\gamma_{21} \\ \gamma_{11}\gamma_{21} & \gamma_{21}^2 + \gamma_{22}^2 \end{pmatrix}$$

yields:

$$\begin{aligned} \gamma_{11} &= \sqrt{a} \\ \gamma_{22} &= \sqrt{c - \gamma_{21}^2} \\ \gamma_{21} &= \frac{b}{\gamma_{11}} \end{aligned}$$

Example 5.7.3.

$$\begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix} = \begin{pmatrix} \gamma_{11}^2 & \gamma_{11}\gamma_{21} & \gamma_{11}\gamma_{31} \\ \gamma_{11}\gamma_{21} & \gamma_{21}^2 + \gamma_{22}^2 & \gamma_{21}\gamma_{31} + \gamma_{22}\gamma_{32} \\ \gamma_{11}\gamma_{31} & \gamma_{21}\gamma_{31} + \gamma_{22}\gamma_{32} & \gamma_{31}^2 + \gamma_{32}^2 + \gamma_{33}^2 \end{pmatrix}$$

yields:

$$\begin{aligned} &\gamma_{11} = \sqrt{a}, \qquad \gamma_{22} = \sqrt{d - \gamma_{21}^2}, \qquad \gamma_{33} = \sqrt{f - (\gamma_{31}^2 + \gamma_{32}^2)} \\ &\gamma_{21} = \frac{b}{\gamma_{11}}, \qquad \gamma_{31} = \frac{c}{\gamma_{11}} \\ &\gamma_{32} = \frac{e - \gamma_{21}\gamma_{31}}{\gamma_{22}} \end{aligned}$$

As an application, we again solve a system of linear equations

$$Ax = b$$

with A symmetric, positive definit. Using the Cholesky decomposition $A = G \cdot G^\top$ we obtain

$$G(\underbrace{G^\top x}_{=:y}) = b \qquad \text{und} \qquad G^\top x = y$$

The first equation

$$Gy = b$$

can be solved by forward substitution, and the second

$$G^\top x = y$$

by backward substitution (cf. Section 5.4).

5.8 Gauß-Newton Method

Let m functios $r = (r_1, \ldots, r_m)$ in n variables $X = (X_1, \ldots, X_n)$ with $m \geq n$ be given. The aim is to minimise the quantity

$$S(X) = \sum_{i=1}^{m} r_i(X)^2$$

The Gauß-Newton method is iterative. The start is with $X = x_0 \in \mathbb{R}^n$. This is incremented by ϵ:

$$x_{s+1} = x_s + \epsilon$$

with $\epsilon^\top \epsilon$ small. In order to determine the increment, we take a Taylor expansion:

$$S(x_s + \epsilon) \approx S(x_s) + \left[\frac{\partial S}{\partial X_i}\right]^\top \epsilon + \frac{1}{2}\epsilon^\top \left[\frac{\partial^2 S}{\partial X_i \partial X_j}\right] \epsilon$$

with

$$\left[\frac{\partial S}{\partial X_i}\right] = 2J_r(X)^\top r$$

where

$$J_r(X) = \left[\frac{\partial r_i}{\partial X_j}\right]$$

is the Jacobi matrix ist, and the Hesse matrix is approximated:

$$\left[\frac{\partial^2 S}{\partial X_i \partial X_j}\right] \approx 2J_r(X)^\top J_r(X)$$

for $r^\top r$ small. This yields

$$S(x_s + \epsilon) \approx S(x_s) + 2r^\top J_r(X)\epsilon + \epsilon^\top J_r(X)^\top J_r(X)\epsilon$$

Then we need to minimise

$$\frac{\partial S}{\partial \epsilon}(x_s + \epsilon) \approx 2J_r(X)^\top r + 2J_r(X)^\top J_r(X)\epsilon \stackrel{!}{=} 0$$

which leads to the *normal equations*:

$$J_r^\top(X) J_r(X) \cdot \epsilon = -J_r(X)^\top r \tag{5.4}$$

The background is:

- In data modelling, $X = \beta$ is a parameter vector, for which a model function

  $$y = f(x, \beta)$$

  is fitted to the data (x_i, y_i).

- The functions

  $$r_i(\beta) = y_i - f(x_i, \beta)$$

  are called the *residues*.

- The increment solves the normal equations (5.4) with $X = \beta$.

- In general, the Cholesky decomposition can be applied.

Comparison with the Newton method

In the Newton method, the Hesse matrix $H(S)$ is used instead of its approximation with twice the square of the Jacobi matrix. The increment here is:

$$\epsilon = -H(S)^{-1}\nabla S$$

This implies that the convergence of the Gauß-Newton method is at most quadratic.

5.9 Lisa and Baseball

When in *The Lisa Series* (2010) Bart's baseball team *The Isotots* loose their trainer, Lisa seizes her chance and becomes their new trainer. However, she has no idea about baseball. But she meets, by chance, Professor Frink who supports the opinion that baseball can be understood only by deep mathematical analysis, and gives her a stack of books which she should work through.

One of these books is *The Bill James Historical Baseball Abstract*, a collection of the most important baseball statistics from the real world, compiled by Bill James. Through studying the books, she manages to lead the Isotots from the bottom of the table upto the second place. But when she tells Bart in one game not to bat, he disregards her order and wins the game through a homerun. Consequently, she removes Bart from the team as he believes he is "better than the laws of probability". The Isotots continue their winning stroke also without Bart. In the final game of the *Little League State Championship*, however, a player drops out. So, she asks Bart to replace him. He hesitates, as he knows he is facing a dilemma: statistics or instinct. In the last inning, Bart defys Lisa's order again. But this time he goes out, and the Isotots loose the game.

5.10 Inner product spaces

Let V be a K-vector space with $K = \mathbb{R}$ or $\mathbb{C}$.

Definition 5.10.1. *A map*

$$\langle\cdot,\cdot\rangle\colon V \times V \to K$$

with

1. $\langle x, y\rangle = \overline{\langle y, x\rangle}$ *(conjugate symmetric)*
2. $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$
   $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$ *(left linear)*
3. $\langle x, x\rangle \geq 0 \quad mit = 0 \text{ nur für } x = 0$ *(positive definite)*

for $x, y, z \in V$, $\alpha \in K$ is called an inner product. *The pair $(V, \langle\cdot,\cdot\rangle)$ is called an* inner product space.

Remark 5.10.2. *1. $\langle x, x\rangle$ is always real.*

2. The following holds true:

$$\langle x, \alpha y\rangle = \bar{\alpha}\langle x, y\rangle$$
$$\langle x, y + z\rangle = \langle x, z\rangle + \langle y, z\rangle$$

Hence, an inner product is sesqui-linear.

3. For $K = \mathbb{R}$, an inner product is symmetric and linear.

Example 5.10.3. *Let $V = \mathbb{R}^n$. Then*

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

is the standard inner product.

Example 5.10.4. *Let $V = \mathbb{C}^n$. Then*

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i \overline{y_i} = y^* x$$

is the standard inner product.

Example 5.10.5. *Let $V = C[a, b] = \{f\colon [a, b] \to \mathbb{K} \mid f \text{ continuous}\}$. Then an inner product is given by*

$$\langle f, g\rangle = \int_a^b f(t)\overline{g(t)}\, dt$$

The first two axioms of inner product follow from the calculation rules of integrals. As for positive-definiteness: if $f \neq 0$, then $f(x_0) \neq 0$ for some $x_0 \in [a, b]$. Then there exists an ϵ-neighbourhood U of x_0, such that

$$f(x) \neq 0$$

for all $x \in U$. Then

$$\langle f, f\rangle = \int_a^b |f(t)|^2\, dt \geq \int_U |f(t)|^2\, dt > 0$$

Let $(V, \langle\cdot,\cdot\rangle)$ be an inner product space.

Definition 5.10.6. *The function*

$$\|\cdot\|\colon V \to \mathbb{R}, \quad x \mapsto \sqrt{\langle x, x\rangle}$$

is called a Norm *on V.*

Properties of an inner product norm

1. **Cauchy-Schwarz inequality.**

   $$|\langle x, y\rangle| \leq \|x\| \cdot \|y\| \tag{5.5}$$

   Proof. If $y = 0$, then the inequality holds true. If $y \neq 0$, then let $\lambda = \frac{\langle x,y\rangle}{\langle y,y\rangle}$. Then:

   $$\begin{aligned} 0 \leq \langle x - \lambda y, x - \lambda y\rangle &= \langle x, x\rangle - \lambda\langle y, x\rangle - \bar{\lambda}\langle y, x\rangle + |\lambda|^2\langle y, y\rangle \\ &= \langle x, x\rangle - \frac{\langle x, y\rangle\langle y, x\rangle}{\langle y, y\rangle} - \frac{\langle y, x\rangle\langle x, y\rangle}{\langle y, y\rangle} + \frac{|\langle x, y\rangle|^2}{\langle y, y\rangle} \\ &= \langle x, x\rangle - \frac{|\langle x, y\rangle|^2}{\langle y, y\rangle} \end{aligned}$$

   This yields

   $$|\langle x, y\rangle|^2 \leq \langle x, x\rangle\langle y, y\rangle$$

   from which the assertion follows. □

2. Because of the Cauchy-Schwarz inequality (5.5), the *angle* between two vectors $x, y \in V \setminus \{0\}$ can be defined:

$$w(x,y) := \arccos \frac{\langle x, y\rangle}{\|x\| \cdot \|y\|}$$

We say that x and y are *orthogonal* ($x \perp y$), if the angle equals π:

$$x \perp y \quad :\Leftrightarrow \quad w(x,y) = \pi \quad \Leftrightarrow \quad \langle x, y\rangle = 0$$

3. **Homogeneity.** For $\alpha \in K$, $x \in V$ it holds true that:

$$\|\alpha \cdot x\| = |\alpha| \|x\|$$

4. **Triangle inequality.** For $x, y \in V$ we have:

$$\|x + y\| \leq \|x\| + \|y\|$$

Proof. We have

$$\begin{aligned}
\|x+y\|^2 = \langle x+y, x+y\rangle &= \langle x, x\rangle + \underbrace{\langle x, y\rangle + \langle y, x\rangle}_{=2\Re(\langle x,y\rangle) \leq 2|\langle x,y\rangle|} + \langle y, y\rangle \\
&\leq \|x\|^2 + 2|\langle x, y\rangle| + \|y\|^2 \\
&\overset{(*)}{\leq} \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 \\
&= (\|x\| + \|y\|)^2
\end{aligned}$$

where in $(*)$ the Cauchy-Schwarz inequality (5.5) was used. □

5. **Theorem of Pythagoras.** If $x_1, \ldots, x_n$ are pairwise orthogonal, then:

$$\sum_{i=1}^{n} \|x_i\|^2 = \left\| \sum_{i=1}^{n} x_i \right\|^2 \tag{5.6}$$

6. **Parallelogramm identity.** For $x, y \in V$ it holds true that:

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2$$

5.11 QR Decomposition

Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space. We consider the projection operator for $u \in V$.

$$\pi_u \colon V \to V, \quad x \mapsto \frac{\langle x, u\rangle}{\langle u, u\rangle} u$$

Remark 5.11.1. *It holds true that*

$$x - \pi_u(x) \perp u$$

Proof. We have

$$\langle x - \pi_u(x), u\rangle = \langle x, u\rangle - \frac{\langle x, u\rangle}{\langle u, u\rangle} \langle u, u\rangle = 0$$

□

Now, we can orthogonalises linearly independent vectors $b_1, \ldots, b_n \in V$ according to Gram-Schmidt:

$$\begin{aligned} u_1 &= b_1, & e_1 &= \frac{u_1}{\|u_1\|} \\ u_2 &= b_2 - \pi_{u_1}(b_2), & e_2 &= \frac{u_2}{\|u_2\|} \\ &\vdots \\ u_n &= b_n - \sum_{i=1}^{n-1} \pi_{u_i}(b_i), & e_n &= \frac{u_n}{\|u_n\|} \end{aligned}$$

Remark 5.11.2. *The $u_1, \ldots, u_n$ span the same linear subspace as $b_1, \ldots, b_n$ and are pairwise orthogonal.*

Proof. The u_i are orthogonal. For $n = 1$ there is nothing to prove. Let $n > 1$. Assume by induction hypothesis that $u_1, \ldots, u_{n-1}$ are orthogonal. Then for $j < n$:

$$\begin{aligned} \langle u_n, u_j \rangle &= \left\langle b_n - \sum_{i=1}^{n-1} \frac{\langle b_n, u_i \rangle}{\langle u_i, u_i \rangle} u_i, u_j \right\rangle \\ &= \langle b_n, u_j \rangle - \sum_{i=1}^{n-1} \frac{\langle b_n, u_i \rangle}{\langle u_i, u_i \rangle} \underbrace{\langle u_i, u_j \rangle}_{= \langle u_i, u_i \rangle \delta_{ij}} \\ &= \langle b_n, u_j \rangle - \langle b_n, u_j \rangle = 0 \end{aligned}$$

The space spanned by the u_i. The $u_1, \ldots, u_n$ are linear combinations of the $b_1, \ldots, b_n$ and are orthogonal, hence, they are also linearly independent: let

$$x = \sum_j \alpha_j u_j = 0$$

Then

$$\langle x, u_i \rangle = \alpha_i \underbrace{\langle u_i, u_i \rangle}_{\neq 0} = 0$$

Hence, all $\alpha_i = 0$. As there are as many u_i as b_i, it follows that the u_i span the same linear subspace as the b_i. □

We have

$$\langle e_i, b_j \rangle e_i = \frac{\langle u_i, b_j \rangle}{\|u_i\|} \frac{u_i}{\|u_i\|} = \frac{\langle u_i, b_j \rangle}{\langle u_i, u_i \rangle} u_i = \pi_{u_i}(b_j)$$

Also,

$$\langle b_k, u_k \rangle = \left\langle u_k + \sum_{i=1}^{k-1} \pi_{u_i}(b_i), u_k \right\rangle = \langle u_k, u_k \rangle + \sum_{i=1}^{k-1} \underbrace{\langle \pi_{u_i}(b_i), u_k \rangle}_{=0} = \langle u_k, u_k \rangle$$

Hence

$$\pi_{u_k}(b_k) = \frac{\langle b_k, u_k \rangle}{\langle u_k, u_k \rangle} u_k = \frac{\langle u_k, u_k \rangle}{\langle u_k, u_k \rangle} u_k = u_k$$

That is why

$$b_k = u_k + \sum_{i=1}^{k-1} \pi_{u_i}(b_i) = \sum_{i=1}^{k} \pi_{u_i}(b_i) = \sum_{i=1}^{k} \langle e_i, b_k \rangle e_i$$

In matrix form this is with $B = (b_1 \mid \cdots \mid b_n)$:

$$B = Q \cdot R \tag{5.7}$$

where

$$Q = (e_1 \mid \cdots \mid e_n)$$

and

$$R = \begin{pmatrix} \langle e_1, b_1\rangle & \langle e_1, b_2\rangle & \langle e_1, b_3\rangle & \dots \\ 0 & \langle e_2, b_2\rangle & \langle e_2, b_3\rangle & \dots \\ 0 & 0 & \langle e_3, b_3\rangle & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} = Q^* B$$

Definition 5.11.3. *(5.7) is called the QR decomposition of B. The columns of Q are orthonormal, and R is an upper triangular matrix.*

Application

Solve a system of linear equations $Ax = b$ with $A \in \mathbb{R}^{m\times n}$, $m \geq n$, where A has full rank. With the QR decomposition of A this can be done as follows:

$$A = QR$$

with $Q \in \mathbb{R}^{m\times n}$, $R \in \mathbb{R}^{n\times n}$. Now, solve

$$Q \cdot \underbrace{Rx}_{=y} = b$$

in this way:

$$y = Q^\top b$$

and

$$Rx = y$$

by backward substitution (cf. Section 5.4).

5.12 Eigenvalue determination using the QR decomposition

Assume that $A \in \mathbb{C}^{n\times n}$ is non-singular and that all eigenvalues are of distinct absolute values. Then the following sequence converges to an upper triangular matrix A_∞:

$$\begin{aligned} A_k &= Q_k R_k && (QR \text{ decomposition}) \\ A_{k+1} &:= R_k Q_k = Q_{k+1} R_{k+1} && (QR \text{ decomposition}) \end{aligned}$$

We have:

1. As
$$A_{k+1} = R_k Q_k = Q_k^* Q_k R_k Q_k = Q_k^* A_k Q_k$$
all A_k have the same eigenvalues.

2. The eigenvalues of A_∞ are the diagonal entries, as the characteristic polynomial is

$$\det \begin{pmatrix} X - a_{11}^{(\infty)} & & * \\ & \ddots & \\ 0 & & X - a_{nn}^{(\infty)} \end{pmatrix} = (X - a_{11}^{(\infty)}) \cdots (X - a_{nn}^{(\infty)})$$

3. If A is symmetric, then the columns of $Q = Q_1 Q_2 \cdots$ are the eigen vectors of A, as

$$A \cdot Q_1 Q_2 \cdots Q_n = Q_1 R_1 Q_1 \cdots Q_n = Q_1 Q_2 R_2 Q_2 \cdots Q_n = \cdots = Q_1 \cdots Q_n A_n$$

and as A is symmetric, it follows that A_∞ is a diagonal matrix $\mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$, and

$$AQ = QA_\infty = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)Q$$

The property of the matrix Q in the QR decomposition also has a name:

Definition 5.12.1. *A matrix $A \in \mathbb{R}^{n \times n}$ is called* orthogonal*, if the columns of A are an orthonormal basis of $\mathbb{R}^n$.*

Remark 5.12.2. *According to the Spectral Theorem II (Theorem 5.5.24), a symmetric matrix $A \in \mathbb{R}^{n \times n}$ is diagonalisable with an orthogonal matrix O:*

$$O^\top AO \quad \textit{is diagonal}$$

as there is an orthonormal basis O of $\mathbb{R}^n$ consisting of eigen vectors of A.

