Continued Fractions

Tomas Sauer

Lehrstuhl für Mathematik mit Schwerpunkt Digitale Bildverarbeitung
FORWISS
University of Passau
Innstr. 43
94032 Passau

Version 1.0
26.3.2019

Die altmodische Tafelvorlesung in der Mathematik ist in der übrigen Wissenschaftslandschaft tatsächlich zur Ausnahmeerscheinung geworden, genießt aber gleichzeitig Kult-Status und wäre durchaus eine würdige Anwärterin für das immaterielle Kulturerbe der UNESCO, ähnlich wie die deutsche Brotkultur oder das mongolische Knöchelschießen.

G. Zöller, *Forschung und Lehre Online, 2.4.2015*

Nothing spoils numbers faster than a lot of arithmetic.

Peppermint Patty, *The Peanuts, 4.12.1968*

Of course she was aware, cognitively, that there was a life outside universities, but she knew nothing about it,

D. Lodge, *Nice Work*

To isolate mathematics from the practical demands of the sciences is to invite the sterility of a cow shut away from the bulls.

P. Chebyshev

…you get to have such a high regard for the truth you can't put courtesy first. You want to, but you haven't the heart.

E. D. Biggers, *Charlie Chan …*

Inhaltsverzeichnis

Continued fractions and what can be done with them | 1

That's the reason they're called lessons,[...] because they lessen from day to day.

(L. Carroll, *Alice's adventures in wonderland*)

The goal of this section is to just provide a coarse overview on continued fractions and to get an idea what are the objects to be considered in this lecture and what can be said about them. It is designed for motivational purposes and not a systematic or structured introduction.

1.1 The first definition

A CONTINUED FRACTION is a fraction, i.e., a ration of integers, whose DENOMINATOR is written as a continued fraction again. This informal version is, however, a somewhat self-referential and recursive definition so that we better give a formal definition immediately.

Definition 1.1.1. For integers[1] $a_0, \dots, a_n \in \mathbb{Z}$ the associated CONTINUED FRACTION is the rational number

$$[a_0; a_1, \dots, a_n] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_{n-1} + \cfrac{1}{a_n}}}}} \tag{1.1.1}$$

Still, this „dot notation“ is neither exact nor gives it rise to a really well–defined object. We just have to look at the cases that $a_n = 0$ or $a_{n-1} = -1/a_n$. In both cases we would divide by zero which is not really welcome in mathematics. A simple recursive definition of the continued fraction results from a closer inspection of (1.1.1) that reveals that the denominator of the „big“ fraction there, represennting the continued fraction, is a continued fraction again, namely $[a_1; a_2, \dots, a_n]$. This way we obtain the recursive definition

$$[a_0; a_1] = a_0 + \frac{1}{a_1}, \qquad [a_0; a_1, \dots, a_n] = a_0 + \frac{1}{[a_1; a_2, \dots, a_n]}, \quad n \in \mathbb{N}. \tag{1.1.2}$$

This definition already shows us what would happen in the degenerate cases mentioned above: if, for example $[a_k; a_{k+1}, \dots, a_n] = 0$, then we have[2]

$$\begin{aligned} [a_{k-1}; a_k, \dots, a_n] &= a_{k-1} + \frac{1}{[a_k; a_{k+1}, \dots, a_n]} = \infty \\ [a_{k-2}; a_{k-1}, \dots, a_n] &= a_{k-2} + \frac{1}{[a_{k-1}; a_k, \dots, a_n]} = a_{k-2}, \end{aligned}$$

[1] They could indeed be signed but normally the sign would only lead to ambiguties.

[2] This is imprecisely written, formally incorrect and for illustrational purposes only. So please do not refer to it.

and as long as we do not have the additional degeneracy $a_{k-2} = 0$ everything proceeds quite normal. Hence, division by zero is not such a sacrilege in continued fractions, at least as long it does not happen too often. Nevertheless, it is even better to avoid all the trouble by choosing[3] $a_0 \in \mathbb{Z}$ and $a_j \in \mathbb{N}$.

On the other hand, we do not have to restrict continued fractions to INTEGER COEFFICIENTS, we could, in the same fashion, even define rational continued fractions of the form $[r_0; r_1, \dots, r_n]$ with[4] $r_j \in \mathbb{Q} \setminus \{0\}$. A simple and immediate formula is

$$\begin{aligned}
[a_0; a_1, ..., a_k, \dots, a_n] &= a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots + \cfrac{1}{\boxed{a_k + \cfrac{1}{\ddots\, a_{n-1} + \cfrac{1}{a_n}}}}}} \\
&= a_0 + \cfrac{1}{a_1 + \cfrac{1}{\ddots\, a_{k-1} + \cfrac{1}{[a_k; a_{k+1}, \dots, a_n]}}} \\
&= [a_0; a_1, \dots, a_{k-1}, [a_k; a_{k+1}, \dots, a_n]] \\
&= [a_0; a_1, \dots, a_{k-1}, r_k]
\end{aligned}$$

making use of the REMAINDER $r_k := [a_k; a_{k+1}, \dots, a_n]$. As long as $a_j \in \mathbb{Z} \setminus \{0\}$ or even $r_j \in \mathbb{Q} \setminus \{0\}$, the continued fraction is a rational number which is quite obvious and can be shown by simple induction over the number of parameters in the formula (1.1.2). All that is needed is the fact that rational numbers form a FIELD and thus are closed under addition and reciprocals.

Every *finite* sequence $a_0, \dots, a_n$ of numbers is the initial sequence of an *infinite* sequence $a = (a_j : j \in \mathbb{N}_0)$ which also enables us to consider INFINITE CONTINUED FRACTIONS of the form

$$[a_0; a_1, \dots] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cdots}}.$$

This is nice as a formal expression but what is the value of such an infinite object? In principle this is clear: it is the LIMIT of the continued fractions associated to the finite initial segments, i.e.,

$$[a_0; a_1, \dots] = \lim_{n \to \infty} [a_0; a_1, \dots, a_n],$$

but we do not really have an idea yet when such a sequence really has a limit, that is, when such an infinite continuouse fraction converges. We will late prove a criterion for that which is not only simple but also very handy and elegant. And it even works for continued fractions with rational coefficients.

Theorem 1.1.2 (Convergence criterion for continued fractions). *For $r_j \in \mathbb{Q}$, $r_j > 0$, $j \in \mathbb{N}$, the coninued fraction $[r_0; r_1, \dots]$ is CONVERGENT if and only if*

$$\sum_{j=0}^{\infty} r_j = \infty.$$

[3]Based on the notation $\mathbb{N} := \{1, 2, \dots\}$ and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ that we will use here.

[4]The notation with „r“ instead of „a“ is supposed to express exactly that.

This is true in the special case that $r_j = a_j \in \mathbb{N}$.

We already see that these infinite continued fractions will be particularly tame if $a_1, a_2, \ldots$ are chosen as POSITIVE INTEGERS. Since in this case the continued fraction $[a_1; a_2, \ldots]$ is positive, we allow $a_0 \in \mathbb{Z}$ to be capable of representing negative numbers as well. And indeed, this approach gives us „everything“.

Theorem 1.1.3. *Any REAL NUMBER $x \in \mathbb{R}$ can be written as a continued fraction $[a_0; a_1, \ldots]$ with $a_0 \in \mathbb{Z}$ and $a_j \in \mathbb{N}_0$, $j \in \mathbb{N}$, and this continued fraction is FINITE if and only if x is a RATIONAL NUMBER.*

Moreover, we will find out that the continued fraction expansion of a real number is *unique*, except a little bit of an ambiguity for rational numbers.

Remark 1.1.4. In some sense, this result is even more elegant than the decimal expansion of a real number where the the number is rational if and only if the expansion is FINITE or PERIODIC.

According to [29, S. 359], who in turn refers to [1], the ancient Greek mathematicians used, after the discovery of IRRATIONAL NUMBERS[5], continued fractions for a first definition of a concept resembling real numbers. They did not use normal fractions and in particular not infinite decimal expansions.

And this was a really good choice since we will see that with the same effort, measured in number of used digits, continued fractions are giving a much better approximation to an irrational number than fractions with just nominator and denominator or decimal expansions. What we will do there will be APPROXIMATION of real numbers by rational numbers. It will turn out tha the BEST APPROXIMATION from $\mathbb{Q}$ to an irrational number among all rational numbers with a certain maximal denominator is a continued fraction.

This theory has the nice side effect that it will tell us that the real number that is hardest to approximate by rational numbers is the GOLDEN RATIO $\frac{1+\sqrt{5}}{2}$; the poor approximation wil be a consequence of its particularly simple representation $[1; 1, 1, \ldots]$ as a continued fraction.

1.2 Continued fractions of polynomials

Continued fractions can be built from various objects. We already saw that for $\mathbb{Z} \setminus \{0\}$ and $\mathbb{Q}$, but it will turn out that most of the concept works whenever we can add and multiply objects, that, is over any RING[6], but the for existence of continued fractions euclidean rings will be preferrable[7]: division with remainder will play a crucial role when obtaining a continued fraction representation. Therefore, we will also consider continued fractions of (univariate) polynomials which are expressions of the form

$$[p_0; p_1, \ldots, p_n], \qquad p_j \in \Pi = \mathbb{K}[x]$$

[5]Those who do not yet know the story about the Pythagoreans, their somewhat religious and rational view of the world and harmony, are recommended to find out about it. As popularized science this can be found, for example, in [24, 42].

[6]Not surprising, a RING is a structure where addition and multiplication are well–defined and interact properly, i.e., according to the DISTRIBUTIVE LAW.

[7]A EUCLIDEAN RING is a ring equipped with a well–defined EUCLIDEAN DIVISION, i.e., division with remainder

for a suitable field $\mathbb{K}$. Such a finite continuoued fraction will then be a RATIONAL FUNCTION

$$[p_0(x); p_1(x), \dots, p_n(x)] = \frac{f(x)}{g(x)}, \qquad f, g \in \Pi. \tag{1.2.1}$$

and their limit objects will be even more special.

Normally each of the p_j is an affine[8] or constant polynomial, in other words, a polynomial of degree at most 1. Also in this case we will have some form of APPROXIMATION THEORY trying to approximate a given function[9], represented by a power series[10] is some best possible way by a rational object which will be the continued fraction. Here „best possible" means that as many terms as possible coincide in the series and the approximation.

Continued fraction with especially simple coefficients in (1.2.1) are those where each p_j is an AFFINE POLYNOMIAL of the form $p_j(x) = \alpha_j x + \beta_j$. These continued fractions will have a close relationship with ORTHOGONAL POLYNOMIALS, polynomial sequences $f_j \in \Pi$, $j \in \mathbb{N}_0$, with the property that

$$\left\langle f_j, f_k \right\rangle = c_j \, \delta_{j,k}, \qquad c_j > 0, \qquad j, k \in \mathbb{N}_0,$$

where $\langle \cdot, \cdot \rangle$ denotes a formal INNER PRODUCT[11]. In fact, orthogonal polynomials can even be characterized and parameterized by means of continued fractions. A result in this direction is as follows.

Theorem 1.2.1. *For each sequence f_j, $j \in \mathbb{N}_0$, of orthogonal polynomials there exist coefficients $\alpha_j < 0$ und β_j, $j \in \mathbb{N}_0$, such that*

$$\left[0; \alpha_1 x + \beta_1, \dots, \alpha_j x + \beta_j\right] = \frac{g_j(x)}{f_j(x)},$$

and vice versa.

Eventually, this theory will even allow us to construct orthogonal polynomials and even quadrature formulas using continued fractions. This actually was the way how Gauß originally constructed what is nowadays known as a GAUSSIAN QUADRATURE FORMULA. In this lecture we will revisit and, hopefully, finally understand this historical approach from [13] and the quite natural idea behind it. The approach relies on the fact that the componentwise limit function of continued functions for an INTEGRAL, that is, an inner product with the property that

$$\langle f, g \rangle = \int_{\mathbb{R}} f(x) \, g(x) \, w(x) \, dx, \qquad w \geq 0,$$

with, for convenience, a compactly supported continuous WEIGHT FUNCTION $w : \mathbb{R} \to \mathbb{R}$, can be described in a rather simple way: it is the Laurent series or GENERATING FUNCTION for the MOMENT SEQUENCE

$$f(x) = \sum_{j=1}^{\infty} \mu_{j-1} \, x^{-j}, \qquad \mu_j = \int_{\mathbb{R}} x^j \, w(x) \, dx, \quad j \in \mathbb{N}_0. \tag{1.2.2}$$

This also connects continued fractions to the classical MOMENT PROBLEM:

> when is a sequence $\mu = \left(\mu_j : j \in \mathbb{N}_0\right)$ a moment sequence with respect to a (nonnegative) weight function in the sense of (1.2.2)?

[8] There is a destinction between an AFFINE POLYNOMIAL of the form $\alpha \, x + \beta$ and a LINEAR POLYNOMIAL of the form $\alpha \, x$ which is sometimes respected and sometimes ignored.

[9] The counterpiece of the real number above

[10] To be precise, a LAURENT SERIES, i.e., a power series in z^{-1}.

[11] Little exercise: recall the axiomatic definitions of an inner product.

1.3 Digital signal processing

We will also be interested in another, more modern, type of problem where, surprising or not, continued fractions play a crucial role. This is DIGITAL SIGNAL PROCESSING, more precisely the construction of digital filters. For the moment it shall suffice that a DIGITAL FILTER can be identified with a RATIONAL FUNCTION[12]

$$f = \frac{p}{q}, \qquad p, q \in \mathbb{C}[z], \tag{1.3.1}$$

and that it can be efficiently realized in a reasonable way as long as the rational function has no poles in the unit circle. This notion, which is clearly equivalent to the fact that the denominator has no zeros in the unit circle is known as STABILITY of a rational filter and we see why this name makes sense. In other words: to be reasonable, a rational filter must not have poles in the unit circle and thus q no zeros there. With the rational linear transformation $z = \frac{w+1}{w-1}$ this is equivalent to the requirement that $q(w)$ has all its zeros in the left half plane which makes it a so–called HURWITZ POLYNOMIAL. In STIELTJES' THEOREM we will characterize Hurwitz polynomials and hence stable rational filter by means of continued fractions, more precisely by continued fractions of the form

$$[c_0; d_1\, x, c_1, \dots, d_n\, x, c_m],$$

where scalar and linear polynomials alternate. Togehter, numerator and denominator of the respective linear function yield, when mixed properly, a Hurwitz polynomial and, conversely, any Hurwitz polynomial can be decomposed in this way.

1.4 And what else?

Of course, the issues presented in this lecture are only partial aspects of the theory of continued fractions. For example, one can find in [28] some measure theory of continued fractions: how are they distributed on the real line. And the two volumes of Perron's book [36, 37] contains a lot that is not even mentioned here, for example the question under which conditions a continued fraction, seen as a power series, converges to an analytic function. But instead of crying over what we are not going to do, let's simply start and see where we get.

[12] We really need a COMPLEX POLYNOMIAL in numerator and denominator, even if the coefficients will be mostly real.

Continued fractions of real numbers | 2

And now I must stop saying what I am not writing about, because there's nothing so special about that; every story one chooses to tell is a kind of censorship, it prevents the telling of other tales ...

(S. Rushdie, *Shame*)

In this chapter we consider the approximation of real numbers by continued fractions whose coefficients are nonnegative numbers[1]. Most of the material here is following the way how it is done in the book of Khinchin [28], since it can hardly be done better.

2.1 Convergents and continuants

Our first step in the direction of understanding continued fractions consists of having a closer look at the expression $[a_0; a_1, \ldots, a_n]$ and its meaning. This leads us to the most fundamental notion in the theory of continued fractions, which is still well–defined even for *rational* coefficients of the continued fraction.

Definition 2.1.1. Given numbers $a_j \in \mathbb{Q}$, $j = 0, 1, \ldots$, the nth CONVERGENT of the infinite continued fraction $[a_0; a_1, \ldots]$ is defined as the finite continued fraction $[a_0; a_1, \ldots, a_n]$.

First note that the nth convergent of a continued fration can always be written as the quotient of two polynomials in the variables $a_0, \ldots, a_n$:

$$[a_0; a_1, \ldots, a_n] = \frac{p_n(a_0, \ldots, a_n)}{q_n(a_0, \ldots, a_n)} \tag{2.1.1}$$

This is trivally true for $n = 0$, as we then only have the constant polynomial r_0, and follows inductively from the definition (1.1.2):

$$\begin{aligned} [a_0; a_1, \ldots, a_{n+1}] &= a_0 + \frac{1}{[a_1; a_2, \ldots, a_{n+1}]} = a_0 + \frac{q_n(a_1, \ldots, a_{n+1})}{p_n(a_1, \ldots, a_{n+1})} \\ &= \frac{a_0\, p_n(a_1, \ldots, a_{n+1}) + q_n(a_1, \ldots, a_{n+1})}{p_n(a_1, \ldots, a_{n-1})} \end{aligned}$$

which immediatly gives a recursive way to obtain p_{n+1} and q_{n+1} as

$$\begin{aligned} p_{n+1}(a_0, \ldots, a_{n+1}) &= a_0\, p_n(a_1, \ldots, a_{n+1}) + q_n(a_1, \ldots, a_{n+1}), \\ q_{n+1}(a_0, \ldots, a_{n+1}) &= p_n(a_1, \ldots, a_{n+1}). \end{aligned} \tag{2.1.2}$$

[1] „Natural numbers“ would be the literal translation from German.