Over the complex numbers, the property has a different name:

Definition 5.12.3. *A matrix $A \in \mathbb{C}^{n \times n}$ is called* unitary*, if the columns of A are an orthonormal basis of $\mathbb{C}^n$.*

Remark 5.12.4. *According to Spectral Theorem II (Theorem 5.5.24), a hermitian matrix $A \in \mathbb{C}^{n \times n}$ is diagonalisable with a unitary matrix U:*

$$U^* AU \quad \textit{is diagonal}$$

as there is an orthonormal basis of $\mathbb{C}^n$ consisting of eigen vectors of A.

5.13 Singular Value Decomposition

Let $K = \mathbb{R}$ or $\mathbb{C}$.

Theorem 5.13.1 (Singular Value Decomposition)**.** *A matrix $M \in K^{m \times n}$ has a decomposition*

$$M = U\Sigma V^*$$

with $U \in K^{m \times m}$ unitary, $\Sigma \in K^{m \times n}$ diagonal with non-negative entries, and $V^* \in K^{n \times n}$ unitary.

Definition 5.13.2. *The diagonal entries of Σ are called the* singular values *of A.*

This has a geometric interpretation. Let $T\colon K^n \to K^m$ be a linear map. $V^* = (v_1 *, \ldots, v_n^*)$ is an orthonormal basis of K^n, and $U = (u_1, \ldots, un_m)$ is an orthonormal basis of K^m, such that

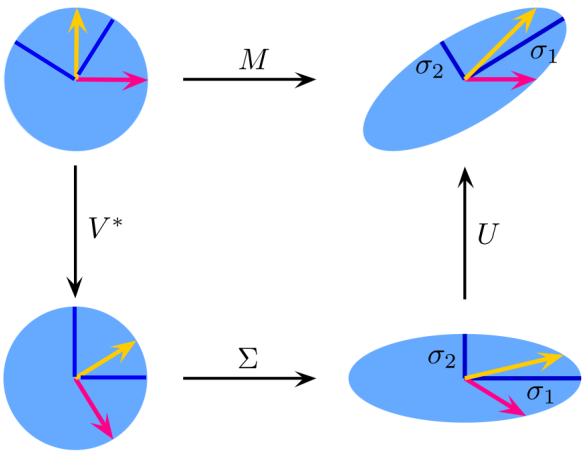
$$T(v_i^*) = \sigma_i u_i$$

for the singular value σ_i.

In the real case, there is the following geometric interpretation: The linear map

$$T\colon \mathbb{R}^n \to \mathbb{R}^m$$

takes the unit sphere in $\mathbb{R}^n$ to an ellipsoid in $\mathbb{R}^m$. The positive singular values are then the lengths of the semi-axes of the ellipsoid (cf. Figure 5.7).



$$M = U \cdot \Sigma \cdot V^*$$

Figure 5.7: Illustration of the singular value decomposition (source: Wikipedia, autor: Georg-Johann).

5.13.1 Best Rank-r approximation

Let $A \in \mathbb{R}^{m\times n}$. Then the singular value decomposition looks thus:

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

with $U \in \mathbb{R}^{m\times m}$ orthogonal, $\Sigma \in \mathbb{R}^{m\times n}$ diagonal with non-negative entries, and $V \in \mathbb{R}^{n\times n}$ orthogonal. Written out, this looks thus:

$$\begin{aligned}
A &= (u_1 \dots u_k \mid u_{k+1} \dots u_m) \left(\begin{array}{ccc|c} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_k & \\ \hline 0 & & & 0 \end{array}\right) \left(\begin{array}{c} v_1^\top \\ \vdots \\ v_k^\top \\ \hline v_{k+1}^\top \\ \vdots \\ v_n^\top \end{array}\right) \\
&= (u_1 \dots u_k) \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_k \end{pmatrix} \begin{pmatrix} v_1^\top \\ \vdots \\ v_k^\top \end{pmatrix} + \underbrace{(u_{k+1} \dots u_m)(0) \begin{pmatrix} v_{k+1}^\top \\ \vdots \\ v_n^\top \end{pmatrix}}_{=0} \\
&= (\sigma_1 u_1 \dots \sigma_k u_k) \begin{pmatrix} v_1^\top \\ \vdots \\ v_k^\top \end{pmatrix} = \sigma_1 u_1 v_1^\top + \dots + \sigma_k u_k v_k^\top
\end{aligned}$$

Remark 5.13.3. *We have*

$$\mathrm{Rank}(A) = \mathrm{Rank}(\Sigma) = k$$

and

$$\mathrm{Rank}(u_i v_i^\top) = 1$$

because every column of $u_i v_i^\top$ is a multiple of u_i.

For the *matrix inner product*

$$A \odot B := \sum_{i,j} \alpha_{ij}\beta_{ij}$$

with $A = (\alpha_{ij}), B = (\beta_{ij}) \in \mathbb{R}^{m\times n}$ the matrices $u_i v_i^\top$ are pairwise orthogonal. This is because for $x, u \in \mathbb{R}^m$, $y, v \in \mathbb{R}^m$ we have

$$\begin{aligned} xy^\top \odot uv^\top &= (xy_1 \mid \cdots \mid xy_n) \odot (uv_1 \mid \cdots \mid uv_n) = \sum_{i,j} x_i y_j u_i v_j \\ &= \sum_j xy_j \cdot uv_j = (x \cdot u) \sum_j y_j v_j = (x \cdot u)(y \cdot v) \end{aligned}$$

i.e. in case $x \perp u$ or $y \perp v$, then $xy^\top \perp uv^\top$. We have denoted the standard inner product here with $\cdot$.

We also have:

Theorem 5.13.4. *The singular value decomposition decomposes $A \in \mathbb{R}^{m\times n}$ into a linear combination*

$$A = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$$

of pairwise orthogonal matrices $u_i v_i^\top$ of rank 1.

For the *Frobenius norm*

$$\|A\|_F := \sqrt{A \odot A}$$

it holds true that

$$\left\| u_i v_i^\top \right\|_F^2 = u_i v_i^\top \odot u_i v_i^\top = (u_i \cdot u_i)(v_i \cdot v_i) = 1$$

Hence, by the Theorem of Pythagoras (5.6):

$$\|A\|_F^2 = \sum_{i=1}^{k} \left\| \sigma_i u_i v_i^\top \right\|_F^2 = \sum_{i=1}^{k} |\sigma_i|^2 = \sum_{i=1}^{k} \sigma_i^2$$

Let Σ be ordered in such a way that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$. Then $\sigma_1 u_1 v_1^\top$ is the best rank-1 approximation of A. The squared error is:

$$\left\| A - \sigma_1 u_1 v_1^\top \right\|_F^2 = \sum_{i=2}^{k} \sigma_i^2$$

In general, $\sum_{i=1}^{r} \sigma_i u_i v_i^\top$ is the best rank-r approximation of A for $r \leq k$. The squared error is:

$$\left\| A - \sum_{i=1}^{r} \sigma_i u_i v_i^\top \right\|_F^2 = \sum_{i=r+1}^{k} \sigma_i^2$$

5.13.2 Data compression as best rank-r approximation

A grayscale image with $m \cdot n$ pixels can be viewed as a matrix $A \in \mathbb{R}^{m \times n}$. In order to store the full image, all $m \cdot n$ grayvalues must be stored. If the best rank-1 approximation

$$\sigma_1 u_1 v_1^\top$$

is used, then only $m + n + 1$ values need to be stored. For the best rank-r approximation

$$\sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

there are $r(m + n + 1)$ values. For r sufficiently small, this number is smaller than $m \cdot n$.

The method is as follows: Let

$$E_r := A - \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

Now, choose r such that

$$\frac{\|E_r\|_F}{\|A\|_F} = \sqrt{\frac{\sum_{i=r+1}^{k} \sigma_i^2}{\sum_{i=1}^{k} \sigma_i^2}} < \epsilon$$

for a given threshold $\epsilon > 0$. Then

$$\sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

is the *compression* of A as best rank-r approximation.

Basic assumption of compression

In this method, it is assumed that the terms $\sigma_i u_i v_i^\top$ for small singular values σ_i do not contain relevant information, i.e. they consist of noise.

5.13.3 Linear least squares

Let $V = \mathbb{R}^m$ and $\{a_1, \ldots, a_n\} \subseteq V$ linearly independent, and $b \in V$. The task is to find coefficients $\xi_1, \ldots, \xi_n \in \mathbb{R}$, such that the error

$$\left\| b - \sum_{i=1}^{n} \xi_i a_i \right\|$$

is minimal. Written out as matrices, this means with $A = (a_1, \ldots, a_n) \in \mathbb{R}^{m \times n}$ the minimisation of

$$\|b - Ax\|$$

where $x = (\xi_1, \ldots, \xi_n)$. We consider the situation that the system of linear equations

$$Ax = b$$

is overdetermined. Then a best possible solution of this system is to be found. Geometrically, this means: find an element of the subspace $S \subseteq \mathbb{R}^n$ spanned by $a_1, \ldots, a_n$ which has minimal distance to b.

The solution of this problem can be found by orthogonal projection $\pi_S(b)$ of b onto S:

$$\pi_S(b) = Ax$$

For the error vector $b - \pi_S(b)$ it holds true that

$$b - \pi_S(b) \perp S$$

This is equivalent to

$$\begin{aligned} & a_i \perp Ax - b && i = 1, \ldots, n \\ \Leftrightarrow \quad & A^\top (Ax - b) = 0 && \\ \Leftrightarrow \quad & A^\top Ax = A^\top b && \textit{(Normal equations)} \end{aligned}$$

Remark 5.13.5. *$A^\top A$ is invertible, as $a_1, \ldots, a_n$ are linearly independent. But the calculation of $\left(A^\top A\right)^{-1}$ should be avoided, as the speed and precision suffers.*

With the singular value decomposition $A = U\Sigma V^\top$ it holds true that

$$Ax - b = U\Sigma V^\top x - b = U(\Sigma \underbrace{V^\top x}_{=:y} - \underbrace{U^\top b}_{=:c})$$

Hence,

$$\|Ax - b\| = \|\Sigma y - c\|$$

because

$$\|Uz\|^2 = \langle Uz, Uz \rangle = (Uz)^\top Uz = z^\top \underbrace{U^\top U}_{=I} z = z^\top z = \|z\|^2 \tag{5.8}$$

This means that we need to find y such that $\|\Sigma y - c\|$ is minimal. As Σ is diagonal, we have:

$$\Sigma y = (\sigma_1 y_1, \ldots, \sigma_k y_k, 0, \ldots, 0)$$

Hence,

$$\Sigma y - c = (\sigma_1 y_1 - c_1, \ldots, \sigma_k y_k - c_k, -c_{k+1}, \ldots, -c_m)$$

And its norm is minimal for

$$y_i = \frac{c_i}{\sigma_i}, \quad i = 1, \ldots, k \tag{5.9}$$

and $y_{k+1}, \ldots, y_n$ can be choosen freely. Namely, we have

$$\|\Sigma y - c\|^2 = \sum_{i=k+1}^{m} c_i^2$$

Hence, the sought for x is

$$x = Vy$$

with y according to (5.9).

5.13.4 Condition number of square matrices

Let $A \in \mathbb{R}^{n\times n}$, and let the singular value decomposition be

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

with $\Sigma = \operatorname{Diag}(\sigma_1, \dots, \sigma_n)$ such that $\sigma_1 \geq \cdots \geq \sigma_n$.

A is invertible, if and only if $\sigma_n > 0$. Then

$$A^{-1} = V \operatorname{Diag}\left(\sigma_1^{-1}, \dots, \sigma_n^{-1}\right) U^\top$$

Definition 5.13.6.

$$\operatorname{cond}(A) := \frac{\sigma_1}{\sigma_n}$$

is the condition number *of* A.

The condition number indicates, how near A is to a singular matrix. If $\operatorname{cond}(A) = \infty$, then A is indeed singular, if $\operatorname{cond}(A) >> 1$, then A is "almost" singular.

The problem

$$Ax = b$$

is called *ill-conditioned* if $\operatorname{cond}(A) >> 1$, *ill-posed* if $\operatorname{cond}(A) = \infty$, and otherwise *well-conditioned*.

5.13.5 Kabsch algorithm

The task is to find the optimal rotation matrix between paired pointsets in $\mathbb{R}^3$. Let the sets be

$$P = \{p_1, \dots, p_n\}, \quad Q = \{q_1, \dots, q_n\}$$

First, form the centroids:

$$C_P = \frac{1}{n}\sum_{i=1}^{n} p_i, \quad C_Q = \frac{1}{n}\sum_{i=1}^{n} q_i$$

Then replace P, Q with

$$\{p_1 - C_P, \dots, p_n - C_P\}, \quad \{q_1 - C_Q, \dots, q_n - C_Q\}$$

and call these points again p_i and q_i, respectively.

The task is to minimise the quantity

$$E(U) = \frac{1}{n}\sum_{i=1}^{n} \left\| \underbrace{Up_i}_{=:p_i'} - q_i \right\|^2$$

For this, write P, Q as $3 \times n$-matrices. We have

$$\begin{aligned} nE &= \sum_{i=1}^{n} \left\|p_i' - q_i\right\|^2 = \operatorname{trace}\left((P' - Q)^\top (P' - Q)\right) \\ &= \operatorname{trace}\left(P'^\top P'\right) + \operatorname{trace}\left(Q^\top Q\right) - 2\operatorname{trace}\left(Q^\top P'\right) \\ &= \sum_{i=1}^{n} \left(\|p_i\|^2 + \|q_i\|^2\right) - 2\operatorname{trace}\left(Q^\top P'\right) \end{aligned}$$

Observe that $\|p_i'\| = \|p_i\|$ because of (5.8), as U is orthogonal. Maximise

$$\operatorname{trace}\left(Q^\top P'\right) = \operatorname{trace}\left(Q^\top U P\right) = \operatorname{trace}\left(PQ^\top \cdot U\right)$$

where $PQ^\top \in \mathbb{R}^{3\times 3}$. The latter equation holds true in general as a rule for the trace. The singular value decomposition is

$$PQ^\top = V\Sigma W^\top$$

Then the optimal rotation is

$$U = W \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{pmatrix} V^\top$$

with

$$d = \operatorname{sgn}\left(\det\left(PQ^\top\right)\right)$$

An application of the Kabsch algorithm is e.g. in the orientation of satellites.

5.14 Hilbert Spaces

Let $(V, \langle\cdot,\cdot\rangle)$ be an inner product space. The associated norm is

$$\|\cdot\|\colon V \to \mathbb{R}, \quad x \mapsto \sqrt{\langle x, x\rangle}$$

Definition 5.14.1. *If $(V, \|\cdot\|)$ is complete, then $(V, \langle\cdot,\cdot\rangle)$ is called a* Hilbert space.

Definition 5.14.2. *If $b_1, b_2, \dots \in V$ with $\|b_\nu\| = 1$ and $b_\mu \perp b_\nu$ for $\mu \neq \nu$, and if for every $x \in V$:*

$$x = \sum_{\nu=1}^{\infty} \alpha_\nu b_\nu$$

for certain $\alpha_\nu \in K$, then the sequence $b_1, b_2, \dots$ is called an orthonormal basis *of V.*

Remark 5.14.3. *Such an orthonormal basis is in general not a basis in the sense of linear algebra!*

Theorem 5.14.4. *Every Hilbert space has an orthonormal basis.*

Remark 5.14.5. *For $x = \sum_{\nu=1}^{\infty} \alpha_\nu b_\nu$ we have*

$$\langle x, b_\mu\rangle = \sum_{\nu=1}^{\infty} \alpha_\nu \underbrace{\langle b_\nu, b_\mu\rangle}_{=\delta_{\nu\mu}} = \alpha_\mu$$

Definition 5.14.6. *The expression*

$$\langle x, b_\mu\rangle$$

is called Fourier coefficient *of x with respect to the orthonormal basis $b_1, b_2, \dots$.*

Example 5.14.7. *Let $V = C[-\pi, \pi]$ with the inner product*

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(t)\overline{g(t)}\, dt$$

Then

$$e_k(t) = \frac{1}{\sqrt{2\pi}} e^{ikt}$$

is an orthonormal basis of V. The orthogonality relations hold true because

$$\langle e_k, e_\ell \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i(k-\ell)t}\, dt = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(k-\ell)t\, dt + i\frac{1}{2\pi} \int_{-\pi}^{\pi} \sin(k-\ell)t\, dt = \delta_{k,\ell}$$

The k-th Fourier coefficient is

$$\langle f, e_k \rangle = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} f(t) e^{-ikt}\, dt$$

The orthogonal basis property of the e_k says furthermore that every signal is a superposition of pure sine waves which occur as harmonics. In a vibrating string, the k-th harmonic can be made audible by touching the string at the place $\frac{1}{k-1}$ (cf. Figure 5.8).