Since $[a_0; a_1, \dots, a_{n+1}] = \dfrac{p_{n+1}(a_0, \dots, a_{n+1})}{q_{n+1}(a_0, \dots, a_{n+1})}$, the second identity in (2.1.2) yields that

$$q_n(a_0, \dots, a_n) = p_{n-1}(a_1, \dots, a_n) \tag{2.1.3}$$

and allows us to conclude that

$$[a_0; a_1, \dots, a_n] = \frac{p_n(a_0, \dots, a_n)}{p_{n-1}(a_1, \dots, a_n)}. \tag{2.1.4}$$

Moreover, from the first identity of (2.1.2) we have the recurrence relation

$$p_{n+1}(a_0, \dots, a_{n+1}) = a_0\, p_n(a_1, \dots, a_{n+1}) + p_{n-1}(a_2, \dots, a_{n+1}), \qquad p_{-2} := 0,\ p_{-1} := 1, \tag{2.1.5}$$

for the numerator as well.

Definition 2.1.2. The polynomials $p_n(x_0, \dots, x_n) : \mathbb{Q}^{n+1} \to \mathbb{Q}$ are called CONTINUANTS.

Remark 2.1.3. Continuants have been considered, if not introduced, already by Euler.

Let us consider some first examples:

$$\begin{aligned} [a_0;] &= a_0 \\ [a_0; a_1] &= a_0 + \frac{1}{a_1} = \frac{a_0 a_1 + 1}{a_1} \\ [a_0; a_1, a_2] &= a_0 + \frac{1}{[a_1; a_2]} = a_0 + \frac{a_2}{a_1 a_2 + 1} = \frac{a_0 a_1 a_2 + a_0 + a_2}{a_1 a_2 + 1}, \end{aligned}$$

which all looks nicely symmetric in the variables.

Exercise 2.1.1 Prove the symmetry property

$$p_n(x_0, \dots, x_n) = p_n(x_n, \dots, x_0)$$

of continuants, see [29, S. 357]. $\diamond$

The next result is an explicit recurrence relation for the numerator and the denominator of the convergent.

Theorem 2.1.4. *For $k \geq 1$ the kth convergent can written with numerator and denominator satisfying the following* RECURRENCE RELATION[2]*:*

$$\begin{aligned} p_k &= a_k p_{k-1} + p_{k-2} \\ q_k &= a_k q_{k-1} + q_{k-2} \end{aligned}, \qquad \begin{aligned} p_{-1} &= 1, \\ q_{-1} &= 0, \end{aligned} \qquad \begin{aligned} p_0 &= a_0 \\ q_0 &= 1. \end{aligned} \tag{2.1.6}$$

Proof: The case $k = 1$ has been computed explicitly in the above examples. To advance the induction hypothesis from k to $k+1$, we use the canonical representation

$$[a_1; a_2, \dots, a_{k+1}] =: \frac{\widetilde{p}_k}{\widetilde{q}_k}$$

of the „shifted“ continued fraction and obtain by definition of continued fractions that

$$\frac{p_{k+1}}{q_{k+1}} = a_0 + \frac{1}{[a_1; a_2, \dots, a_{k+1}]} = a_0 + \frac{\widetilde{q}_k}{\widetilde{p}_k} = \frac{\widetilde{p}_k\, a_0 + \widetilde{q}_k}{\widetilde{p}_k}.$$

[2] A déjà-vu for everyone who already encountered orthogonal polynomials. From that point of view it is no surprise that we will encounter them later.

Using the induction hypothesis (2.1.6) for $\widetilde{p}_k$ und $\widetilde{q}'_k$ and taking into account the shift of the indices there, we get that we can choose p_{k+1} and q_{k+1} as

$$\begin{aligned} p_{k+1} &= a_0 \left(a_{k+1} \widetilde{p}_{k-1} + \widetilde{p}_{k-2} \right) + \left(a_{k+1} \widetilde{q}_{k-1} + \widetilde{q}_{k-2} \right) \\ &= a_{k+1} \left(a_0 \widetilde{p}_{k-1} + \widetilde{q}_{k-1} \right) + \left(a_0 \widetilde{p}_{k-2} + \widetilde{q}_{k-2} \right) = a_{k+1} p_k + p_{k-1}, \\ q_{k+1} &= a_{k+1} \widetilde{p}_{k-1} + \widetilde{p}_{k-2} = a_{k+1} q_k + q_{k-1}, \end{aligned}$$

which completes the induction. □

It is well known that the representation of a fraction as a quotient of integers is not unique, $\frac{1}{2} = \frac{2}{4} = \frac{3}{6} = \dots$ and only the normal form with coprime numerator and denominator is unique. The same holds true for the representation of a convergent which we make unique by means of the above recurrence. It will turn out later that under additional assumptions this representation with integer parameters is even IRREDUCIBLE, but at the moment, we take it as it is and use the following definition.

Definition 2.1.5. The values defined in (2.1.6) are called the numerator and denominator in the CANONICAL REPRESENTATION of the kth convergent

$$[a_0; a_1, \dots, a_k] = \frac{p_k}{q_k}$$

of a continued fraction with arguments $a_j \in \mathbb{Q}$, $j \in \mathbb{N}_0$.

Corollary 2.1.6. *For $k \geq 0$ we have that*

$$q_k \, p_{k-1} - p_k \, q_{k-1} = (-1)^k, \tag{2.1.7}$$

or

$$\frac{p_{k-1}}{q_{k-1}} - \frac{p_k}{q_k} = \frac{(-1)^k}{q_{k-1} \, q_k}, \tag{2.1.8}$$

respectively.

Proof: We multiply the first line of the recurrence (2.1.6) by $-q_{k-1}$ and the second by p_{k-1} to get

$$\begin{aligned} q_k \, p_{k-1} - p_k \, q_{k-1} &= -a_k p_{k-1} q_{k-1} - q_{k-1} \, p_{k-2} + a_k p_{k-1} q_{k-1} + q_{k-2} p_{k-1} \\ &= -\left(q_{k-1} \, p_{k-2} - q_{k-2} p_{k-1} \right) = \cdots = (-1)^k \left(q_0 \, p_{-1} - q_{-1} \, p_0 \right) = (-1)^k, \end{aligned}$$

which is (2.1.7). If we divide that by $q_{k-1} \, q_k$, we end up with (2.1.8). □

And there is one more cute formula.

Theorem 2.1.7. *For $k \geq 2$ one has*

$$p_k \, q_{k-2} - q_k \, p_{k-2} = (-1)^k \, a_k \qquad \text{or} \qquad \frac{p_k}{q_k} - \frac{p_{k-2}}{q_{k-2}} = \frac{(-1)^k \, a_k}{q_{k-2} \, q_k}, \tag{2.1.9}$$

respectively.

Proof: The proof is not particularly surprising: we multiply the two lines of (2.1.6) by q_{k-2} and $-p_{k-2}$, respectively, add the expressions and end up with

$$q_k \, p_{k-2} - p_k \, q_{k-2} = a_k \left(p_{k-1} \, q_{k-2} - q_{k-1} \, p_{k-2} \right) = -a_k (-1)^{k-1} = (-1)^k a_k$$

because of (2.1.7). □

This apparently innocent theorem already provides information on the convergence of convergents for infinite continued fractions, at least in the case that $a_j \in \mathbb{Q}_+$, $j \in \mathbb{N}$, where $\mathbb{Q}_+$ stands for the set of all nonnegative rational numbers.

Corollary 2.1.8. *If $a_j \in \mathbb{Q}_+$, $j \in \mathbb{N}$, then the sequence of convergents of even order, $[a_0; a_1, \dots, a_{2k}]$, is monotonicall increasing, the convergents of odd order, $[a_0; a_1, \dots, a_{2k+1}]$ decrease montonically. Moreover,*

$$\inf_{k\in\mathbb{N}} [a_0; a_1, \dots, a_{2k-1}] \geq \sup_{k\in\mathbb{N}} [a_0; a_1, \dots, a_{2k}] . \tag{2.1.10}$$

Proof: A view on the recurrence (2.1.6) shows that $q_k > 0$, $k \in \mathbb{N}$, as long as all a_j are strictly positive[3]. Then (2.1.9) yields that

$$\frac{p_{2k}}{q_{2k}} - \frac{p_{2(k-1)}}{q_{2(k-1)}} = \frac{(-1)^{2k}\, a_{2k}}{q_{2(k-1)}\, q_{2k}} > 0$$

or

$$\frac{p_{2k+1}}{q_{2k+1}} - \frac{p_{2k-1}}{q_{2k-1}} = \frac{(-1)^{2k+1}\, a_{2k+1}}{q_{2k-1}\, q_{2k+1}} < 0,$$

respectively. Next, we show that any convergent of even order is smaller than any convergent of odd order. To that end, let $m, m' \in \mathbb{N}$ and $\ell \geq \max\{m, m'\}$. From (2.1.8) with $k = 2\ell + 1$ it follows that

$$\frac{p_{2\ell}}{q_{2\ell}} = \frac{p_{2\ell+1}}{q_{2\ell+1}} + \frac{(-1)^{2\ell+1}}{q_{2\ell}\, q_{2\ell+1}} < \frac{p_{2\ell+1}}{q_{2\ell+1}}$$

and the already proven monotonicity property of convergents yields

$$\frac{p_{2m}}{q_{2m}} < \frac{p_{2\ell}}{q_{2\ell}} < \frac{p_{2\ell+1}}{q_{2\ell+1}} < \frac{p_{2m'+1}}{q_{2m'+1}}$$

as claimed. From this, (2.1.10) is immediate. □

Let us make clear what Corollary 2.1.8 means. Even order convergents form a monotonically *increasing* sequence, odd order convergents, on the other hand, a monotonically *decreasing* sequence. Moreover, the decreasing one is bounded from below[4] and thus has to be CONVERGENT. In the same way, the increasing sequence of odd order convergents, being bounded from above, must converge as well. From this we conclude the following.

Corollary 2.1.9. *The sequence of convergents, $[a_0; a_1, \dots, a_k]$, $k \in \mathbb{N}$, has at most two accumulation points, namely*

$$\lim_{k\to\infty} [a_0; a_1, \dots, a_{2k}] \qquad \text{and} \qquad \lim_{k\to\infty} [a_0; a_1, \dots, a_{2k+1}]$$

and converges if and only if equality holds in (2.1.10).

Moreover, this enclosing convergence, is also welcome since at any finite step it gives us an upper and a lower estimate for the limit – provided it exists, of course.

We close this section by extending our toolbox by two more formulas for continued fractions and their convergents.

Proposition 2.1.10. *For $1 \leq k \leq n$ we have that*

$$[a_0; a_1, \dots, a_n] = \frac{p_{k-1}\, r_k + p_{k-2}}{q_{k-1}\, r_k + q_{k-2}}, \qquad r_k := [a_k; a_{k+1}, \dots, a_n], \tag{2.1.11}$$

as well as[5]

$$\frac{q_k}{q_{k-1}} = [a_k; a_{k-1}, \dots, a_1] . \tag{2.1.12}$$

[3] Even some zeros would not hurt as soon as we once reached a positive value.
[4] By any member of the increasing one ...
[5] Note that here the order of the digits in the continued fraction expansion is reversed.

Proof: From the recurrence (1.1.2) for the definition of continued frations it follows[6] that

$$\begin{aligned}
[a_{k-1}; a_k, \dots, a_n] &= a_{k-1} + \frac{1}{[a_k; a_{k+1}, \dots, a_n]} = a_{k-1} + \frac{1}{r_k} = [a_{k-1}; r_k], \\
[a_{k-2}; a_{k-1}, \dots, a_n] &= a_{k-2} + \frac{1}{[a_{k-1}; r_k]} = [a_{k-2}; a_{k-1}, r_k], \\
&\vdots \\
[a_0; a_1, \dots, a_n] &= [a_0; a_1, \dots, a_{k-1}, r_k].
\end{aligned}$$

If p_{k-1}, q_{k-1} are numerator and denominator of the $(k-1)$st convergent and p_k, q_k the components of the kth convergent of $[a_0; a_1, \dots, a_{k-1}, r_k]$, then (2.1.6) yields that

$$[a_0; a_1, \dots, a_n] = [a_0; a_1, \dots, a_{k-1}, r_k] = \frac{p_k}{q_k} = \frac{r_k\, p_{k-1} + p_{k-2}}{r_k\, q_{k-1} + q_{k-2}},$$

which is precisely (2.1.11).

Formula (2.1.12) will be proved by induction on k. Since the continued fraction only starts at a_1, the case $k = 1$ takes the form

$$[a_1;] = a_1 = \frac{q_1}{q_0} = q_1 = p_0 = a_1.$$

Having verified (2.1.12) for some $k \geq 1$, we simply substitute the induction hypotheis (2.1.12) into (2.1.6) and get

$$\begin{aligned}
q_{k+1} &= a_{k+1} q_k + q_{k-1} = q_k \left(a_{k+1} + \frac{q_{k-1}}{q_k} \right) = q_k \left(a_{k+1} + \frac{1}{[a_k; a_{k-1}, \dots, a_1]} \right) \\
&= q_k\, [a_{k+1}; a_k, \dots, a_1],
\end{aligned}$$

which is exactly what we wanted. □

2.2 Infinite continued fractions and their convergence

In this section we consider INFINITE CONTINUED FRACTIONS of the form $[a_0; a_1, \dots]$ and their convergence. To that end, we will assume that

$$a_j > 0, \qquad j = 1, 2, \dots \tag{2.2.1}$$

We still do not (yet) assume that the coefficients are integers, as we will motivate why it is a good choice to select them as integers. Indeed, inspection of the proofs will show that everything works for $a_1, a_2, \dots \in \mathbb{Q}_+$. However, we will show in the next section that continued fractions with integer entries are „sufficient“ anyway and we can make our lives significantly easier by not enforcing ultimate generality, especially since we will get convergence for free then.

Our goal here is to collect information about the CONVERGENCE of infinite continued fractions and, in particular, to prove Theorem 1.1.2. We start with some preliminary remarks that will clarify the real meaning of convergence.

[6]As we have already seen in the introduction on page 4.

Definition 2.2.1. The infinite continued fraction $[a_0; a_1, \dots]$ is called CONVERGENT if the limit

$$[a_0; a_1, \dots] := \lim_{n\to\infty} [a_0; a_1, \dots, a_n]$$

exists and is **finite**[7]. Otherwise the continued fraction is called DIVERGENT.

Remark 2.2.2. From now on we will no more emphasize that the infinite continued fraction is infinite and just speak of a continued fraction. The „... " notation should speak for itself.

Proposition 2.2.3. *If the continued fraction $a = [a_0; a_1, \dots]$ converges, then also all the remainders*[8] *$r_k = [a_k; a_{k+1}, \dots]$ converge. Conversely, if at least one r_k converges, then so does a and hence all r_k.*

Proof: We choose any $k, n \in \mathbb{N}$ and consider the nth convergent

$$r_{k,n} := \frac{p'_n}{q'_n} = [a_k; a_{k+1}, \dots, a_{k+n}]$$

of the remainder r_k. Using (2.1.11) we get that

$$\frac{p_{k+n}}{q_{k+n}} = [a_0; a_1, \dots, a_{k+n}] = \left[a_0; a_1, \dots, a_{k-1}, r_{k,n}\right] = \frac{p_{k-1}\, r_{k,n} + p_{k-2}}{q_{k-1}\, r_{k,n} + q_{k-2}}. \tag{2.2.2}$$

Solving this rational equation for $r_{k,n}$ yields

$$r_{k,n} = \frac{p_{k-2}\, q_{k+n} - q_{k-2}\, p_{k+n}}{q_{k-1}\, p_{k+n} - p_{k-1}\, q_{k+n}} = \frac{p_{k-2} - q_{k-2}\, \dfrac{p_{k+n}}{q_{k+n}}}{q_{k-1}\, \dfrac{p_{k+n}}{q_{k+n}} - p_{k-1}},$$

and thus, due to the convergence of $[a_0; a_1, \dots]$ to a,

$$r_k := \lim_{n\to\infty} r_{k,n} = \frac{p_{k-2} - q_{k-2}\, a}{q_{k-1}\, a - p_{k-1}}.$$

If the limit of the denominator were zero and the sequence $r_{k,n}$, $n \in \mathbb{N}_0$, divergent, we only have to look at the values $r_{k,2n+1}$ to see that something is wrong[9]: by Corollary 2.1.8 they would form a *monotonically decreasing* sequence that diverges to $+\infty$.

For the converse assume that the limit $r_{k,n} \to r_k$ for $n \to \infty$ exists, then we have

$$\lim_{n\to\infty} [a_0; a_1, \dots] = \frac{p_{k-1} \lim\limits_{n\to\infty} r_{k,n} + p_{k-2}}{q_{k-1} \lim\limits_{n\to\infty} r_{k,n} + q_{k-2}} = \frac{p_{k-1}\, r_k + p_{k-2}}{q_{k-1}\, r_k + q_{k-2}} =: a$$

and the continued fraction converges which implies, by the first part of the proof, that *all* remainders converge. □

Next, we will get a *quantitative* Approximation about convergence which will turn out to be one of the central results in continued fraction theory with plenty of consequences.

[7] There is no notion like CONVERGENCE TO ∞. Go and check your basic analysis class if this is not clear to you.

[8] Which are infinite continued fractions as well!

[9] To say it in proper mathematical terms: this leads to a contradiction.