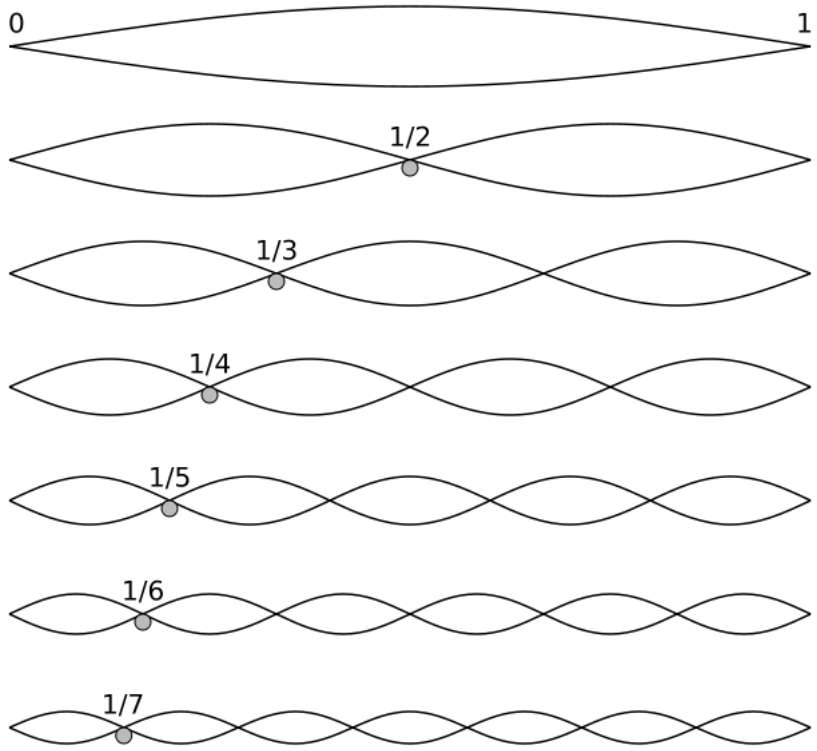


Figure 5.8: Harmonics of a vibrating string (Source: Wikipedia, author: Qef).

5.14.1 40000 decimals of π

In the episode *Marge in Chains* (1993), Marge faces a court trial for shop-lifting. Her attorney wants to cast doubt on the supposed witness Apu Nahasapeemapetilon by hinting on the possibility that his memory might be wrong. Apu replies that he can tell the π up to the $40,000$-th place after the decimal point. This digit is a 1.

If Apu had had a time machine, then he could have checked the Bailey-Borwein-Plouffe formula in the year 1995, which gives any decimal of π without knowledge of the decimals coming before it. However, the formula uses the hexa-decimal system.

Theorem 5.14.8 (Bailey-Borwein-Plouffe formula, 1995)**.** *It holds true that*

$$\pi = \sum_{k=0}^{\infty} \frac{1}{16^k} \left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6} \right)$$

In the hexa-decimal representation of π this means

$$\pi = \sum_{k=0}^{\infty} \frac{z_k}{16^k}$$

with

$$z_k = \left\lfloor \left(16^{k-1}\pi \mod 1 \right) \cdot 16 \right\rfloor$$

Then, according to Theorem 5.14.8,

$$16^{n-1}\pi = 4\sigma_1 - 2\sigma_4 - \sigma_5 - \sigma_6$$

with

$$\sigma_\ell = \sum_{k=0}^{\infty} \frac{16^{n-k-1}}{8k-\ell}$$

From each summand, the integer part needs to be removed. This can be done as follows: change σ_ℓ to

$$\sigma'_\ell = \sum_{k=0}^{n-1} \frac{\left(16^{n-k-1} \mod (8k+\ell)\right)}{8k+\ell} + \sum_{k=n}^{\infty} \frac{16^{n-k-1}}{8k+\ell}$$

Then

$$16^{n-1}\pi \equiv 4\sigma'_1 - 2\sigma'_4 - \sigma'_5 - \sigma'_6 \equiv \theta_n \mod 1$$

where $\theta_n \in [0,1)$. Then

$$z_n = \lfloor 16 \cdot \theta_n \rfloor$$

is the wanted place in the decimal system. We used

$$\lfloor x \rfloor := n \in \mathbb{Z} \quad \text{with} \quad x - n \in [0,1)$$

Then one can calculate on a machine that $z_{40\,000} = 1$.

Chapter 6

Trigonometric Functions

6.1 Discrete Fourier Transformation

Let $\zeta = e^{2\pi i/N} \in \mathbb{C}$ be a primitive N-th root of unity. This is a complex solution to the equation

$$X^N = 1$$

All solutions of this equation are powers of ζ:

$$\zeta^0, \zeta^1, \ldots, \zeta^{N-1}$$

From these we form the vectors

$$z = \left(1, \zeta, \zeta^2, \ldots, \zeta^{N-1}\right) \in \mathbb{C}^N$$
$$z^k = \left(1, \zeta^k, \zeta^{2k}, \ldots, \zeta^{(N-1)k}\right) \in \mathbb{C}^N$$

By indexing the vectors in $\mathbb{C}^N$ as follows:

$$f = (f_0, \ldots, f_N) \in \mathbb{C}^N$$

we obtain a notation for the standard inner product on $\mathbb{C}^N$ as follows:

$$\langle a, b \rangle = \sum_{\nu=0}^{N-1} a_\nu \bar{b}_\nu$$

Lemma 6.1.1. *The vectors $z^0, \ldots, z^{N-1}$ form an orthogonal basis of $\mathbb{C}^N$.*

Proof. This follows from the orthogonality relations:

$$\langle z^k, z^\ell \rangle = \sum_{\nu=0}^{N-1} \zeta^{\nu k} \zeta^{-\nu \ell} = \sum_{\nu=0}^{N-1} e^{\frac{2\pi i}{N}(k-\ell)\nu} = N\delta_{k\ell} \tag{6.1}$$

Latter equality holds true, as $\xi = \zeta^{k-\ell}$ for $k \neq \ell$ is an N-te root of unity, and

$$0 = \frac{1-\xi^N}{1-\xi} = 1 + \xi + \cdots + \xi^{N-1}$$

Hence, we have N pairwise orthogonal vectors in $\mathbb{C}^N$. □

A consequence is that a vector $f \in \mathbb{C}^N$ has a coordinate representation with respect to $z^0, \ldots, z^{N-1}$:

$$f = \sum_{k=0}^{N-1} \alpha_k z^k$$

The coefficient α_k can be calculated as

$$\alpha_k = \frac{1}{N} \sum_{\nu=0}^{N-1} \alpha_\ell N \delta_{k\ell} = \frac{1}{N} \sum_{\ell=0}^{N-1} \alpha_\ell \langle z^\ell, z^k \rangle = \frac{1}{N} \left\langle \sum_{\ell=0}^{N-1} \alpha_\ell z^\ell, z^k \right\rangle = \frac{1}{N} \langle f, z^k \rangle$$

The quantity $\langle f, z^k \rangle$ is called *discrete Fourier coefficient* of f. It has the representation

$$b_k := \langle f, z^k \rangle = \sum_{\nu=0}^{N-1} f_\nu \zeta^{-k\nu} = \Phi(\zeta^{-k})$$

with the polynomial

$$\Phi(X) = \sum_{\nu=0}^{N-1} f_\nu X^\nu \in \mathbb{C}[X]$$

Remark 6.1.2. *The discrete Fourier coefficient b_k is periodic:*

$$b_k = b_{k+N}$$

This follows from

$$\Phi\left(\zeta^{-k}\right) = \Phi\left(\zeta^{-k-N}\right)$$

as

$$\zeta^{-k-N} = \zeta^{-k}\zeta^{-N} = \frac{\zeta^{-k}}{\zeta^N} = \zeta^{-k}$$

since $\zeta^N = 1$.

Definition 6.1.3. *The vector*

$$\mathcal{F}(f) = (b_0, \ldots, b_{N-1})$$

is called the discrete Fourier transform (DFT) *of f.*

Remark 6.1.4. *$\mathcal{F}$ is multiplication with F^*, where F is the Vandermonde matrix:*

$$F = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & \zeta & \ldots & \zeta^{N-1} \\ \vdots & \vdots & & \vdots \\ 1 & \zeta^{N-1} & \ldots & \zeta^{(N-1)(N-1)} \end{pmatrix}$$

The orthogonality relations say:

$$F^{-1} = \frac{1}{N} F^*$$

A consequence is

$$f_k = \langle \mathcal{F}(f), z^{-k} \rangle = \frac{1}{N} \sum_{\nu=0}^{N-1} b_\nu e^{\frac{2\pi i}{N} k\nu} = \frac{1}{N} \tilde{\Phi}(\zeta^k)$$

with the polynomial

$$\tilde{\Phi}(X) = \sum_{\nu=0}^{N-1} b_\nu X^\nu \in \mathbb{C}[X]$$

for the inverse discrete Fourier transformation (IDFT).

6.1.1 Fast Fourier Transformation

We use the notations of the previous section, only that now an additional index N is used. I.e. $\zeta_N = e^{2\pi i/N}$ is the primitve N-th root of unity, and let

$$z_N^k = \left(1, \zeta_N^k, \zeta_N^{2k}, \ldots, \zeta_N^{(N-1)k}\right) \in \mathbb{C}_N$$

Further, we assume that $N = 2^n$ is a power of two. We decompose the vector

$$f = (f_0, \ldots, f_{N-1}) \in \mathbb{C}^N$$

into an even and an odd part:

$$\begin{aligned} f &= \tilde{g} + \tilde{u} \\ \tilde{g} &= (f_0, 0, f_2, 0, \ldots, f_{N-2}, 0) \\ \tilde{u} &= (0, f_1, 0, f_3, \ldots, 0, f_{N-1}) \end{aligned}$$

Removing the zeros yields:

$$\begin{aligned} g = (g_\mu) \in \mathbb{C}^{N/2}, \quad g_\mu = \tilde{g}_{2\mu} \\ u = (u_\mu) \in \mathbb{C}^{N/2}, \quad u_\mu = \tilde{u}_{2\mu+1} \end{aligned}$$

The discrete Fourier coefficient $b_{k,N}$ also has a decomposition:

$$b_{k,N} := \langle f, z_N^k \rangle = \langle \tilde{g}, z_N^k \rangle + \langle \tilde{u}, z_N^k \rangle$$

Here,

$$\begin{aligned} \langle \tilde{g}, z_N^k \rangle &= \sum_{\mu=0}^{N/2-1} \tilde{g}_{2\mu} e^{-\frac{2\pi i}{N}(2\mu)k} = \sum_{\mu=0}^{N/2-1} \tilde{g}_{2\mu} e^{-\frac{2\pi i}{N/2}\mu k} = \sum_{\mu=0}^{N/2-1} g_\mu \zeta_{N/2}^{-\mu k} = \langle g, z_{N/2}^k \rangle \\ \langle \tilde{u}, z_N^k \rangle &= \sum_{\mu=0}^{N/2-1} \tilde{u}_{2\mu+1} e^{-\frac{2\pi i}{N}(2\mu+1)k} = e^{-\frac{2\pi i}{N}k} \sum_{\mu=0}^{N/2-1} \tilde{u}_{2\mu+1} e^{-\frac{2\pi i}{N/2}k} = \zeta_N^k \langle u, z_{N/2}^k \rangle \end{aligned}$$

Hence,

$$b_{k,N} = \langle g, z_{N/2}^k \rangle + \zeta_N^k \langle u, z_{N/2}^k \rangle, \quad k = 0, \ldots, N-1$$

i.e. the Fourier coefficient for N decomposes into a Fourier coefficient for $N/2$ and a "twisted" Fourier coefficient for $N/2$. Further:

$$z_{N/2}^k = z_{N/2}^{k+N/2}$$

and

$$\zeta_N^{k+N/2} = \zeta_N^{N/2} \zeta_N^k = -\zeta_N^k$$

Hence:

$$b_{k,N} = \begin{cases} \langle g, z_{N/2}^k \rangle + \zeta_N^k \langle u, z_{N/2}^k \rangle, & k < N/2 \\ \langle g, z_{N/2}^{k-N/2} \rangle + \zeta_N^{k-N/2} \langle u, z_{N/2}^{k-N/2} \rangle, & k \geq N/2 \end{cases}$$

Hence, the DFT at length N reduces to the DFT at length $N/2$, and we can continue in this way, until we arrive at length 2. This yields a *divide-and-conquer algorithm* for computing the discrete Fourier coefficient at length $N = 2^n$.

6.1.2 Fourier series

It is known that every integrable function $f\colon [0,L] \to \mathbb{R}$ with $f(0) = f(L)$[1] can be expanded into a Fourier series:

$$f(x) = \sum_{k=-\infty}^{\infty} \beta_k e^{2\pi i k x/L}$$

For the Fourier coefficient β_k ($k \in \mathbb{Z}$) it holds true that

$$\beta_k = \frac{1}{L} \int_0^L f(x) e^{-2\pi i k x/L}\, dx = \langle f, e_k \rangle \tag{6.2}$$

with

$$e_k\colon [0,L] \to \mathbb{C}, \quad x \mapsto \frac{1}{L} e^{2\pi i k x/L}$$

Discrete Approximation of Fourier Coefficients

By assuming an equidistant sampling of the periodic function f, we first obtain the partitioning of the interval $[0,L]$:

$$h = \frac{L}{N}, \quad x_\nu = \nu h, \quad \nu = 0, \ldots, N-1$$

and the values

$$f_\nu = f(x_\nu)$$

Then, using the left Riemann sum approximation of the integral:

$$\beta_k = \frac{1}{L} \int_0^L f(x) e^{-2\pi i k x/L}\, dx = \frac{1}{L} \sum_{\nu=0}^{N-1} \int_{\nu L/N}^{(\nu+1)L/N} f(x) e^{-2\pi i k x/L}\, dx \approx \frac{1}{L} \frac{L}{N} \sum_{\nu=0}^{N-1} f_\nu e^{-\frac{2\pi i}{N}\nu k} = \frac{1}{N} b_k$$

where b_k is the discrete Fourier coefficient. The *left Riemann sum* approximation is as follows:

$$f(x) \approx f_\nu, \quad x \approx \frac{\nu L}{N}$$

where $x \in \left[\frac{\nu L}{N}, \frac{(\nu+1)L}{N}\right]$.

Remark 6.1.5. *The approximation*

$$\beta_k \approx \frac{b_k}{N}$$

is good only for small $|k|$, because b_k is periodic, while often $\beta_k \to 0$ for $|k| \to \infty$.

Approximiated Fourier series

Let N from Remark 6.1.5 be even and $\alpha_k = \frac{b_k}{N}$. Then

$$\begin{aligned} \alpha_k &\approx \beta_k, & k &= 0, \ldots, N/2-1 \\ \alpha_k = \alpha_{k+N} &\approx \beta_k, & k &= -N/2, \ldots, -1 \end{aligned}$$

and we approximate:

$$f(x) \approx \sum_{k=-N/2}^{N/2} \alpha_k e^{2\pi i k x/L}$$

[1] The condition means that f can be extended to a periodic function on $\mathbb{R}$.

Digital Signal Transmission

In order to transmit an anlog signal $f\colon [0,L] \to \mathbb{R}$, one samples N values equidistantly and obtains (normalised, periodic) discrete Fourier coefficients α_k ($k \in \mathbb{Z}$). Transmit these for $k = -N/2, \ldots, N/2$ to the desired place and there reconstruct the analog signal

$$\sum_{k=-N/2}^{N/2} \alpha_k e^{2\pi i k x/L}$$

The result is a smoothing or compression of the signal, where the high-frequency parts are omitted.

6.2 Trigonometric Interpolation

We now pose ourselves the problem to interpolate a 2π-peridic function f at equidistant places with trigonometric sums. I.e. for our function f it holds true that

$$f(x + 2\pi) = f(x)$$

and the trigonometric sums have the form:

$$T_n(x) = \sum_{k=0}^{n} \gamma_k e^{ikx}$$

The interpolation interval is $[0, 2\pi]$ with equidistant places

$$x_k = \frac{2\pi}{n+1} k, \quad k = 0, \ldots, n$$

Theorem 6.2.1. *The trigonometric interpolation problem has a unique solution. I.e. for $y_0, \ldots, y_n \in \mathbb{C}$ there is one and only one function*

$$T_n(x) = \sum_{k=0}^{n} \gamma_k e^{ikx}$$

with $T_n(x_\nu) = y_\nu$ for $\nu = 0, \ldots, n$.

Proof. Set $\omega = e^{ix}$, $\omega_k = e^{ix_k} = e^{\frac{2\pi i}{n+1}k}$ and

$$P_n(X) = \sum_{k=0}^{n} \gamma_k X^k$$

Then

$$\begin{aligned} T_n(x) &= P_n(\omega) \\ y_\nu = T_n(x_\nu) &= P_n(\omega_\nu) \end{aligned}$$

As the interpolation polynomial $P_n(X)$ is uniquely determined by Theorem 4.1.3, it follows that also $T_n(x)$ is uniquely determined. □

Calculating the coefficients

The coefficients γ_k of the trigonomatric interpolation polynomial $T_n(x)$ can be calculated as follows:

$$\gamma_k = \frac{1}{n+1}\sum_{\nu=0}^{n} y_\nu e^{-i\nu x_k} = \frac{1}{n+1}\sum_{\nu=0}^{n} y_\nu \omega_k^{-\nu}$$

Proof.

$$\sum_{\nu=0}^{n} y_\nu \omega_k^{-\nu} = \sum_{\nu=0} P_n(\omega_\nu)\omega_k^{-\nu} = \sum_{\nu=0}^{n}\sum_{\ell=0}^{n} \gamma_\ell \omega_\nu^\ell \omega_k^{-\nu} = \sum_{\nu=0}^{n}\sum_{\ell=0}^{n} \gamma_\ell \omega_\nu^{\ell-k}$$

where the last equality holds true because

$$\omega_k^\nu = e^{\frac{2\pi i}{n+1}k\nu} = \omega_\nu^k$$

Hence,

$$\sum_{\nu=0}^{n} y_\nu \omega_k^{-\nu} = \sum_{\nu=0}^{n}\sum_{\ell=0}^{n} \gamma_\ell \omega_\nu^{\ell-k} = \sum_{\ell=0}^{n} \gamma_\ell \sum_{\nu=0}^{n} \omega_\nu^{\ell-k} \overset{(*)}{=} \sum_{\ell=0}^{n} \gamma_\ell \cdot (n+1)\delta_{k,\ell} = \gamma_k(n+1)$$

which implies the assertion. Here $(*)$ holds true because of the orthogonality relations (6.1). □

6.3 Multiplication of Large Numbers

An application of fast multiplication of large numbers occurs in the encryption of data in the internet. This will be treated in Section 7.1.