Theorem 2.2.4. *If* $a = [a_0; a_1, \dots]$ *is convergent, then we have for any* $k > 0$ *the estimate*

$$\left| a - \frac{p_k}{q_k} \right| < \frac{1}{q_k\, q_{k+1}}. \tag{2.2.3}$$

Proof: The strikingly short and simple proof relies on the MONOTONIC CONVERGENCE of convergents: if k is even, then, by Corollary 2.1.8

$$\frac{p_k}{q_k} < a < \frac{p_{k+1}}{q_{k+1}},$$

and (2.1.8) yields that

$$0 < a - \frac{p_k}{q_k} < \frac{p_{k+1}}{q_{k+1}} - \frac{p_k}{q_k} = \frac{1}{q_k\, q_{k+1}},$$

whereas for odd k the estimate

$$0 > a - \frac{p_k}{q_k} > \frac{p_{k+1}}{q_{k+1}} - \frac{p_k}{q_k} = -\frac{1}{q_k\, q_{k+1}}$$

holds. Together this gives (2.2.3). □

Now we have already provided all the tools we need to prove our convergence criterion. Let us first recall it for the sake of completeness[10].

Theorem 2.2.5 (Theorem 1.1.2 on page 4). *For any choice of* $a_0 \in \mathbb{Q}$, $a_j \in \mathbb{Q}_+$, $j \in \mathbb{N}$, *the* INFINITE CONTINUED FRACTION[11] $[a_0; a_1, \dots]$ *converges if and only if*

$$\sum_{j=0}^{\infty} a_j = \infty. \tag{2.2.4}$$

Since (2.2.4) trivally holds true whenever $a_j \geq 1$, we can immediately state the following consequence of Theorem 2.2.5.

Corollary 2.2.6. *Any continued fraction* $[a_0; a_1, \dots]$ *with* $a_0 \in \mathbb{Z}$ *and* $a_j \in \mathbb{N}$, $j \in \mathbb{N}$, *converges.*

Proof of Theorem 2.2.5: By Corollary 2.1.8 we have to show that the sequences of the even and odd convergents have *the same* limit as we already know that individually they converge. If all the convergents converge[12], (2.1.8) implies that $(q_k\, q_{k-1})^{-1}$ converge to zero which is by (2.2.3) necessary for convergence. In other words, the continued fraction converges if and only if

$$\lim_{k \to \infty} q_k\, q_{k+1} = \infty. \tag{2.2.5}$$

Let us now assume that the sequence in (2.2.4) converges. That means that $a_k \to 0$ for $k \to \infty$ and there exists $k_0 \in \mathbb{N}$ such that $a_k < 1$ for $k \geq k_0$. The recurrence (2.1.6) for the q_k tells us that these values have to be positive for $k \geq 1$ and, consequently, that

$$q_k = a_k\, q_{k-1} + q_{k-2} > q_{k-2}$$

[10] And noone likes to go to the page where it originally appeared, even if the reference is there. Not even in the wonderful new age of hyperlinks.

[11] It is a theorem, so we are very precise and mention that it is infinite.

[12] And hence are faithful to their name.

holds. Hence either $q_{k-1} \le q_{k-2}$ and thus $q_{k-1} < q_k$, or $q_{k-1} > q_{k-2}$. In the first case another application of (2.1.6) yields that

$$q_k < a_k\, q_k + q_{k-2} \qquad \Longrightarrow \qquad q_k < \frac{q_{k-2}}{1-a_k}, \quad k \ge k_0,$$

while in the second case we have that

$$q_k < (1+a_k)\, q_{k-1} = \frac{1-a_k^2}{1-a_k}\, q_{k-1} < \frac{q_{k-1}}{1-a_k}, \qquad k \ge k_0.$$

Since one of these case has to be true, there exists $\ell \in \{k-1, k-2\}$ such that

$$q_k < \frac{q_\ell}{1-a_k}.$$

If $\ell \ge k_0$, we can repeat the argument and obtain that

$$q_k < \frac{q_m}{(1-a_k)\,(1-a_\ell)}$$

for some $m \in \{k-2, k-3, k-4\}$ and that eventually[13]

$$q_k < \frac{q_{\ell_m}}{(1-a_k)\left(1-a_{\ell_1}\right)\cdots\left(1-a_{\ell_{m-1}}\right)}, \qquad \ell_m < k_0 \le \ell_{m-1}, \tag{2.2.6}$$

where $\ell_j \in \{k-j, \dots, k-2j\}$. Since the series in (2.2.4) converges, the same also holds true for the infinite product [14]

$$0 < \lambda := \prod_{j=k_0}^{\infty} (1-a_j) \le \prod_{j=0}^{m-1} \left(1 - a_{\ell_j}\right), \qquad \ell_0 = k. \tag{2.2.7}$$

Setting $Q := \max\left\{q_j \;:\; j < k_0\right\}$ we deduce from (2.2.6), that $q_k < Q/\lambda$ for $k \ge k_0$ and that the sequence $q_k\, q_{k+1}$ is bounded by

$$q_k\, q_{k+1} \le \frac{Q^2}{\lambda^2}, \qquad k \ge k_0,$$

hence cannot diverge. Since this divergence was necessary for the convergence of the continued fraction, however, (2.2.4) is also a necessary condition for convergence.

For the converse, we suppose that the series diverges and therefore satisfies (2.2.4). Since we still have $q_k > q_{k-2}$, $k \ge 2$, we define $q := \min\{q_0, q_1\}$ and find that $q_k > q$ for any $k \ge 2$. Once more, we use the recurrence relation, this time to get the estimate

$$q_k \ge a_k\, q + q_{k-2} \ge (a_k + a_{k-2})\, q + q_{k-4} \ge \cdots,$$

from which

$$q_{2k+\epsilon} \ge q_\epsilon + q \sum_{j=1}^{k} a_{2j+\epsilon} \qquad \epsilon \in \{0,1\},$$

[13] After iterating to the „bitter end“.

[14] To quote Khinchin [28]: *“... the infinite product [...], as we know, converges: that is, it has positive value ...”* To be complete, we give a proof of this folklore result in Lemma 2.2.7 later.

and thus

$$q_{2k} + q_{2k+1} \geq q_0 + q_1 + q \sum_{j=2}^{2k+1} a_j \qquad \Rightarrow \qquad q_k + q_{k+1} > q \sum_{j=0}^{k+1} a_j$$

follows. This, in turn, implies that

$$\max\{q_k, q_{k+1}\} \geq \frac{q}{2} \sum_{j=0}^{k+1} a_j,$$

and we can use the above estimate for the larger of these values and $q_k > q$ or $q_{k+1} > q$, respectively, for the smaller, we can conclude that

$$q_k\, q_{k+1} > \frac{q^2}{2} \sum_{j=0}^{k+1} a_j \to \infty, \qquad k \to \infty,$$

which yields convergence. □

To complete the proof and to be self-contained, we recall some folklore result which is useful in various situations.

Lemma 2.2.7. *For $a_j \in [0,1)$, $j \in \mathbb{N}$, the infinite product*

$$\prod_{j=1}^{\infty} (1 - a_j)$$

*has a **positive**[15] limit if and only if the infinite series*

$$\sum_{j=1}^{\infty} a_j$$

converges.

Proof: Since $a_j \in [0,1)$, the PARTIAL PRODUCTS $(1-a_1)\cdots(1-a_n)$, $n \in \mathbb{N}$, form a monotonically decreasing sequence of positive numbers, the limit

$$0 \leq \lambda = \prod_{j=1}^{\infty} (1 - a_j) = \lim_{n\to\infty} \prod_{j=1}^{n} (1 - a_j)$$

has to exist and the only question is whether it is zero or not. It is easy to see that $\lambda = 0$ if a_j is not converging to zero as we then have infinitely many factors smaller than $1 - \varepsilon$ for some $\varepsilon > 0$ and their infinite product is already zero. Thus, we only have to work in the proof of Lemma 2.2.7 only in the case $a_j \to 0$ for $j \to \infty$.

The simple idea[16] is based on the estimate[17]

$$e^{-2x} \leq 1 - x \leq e^{-x}, \qquad 0 \leq x \leq \frac{1}{2}\log 2. \tag{2.2.8}$$

[15] Convergence of an infinite product implies that its „limit" is neither $\pm\infty$ nor 0.

[16] Based on the fact that exponentiation/logarithm connect sums and products.

[17] Fig. 2.2.1 shows that this is satisfied on an even larger region that $[0, \frac{1}{2}\log 2]$, but that's enough for the proof.

Indeed, at $x = 0$ all three expressions have the value 1 and their derivatives satisfy

$$-2e^{-2x} \le -1 \le -e^{-x}, \qquad 0 \le x \le \frac{1}{2} \log 2,$$

so that a simple Taylor argument of order zero with integral remainder verifies (2.2.8). If $a_j \to 0$ then there exists some n_0 such that $a_j < \frac{1}{2} \log 2$, $j \ge n_0$ and we get

$$\prod_{j=n_0}^{\infty} (1 - a_j) \ge \prod_{j=n_0}^{\infty} e^{-2a_j} = \exp\left(-2 \sum_{j=n_0}^{\infty} a_j\right) \tag{2.2.9}$$

as well as

$$\prod_{j=n_0}^{\infty} (1 - a_j) \le \prod_{j=n_0}^{\infty} e^{-a_j} = \exp\left(-\sum_{j=n_0}^{\infty} a_j\right). \tag{2.2.10}$$

If the series converges, then so does the subsequence starting at n_0, say to a limit a, and (2.2.9) yields that

$$\lambda \ge e^{a} \prod_{j=0}^{n_0-1} (1 - a_j) > 0.$$

If, on the other hand, the series diverges, we get from (2.2.10) that

$$\lambda \le e^{-\infty} \prod_{j=0}^{n_0-1} (1 - a_j) = 0,$$

as claimed. □

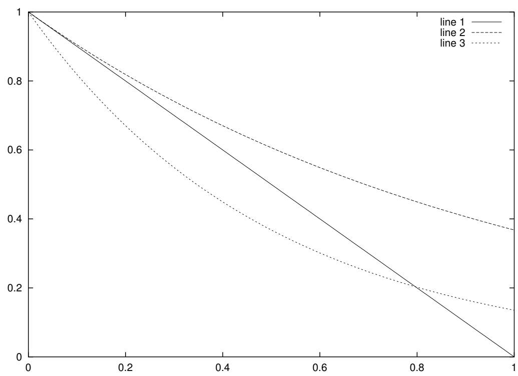


Abbildung 2.2.1: The three functions from estimate (2.2.8) which also holds for values of x larger than $\frac{1}{2} \log 2 \approx .35$.

2.3 Continued fractions with integer coefficients

Having realized in Corollary 2.2.6 that continued fractions with positive integer coefficients[18] behave nicely and always converge, we will next convince ourselves that this class

[18]Except a_0, of course.

of continued fractions is completely *sufficient* for studying real and rational numbers. Since then we get convergence for free, continued fractions with integer coefficients give us an easy and direct access to real numbers, provided we can indeed represent any such numbers in this way. This is in analogy to a DIGIT EXPANSION witb BASIS B

$$x = \sum_{j=1}^{\infty} x_j \, B^{-j}, \qquad x_j \in \{0, \dots, B-1\},$$

which also converges since the partial sums are bounded:

$$\sum_{j=n}^{\infty} x_j \, B^{-j} \le B^{1-n}, \qquad n \in \mathbb{N}, \tag{2.3.1}$$

hence, in this case we do not have to worry about convergence issues as well.

Exercise 2.3.1 Prove (2.3.1). ◊

Theorem 2.3.1. *Any nonnegative* RATIONAL NUMBER $x = \frac{p}{q}$ *can be represented by a* ***finite*** *continued fraction with* ***positive integer*** *coefficients.*

Proof: We assume that p/q is the NORMALIZED FORM of the fraction, that is $\gcd(p,q) = 1$ and $p \ge 0$, $q > 0$, otherwise we could just divide by $\gcd(p,q)$ and multiply both by -1 if needed. Next we define, as in the Euclidean algorithm, cf. [12], a_0 and r by DIVISION WITH REMAINDER:

$$p = a_0 q + r, \qquad 0 \le r < q.$$

If $r = 0$, then we have the simple form $x = \frac{p}{q} = a_0 = [a_0;]$, otherwise we get

$$\frac{p}{q} = \frac{a_0\, q + r}{q} = a_0 + \frac{r}{q} = a_0 + \frac{1}{\frac{q}{r}} = \left[a_0; \frac{q}{r}\right]. \tag{2.3.2}$$

Now we do induction[19] on the numerator q and get, since $r < q$, by the induction hypothesis that

$$\frac{q}{r} = [a_1; a_2, \dots, a_k], \qquad a_j \in \mathbb{N},$$

which we substitute into (2.3.2) to get

$$\frac{p}{q} = a_0 + \frac{1}{[a_1; a_2, \dots, a_k]} = [a_0; a_1, \dots, a_k],$$

which is a finite continued fraction expression. That, conversely, any finite continued fraction defines a rational number, has been mentioned several times and lies in the nature of the definition (1.1.1). □

From the proof we get the following estimate for the LENGTH of a continued fraction

Corollary 2.3.2. *If* $\frac{p}{q} = [a_0; a_1, \dots, a_k]$, *then* $k \le q$.

Exercise 2.3.2 Can the case $k = q$ happen in Corollary 2.3.2? ◊

[19]The case $q = 0$ is nonsense, the case $q = 1$ trivial – so much about the initialization of the induction.

Remark 2.3.3 (Continued fractions with positive integer components).

1. Formally, Theorem 2.3.1 holds only for *nonnegative* rational numbers $x \in \mathbb{Q}_+$, but it is easily extended to $\mathbb{Q}$. Indeed, we only have to set

$$a_0 := \lfloor x \rfloor := \max\{k \in \mathbb{Z} : k \le x\}, \qquad \text{hence,} \qquad r_0 := x - a_0 \in [0,1),$$

   and then proceed by determining the other components by the rule

$$a_j = \left\lfloor \frac{1}{r_{j-1}} \right\rfloor \in \mathbb{N}, \qquad r_j = r_{j-1} - \frac{1}{a_j}, \qquad j \in \mathbb{N}, \tag{2.3.3}$$

   which gives

$$r_{j-1} = a_j + r_j = \left[a_j; \frac{1}{r_j} \right],$$

   so that the iteration (2.3.3) determines the coefficients of

$$x = [a_0; a_1, \dots, a_k], \qquad a_0 \in \mathbb{Z}, \quad a_j \in \mathbb{N},\ j \ge 1,$$

   where Theorem 2.3.1 ensures that the expansion is finite and the iteration terminates after finitely many steps.

2. The above procedure also gives a way to define a NORMAL FORM for the continued fraction expansion of any given rational number, and this normal form consists, except perhaps a_0, always on **positive** integers only.

3. The same process, when applied to a REAL NUMBER will eventually lead to an always convergent CONTINUED FRACTION EXPANSION of that number that automatically converges. We will see that this expansion will even enable us to do number theory and that the expansions of rational, algebraic and transcendental numbers will be easily distinguishable.

Exercise 2.3.3 Implement the routine to compute continued fractions in `Matlab` or `Octave`. ◇

The recurrence (2.1.6) for the canonical representation of the kth CONVERGENT shows us that for the representation according to Theorem 2.3.1 its numerator p_k is always an integer and that its denominator is a even a positive integer. The natural question is whether this representation is already optimal, i.e., NORMALIZED or if the two have a nontrivial common divisor. The answer is „irreducible“ and the proof strikingly simple.

Theorem 2.3.4. *The canonical representation $\frac{p_k}{q_k}$ of the kth convergent is* IRREDUCIBLE.

Proof: Any common divisor of p_k and q_k would also divide the expression

$$q_k \, p_{k-1} - p_k \, q_{k-1} = (-1)^k$$

from (2.1.7) and thus can only be ± 1. □

The RECURRENCE RELATION immediately implies that the denominators of the canonical representations

$$q_k = a_k \, q_{k-1} + q_{k-2} > a_k \, q_{k-1} \ge q_{k-1}$$

and that obviously the growth is related to the components: the larger a_k, the faster the grow. But in any case the growth rate is at least exponential, namely

$$q_k \geq 2^{(k-1)/2}, \qquad k \geq 1. \tag{2.3.4}$$

This is yet another consequence of the recurrence relation, which yields, together with the monotonic growth

$$q_k > (a_k + 1)\, q_{k-2} \geq 2\, q_{k-2}$$

from which (2.3.4) follows[20] with the initial conditions $q_0 = 1$ and $q_1 = a_0 \geq 1$.

Definition 2.3.5. For $k \geq 2$ the fractions

$$\frac{p_{k-2} + j\, p_{k-1}}{q_{k-2} + j\, q_{k-1}}, \qquad j = 0, \ldots, a_k,$$

are called INTERMEDIATE FRACTIONS between the $(k-2)$nd and $k$the convergent of the continued fraction.

The name *intermediate fraction* is easily explained: setting $j = 0$ we get the canonical representation of the $(k-2)$nd convergent while the other extreme case, $j = a_k$, gives the kth convergent of the continued fraction. This is once more a direct consequence of the recurrence relation (2.1.6).