6.3.1 Multiplication via complex DFT

The multiplication of two m-digit natural numbers by using the school method reduces to m^2 multiplications of 1-digit numbers. For large m, this is very inefficient.

A first idea for dealing with this problem is to view numbers as polynomials. In this way, e.g. the numebr $q = 5821$ in its decimal representation is

$$q = 1 + 2 \cdot 10 + 8 \cdot 10^2 + 5 \cdot 10^3 = Q(10)$$

for the polynomial

$$Q(X) = 1 + 2X + 8X^2 + 5X^3 \in \mathbb{Z}[X]$$

The product of two numbers is thus given by the product $R(X) = P(X) \cdot Q(X)$ of two polynomials with subsequent evaluation

$$pq = R(10)$$

This works not only for decimal numbers, but also for any basis g of a g-adic representation of numbers:

$$a = \sum_{i=0}^{n} a_i g^i$$

with $a_i \in \{0, \ldots, g-1\}$.

The problem with this approach is that the direct multiplication of two polynomials of degree $n-1$ also uses n^2 multiplications. The goal is now to accelerate these multiplications. Assume that $\deg R = m-1$ ist. Then we consider the following method:

1. Evaluate P and Q at m places $x_0, \ldots, x_{m-1}$.
2. $R(x_s) = P(x_s) \cdot Q(x_s)$ evaluates R at m places.
3. Determine from this the coefficients of R (i.e. interpolate).

Step 2 uses only m multiplications. Hence, we need efficient ways to realise Steps 1 and 3.

It helps to take for x_s m-th roots of unity, i.e. solutions of the equation $X^m = 1$. Then

$$x_s = e^{-2\pi is/m} = \omega^{-s}$$

with the primitive m-th root of unity $\omega = e^{2\pi i/m}$. The evaluation of a polynomial $A(X) = \sum_{t=0}^{m-1} a_t X^t$ at the place x_s is thus

$$\tilde{a}_s = A(\omega^{-s}) = \sum_{t=0}^{m-1} a_t e^{-2\pi ist/m}$$

which is nothing else than the s-th coefficient of the discrete Fourier transformation (DFT). These Fourier coefficients can be computed efficiently with the Fast Fourier Transformation (FFT), as seen in Section 6.1.1. Now, we have

Theorem 6.3.1 (Convolution Theorem)**.** *The Fourier coefficients of a product $R(X) = P(X) \cdot Q(X)$ are the products of the Fourier coefficients of the polynomials $P(X)$ and $Q(X)$.*

A consequence is that the Fourier coefficients of the product $R(X)$ can be obtained via the inverse DFT (IDFT). More precisely, we have

$$a_t = \frac{1}{m}\tilde{A}(\bar{\omega}^{-t}) = \frac{1}{m}\sum_{s=0}^{m-1} a_s e^{2\pi ist/m} \tag{6.3}$$

where $\tilde{A}(X) = \sum_{s=0}^{m-1} \tilde{a}_s X^s$ is the Fourier-transformed polynomial of $A(X)$. (6.3) says that the IDFT can also be computed with FFT, where the root of unity ω is replaced by its complex conjugate $\bar{\omega}$.

This method is a lot more efficient for large numbers than the naive multiplication. However, by using the DFT over the complex numbers, one can have rounding errors.

6.3.2 Multiplication via modular DFT

In order to use the FFT method for multiplying polynomials in order to multiply large numbers without obtaining rounding errors, one can work in congruences modulo a large number of the form $N = 2^{2^w} + 1$. For large N, a product of integers $p \cdot q$ is the same thing as $p \cdot q \mod N$. The advantage of this choice of N is that, because

$$2^{2^w} \equiv -1 \mod N \quad \text{und} \quad 4^{2^w} \equiv 2^{2\cdot 2^w} \equiv \left(2^{2^w}\right)^2 \equiv 1 \mod N$$

4 is a primitve 2^w-th root of unity. This means that 2^k-th roots of unity for $k \leq w$ are powers of 2. The Fourier transformation with 2^k-th roots of unity N can be computed efficiently by

using shift operations on binary numbers. This can be exploited to compute the multiplication of large integers with n bits via modular arithmetic with a time-complexity of

$$O(n\log(n)\log(\log(n)))$$

which is a lot mor efficient than the school method with time-complexity $O(n^2)$. It is conjectured that $O(n\log(n))$ is a lower complexity bound for the multiplication of two large integers.

Both, the complex as well as the modular version of the DFT-method for multiplication of large numers are known as the *Schönhage-Strassen algorithm.*

6.4 Euler's Formula and the existence of God

The equation

$$e^{i\pi}+1=0$$

by Leonhard Euler appears as a booktitle in Lisa's collection with which she prepares for her career as baseball trainer.

Another appearance of this equation is in *Homer*[3], where it appears to Homer Simpson in the third dimension.

For some, this formula is a proof of the existence of God, as in it are united the different mathematical disciplines: arithmetic (0 and 1), algebra (i), geometry (π), and analysis (e), a fact which cannot be a coincidence.

However, the nature of God cannot be grasped with our limited human mind.

Although rational arguments for the existence of God are remarkable and also used by the Fathers of the Church, knowledge of God coming from *personal spiritual experience* is of much greater significance. This is expressed in the beatitudes:

> Blessed are the pure in heart: for they shall see God. (Matthew 5:8)

The Holy Fathers of the Church confess the truth of this statement, as they were granted the vision of God after a purification process in which they cleansed the eye of the soul[2] from all stains produced by sin such that they could see with it the Uncreated Light. They say that anyone can reach this state of soul by Holy Baptism into the Orthodox Church, and afterwards (as we keep falling into sin) by struggling against the passions and by meticulously examining one's conscience, and in repentance exposing before a priest during Holy Confession all impurities of the soul.

Orthodox Tradition is not about speculative reasonings *about* God, the aim is rather the participation in divine life. The prophets, apostles and saints experienced this participation and showed the way to it. They speak thereby of experiences which are not expressible in human words. The path to this goal of spiritual life is first of all the knowledge that I, in my present fallen state, am not able to fulfill God's commandments. Then I walk on the path of repentance[3] and am ready to cooperate with God on the healing of my *nous*. This is called *synergy*. The

[2]Greek: $\nu o\upsilon\varsigma$ (nous). this word does not have an adequate expression in Western languages. It describes the organ of the soul with which a person can communicate with God. Due to the fall into sin, this organ has become ill and needs healing. In a spiritually healthy person, it works correctly and enables him to participate in divine life.

[3]gr. metanoia

Church sees itself as a hospital and applies the medication (the *mysteries*) inspired by the Holy Spirit. Christ himself is the doctor. The person healed in the Church is able to participate in divine life and is also capable of selfless love according to the divine commandment.

Thus Orthodoxy is the path

$$\text{purification} \preceq \text{illumination} \preceq \text{glorification}$$

with $\preceq$ as in Example 5.1.22, which stands ready for every person and begins with the purification of the nous. If during this process one has the impression of being already on the stage of illumination, then this is a sure sign of illusion and deep fallenness. For in this case one has not reached true humility. In any case, it is highly recommended to speak with a spiritual father about your spiritual state. The spiritual father should be chosen according to what you know about his own progress on the path of spiritual struggle, preferably already having reached the state of illumination.

Chapter 7

Cryptography

7.1 RSA Cryptography

RSA[1] is an asymmetric cryptographic method for encryption or for digital signatures.

7.1.1 Euler's Phi function

Euler's Phi function phi(n) gives for every natural number n the amount of numbers between 1 and n which are prime to n:

$$\phi(n) := |\{a \in \mathbb{N} \mid 1 \leq a \leq n \text{ and } \operatorname{lcd}(a, n) = 1\}|$$

The Phi function is *weakly multiplikative*, i.e. for coprime m, n the following holds true:

$$\phi(m \cdot n) = \phi(m) \cdot \phi(n)$$

E.g.

$$\phi(18) = \phi(2) \cdot \phi(9) = 1 \cdot 6 = 6$$

Remark 7.1.1. *$\phi(n)$ is the number of invertible elements modulo n.*

In order to calculate $\phi(n)$ we can say:

Lemma 7.1.2. *If p is a prime number, then:*

1. $\phi(p) = p - 1$
2. $\phi\left(p^k\right) = p^k \cdot \left(1 - \frac{1}{p}\right)$ *for* $k \geq 1$.

Proof. 1. p is coprime to all numbers between 1 and $p - 1$, but not to p. These are precisely $p - 1$ numbers.

2. p^k is coprime to precisely the numbers $p \cdot 1, p \cdot 2, \ldots, p \cdot p^{k-1}$ between 1 and p^k. These are precisely

$$p^k - p^{k-1} = p^k \cdot \left(1 - \frac{1}{p}\right)$$

numbers. □

[1] named after R.L. Rivest, A. Shamir und L. Adleman

From the prime number decomposition of n:

$$n = \prod_{p|n} p^{\alpha_p}$$

and the weak multiplicativity of Φ we obtain the formula

$$\phi(n) = \prod_{p|n} p^{\alpha_p} \left(1 - \frac{1}{p}\right) = n \prod_{p|n} \left(1 - \frac{1}{p}\right)$$

Important for cryptography is the Theorem of Fermat-Euler:

Theorem 7.1.3 (Fermat-Euler)**.** *If* $\operatorname{lcd}(a, n) = 1$*, then*

$$a^{\phi(n)} \equiv 1 \mod n$$

7.1.2 RSA crypto system

RSA is an asymetric cryptographic method which uses pairs of keys. The private key is used for decryption or for signing data, and the public key is for encryption or checking a signature. The private key is kept secret, and is difficult to compute from the public key.

The public key is a pair (e, N), and the private key is a pair (d, N). N is called the *RSA module*, e the *private exponent* and d the *public exponent*. The keys are generated as follows:

1. Randomly choose two stochastically independent prime numbers $p \neq q$, for which holds true that
$$0.1 < |\log_2 p - \log_2 q| < 30$$
In practice, numbers of corresponding lengths are generated and checked with a prime number test.

2. Calculate the RSA module $N = p \cdot q$ and Euler's Phi function
$$\phi(N) = (p - 1) \cdot (q - 1)$$

3. Choose a number e prime to $\phi(N)$ with $1 < e < \phi(N)$. This is the public exponent.

4. Calculate the private exponent d as the solution of
$$e \cdot d \equiv 1 \mod \phi(N) \tag{7.1}$$

Remark 7.1.4. *The congruence (7.1) is solved with the extended Euclidean algorithm (Theorem 5.2.3). For reasons of efficiency, e is not chosen too big. A usual choice is the fourth Fermat number:*

$$e = 2^{16} + 1 = 65537$$

e should not be smaller, in order to not give further possibilities for attack.

The encryption of a message m is done thus:

$$c \equiv m^e \mod N$$

The secret code c is then sent to the receiver whose public key is (e, N). It must be made sure that $1 < m < N$.

The secret code is decrypted with the private key (d, N) as follows:

$$m \equiv c^d \mod N$$

This works, because

$$1 = \text{lcd}(e, \phi(N)) = d \cdot e + k \cdot \phi(N)$$

Namely,

$$c^d \equiv m^{d \cdot e} \overset{(*)}{\equiv} m^{d \cdot e + k \cdot \phi(N)} \equiv m^1 \equiv m \mod N$$

where $(*)$ holds true because $m^{\phi(N)} \equiv 1 \mod N$ according to Fermat-Euler (Theorem 7.1.3).

Example 7.1.5. *Key generation for person B.*

1. *Choose $p = 11$ and $q = 13$ as prime numbers.*
2. *The RSA module is $N = p \cdot q = 143$. $\phi(N) = 10 \cdot 12 = 120$.*
3. *Choose $e = 23$: e is prime to N and smaller than N.*
4. *The extended Euclidean algorithm (Theorem 5.2.3) yields:*

$$1 = \text{lcd}(23, 120) = 23 \cdot d + k \cdot 120$$

with $d = 47$ and $k = -9$. Hence, $d = 47$ is the private exponent.

The sender A wants to send to B an encrypted message $m = 7$. For this, A calculates:

$$7^{23} \equiv 2 \mod 143$$

B decrypts the secret code $c = 2$:

$$2^{47} \equiv 7 \mod 143$$

Hence, the plain text message is $m = 7$.

7.1.3 Binary Exponentiation

Encryption and decryption of a message m is done by exponentiaton. Integer powers can be computed efficiently by "continued squaring and occasional multiplication". This works for real numbers, matrices, elliptic curves, or in general in any *semi-group*, i.e. when an operation on a set satisfies the associative law.

Algorithmus 7.1.6. *1. Exponent k is transformed to its binary representation.*

2. *Replace each 0 with Q and each 1 with QM.*
3. *Q means "sqare", and M means "multiply with x".*
4. *Apply the resulting string from left to right on 1.*

For $k > 0$, the binary representation always begins with 1. Hence, the first command is QM, i.e. $1^2 \cdot x = x$. That is why the first step can QM can be replaced with x.

Example 7.1.7. *Let $k = 23$. Its binary representation is $k = 10111$. This yields $QM\,Q\,QM\,QM\,QM$. By the simplification rule, we obtain: $Q\,QM\,QM\,QM$ applied to x. I.e.*

$$x^{23} = \left(\left(\left(x^2\right)^2 \cdot x\right)^2 \cdot x\right)^2 \cdot x$$

Remark 7.1.8. *When calculating modulo N, reduce modulo N after each step Q or M.*

7.1.4 Padding

In practice, the RSA method described above is not used, because it has several weaknesses.

First of all, the method is deterministic. Hence, an attacker can guess some plain text and encrypt it with the public key, and then compare with the code. If the attacker sets up a large table of plaintext-code pairs, then he has a "dictionary" which helps him in the analysis of encrypted messages.

If $c = m^e < N$, then an attacker can compute the integer e-th root of c and obtains the plaintext m.

As the product of two encrypted messages is itself an encrypted message:

$$m_1^e \cdot m_2^e \equiv (m_1 \cdot m_2)^e \mod N$$

an attacker can modify an encrypted message $c \equiv m^e \mod N$ to $c' \equiv c \cdot r^e \mod N$ and ask the receiver to decrypt the innocuous text c', which yields $m' \equiv m \cdot r \mod N$. With the extended Euclidean algorithm (Theorem 5.2.3), the attacker then has the plaintext $m \equiv m' \cdot r^{-1} \mod N$.

Suppose, the same message m is sent to e different receivers having the same public exponent e, but pairwise distinct moduli N_i. Then the attacker has only to solve the congruences

$$x \equiv m^e \mod N_i, \quad i = 1, \ldots, e$$

simultaneously, which can be done with the Chinese Remainder Theorem (Theorem 7.1.9) and yields an $x \equiv m^e \mod \prod N_i$. As $x < \prod N_i$, the integer e-th root can now be calculated in order to compute the plaintext.

Theorem 7.1.9 (Chinese Remainder Theorem)**.** *Let $m_1, \ldots, m_e$ be pairwise coprime integers. Then there exists for each tuple $a_1, \ldots, a_e$ of integers an integer x which solves the simultaneous congruece*

$$x \equiv a_i \mod m_i, \quad i = 1, \ldots, e$$

All solutions of this congruence are congruent modulo $M := m_1 \cdots m_e$.

Proof. For each i, the numbers m_i and $M_i := M/m_i$ are coprime. Hence, according to Theorem 5.2.3, there exist two numbers r_i and s_i with

$$1 = r_i m_i + s_i M i$$

Set $e_i := s_i M_i$. Then

$$\begin{aligned} e_i &\equiv 1 \mod m_i \\ e_i &\equiv 0 \mod m_j, \quad j \neq i \end{aligned}$$

The number $x := \sum_{i=1}^{e} a_i e_i$ is a solution of the simultaneous congruence. □

In order to prevent such attacks, the plaintext is extended by a sequence of characters R with some structure having a randomisation *(Padding)*. Hence, not the message M is transmitted, but the plaintext M extended by R is encrypted. This makes, for a suitable choice of padding method, attacks more difficult. For computing R, often pseudo-random numbers are used.

7.1.5 Security of RSA

The security of the RSA crypto system relies on two mathematical problems:

1. Factorisation of large numbers
2. RSA problem

The RSA problem says: For given $m^e \mod N$ and pair (e, N), determine m. Hence, calculate the e-th root modulo composite N. The most promising approach seems to be to factorise N: If an attacker has the factorisation $N = p \cdot q$, then he computes $\phi(N) = (p-1)(q-1)$ and can calculate e efficiently with the extended Euclidean algorithm from d. However, to date there is no known algorithm on conventional computers which can factorise an integer number in polynomial time. Momentarily, it is recommended to have N at least of length 2048 bits, in order to keep the running time of a factorisation sufficiently long.

On a quantum computer, things look differently: in 1994, Peter Shor developped a quantum algorithm which can factorise natural numbers in polynomial time. This method, if it can be implemented some time in the future, makes RSA insecure.

For reasons of efficiency, RSA is often used as part of a hybrid crypto system. The actual message is encrypted with a *symmetric encryption method*, in which the same key is used for encryption and decryption. RSA is then used for exchanging the key. E.g. the TLS protocal in the internet uses this method.

As key lengths need to become bigger and bigger for retaining the same security, RSA is gradually being replaced with *elliptic curve cryptography*.