Proposition 2.3.6. *For even k the intermediate fractions form a monotonically* increasing *sequence, for odd k a monotonically* decreasing *one.*

Proof: For $j \geq 0$ we consider the difference

$$\begin{aligned}
&\frac{(j+1)\, p_{k-1} + p_{k-2}}{(j+1)\, q_{k-1} + q_{k-2}} - \frac{j\, p_{k-1} + p_{k-2}}{j\, q_{k-1} + q_{k-2}} \\
&= \frac{(j+1)p_{k-1}q_{k-2} + jp_{k-2}q_{k-1} - jp_{k-1}q_{k-2} - (j+1)p_{k-2}q_{k-1}}{((j+1)q_{k-1} + q_{k-2})\,(jq_{k-1} + q_{k-2})} \\
&= \frac{p_{k-1}q_{k-2} - q_{k-1}p_{k-2}}{((j+1)q_{k-1} + q_{k-2})\,(jq_{k-1} + q_{k-2})} = \frac{(-1)^k}{((j+1)q_{k-1} + q_{k-2})\,(jq_{k-1} + q_{k-2})},
\end{aligned}$$

which is positive for even k and negative for odd k. □

The next concepts adds fractions in a way that was forbidden in school and still makes meaning out of it.

Definition 2.3.7. The MEDIANT between the fractions a/b und c/d is defined as

$$\frac{a}{b} \oplus \frac{c}{d} := \frac{a+c}{b+d}. \tag{2.3.5}$$

Remark 2.3.8. Definition 2.3.7 is a nice example that even „forbidden“ mathematical operations like a too naive addition of fractions can be meaningful when considered properly in the right context. Another example for that is the HADAMARD PRODUCT of two matrices, cf. [25].

[20] Writing this as a formally correct and complete induction is a nice exercise.

An INTERMDIATE FRACTION is the mediant between two successive convergents of two consecutive fraction, more precisely, the kth intermediate fraction is the mediant between the kth and the $(k-1)$st convergent.

As we know from Proposition 2.3.6, the value of the mediant and thus of the intermediate fraction depends on the *representation* of the fraction itself: the intermediate fractions are mediants of

$$\frac{p_{k-2}}{q_{k-2}} \qquad \text{and} \qquad \frac{j\,p_{k-1}}{j\,q_{k-1}} = \frac{p_{k-1}}{q_{k-1}},$$

and have, for different j different values, that can be monotonically increasing or decreasing with respect to k. We can also view it differently:

> *The jth intermediate fraction is the mediant between the $(j-1)$st intermediate fraction and the the $(k-1)$st convergent, i.e.,*
>
> $$\frac{j\,p_{k-1}+p_{k-2}}{j\,q_{k-1}+q_{k-2}} = \frac{(j-1)\,p_{k-1}+p_{k-2}}{(j-1)\,q_{k-1}+q_{k-2}} \oplus \frac{p_{k-1}}{q_{k-1}}$$

In general, the value of the mediant always lies between the values of the two fractions, more precisely,

$$b, d > 0, \quad \frac{a}{b} < \frac{c}{d} \qquad \Rightarrow \qquad \frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d}, \tag{2.3.6}$$

The assumption $a/b < c/d$ or, equivalently, $bc - ad > 0$ is no restriction as long as the two rational numbers[21] are not equal. The inequalities in (2.3.6) now follow from the observation that

$$\frac{a+c}{b+d} - \frac{a}{b} = \frac{ab+bc-ab-ad}{(b+d)b} = \frac{bc-ad}{b^2+bd} > 0$$

and

$$\frac{c}{d} - \frac{a+c}{b+d} = \frac{bc+cd-ad-cd}{d(b+d)} = \frac{bc-ad}{bd+d^2} > 0.$$

Hence any intermediate fraction is enclosed by to successive convergents. To that end, consider the sequence of potential intermediate fractions b_j, defined by

$$b_j := \frac{j\,p_{k-1}+p_{k-2}}{j\,q_{k-1}+q_{k-2}} = b_{j-1} \oplus \frac{p_{k-1}}{q_{k-1}}, \qquad b_0 := \frac{p_{k-2}}{q_{k-2}}. \tag{2.3.7}$$

Exercise 2.3.4 Show that the representation

$$b_j = \frac{j\,p_{k-1}+p_{k-2}}{j\,q_{k-1}+q_{k-2}}$$

is irreducible. ◇

Being defined as a mediant in (2.3.7), b_j lies between b_{j-1} and p_{k-1}/q_{k-1}. Since the kth convergent is just b_{a_k} and since the limit $a = [a_0; a_1, \dots]$ of an infinite continued fraction is enclosed by the $(k-1)$st and kth convergent, we alway find the limit between b_1 and p_{k-1}. On the other hand, b_1 is the mediant between the $(k-2)$nd and $(k-1)$st convergent.

[21] One still has to distinguish between the FRACTION, i.e., the pair of numerator and denominator and the RATIONAL NUMBER represented by the fraction as this makes a difference for mediants as we already saw above. And $\frac{1}{2}$ and $\frac{2}{4}$ are different fractions representing the same rational number.

Before this gets too confusing, we illustrate the situation for even k:

$$\frac{p_{k-2}}{q_{k-2}} = b_0 < b_1 = \frac{p_{k-2}}{q_{k-2}} \oplus \frac{p_{k-1}}{q_{k-1}} < \cdots < b_{a_k} = \frac{p_k}{q_k} < a < \frac{p_{k-1}}{q_{k-1}}, \tag{2.3.8}$$

for odd k simply all the inequality signs have to be reversed. If we replace k by $k+2$ in (2.3.8), we conclude that for any even k the relation

$$\frac{p_k}{q_k} < \frac{p_k}{q_k} \oplus \frac{p_{k+1}}{q_{k+1}} < a < \frac{p_{k+1}}{q_{k+1}}, \tag{2.3.9}$$

holds, while for odd k we have the same with reversed inequality signs[22]. This simple observation has a very interesting consequence for the APPROXIMATION QUALITY of continued fractions.

Theorem 2.3.9. *For $a = [a_0; a_1, \ldots]$ and $k \geq 0$ we have that*

$$\frac{1}{q_k\,(q_{k+1} + q_k)} < \left| a - \frac{p_k}{q_k} \right| < \frac{1}{q_k\, q_{k+1}}. \tag{2.3.10}$$

This theorem tells us that the upper estimate for the CONVERGENCE RATE of continued fractions is practically optimal. Since the denominators q_k of the convergents are monotonically increasing, that is, $q_{k+1} > q_k$ and therefore also $q_{k+1} + q_k < 2q_{k+1}$, we get that

$$\frac{1}{q_k\,(q_{k+1} + q_k)} > \frac{1}{2\, q_k\, q_{k+1}},$$

which gives us the slightly coarser but more illustrating enclosure

$$\frac{1}{2\, q_k\, q_{k+1}} < \left| a - \frac{p_k}{q_k} \right| < \frac{1}{q_k\, q_{k+1}}. \tag{2.3.11}$$

Since the q_k grow like $2^{k/2}$, the factor 2 in (2.3.11) is more or less irrelevant and we can say that the kth convergents converge like 2^{-k}. In other words: any convergent determines approximately one BINARY DIGIT of the the fraction.

Proof of Theorem 2.3.9: The upper estimate in (2.3.10) is precisely Theorem 2.2.4, for the lower estimates we have a closer look at the mediants[23]; indeed, (2.3.9) says that the mediant of the kth and $(k+1)$st convergent is closer to the value of the continued fraction than the kth convergent, yielding

$$\begin{aligned} \left| a - \frac{p_k}{q_k} \right| &> \left| \left(\frac{p_k}{q_k} \oplus \frac{p_{k+1}}{q_{k+1}} \right) - \frac{p_k}{q_k} \right| = \left| \frac{p_{k+1} + p_k}{q_{k+1} + q_k} - \frac{p_k}{q_k} \right| = \left| \frac{p_{k+1}\, q_k - p_k\, q_{k+1}}{q_k\,(q_{k+1} + q_k)} \right| \\ &= \left| \frac{(-1)^k}{q_k\,(q_{k+1} + q_k)} \right| = \frac{1}{q_k\,(q_{k+1} + q_k)}, \end{aligned}$$

as claimed. □

Now we come to a fundamental result on continued fractions for real numbers.

Theorem 2.3.10. *Any REAL NUMBER $x \in \mathbb{R}$ can be written in exactly one way as a continued fraction $[a_0; a_1, \ldots]$ with $a_0 \in \mathbb{Z}$ and positive integer entries $a_j \in \mathbb{N}$, $j \in \mathbb{N}$. This continued fraction is finite if the number is rational and infinite if it is irrational.*

[22] **Exercise**: verify that.

[23] There must have been some reason for their introduction.

Remark 2.3.11. The way it is stated, Theorem 2.3.10 is not correct as finite continued fractions cannot be unique without an additional assumption! This can be seen from the simple example

$$[a_0;] = a_0 = a_0 - 1 + 1 = a_0 - 1 + \frac{1}{1} = [a_0 - 1; 1] \, .$$

This implies that always

$$[a_0; a_1, \ldots, a_n] = [a_0; a_1, \ldots, a_n - 1, 1], \qquad [a_0; a_1, \ldots, a_n, 1] = [a_0; a_1, \ldots, a_n + 1] \, . \tag{2.3.12}$$

Hence, finite continued fractions that end on „1"have a builtin AMBIGUITY. This enforces the convention from the following definition.