7.1.6 A One-Million-Dollar Problem

Another equation which appears to Homer Simpson in the third dimension in *Homer*^{3}, is

$$P = NP$$

This is an answer to the *P-versus-NP problem* of computer science. P is the class of problems, for which there exists an algorithm which can solve it in polynomial time. NP is the class of all problems, for which an answer can be verified in polynomial time. NP stands for *non-deterministic polynomial time*.

An algorithm is said to run in *polynomial time*, if its running time is bounded from above by a polynomial in the size n of the input data. E.g.

$$T(n) = O(n^2)$$

means that the running time is in the worst case quadratic in n.

An example for a problem in NP is the factorisation of natural numbers. It can be checked in polynomial time if a given factorisation of a number is correct. However, there is no known polynomial-time factorisation algorithm on a conventional computer. If $P = NP$ holds true, then every problem that can be verified in polynomial time, can also be solved in polynomial time. Then there must also exist an algorithm for factorising natural numbers which runs in polynomial time. If $P \neq NP$, then this is not necessarily the case.

The P-versus-NP problem is one of the seven Millenium Problems, announced by the Clay Mathematics Institute in the year 2000. However solves one of these problems first, obtains a prize of one million dollars. To date, one of the problems was solved, namely the *Poincaré conjecture* was solved in the year 2002 by G.J. Perelman who declined the prize.

7.2 Elliptic Curve Cryptography

RSA cryptography is at the moment gradually being replace with *elliptic curve cryptography*, as these guarantee more security for equal key length. This kind of cryptography relies on the *Diffie-Hellman key exchange* which will be introduced in the following subsection.

7.2.1 Diffie-Hellman Key Exchange

For the key exchange by Diffie and Hellman, an *abelian group* $(G, +)$ is used, for which the *discrete logarithm problem* (DLP) is difficult to solve. The meaning of these two terms is now going to be explained:

The discrete logarithm problem (DLP). *Let in some abelian group be given the elements P and $n \cdot P$, where n is a natural number. Determine n.*

Definition 7.2.1. *An* abelian group *is a pair $(G, +)$, where G is a set and*

$$+ \colon G \times G \to G$$

is a map (the group addition*, which satisfy the following conditions:*

1. $(a + b) + c = a + (b + c)$ *(associativity)*
2. *There exists an element $0 \in G$, such that always $a + 0 = 0 + a = a$ holds true. (zero or neutral element)*
3. *For every $a \in G$ there exists $a\ -a \in G$ with $a + (-a) = (-a) + a = 0$.* *(inverse)*
4. $a + b = b + a$ *(commutativity)*

Here, a, b, c are arbitrary elements of G.

Example 7.2.2. *Examples for abelian groups are $(\mathbb{Z}, +)$ and $(K, +)$, as well as $(K \setminus \{0\}, \cdot)$ with $K \in \{\mathbb{Q}, \mathbb{R}, \mathbb{C}, \mathbb{F}_p\}$. The latter abelian groups are written multiplicatively: the neutral element is 1, and the inverse of a is a^{-1}.*

On notation. The group addition is often denoted with the symbol $+$. Then one simply writes:

$$nP := \underbrace{P + \cdots + P}_{n \text{ mal}}$$

where $n \in \mathbb{N}$.

The key exchange. There is a public element $P \in G$.

Alice chooses a secret number $n \in \mathbb{N}$ and publishes nP.

Bob chooses a secret number $m \in \mathbb{N}$ and publishes mP.

Alice computes $n(mP) = nmP = Q$.

Bob computes $m(nP) = mnP = Q$.

Q is the secret key.

Security. The security of the Diffie-Hellman key exchange relies on the difficulty of the discrete logarithm problem. Namely, if an attacker discovers the numbers n and m, then he also has the key

$$Q = nmP$$

Example 7.2.3. *The abelian group* $(\mathbb{F}_p^\times, \cdot)$ *with*

$$\mathbb{F}_p^\times := \{1, \ldots, p-1\}$$

for p *prime has the special property that it is* cyclic. *This means that every* $a \in \mathbb{F}_p^\times$ *is the power of a fixed element* $t \in \mathbb{F}_p^\times$*:*

$$a = t^n$$

The element t *is called a* generator *of the cyclic group. The discrete logarithm problem means here: given a generator* t *and* $a = t^n$*, determine* n*.*

Finite fields. Other than the finite fields $\mathbb{F}_p$ there are more finite fields. These are constructed in the following way: Let

$$\mathbb{F}_p[t] := \{\text{Polynome in } t \text{ mit Koeffizienten aus } \mathbb{F}_p\}$$

A polynomial $\pi \in \mathbb{F}_p[t]$ is called *irreducible*, if $\deg(\pi) > 0$ and π has only trivial factorisations:

$$\pi = f \cdot g \quad \Rightarrow \quad f \in \mathbb{F}_p^\times \quad \text{oder} \quad g \in \mathbb{F}_p^\times$$

Let an irreducible polynomial $\pi \in \mathbb{F}_p[t]$ be given, and let $n := \deg(\pi)$. Then

$$\mathbb{F}_{p^n} := \mathbb{F}_p[t]/\pi\mathbb{F}_p[t] := \{\text{Reste von } f \in \mathbb{F}_p[t] \text{ modulo } \pi\}$$

with addition and multiplication of polynomials modulo π, is a field.

The first observation is that every element $a \in \mathbb{F}_{p^n}$ is a linear combination of the elements $1, t, \ldots, t^{n-1}$. This means that $\mathbb{F}_{p^n}$ is a vector space over $\mathbb{F}_p$ of dimension n. It follows that the field $\mathbb{F}_{p^n}$ has precisely p^n elements.

The construction also yields all possible finite fields. Namely:

Theorem 7.2.4. *Every finite field is isomorphic to a field* $\mathbb{F}_{p^n}$*.*

Example 7.2.5. *The polynomial* $\pi = t^2 + t + 1 \in \mathbb{F}_2[t]$ *is irreducible, as it has no zeros in* $\mathbb{F}_2$ *(having a zero is equivalent to having a linear factor). Hence,*

$$\mathbb{F}_{2^2} = \mathbb{F}_2 \cdot 1 + \mathbb{F}_2 t = \{0, 1, t, t+1\}$$

The multiplication table is

$\cdot$	1	t	$t+1$
1	1	t	$t+1$
t	t	$t+1$	1
$t+1$	$t+1$	1	t

Charakteristic of a field. The *charakteristic* of a field K is the smallest positive integer n, for which

$$n \cdot 1_K = 0_K$$

where 1_K and 0_K are the unity element, respectively the zero element of K. If no such number exists, then we define the characteristic of K to be zero. The characteristic of K is denoted as $\mathrm{char}(K)$.

Satz. *The characteristic of a field is either zero or a prime number.*

Proof. Let $n = \mathrm{char}\, K > 0$, and let $n = a \cdot b$ be a factorisztion. Observe that $a \neq 0$ and $b \neq 0$. Then it holds true that

$$a \cdot 1_K \cdot b \cdot 1_K = 0_K$$

where $a \cdot 1_K \neq 0_K$ and $b \cdot 1_K \neq 0_K$, which is impossible in a field, unless $a = 1$ or $b = 1$. Otherwise, we would have in K:

$$b \cdot 1_K = baa^{-1} \cdot 1_K = b \cdot 1_K \cdot a \cdot 1_K \cdot a^{-1} \cdot 1_K = 0_K \cdot a^{-1} \cdot 1_K = 0_K$$

a contradiction. □

Example. *It holds true that*

$$\mathrm{char}\left(\mathbb{F}_{p^n}\right) = p$$

Remark 7.2.6. *The multiplicative group $(\mathbb{F}_{p^n}^{\times}, \cdot)$ is cyclic. This means that it is also interesting for being used in the Diffie-Hellman key exchange.*

7.2.2 Elliptic Curves

Historically, elliptic curves arose from the attempt to calculate the arc length of an ellipse. We will follow this path in the following.

Elliptic Integrals. The arc length L of an ellips can be represented as the following integral:

$$L = 4a \int_0^{\frac{\pi}{2}} \sqrt{1 - k^2 \sin^2 t}\, dt$$

with $k = \frac{\sqrt{a^2 - b^2}}{a}$ and the semi-axes a and b. This integral is an example of an *elliptic integral of the first kind*:

$$E(\phi) = \int_0^{\phi} \sqrt{1 - k^2 \sin^2 t}\, dt$$

Such an integral also appears in the courses *Kartenprojektionslehre* (Master) and *Grundlagen kinematischer und dynamischer Modelle der Geodäsie* (Bachelor). With the substitution $x = \sin t$, this becomes

$$E(u) = \int_0^{u} \frac{1 - k^2 x^2}{\sqrt{(1 - x^2)(1 - k^2 x^2)}}\, dx$$

A general *elliptic integral* is given as

$$f(x) = \int \frac{A(x) + B(x)}{C(x) + D(x)\sqrt{S(x)}}\, dx$$

where A, B, C, D are polynomials, and S is a polynomial of degree 3 or 4.

Example.

$$u = f(x) = \int_0^x \frac{dt}{\sqrt{1-t^2}} = \arcsin(x)$$

is an elliptic integral. Abel saw that it is better to consider the inverse function. In this example, it is $\sin(x)$*, a periodic function.*

Elliptic function. An elliptic function p is the inverse function of an elliptic integrals of the second kind. For $k \neq 0$, elliptic functions are *doubly periodic*:

$$p(u + m\alpha) = p(u + n\beta) = p(u)$$

for certain $\alpha, \beta \in \mathbb{C}$ with $\frac{\alpha}{\beta} \notin \mathbb{R}$. Eisenstein, on the other hand, noticed that doubly periodic functions are elliptic.

The general form of an elliptic function is

$$f(z) = \sum_{m,n \in \mathbb{Z}} (z + m\omega_1 + n\omega_2)^{-2}$$

with the periods $\omega_1, \omega_2 \in \mathbb{C}$, $\frac{\omega_1}{\omega_2} \notin \mathbb{R}$. The function

$$y(z) = \sum_{m,n \in \mathbb{Z}} (z + m\omega_1 + n\omega_2)^{-2} - \sum_{m,n \in \mathbb{Z}\setminus\{0\}} (m\omega_1 + n\omega_2)^{-2}$$

satisfies a differential equation of the form

$$y'(z)^2 = p(y(z))$$

where $p(X)$ is a polynomial of degree 3 with only simple zeros.

Weierstraß $\wp$-function. The Weierstraß $\wp$-function is

$$\wp(z) = z^{-2} + \sum_{m,n \in \mathbb{Z}\setminus\{0\}} (z + m\omega_1 + n\omega_2)^{-2} - (m\omega_1 + n\omega_2)^{-2}$$

It satisfies the differential equation

$$\wp'(z)^2 = 4\wp(z)^3 - g_2\wp(z) - g_3$$

with coefficients g_2, g_3. By setting $x := \wp(z)$ and $y := \wp'(z)$, one obtains the equation

$$E\colon y^2 = 4x^3 - g_2 x - g_3$$

This is the equation of an *elliptisc curve*. Due to the construction, one can see that the elliptic curve E is isomorphic to $\mathbb{C}/\Lambda$ with the lattice

$$\Lambda = \{m\omega_1 + n\omega_2 \mid m, n \in \mathbb{Z}\}$$

Hence, E is an abelian group.

Elliptic curves over a field K. Formally, an elliptic curve over the field K is a non-singular projective algebraic curve of genus 1. The equation

$$E\colon y^2 = x^3 + ax + b$$

with $a, b \in K$ is the *Weierstraß normal form* and is valid in case $\mathrm{char}(K) \neq 2, 3$. The polynomial $f(x) = x^3 + ax + b$ has only simple zeros. On an elliptic curve, one can define geometrically certain things. First the *K-rational* points are given as

$$E(K) = \left\{(x, y) \in K^2 \mid y^2 = x^3 + ax + b\right\} \cup \{O\}$$

with the point O "at infinity". This point can be found in the projective plane $\mathbb{P}^2$ via homogenising the equation for E:

$$y^2 z = x^3 + axy^2 + bz^3$$

and for $z = 0$ it follows that $x = 0$ and $y = 1$ (projective coordinates!). Hence, $O = (0 : 1 : 0)$.

The next step is to observe that E is symmetric with respect to the x-axis:

$$P = (x, y) \in E(K) \Rightarrow -P := (x, -y) \in E(K)$$

Define $-O := O$. Further, let $P, Q \in E(K)$. Then the straight line L through P and Q intersects the curve E in a third point $R \in E(K)$. Hence, we can define:

$$\begin{aligned} P + Q &:= Q + P := -R \\ P + O &:= O + P := P \\ P + (-P) &:= O \end{aligned}$$

If $P = Q$, then let L be the tangent in P. If L intersects E in a second point $R \in E(K)$, then let

$$2P := -R$$

otherwise, let

$$2P = -P$$

We have

Satz. *$(E(K), +)$ is an abelian group with zero element O.*

Elliptic curves over $\mathbb{F}_{p^n}$. Now, let $K = \mathbb{F}_{p^n}$. With $q = p^n$, we have:

Satz (Hasse-bound)**.** *For the number $|E(K)|$ of K-rational points on an elliptic curve E, it holds true that:*

$$||E(K)| - (q + 1)| \leq 2\sqrt{q}$$

It follows that for large n, there are about $q = p^n$ K-rational points on the elliptic curve E. This is interesting for generating keys. Further: $E(K)$ is cyclic or a product of two cyclic groups. This is important for the key exchange, as the public point P should be an element of high order:

$$nP = O \quad \text{with minimal } n > 0 \text{ as large as possible}$$

ECDLP.

- There are numerous suitable elliptic curves over finite fields.
- The elliptic curve discrete logarithm problem (ECDLP) is more difficult than the factorisation of natural numbers or the DLP in $\mathbb{F}_q^\times$ with $q = p^n$.
- Best possible fields are $K = \mathbb{F}_p$ (p prime) or $K = \mathbb{F}_{2^n}$.

7.3 Quantum Cryptography

As in the near future, there will be powerful quantum computers, RSA and elliptic curve cryptography will become insecure. A way out is *quantum cryptography*. There, polarised photons are used, and the laws of quantum mechanics guarantee that a secure key exchange can be managed.

In the BB84 protocol, two pairs of orthogonal polarisation states are used:

- die *rectilinear basis*: 0° and 90°
- die *diagonal basis*: 45° and 135°

From the laws of quantum mechanics, it follows that no measurement can distinguish all four different states, as they are not all pairwise orthogonal. The reason is that a measurement chooses an orthonormal basis, and the measurement result is one of these orthogonal states. For example, in the rectilinear basis, only the states "horizontal" and "vertical" can be measured. After a measurement, the photon is in the measured state, i.e. a measurement changes the state of a particle.

The BB84 protocol.

1. Alice sets up a coding table, e.g.

	0	1
+	↑	→
×	↗	↘

2. Alice generates a random bit (0 or 1) and randomly chooses an onb (rectilinear or diagonal), and sends to Bob a photon in the corresponding state. She repeats this process several times.
3. Bob randomly chooses a basis and measures the state of the photon.
4. Alice and Bob compare their sequences of bases. If the bases are equal, they retain the corresponding bit, otherwise they discard it.

In about 50% of the cases, Alice and Bob have a common bit. The sequence of retained bits is the common key. In order to check if there was an eavesdropper, Alice and Bob compare a chosen subsequence of their versions of the key. If Eve has obtained information about the polarisations, transmission errors must have occurred. If too many bits are different, they discard their key and repeat the procedure on a different channel.

Example 7.3.1. *Assume that Alice chooses the basis + and her photon has the polarisation →. If Bob also chooses +, then he will measure →, and both agree on this bit. If, on the other hand, Bob chooses ×, then he will measure either ↗ or ↘, each with probability $\frac{1}{2}$. So, there is a 50% chance that their bits will disagree.*

Remark 7.3.2. *Assume that Eve is an eavesdropper who manages to catch a photon. She does not know which onb was used for its polarisation. So, she chooses one at random and performs a measurement. If the onb's do not agree, then her measurement will alter the state of the photon, and there is a chance that an error is introduced which will be observed by Alice and Bob.*

Chapter 8

Approximation

The task of approximation is to approximate a function f (which may be unknown) with a simpler function (e.g. a polynomial). We will consider *linear approximation*, in which f is to be approximated with a linear combination of predetermined linearly independent functions $f_1, \ldots, f_n$:

$$f \approx \sum_{i=1}^{n} \gamma_i f_i$$

Here, the functions $f_1, \ldots, f_n$ span a linear subspace U of the space $V = C[a, b]$ of continuous functions on the interval $[a, b]$. The approximation task is to approximate f as well as possible by an element of the given subspace U of V.

Example. *Examples for given subspaces of $V = C[a, b]$ are such that are spanned by polynomials, trigonometric functions, exponential functions or rational functions. E.g.*

1. $f_1 = 1,\ f_2 = x,\ f_3 = x^2, \ldots,\ f_n = x^{n-1}$.
2. $f_1 = 1,\ f_2 = \cos x,\ f_3 = \sin x,\ f_4 = \cos 2x,\ f_5 = \sin 2x,\ \ldots$
3. $f_1 = 1,\ f_2 = e^{\alpha_1 x},\ f_2 = e^{\alpha_2 x}, \ldots$
4. $f_1 = 1,\ f_2 = \frac{1}{(x-a_1)^{p_1}},\ f_3 = \frac{1}{(x-a_2)^{p_2}}, \ldots$

8.1 Best Approximation

In order to evaluate an approximation or to determine an approximation error, a *norm* on the vector space V is used.

Let $K = \mathbb{R}$ or $K = \mathbb{C}$, and let V be a K-vector space.

Definition 8.1.1. *A* Norm *on V is a function $\|\cdot\| \colon V \to \mathbb{R}_{\geq 0}$ with the following properties:*

1. $\|f\| = 0$, *if and only if* $f = 0$.
2. $\|\alpha f\| = |\alpha| \|f\|$ *for* $\alpha \in K$
3. $\|f + g\| \leq \|f\| + \|g\|$

A norm always defines a metric on V by:

$$d(f, g) := \|f - g\|$$

Let us check that d is indeed a metric:

Proof. 1. $d(f,g) = 0$ holds true, if and only if $\|f-g\| = 0$. This holds true, if and only if $f - g = 0$. This is the case, if and only if $f = g$. This shows positivity.

2. Symmetry follows from:

$$d(f,g) = \|f-g\| = \|(-1)(g-f)\| = |-1|\|g-f\| = \|g-f\| = d(g,f)$$

3. The triangle inequality holds true because:

$$d(f,h) = \|f-h\| = \|f-g+g-h\| \leq \|f-g\| + \|g-h\| = d(f,g) + d(g,h)$$

□

Example 8.1.2. *Examples for norms on $V = C[a,b]$ are:*

1. *L^1-norm:*
$$\|f\|_1 := \int_a^b |f(t)|\,dt$$

2. *L^2-norm:*
$$\|f\|_2 := \left(\int_a^b |f(t)|^2\,dt\right)^{\frac{1}{2}}$$

3. *L^∞-norm:*
$$\|f\|_\infty := \max_{t\in[a,b]} |f(t)|$$

Now, we can define a *best approximation* to f from the subspace U. This is a function $\hat{\phi} \in U$ with

$$d(f,\hat{\phi}) = \min_{\phi\in U} d(f,\phi)$$

Theorem 8.1.3 (Existence Theorem)**.** *For every function $f \in V = C[a,b]$ and every finite-dimensional linear subspace U of V and every norm $\|\cdot\|$ on V, there exists at least one best approximation $\hat{\phi} \in U$ to f.*

8.2 Gauß approximation

We now endow $V = C[a,b]$ with the L^2-norm. A best approximation with respect to the L^2-norm is called *best L^2-approximation*.

In the following, we exploit the fact that the L^2-norm comes from an inner product on V. Namel, we have

$$\|f\|_2 = \sqrt{\langle f, f\rangle}$$

where $\langle\cdot,\cdot\rangle$ is the standard inner product on V:

$$\langle f,g\rangle = \int_a^b f(t)\overline{g(t)}\,dt$$

Let U be a finite-dimensional linear subspace of V, and $\hat{\phi} \in U$. We now consider the approximation error for approximating $f \in V$ with $\hat{\phi}$. We have

Lemma 8.2.1. *$\hat{\phi} \in U$ is best L^2-approximation to f, if and only if $f - \hat{\phi}$ is orthogonal to U.*

Proof. $\Leftarrow$. Let $f - \hat{\phi}$ be orthogonal to U. Then with $\phi \in U$ arbitrary:

$$\left\|f - \hat{\phi}\right\|^2 = \langle f - \hat{\phi}, f - \hat{\phi}\rangle = \langle f - \hat{\phi}, f - \phi + \phi - \hat{\phi}\rangle = \langle f - \hat{\phi}, f - \phi\rangle + \underbrace{\langle f - \hat{\phi}, \phi - \hat{\phi}\rangle}_{=0}$$
$$= \langle f - \hat{\phi}, f - \phi\rangle \leq \left\|f - \hat{\phi}\right\| \|f - \phi\|$$

Here, we have $\langle f - \hat{\phi}, \phi - \hat{\phi}\rangle = 0$, as $\phi - \hat{\phi} \in U$. The last inequality is the Cauchy-Schwarz inequality. It follows that

$$\left\|f - \hat{\phi}\right\| \leq \min_{\phi \in U} \|f - \phi\|$$

$\Rightarrow$. Let $K = \mathbb{R}$. If $\hat{\phi} \in U$ is best L^2-approximation, and $\phi \in U$, then

$$F_\phi(t) := \left\|f - \hat{\phi} - t\phi\right\|^2$$

has a minimum in $t = 0$. Then

$$0 = \frac{d}{dt} F_\phi(t)\Big|_{t=0} = \frac{d}{dt}\left\|f - \hat{\phi} - t\phi\right\|^2\Big|_{t=0} = 2\,\langle f - \hat{\phi} - t\phi, \phi\rangle\Big|_{t=0}$$

as

$$\frac{d}{dt}\langle a + bt, a + bt\rangle = \frac{d}{dt}\left(\langle a, a\rangle + 2t\langle a, b\rangle + t^2\langle b, b\rangle\right) = 2\langle a, b\rangle + 2t\langle b, b\rangle = 2\langle a + bt, b\rangle$$

It follows that for every $\phi \in U$ we have:

$$\langle f - \hat{\phi}, \phi\rangle = 0$$

i.e. $f - \hat{\phi}$ is orthogonal to U.

Let now $K = \mathbb{C}$. If $f - \hat{\phi}$ is not orthogonal to U, then there exists some $\psi \in U$ with

$$\langle f - \hat{\phi}, \psi\rangle \neq 0$$

Without loss of generality let $\langle f - \hat{\phi}, \psi\rangle < 0$. Otherwise, replace ψ with $e^{i\alpha}\psi$ for a suitable argument α. For $0 < t << 1$ we now have

$$\begin{aligned}\left\|f - \hat{\phi} + t\psi\right\|^2 &= \langle f - \hat{\phi} + t\psi, f - \hat{\phi} + t\psi\rangle \\ &= \langle f - \hat{\phi}, f - \hat{\phi}\rangle + \underbrace{t\langle f - \hat{\phi}, \psi\rangle + t\langle \psi, f - \hat{\phi}\rangle + t^2\langle \psi, \psi\rangle}_{<0} \\ &< \langle f - \hat{\phi}, f - \hat{\phi}\rangle = \left\|f - \hat{\phi}\right\|^2\end{aligned}$$

This means that $\hat{\phi} - t\psi \in U$ is a better L^2-approximation than $\hat{\phi}$. Hence, $\hat{\phi}$ is not best L^2-approximation. □

From this, we obtain uniqueness:

Theorem 8.2.2 (Gauß approximation)**.** *For $f \in V = C[a, b]$ and finite-dimensional linear subspace U of V, there exis precisely one best L^2-approximation $\hat{\phi} \in U$.*

Proof. Let $\hat{\phi}_1, \hat{\phi}_2 \in U$ be best L^2-approximations to f. Then for all $\phi \in U$:

$$\langle f - \hat{\phi}_1, \phi \rangle = \langle f - \hat{\phi}_2, \phi \rangle = 0$$

Then also

$$0 = \langle f - \hat{\phi}_2 - (f - \hat{\phi}_1), \phi \rangle = \langle \hat{\phi}_1 - \hat{\phi}_2, \phi \rangle$$

Hence, because $\hat{\phi}_1 - \hat{\phi}_2 \in U$:

$$0 = \langle \hat{\phi}_1 - \hat{\phi}_2, \hat{\phi}_1 - \hat{\phi}_2 \rangle = \left\| \hat{\phi}_1 - \hat{\phi}_2 \right\|^2$$

Hence, $\phi_1 = \phi_2$. This proves uniqueness.

Existence already follows from Theorem 8.1.3. □

What is still missing, is a method for calculating the best L^2-approximation. For this, we start with a basis $f_1, \ldots, f_n$ of U. The best L^2-approximation $\hat{\phi}$ has the form

$$\hat{\phi} = \sum_{k=1}^{n} \gamma_k f_k$$

For the approximation error $f - \hat{\phi}$ it follows with $\phi \in U$ that

$$0 = \langle f - \hat{\phi}, \phi \rangle = \left\langle f - \sum_{k=1}^{n} \gamma_k f_k, \phi \right\rangle = \langle f, \phi \rangle - \sum_{k=1}^{n} \gamma_k \langle f_k, \phi \rangle$$

This means that the coefficients $\gamma_1, \ldots, \gamma_n$ solve the system of linear equations

$$\sum_{k=1}^{n} \langle f_k, f_\ell \rangle \gamma_k = \langle f, f_\ell \rangle, \quad \ell = 1, \ldots, n \tag{8.1}$$

Here, the coefficient matrix

$$A = (\langle f_k, f_\ell \rangle)$$

is hermitean, respectively, symmetric. A is even positive definite. Namely, for $g = (\gamma_k)$ we have

$$g^* A g = \sum_{k,\ell=1}^{n} \bar{\gamma}_\ell \gamma_k \langle f_k, f_\ell \rangle = \langle \hat{\phi}, \hat{\phi} \rangle = \|\phi\|^2$$

The equations (8.1) are called *normal equations*. Their uniquely determined solution is the best L^2-approximation. Ideally, one takes an orthonormal basis $f_1, \ldots, f_n$ of U. Then from the normal equations, it follows that

$$\gamma_k = \langle f, f_k \rangle$$

i.e. the best L_2-approximation is given in this case by

$$\hat{\phi} = \sum_{k=1}^{n} \langle f, f_k \rangle f_k$$

Example 8.2.3. *We want to find the best L^2-approximation to*

$$f \colon [-1, 1] \to \mathbb{R}, \quad x \mapsto \frac{1}{1 + x^2}$$

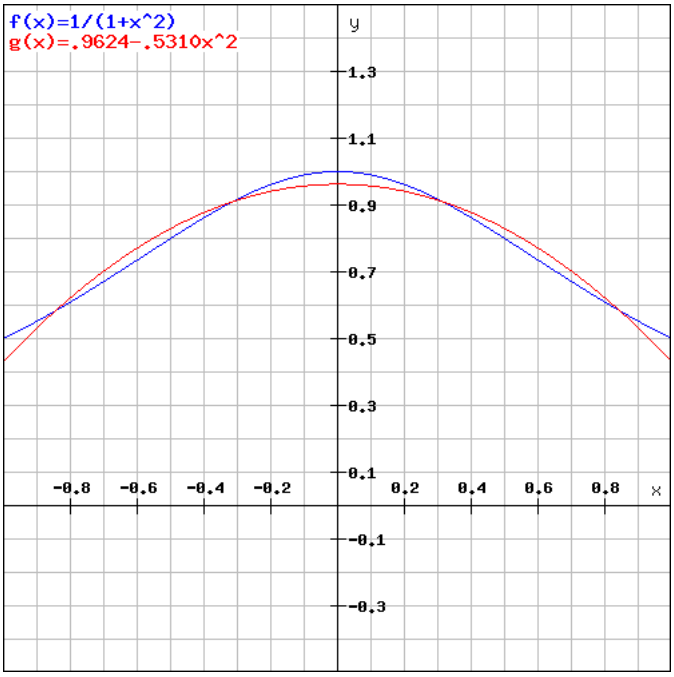


Figure 8.1: A function and its best L^2-approximation with quadratic polynomials.

with quadratic polynomials.

A basis for the linear subspace U of $C[-1,1]$ consisting of quadratic polynomials is given by

$$f_1 = 1,\ f_2 = x,\ f_3 = x^2$$

The normal equations are given by the following system of linear equations:

$$\begin{pmatrix} \int_{-1}^{1} dx & \int_{-1}^{1} x\,dx & \int_{-1}^{1} x^2\,dx \\ \int_{-1}^{1} x\,dx & \int_{-1}^{1} x^2\,dx & \int_{-1}^{1} x^3\,dx \\ \int_{-1}^{1} x^2\,dx & \int_{-1}^{1} x^3\,dx & \int_{-1}^{1} x^4\,dx \end{pmatrix} \begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \hat{\gamma}_3 \end{pmatrix} = \begin{pmatrix} \int_{-1}^{1} \frac{dx}{1+x^2} \\ \int_{-1}^{1} \frac{x\,dx}{1+x^2} \\ \int_{-1}^{1} \frac{x^2\,dx}{1+x^2} \end{pmatrix}$$

o, equivalently:

$$\begin{pmatrix} 2 & 0 & 2/3 \\ 0 & 2/3 & 0 \\ 2/3 & 0 & 2/5 \end{pmatrix} \begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \hat{\gamma}_3 \end{pmatrix} = \begin{pmatrix} 2\arctan 1 \\ 0 \\ 2 - 2\arctan 1 \end{pmatrix}$$

Its solution is

$$\begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \hat{\gamma}_3 \end{pmatrix} = \begin{pmatrix} 0.9624 \\ 0 \\ -0.5310 \end{pmatrix}$$

Hence, $\hat{\phi} = 0.9624 - 0.5310x^2$ is the best L^2-approximation from U to f. Figure 8.1 shows f and the best L^2-approximating quadratic polynomial to f. More efficient is the use of orthogonal polynome *from the following section.*

Trigonometric functions

Let

$$f_k = e^{ikx}, \quad k = -n, \dots, n$$

These functions are an orthonormal system for the linear subspace of $C[0, 2\pi]$, spanned by them. If

$$f = \sum_{\nu=-\infty}^{\infty} \gamma_\nu e^{i\nu x}$$

is a Fourier series, then

$$\gamma_k = \frac{1}{2\pi}\langle f, f_k\rangle = \frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-ikx}\,dx$$

Hence,

$$\hat{\phi} = \sum_{k=-n}^{n} \gamma_k e^{ikx} \in U$$

is the best L^2-approximation to f. Smoothing by filtering the high-frequent parts is nothing other than using the best L^2-approximation.

8.3 Orthogonal Polynomials

Here, we use a continuous weight function

$$w\colon [a, b] \to \mathbb{R}_{>0}$$

and define a weighted inner product

$$\langle f, g\rangle_w = \int_a^b f(t)\overline{g(t)}w(t)\,dt$$

on $V = C[a, b]$. This is indeed an inner product:

Proof. 1. If $f \neq 0$, then there exists an interval I_ϵ with $f(x) \neq 0$ for all $x \in I_\epsilon$. Then

$$\langle f, f\rangle_w = \int_a^b |f(t)|^2 w(t)\,dt \geq \int_{I_\epsilon} |f(t)|^2 w(t)\,dt > 0$$

I.e. $\langle f, f\rangle_w = 0$ implies $f = 0$.

2. $\langle f, g\rangle_w = \overline{\langle g, f\rangle_w}$ follows from inserting into the integral.
3. It also holds true that

$$\begin{aligned}\langle \alpha f + g\rangle_w &= \alpha\langle f, g\rangle_w \\ \langle f + g, h\rangle_w &= \langle f, h\rangle_w + \langle g, h\rangle_w\end{aligned}$$

□

The inner product $\langle \cdot, \cdot \rangle_w$ induces a norm $\|\cdot\|_w$ on V via

$$\|f\|_w := \sqrt{\langle f, f \rangle_w} = \left(\int_a^b |f(t)|^2 w(t)\, dt \right)^{\frac{1}{2}}$$

Orthogonal polynomials arise by orthogonalisation of $1, X, X^2, \ldots$:

$$\begin{aligned} p_0 &= 1 \\ p_n &= X^n - \sum_{\mu=0}^{n-1} \frac{\langle X^n, p_\mu \rangle_w}{\langle p_\mu, p_\mu \rangle_w} p_\mu, \qquad\qquad n = 1, 2, 3, \ldots \end{aligned}$$

according to the method of Gram-Schmidt (cf. Section 5.11). It holds true that p_n is a monic polynomial and is orthogonal to $K[X]_{n-1}$, the polynomials of degree $\leq n-1$.

Orthogonal polynomials satisfy a 3-term recursion:

Theorem 8.3.1 (3-term recursion)**.** *For orthogonal polynomials $p_0, p_1, \ldots$ it holds true that:*

$$p_0 = 1, \quad p_1 = X - \beta_0, \quad p_{n+1} = (X - \beta_n) p_n - \gamma_n^2 p_{n-1}$$

with $n = 1, 2, \ldots$. Here, we have

$$\beta_n = \frac{\langle X p_n, p_n \rangle}{\langle p_n, p_n \rangle}, \quad \gamma_n^2 = \frac{\langle p_n, p_n \rangle}{\langle p_{n-1}, p_{n-1} \rangle}$$

Proof. $p_0 = 1$ by construction (Gram-Schmidt). Also:

$$p_1 = X - \frac{\langle X, p_0 \rangle}{\langle p_0, p_0 \rangle} p_0 = X - \beta_0$$

Let $n \geq 1$, and let

$$q_{n+1} := (X - \beta_n) p_n - \gamma_n^2 p_{n-1}$$

We need to show that $q_{n+1} = p_{n+1}$. First, q_{n+1} and p_{n+1} are monic polynomials of degree $n+1$. Hence,

$$r := p_{n+1} - q_{n+1} \in K[X]_n$$

Now, we show that q_{n+1} is orthogonal to $K[X]_n$. Then also $r = p_{n+1} - q_{n+1}$ is orthogonal zu $K[X]_n$, in particular,

$$\langle r, r \rangle = 0$$

Hence: $q_{n+1} = p_{n+1}$. In order to prove our orthogonality claim, we show one after the other that q_{n+1} is orthogonal to p_n, to p_{n-1} and to $K[X]_{n-2}$. We have

$$\langle q_{n+1}, p_n \rangle = \langle X p_n, p_n \rangle - \beta_n \langle p_n, p_n \rangle - \gamma_n^2 \underbrace{\langle p_{n-1}, p_n \rangle}_{=0} = 0$$

by definition of β_n. Hence, q_{n+1} is orthogonal to p_n. Further:

$$\begin{aligned} \langle q_{n+1}, p_{n-1} \rangle &= \langle X p_n, p_{n-1} \rangle - \beta_n \underbrace{\langle p_n, p_{n-1} \rangle}_{=0} - \underbrace{\gamma_n^2 \langle p_{n-1}, p_{n-1} \rangle}_{= \langle p_n, p_n \rangle} \\ &= \underbrace{\langle X p_n, p_{n-1} \rangle}_{\overset{(*)}{=} \langle p_n, X p_{n-1} \rangle} - \langle p_n, p_n \rangle = \langle p_n, \underbrace{X p_{n-1} - p_n}_{\in K[X]_{n-1}} \rangle = 0 \end{aligned}$$

interval	$w(x)$	orthogonal polynomials
$[-1,1]$	1	Legendre polynomials
$[-1,1]$	$\frac{1}{\sqrt{1-x^2}}$	Chebyshev polynomials
$[-1,1]$	$(1-x)^\alpha(1+x)^\beta,\ \alpha,\beta>-1$	Jacobi polynomials
$(-\infty,\infty)$	e^{-x^2}	Hermite polynomials
$(0,\infty)$	$e^{-x}x^\alpha,\ \alpha>-1$	Laguerre polynomials

Table 8.1: Some classes of orthogonal polynomials.

where $(*)$ holds true, as X takes only real values:

$$\langle Xp_n, p_{n-1}\rangle = \int_a^b tp_n(t)\overline{p_{n-1}(t)}w(t)\,dt = \int_a^b p_n(t)\overline{tp_{n-1}(t)}w(t)\,dt$$

Hence, q_{n+1} is orthogonal to p_{n-1}. Now, let $q \in K[X]_{n-2}$. Then

$$\langle q_{n+1}, q\rangle = \underbrace{\langle Xp_n, q\rangle}_{=\langle p_n, Xq\rangle=0} - \beta_n \underbrace{\langle p_n, q\rangle}_{=0} - \gamma_n^2 \underbrace{\langle p_{n-1}, q\rangle}_{=0} = 0$$

Hence, q_{n+1} is orthogonal to $K[X]_{n-2}$. As p_n, p_{n-1} and $K[X]_{n-2}$ span the space $K[X]_n$, the claim follows. □

Example 8.3.2. *Special weight functions on special intervals yield orthogonal polynomials with certain names. Some are listed in table 8.1.*

Example 8.3.3. *We want to find again the best L^2-approximation to $f = \frac{1}{1+x^2}$ by quadratic polynomials. This time, orthogonal polynomials are to be used.*

First, we construct the orthogonal polynomials. The weight function is $w = 1$, the interval is $[-1,1]$. By table 8.1, these are the Legendre polynomials. In the 3-terme recursion (Theorem 8.3.1) the following parameters arise:

$$\beta_0 = \beta_1 = 0, \quad \gamma_1^2 = \frac{1}{3}$$

This means that the firs three Legendre polynomials are

$$p_0 = 1, \quad p_1 = X, \quad , p_2 = X^2 - \frac{1}{3}$$

In order to obtain an orthonormal basis of $K[X]_2$, we find the normalising factors:

$$\langle p_0, p_0\rangle = 2$$

$$\langle p_1, p_1\rangle = \int_{-1}^1 t^2\,dt = \frac{2}{3}$$

$$\langle p_2, p_2\rangle = \int_{-1}^1 \left(t^4 - \frac{2}{3}t^2 + \frac{1}{9}\right)\,dx = \frac{8}{45}$$

Thus, we obtain the polynomials:

$$P_0 = \frac{1}{\sqrt{2}}, \quad P_1 = \sqrt{\frac{3}{2}}X, \quad P_2 = \frac{3}{2}\sqrt{\frac{5}{2}}\left(X^2 - \frac{1}{3}\right)$$

The coefficients are:

$$\begin{aligned}\gamma_0 &= \langle f, P_0\rangle = \sqrt{2}\arctan 1 \approx 1.111\\ \gamma_1 &= \langle f, P_1\rangle = 0\\ \gamma_2 &= \langle f, P_2\rangle = \frac{2}{3}\sqrt{\frac{5}{2}}\left(2 - \frac{8}{3}\arctan 1\right) \approx -0.2239\end{aligned}$$

Hence,

$$\hat{\phi} = \gamma_0 P_0 + \gamma_1 P_1 + \gamma_2 P_2 \approx 0.9624 - 0.5310x^2$$

is the best L^2-approximation to f with quadratic polynomials.

8.4 Chebyshev approximation

Here, let $K = \mathbb{R}$, and

$$C[a,b] = \{f\colon [a,b] \to \mathbb{R} \mid f \text{ stetig}\}$$

this time endowed with the L^∞-norm $\|\cdot\|_\infty$:

$$\|f\|_\infty = \max_{t\in[a,b]} |f(t)|$$

Let U be a finite-dimensional linear subspace of $V = C[a,b]$, and let $f \in V$. Here, the best L^∞-approximation $\hat{\phi} \in U$ with

$$\left\|f - \hat{\phi}\right\|_\infty = \min_{\phi\in U} \|f - \phi\|_\infty$$

is in general not unique.

Example 8.4.1. *Let $[a,b] = [0,1]$, $f = 1$ and $U = \mathbb{R}x$. Now, for $\phi \in U$:*

$$\|f - \phi\|_\infty \geq 1$$

and for all ϕ of the form $\phi = \alpha x$ with $0 \leq \alpha \leq 2$:

$$\|f - \phi\|_\infty = 1$$

Here, there is a non-uniqueness of the best L^∞-approximation.

Uniqueness is given by the

Haar condition. *Let $\dim U = n$ and let the interpolation problem*

$$\phi(x_i) = y_i, \quad i = 1, \ldots, n$$

with arbitrary places $a \leq x_1 < \cdots < x_n \leq b$ and values $y_1, \ldots, y_n$ always have a solution $\phi \in U$.

Proof. Let $f_1, \ldots, f_n$ be a basis of U. An interpolating $\phi = \sum_{i=1}^{n} \gamma_i f_i$ exists, if and only if the system of linear equations

$$\sum_{i=1}^{n} \gamma_i f_i(x_j) = y_j, \quad j = 1, \ldots, n \tag{8.2}$$

has a solution for $g = (x_i) \in \mathbb{R}^n$. The Haar condition says that the linear system of equations has a unique solution. If one rewrites (8.2) as

$$Ag = y \tag{8.3}$$

with $A = (f_i(x_j)) \in \mathbb{R}^{n \times n}$ and $y = (y_j) \in \mathbb{R}^n$, then, by the Haar condition, (8.3) has a solution for every right-hand side y. If we choose as right-hand side every column of the unity matrix I, then it follows that the matrix equation

$$AX = I$$

has a solution. Hence, A is invertible, and thus (8.3) has a unique solution. □

8.5 Chebyshev polynomials of the first kind

The Chebyshev polynomials of the first kind can be defined directly as

$$T_n(x) = \cos(n \arccos(x)), \quad x \in [-1, 1],\ n = 0, 1, 2, \ldots$$

We have for $\theta \in [0, \pi]$

$$T_n(\cos\theta) = \cos(n\theta)$$

These polynomials satisfy the recursion:

$$\begin{aligned} T_0(X) &= 1 \\ T_1(X) &= X \\ T_{n+1}(X) &= 2XT_n(X) - T_{n-1}(X), \quad n = 1, 2, 3, \ldots \end{aligned}$$

Proof. From the addition thereom for the cosine:

$$\cos x + \cos y = 2 \cos\left(\frac{x+y}{2}\right) \cos\left(\frac{x-y}{2}\right)$$

it follows that

$$2\cos\theta\cos(n\theta) = \cos((n+1)\theta) + \cos((n-1)\theta)$$

Hence, with $t = \cos\theta$:

$$2tT_n(t) - T_{n-1}(t) = 2\cos\theta\cos(n\theta) - \cos((n-1)\theta) = \cos((n+1)\theta) = T_{n+1}(t)$$

□

It follows that $T_n(X) \in \mathbb{Z}[X]_n$ is a polynomial with integer coefficients. The leading coefficient is 2^{n-1} and $\deg(T_n) = n$.

Properties of the Chebyshev polynomials of the first kind

1. We have
$$\max_{t\in[-1,1]} |T_n(t)| = 1$$

2. T_n has in $[-1,1]$ in total $n+1$ extrema:
$$s_k^{(n)} = \cos\left(\frac{k\pi}{n}\right), \quad T_n\left(s_k^{(n)}\right) = (-1)^k, \quad k = 0,1,\ldots,n$$

3. T_n has in $[-1,1]$ in total n simple zeros
$$t_k^{(n)} = \cos\left(\frac{(2k-1)\pi}{2n}\right), \quad T_n\left(t_k^{(n)}\right) = 0, \quad k = 1,\ldots,n$$

4. It holds true that
$$\max_{t\in[-1,1]} \prod_{k=1}^{n+1} \left|t - t_k^{(n+1)}\right| = 2^{-n}$$

5. Orthogonality relations:
$$\int_{-1}^{1} T_n(t)T_m(t)\,\frac{dt}{\sqrt{1-t^2}} = \begin{cases} 0, & n \neq m \\ \pi, & n = m = 0 \\ \frac{\pi}{2}, & n = m \neq 0 \end{cases}$$

   Hence, Chebyshev polynomials are orthogonal polynomials.

Proof of 4. This follows from:
$$\frac{1}{2^n}T_{n+1}(X) = \prod_{k=0}^{n+1}\left(X - t_k^{(n+1)}\right)$$
and property 1. □

Example 8.5.1. *In Figure 8.2, the Chebyshev polynomials of the first kind T_2 to T_7 are depicted.*

8.6 Optimal Lagrange interpolation

Let $f \in C[a,b]^{n+1}$. We want to approximate f as well as possible by polynomial interpolation in $n+1$ places. If $P_n(X) \in \mathbb{R}[X]_n$ is the Lagrange interpolation polynomial, then the interpolation error is given by:
$$f(t) - P_n(t) = \frac{f^{(n+1)}(\xi)}{(n+1)!}N_{n+1}(t)$$
(4.1.8), where $\xi \in I_t$ and
$$N_{n+1}(X) = \prod_{\nu=0}^{n}(X - x_\nu)$$
and I_t is the smallest interval containing the places $x_0 < \cdots < x_n$ (in the interval $[a,b]$) and t.

The task is now to choose the places $x_0,\ldots,x_n$ in such a way that $\|N_{n+1}\|_\infty$ becomes minimal.

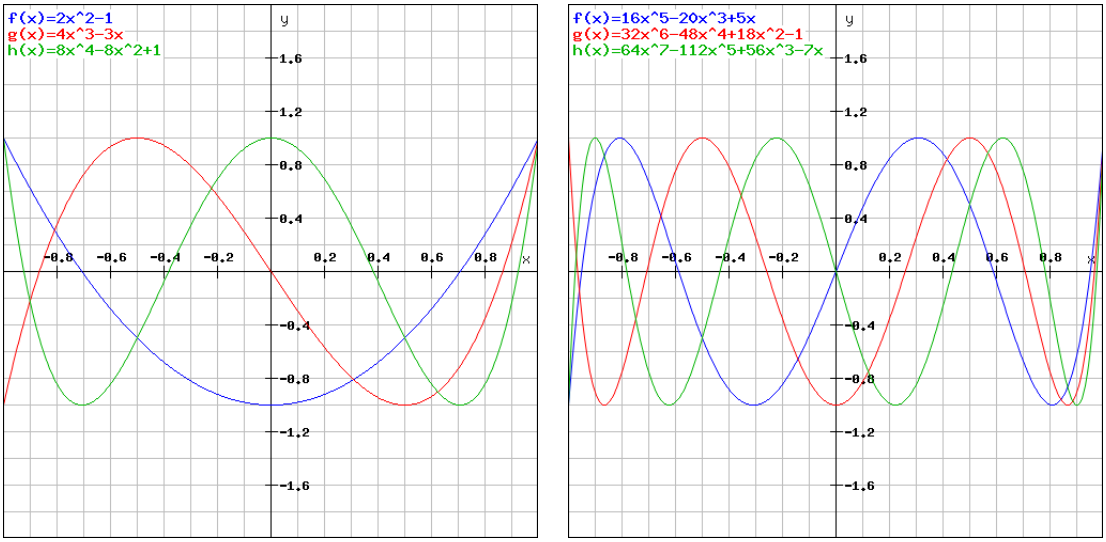


Figure 8.2: The Chebyshev polynomials of the first kind T_2 to T_7.

The monic polynomial $N_{n+1}(X)$ has the form

$$N_{n+1} = X^{n+1} - \phi$$

with $\phi \in \mathbb{R}[X]_n$. Wanted is the best L^∞-approximation $\hat{\phi} \in U = \mathbb{R}[X]_n$ to $f = X^{n+1}$. As U satisfies the Haar condition (Section 8.4), the best L^∞-approximation is unique. The following holds true:

Theorem 8.6.1. *On $[a, b] = [-1, 1]$, the best L^∞-approximation $\hat{\phi} \in \mathbb{R}[X]_n$ to $f = X^{n+1}$ is given by*

$$\hat{\phi} = X^{n+1} - 2^{-n} T_{n+1}(X)$$

where T_{n+1} is the $n+1$-th Chebyshev polynomial of the first kind. The zeros of T_{n+1} are the optimal places for Lagrange interpolation on $[-1, 1]$.

Chapter 9

Numerical Integration

Numerical Integration deals with the approximation of definite integrals:

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n} \alpha_i f(x_i)$$

with places $a \leq x_0 < \cdots < x_n \leq b$ and weights $\alpha_i \in \mathbb{R}$.

Example. *The* left Riemann sum

$$\int_a^b f(x)\,dx \approx \sum_{i=0}^{n-1} (x_{i+1} - x_i) f(x_i)$$

is a form of numerical integration. This is depicted in Figure 9.1.

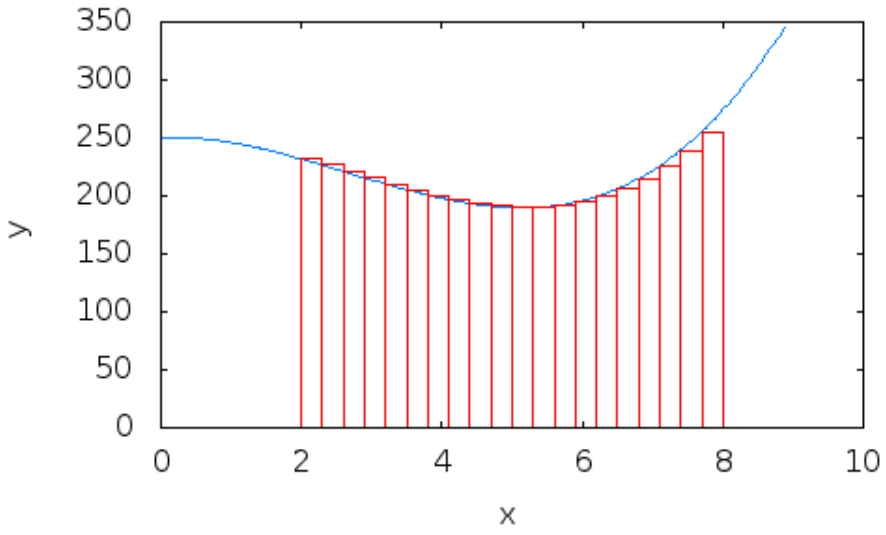


Figure 9.1: Left Riemann sum (source: Wikipedia, author: Mkwadee).

9.1 Interpolatoric Quadrature

9.1.1 Trapezoidal rule

In the *Trapezoidal rule*, the area under $y = f(x)$ from $x = 0$ to $x = h$ is approximated by a trapezium $ABCD$ (cf. Figure 9.2). Then

$$\int_a^b f(x)\,dx \approx h\frac{f(0)+f(h)}{2}$$

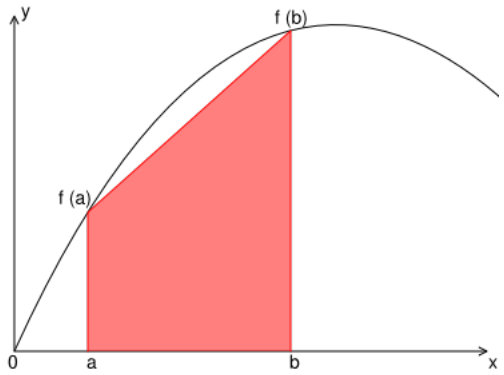


Figure 9.2: Trapezoidal rule (source: Wikipedia, author: Boris23).

The idea behind the trapezoidal rule is to interpolate f in 0 and h with a linear polynomial $\ell(x)$:

$$\int_a^b f(x)\,dx \approx \int_a^b \ell(x)\,dx$$

Here, the interpolant is

$$\ell(x) = f(0) + \frac{f(h)-f(0)}{h}x$$

It follows that

$$\int_a^b \ell(x)\,dx = f(0)x + \frac{1}{2}\frac{f(h)-f(0)}{h}x^2\Big|_0^h = h\frac{f(0)+f(h)}{2}$$

i.e. the trapezoidal rule.

The quadrature error is derived from the interpolation error

$$f(x) - \ell(x) = \frac{f''(\xi_x)}{2}x(x-h), \quad \xi_x \in [0,h]$$

(4.1.8). Hence,

$$\int_0^h f(x)\,dx - \int_0^h \ell(x)\,dx = \int_0^h (f(x)-\ell(x))\,dx = \frac{1}{2}\int_0^h f''(\xi_x)x(x-h)\,dx$$

$$\overset{\text{MVTDI}}{=} \frac{f''(\eta)}{2}\int_0^h x(x-h)\,dx = -\frac{f''(\eta)}{12}h^3, \quad \eta \in [0,h]$$

With MVTDI is meant the *Mean value theorem for definite integrals* gemeint:

Theorem 9.1.1 (Mean value theorem for definite integrals)**.** *Let $f\colon [a,b] \to \mathbb{R}$ be a continuous function, and $g\colon [a,b] \to \mathbb{R}$ integrable with either $g \geq 0$ or $g \leq 0$. Then there exists an $\eta \in [a,b]$, such that*

$$\int_a^b f(x)g(x)\,dx = f(\eta)\int_a^b g(x)\,dx$$

Notice that $x(x-h) \leq 0$ for $x \in [0,h]$.

9.1.2 Chained trapezoidal rule

If the interval is large, then the simple trapezoidal rule becomes imprecise by the quadrature error consideration of the previous section. A way out can be found through an equidistant subdivision of the interval $[a,b]$:

$$\begin{aligned} a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b \\ x_i = a + ih, \quad h = \frac{b-a}{n}, \quad i = 0,\ldots,n \end{aligned}$$

Now, apply the trapezoidal rule to each subinterval $[x_{i-1}, x_i]$:

$$\int_{x_{i-1}}^{x_i} f(x)\,dx \approx h\frac{f(x_{i-1}) + f(x_i)}{2}$$

This yields:

$$\begin{aligned} \int_a^b f(x)\,dx &\approx \sum_{i=1}^{n} h\frac{f(x_{i-1}) + f(x_i)}{2} \\ &= h\left(\frac{f(x_0)}{2} + f(x_1) + \cdots + f(x_{n-1}) + \frac{f(x_n)}{2}\right) =: \mathrm{ZT}_h(f) \end{aligned}$$

The quadrature error is now:

$$\sum_a^b f(x)\,dx - \mathrm{ZT}_h(f) = \sum_{i=1}^{n} -\frac{h^3}{12}f''(\eta_i) = -\frac{h^2}{12}\frac{b-a}{n}\sum_{i=1}^{n} f''(\eta_i)$$

with $\eta_i \in [x_{i-1}, x_i]$. The quantity $\frac{1}{n}\sum\limits_{i=1}^{n} f''(\eta_i)$ is the arithmetic mean of the values $f''(\eta_i)$. Hence, it lies between the largest and the smallest value. By the mean value theorem, it follows that there exists $\eta \in [a,b]$ with

$$f''(\eta) = \sum_{i=1}^{n} f''(\eta_i)$$

Hence,

$$\int_a^b f(x)\,dx - \mathrm{ZT}_h(f) = -\frac{(b-a)f''(\eta)}{12}h^2$$

Consequence. *Approximation by the chained trapezoid rule becomes arbitrarily precise by adding more places (i.e. by decreasing h). The quadrature error is quadratic in h.*

9.1.3 Newton-Cotes formulae

The trapezoidal rule interpolates with a linear polynomial. Now, we interpolate with a polynomial of degree at most n at places $x_0, \dots, x_n \in [a, b]$. Now, find weights $\alpha_0, \dots, \alpha_n \in \mathbb{R}$, such that polynomials $f \in \mathbb{R}[X]_n$ are integrated exactly, i.e.

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} f(x_i)\alpha_i, \quad \text{falls } f \in \mathbb{R}[X]_n$$

The solution is given by the Lagrange basis polynomials

$$\ell_i(X) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{X - x_j}{x_i - x_j}$$

Namely,

$$\alpha_i := \int_a^b \ell_i(x)\,dx$$

solves this task.

Proof. The quadrature is exact for $f \in \mathbb{R}[X]_n$: Namely, we have

$$f(X) = \sum_{i=0}^{n} f(x_i)\ell_i(X)$$

Hence,

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} f(x_i) \int_a^b \ell_i(x)\,dx = \sum_{i=0}^{n} f(x_i)\alpha_i$$

□

Closed Newton-Cotes formulae

Here, the places are chosen to be equidistant:

$$x_i = a + ih, \quad h = \frac{b-a}{n}, \quad i = 0, \dots, n$$

It follows that every $x \in [a, b]$ is of the form

$$x = a + th, \quad t \in [0, n]$$

The Lagrange basis polynomials ℓ_i can then be written as

$$\ell_i(X) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{X - x_j}{x_i - x_j} = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{a + th - a - jh}{a + ih - a - jh} = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{t - j}{i - j}$$

Hence, as $dx = h\,dt$:

$$\alpha_i = h \int_0^n \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{t - j}{i - j}\,dt, \quad i = 0, \dots, n$$

Example 9.1.2. *Let $n = 2$. Then $h = \frac{b-a}{2}$. We have:*

$$\alpha_0 = h \int_0^2 \frac{t-1}{0-1}\frac{t-2}{0-2}\,dt = \frac{h}{3}$$

$$\alpha_1 = h \int_0^2 \frac{t-0}{1-0}\frac{t-2}{1-2}\,dt = \frac{4}{3}h$$

$$\alpha_2 = h \int_0^2 \frac{t-0}{2-0}\frac{t-1}{2-1}\,dt = \frac{h}{3}$$

This yields the Simpson rule[1]

$$\int_a^b f(x)\,dx \approx \frac{h}{3}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$$

Some explicit closed Newton-Cotes formulae

Let $n = 1, 2, 3, 4$ and $h = \frac{b-a}{n}$. Then with $I = \int_a^b f(x)\,dx$:

$n = 1:$ $\quad I \approx \frac{b-a}{2}(f(a) + f(b))$ *Trapezoidal rule*

$n = 2:$ $\quad I \approx \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$ *Simpson rule*

$n = 3:$ $\quad I \approx \frac{b-a}{8}\left(f(a) + 3f(a+b) + 3f(b-h) + f(b)\right)$ *3/8-rule*

$n = 4:$ $\quad I \approx \frac{b-a}{90}\left(7f(a) + 32f(a+h) + 12f\left(\frac{a+b}{2}\right) + 32f(b-h) + 7f(b)\right)$ *Boole rule*

Example 9.1.3. *We approximate $I = \int_0^1 \frac{dx}{1+x^2}$ with the first 4 closed Newton-Cotes formulae.*

1. *Trapezoidal rule: $I \approx 0.75000$.*
2. *Simpson rule: $I \approx 0.78333$.*
3. *3/8-rule: $I \approx 0.78462$.*
4. *Boole rule: $I \approx 0.78553$.*

Error of Simpson rule

The quadrature error of the Simpson rule is

$$\int_a^b f(x)\,dx - \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right) = \frac{(b-a)^5}{2880}f^{(4)}(\eta), \quad \eta \in [a, b]$$

[1] this time named after Thomas Simpson (1710–1761)

9.2 Gauß Quadrature

First, we define

$$\int f := \int_a^b f(x)w(x)\,dx$$

with fixed positive, continuous weight function $w\colon [a,b] \to \mathbb{R}$, and observe that $\int$ is linear:

$$\int \alpha f = \alpha \int f, \quad \alpha \in \mathbb{R}$$
$$\int (f+g) = \int f + \int g, \quad f, g \in C[a,b]$$

We remind that $\int fg$ defines an inner product on $C[a,b]$ (cf. Section 8.3).

Zeros of orthogonal polynomials

For the Gauß quadrature, we will use the zeros of orthogonal polynomials. We have:

Theorem 9.2.1. *Let $p_0, p_1, p_2, \ldots$ be a sequence of orthogonal polynomials in $C[a,b]$ with $\deg p_i = i$. Then their zeros are simple, real and lie in the interval $[a,b]$.*

Proof. Let $x_0, \ldots, x_k$ be the distinct zeros of p_{n+1} inside the interval $[a,b]$. If $k = n$, then the assertion is proven. If, however, $k < n$, then let

$$q(X) := (X - x_0) \cdot (X - x_k)$$

We have $\deg q = k + 1 < n + 1$. Hence,

$$\int p_{n+1} q = 0 \tag{9.1}$$

But $p_{n+1}q$ has no sign change in $[a,b]$, as every zero appears with even multiplicity. Hence,

$$\int p_{n+1} q \neq 0$$

in contradiction to (9.1). □

Indeed, the zeros of orthogonal polynomials are well-suited as places for integration. Let $x_0, \ldots, x_n \in [a,b]$ be the zeros of p_{n+1}, where $p_0, p_1, \ldots$ is a sequence of orthogonal polynomials in $C[a,b]$ mit $\deg p_i = i$. We set

$$\alpha_i := \int \ell_i, \quad i = 1, \ldots, n$$

where

$$\ell_i(X) := \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{X - x_j}{x_j - x_j}$$

is the i-th Lagrange basis polynomial. The *Gauß quadrature formula* is

$$G_n f = \sum_{i=0}^{n} \alpha_i f(x_i)$$

We have:

Theorem 9.2.2. *If $f \in \mathbb{R}[X]_{2n+1}$, then $G_n f$ is exact:*

$$G_n f = \int f$$

Proof. $G_n f$ is exact for polynomials of degree $\leq n$ by Section 9.1.3. Let f be a polynomial with $\deg f \leq 2n+1$. We have: $\deg p_{n+1} = n+1$. Hene, by division with remainder:

$$f = p_{n+1} q + r, \quad \deg q, \deg r \leq n$$

Hence,

$$\begin{aligned} G_n f = \sum_{i=1}^{n} \alpha_i f(x_i) = \sum_{i=1}^{n} \alpha_i (\underbrace{p_{n+1}(x_i)}_{=0} q(x_i) + r(x_i)) = \sum_{i=1}^{n} \alpha_i r(x_i) = G_n r \\ \overset{\deg r \leq n}{=} \int r \overset{(*)}{=} \int (p_{n+1} q + r) = \int f \end{aligned}$$

where $(*)$ holds true, as p_{n+1} is orthogonal to q, because $\deg q \leq n$. □

Consequence. *The weights α_i are all positive and $\leq \int 1$.*

Proof. 1. We have:

$$0 < \int \ell_i^2 \overset{(*)}{=} G_n \ell_i^2 = \sum_j \alpha_j \underbrace{\ell_i^2(x_j)}_{=\delta_{ij}} = \alpha_i$$

where $(*)$ holds true because $\deg \ell_i^2 \leq 2n+1$.

2. We have:

$$\sum \alpha_i 1 = G_n 1 = \int 1$$

If a sum of positive real numbers is at most $\int 1$, then all summands are at most $\int 1$. □

Gauß quadrature error

We have

$$\int f - G_n f = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int p_{n+1}^2, \quad \xi \in [a,b]$$

Convergence

For $f \in C[a,b]$ we have:

$$\lim_{n \to \infty} G_n f = \int f$$

Example 9.2.3.

1. *For $[a,b] = [-1,1]$ and $w = 1$ one speaks of the* Gauß-Legendre quadrature.
2. *For the interval $[0,\infty)$ and $w(x) = e^{-x}$ it is the* Gauß-Laguerre quadrature.
3. *For $(-\infty,\infty)$ and $w(x) = e^{-x^2}$ it is the* Gauß-Hermite quadrature.

Gauß-Legendre formulae

In the Gauß-Legendre quadrature, the zeros of the *Legendre polynomials* $P_0, P_1, \ldots$ are used. They satisfy the orthogonality relations

$$\int_{-1}^{1} P_n(x)P_m(x)\,dx = \frac{2}{2n+1}\delta_{m,n}$$

and the recursion formula

$$(n+1)P_{n+1}(X) + nP_{n-1}(X) = (2n+1)XP_n(X)$$

The first three Legendre polynomials are

$$P_0 = 1,\ P_1 = X,\ P_2 = \frac{3}{2}X^2 - \frac{1}{2}$$

Table 9.1 gives the nodes and weights of the Gauß-Legendre quadrature for $n \leq 5$.

n	node x_i	weight α_i
1	0	2
2	$\pm 1/\sqrt{3}$	1
3	0	$8/9$
	$\pm\sqrt{3/5}$	$5/9$
4	$\pm\sqrt{(3-2\sqrt{6/5})/7}$	$\frac{18+\sqrt{30}}{36}$
	$\pm\sqrt{(3+2\sqrt{6/5})/7}$	$\frac{18-\sqrt{30}}{36}$
5	0	$128/225$
	$\pm\frac{1}{3}\sqrt{5-2\sqrt{10/7}}$	$\frac{322+13\sqrt{70}}{900}$
	$\pm\frac{1}{3}\sqrt{5+2\sqrt{10/7}}$	$\frac{322-13\sqrt{70}}{900}$

Table 9.1: Nodes and weights of the Gauß-Legendre quadrature.

9.3 Interval transformation

Let weights and nodes of a quadrature formula on the interval $[-1, 1]$ be given:

$$\int_{-1}^{1} g(x)\,dx \approx \sum_{i=0}^{n} \alpha_i g(x_i)$$

Suppose that these should be used in order to approximate the integral

$$\int_{a}^{b} f(t)\,dt$$

The the following transformation can be used:

$$t\colon [-1,1] \to [a,b], \quad x \mapsto \frac{b-a}{2}(x+1) + a$$

This is affine-linear and maps $[-1,1]$ bijectively to $[a,b]$. With this substitution, we obtain

$$\int_a^b f(t)\,dt = \int_{-1}^{1} f(t(x)) \underbrace{\frac{b-a}{2}\,dx}_{=dt} = \frac{b-a}{2}\int_{-1}^{1} g(x)\,dx$$

with $g(x) = f(t(x))$. This yields the following quadrature formula for f:

$$\int_a^b f(t)\,dt \approx \frac{b-a}{2}\sum_{i=1}^{n} \alpha_i f(t_i), \quad t_i = t(x_i)$$

The new nodes are $t(x_i)$, and the new weights are

$$\frac{b-a}{2}\alpha_i$$

due to the transformation $t\colon [-1,1] \to [a,b]$.

Example 9.3.1. *We approximate*

$$I = \int_0^1 \frac{dx}{1+x^2}$$

by using the Gauss-Legendre quadrature formulae. We take the transformation

$$t\colon [-1,1] \to [0,1], \quad x \mapsto \frac{1}{2}(x+1)$$

We have $\frac{b-a}{2} = \frac{1}{2}$. *By using table 9.1 we obtain:*

$n = 1$. $t_1 = t(0) = \frac{1}{2}$, $\alpha_1 = \frac{1}{2}\cdot 2 = 1$.

$$I \approx f\left(\frac{1}{2}\right) = 0.8000$$

$n = 2$.

$$\begin{aligned} t_1 &= t\left(1/\sqrt{3}\right) \approx 0.7887, \quad \alpha_1 = \frac{1}{2}\cdot 1 = \frac{1}{2} \\ t_2 &= t\left(-1/\sqrt{3}\right) \approx 0.2113, \quad \alpha_2 = \frac{1}{2} \\ I &\approx \frac{1}{2}f(0.7887) + \frac{1}{2}f(0.2113) \approx 0.7869 \end{aligned}$$

$n = 3$.

$$\begin{aligned} t_1 &= t(0) = \frac{1}{2}, \quad \alpha_1 = \frac{1}{2}\cdot\frac{8}{9} = \frac{4}{9} \\ t_2 &= t\left(\sqrt{3/5}\right) \approx 0.8873, \quad \alpha_2 = \frac{1}{2}\cdot\frac{5}{9} = \frac{5}{18} \\ t_3 &= t\left(-\sqrt{3/5}\right) \approx 0.1127, \quad \alpha_3 = \frac{5}{18} \\ I &\approx \frac{4}{9}f\left(\frac{1}{2}\right) + \frac{5}{18}\left(f(0.8873) + f(0.1127)\right) \approx 0.7853 \end{aligned}$$

9.4 Romberg Integration

Let T_k^0 be the chained trapezoid rule for $n = 2^k$ equidistant subintervals of $[a, b]$:

$$T_k^0 = \mathrm{ZT}_{2^k}(f), \quad k = 0, 1, 2, \ldots$$

with ZT_n as in Section 9.1.2. Then one constructs higher difference quotients:

$$T_k^i = \frac{4^i T_k^{i-1} - T_{k-1}^{i-1}}{4^i - 1}, \quad i = 1, 2, \ldots, k$$

In the tableau

$$\begin{array}{ccccccc}
T_0^0 & & & & & & \\
 & \searrow & & & & & \\
T_1^0 & \to & T_1^1 & & & & \\
 & \searrow & & \searrow & & & \\
T_2^0 & \to & T_2^1 & \to & T_2^2 & & \\
 & \searrow & & \searrow & & \searrow & \\
T_3^0 & \to & T_3^1 & \to & T_3^2 & \to & T_3^3 \\
\ldots & & & & & &
\end{array}$$

each element depends only of its left and upper left neighbours.

Remark 9.4.1. *The error for T_k^n is $O(h_k^{2i+2})$, where $h_k = \frac{b-a}{2^k}$.*

Example 9.4.2. *We compute*

$$\ln 2 = \int_1^2 \frac{dx}{x}$$

with the Romberg method. First:

$$\begin{aligned}
T_0^0 &= \frac{1}{2}\left(1 + \frac{1}{2}\right) = 0.75 \\
T_1^0 &= \frac{1}{2}\left(1 + \frac{1}{2}\cdot\frac{2}{3} + \frac{1}{2}\right) = 0.708333333 \\
T_2^0 &= \frac{1}{4}\left(1 + \frac{1}{2}\cdot\frac{4}{5} + \frac{1}{2}\cdot\frac{2}{3} + \frac{1}{2}\cdot\frac{4}{7} + \frac{1}{2}\right) = 0.69702380952
\end{aligned}$$

Then:

$$\begin{aligned}
T_1^1 &= \frac{4T_1^0 - T_0^0}{3} = 0.694444 \\
T_2^1 &= \frac{4T_3^0 - T_2^0}{3} = 0.693253, \quad T_2^2 = \frac{16T_2^1 - T_1^1}{15} = 0.69317460
\end{aligned}$$

Comparison with $\ln 2 \approx 0.69214718$ shows that T_2^2 already has 4 correct digits after the decimal point.