Definition 2.3.12 (Convention on last digits). Any **finite** continued fraction $[a_0; a_1, \ldots, a_n]$, $a_0 \in \mathbb{Z}$, $a_j \in \mathbb{N}$, $j \in \mathbb{N}$, must always satisfy $a_n \neq 1$.

Note that Definition 2.3.12 is no restriction as any continued fraction that accidentially happens to have last digit $a_n = 1$ can be rewritten and even shortend and simplified by means of (2.3.12) until the last digit is indeed $\neq 1$.

Remark 2.3.13. Theorem 2.3.10 shows that the distinction between rational and irrational numbers is simpler in terms of continued fractions than in terms of DIGIT EXPANSIONS like binary or decimal digits. Recall that rational numbers are characterized by having either finite or periodic digit expansions, independently of the basis.

Proof of Theorem 2.3.10: That rational numbers can be represented by finite continued fractions, we already know from Theorem 2.3.1. So it remains to show the existence of a continued fraction expansion for irrational numbers and, in particular, the UNIQUENESS of the continued fraction expansion.

To that end, we first (re)consider the general method to compute a continued fractions, starting from a number $x \in \mathbb{R} \setminus \mathbb{Q}$. In the first step the only reasonable choice is to set

$$a_0 = \lfloor x \rfloor := \max \{ j \in \mathbb{Z} \ : \ j \leq x \} ,$$

which either gives $x = a_0$ or there exists some $r_1 \neq 0$ such that we can write

$$x = [a_0; r_1] = a_0 + \frac{1}{r_1} \qquad \Rightarrow \qquad r_1 = \frac{1}{x - a_0} > 1,$$

since $0 < x - a_0 < 1$. And now we continue iteratively, setting

$$a_j = \lfloor r_j \rfloor , \qquad r_{j+1} = \frac{1}{r_j - a_j}, \qquad j = 1, 2, \ldots, \tag{2.3.13}$$

and noting that the sequences we obtain this way already satisfy $a_0 \in \mathbb{Z}$ and $a_j \in \mathbb{N}$, $j \in \mathbb{N}$, thus defining a *convergent* continuous fraction. The sequence would terminate only if $a_j = r_j$, but then the continued fraction were finite and $x \in \mathbb{Q}$, i.e., a rational number. So irrational numbers must have an infinite continued fraction expansion[24]. By construction and using an infinite version of (2.1.11), we thus obtain

$$x = [a_0; a_1, \ldots, a_{n-1}, r_n] = \frac{r_n \, p_{n-1} + p_{n-2}}{r_n \, q_{n-1} + q_{n-2}}.$$

[24]Which is less of a surprise, of course.

But then we get[25]

$$\begin{aligned} x - \frac{p_n}{q_n} &= \frac{r_n\, p_{n-1} + p_{n-2}}{r_n\, q_{n-1} + q_{n-2}} - \frac{a_n\, p_{n-1} + p_{n-2}}{a_n\, q_{n-1} + q_{n-2}} \\ &= \frac{r_n\, p_{n-1}\, q_{n-2} + a_n\, q_{n-1}\, p_{n-2} - r_n\, q_{n-1}\, p_{n-2} - a_n\, p_{n-1}\, q_{n-2}}{(r_n\, q_{n-1} + q_{n-2})\,(a_n\, q_{n-1} + q_{n-2})} \\ &= \frac{(p_{n-1}\, q_{n-2} - q_{n-1}\, p_{n-2})\,(r_n - a_n)}{[(r_n - a_n)\, q_{n-1} + q_n]\, q_n} = \frac{(-1)^n\,(r_n - a_n)}{q_n^2 + (r_n - a_n)\, q_{n-1} q_n}, \end{aligned}$$

und thus

$$\left| x - \frac{p_n}{q_n} \right| < \frac{1}{q_n^2}. \tag{2.3.14}$$

Hence the convergents indeed converge to x, and even with the predicted speed. In summary, the infinite continued fraction constructed above is a representation for x.

It remains to show uniqueness where we make use that the representation of $x \in \mathbb{R}$ only contains *positive* integers, except maybe a_0. This already enforces the choice $a_0 = \lfloor x \rfloor$ since

$$x - a_0 = [0; a_1, a_2, \dots] \in (0,1)$$

except when $x \in \mathbb{Z}$, but that is a trivial case anyway. If now $[a_0; a_1, \dots]$ and $\left[a_0'; a_1', \dots\right]$ are two continued fraction expressions of $x \in \mathbb{R}$, then the above reasoning yields that $a_0 = a_0'$. Now suppose that, by induction on $k \ge 0$, we have already shown that

$$a_j = a_j' \qquad \Rightarrow \qquad p_j = p_j',\ q_j = q_j', \qquad j = 0, \dots, k,$$

then

$$x = \frac{r_{k+1}\, p_k + p_{k-1}}{r_{k+1}\, q_k + q_{k-1}} = \frac{r_{k+1}'\, p_k' + p_{k-1}'}{r_{k+1}'\, q_k' + q_{k-1}'} = \frac{r_{k+1}'\, p_k + p_{k-1}}{r_{k+1}'\, q_k + q_{k-1}},$$

yields that

$$[a_{k+1}; a_{k+2}, \dots] = r_{k+1}' = r_{k+1} = \left[a_{k+1}'; a_{k+2}', \dots\right],$$

and therefore, repeating the above argument, that $a_{k+1}' = a_{k+1}$. Regardless of whether the continued fractions are finite or infinite, this yields that they must coincide. □

Example 2.3.14. The construction procedure for continued fractions enables us to easily determine the (irrational) numbers that have a particularly simple inifite continued fraction expansion of the form

$$x = [k; k, \dots], \qquad k \in \mathbb{N}.$$

They have the property that $r_1 = x$ and therefore

$$x = k + \frac{1}{x} \qquad \Rightarrow \qquad x^2 - kx - 1 = 0 \qquad \Rightarrow \qquad x = \frac{k + \sqrt{k^2+4}}{2}.$$

Since the $x > 0$, the negative zero of the quadratic equation can be excluded. In particular, we find that

$$\frac{1+\sqrt{5}}{2} = [1; 1, \dots],$$

which means that the GOLDEN RATIO has the simplest possible continued fraction expansion.

[25]Needless to say that once more the recurrence relation enters the scene.

We can extend the same idea and see, what we can do with 2–PERIODIC continued fractions of the form $x = [k_1; k_2, k_1, k_2, \dots]$. Now the FIX POINT EQUATION is $r_2 = x$, leading to

$$x = k_1 + \cfrac{1}{k_2 + \cfrac{1}{x}} = k_1 + \frac{x}{k_2 x + 1} = \frac{(k_1\, k_2 + 1)\, x + k_1}{k_2 x + 1}$$

and we now look for the zeros of

$$k_2 x^2 - k_1 k_2 x - k_1 = k_2 \left(x^2 - k_1 x - \frac{k_1}{k_2} \right) \qquad \Rightarrow \qquad x = \frac{k_1 + \sqrt{k_1\,(k_1 + 4/k_2)}}{2}.$$

Again, the numbers are rational plus a plain square root.

Exercise 2.3.5 Show: any PERIODIC CONTINUED FRACTION belongs to $\mathbb{Q} + \sqrt{\mathbb{Q}}$, hence can be written as $q + r$, $q, r^2 \in \mathbb{Q}$.
Hint: First show that any $x \in \mathbb{R}$ that can be written as a periodic continued fraction satisfies an equation of the form

$$x = \frac{p(x)}{q(x)}, \qquad p, q \in \mathbb{N}[x],\ \deg p = \deg q = 1.$$

◇

2.4 Convergents as best approximants

Knowing that any real number can be represented as an infinite continued fraction and thus approximated by finite continued fractions, namely its convergent, we will justify their use by showing that continued fractions approximate real numbers *better*[26] than other fractions. Of course, with $\mathbb{Q}$ being dense in $\mathbb{R}$, there are lots of[27] fractions that converge to a given $x \in \mathbb{R}$.

Remark 2.4.1 (Myths and legends, cf. [28]). When Christiaan Huygens built his mechanical planetarium, a model of our solar system, he had to approximate the irrational duration of the time it takes planets to complete their orbit as good as possible by rational numbers. Rational numbers can be implemented mechanically by COGWHEELS and the transmission is simply the ratio of the number of teeth in the gear, hence a rational number[28]. So good approximations by rational numbers were crucial for the mechanical implememtation and, of course knowing the theory of continued fractions, using convergents proved to be the way to go.

A good measure for the COMPLEXITY of a fraction[29] $x \in \mathbb{Q}$ is the size of its denominator: writing x as

$$x = a + \frac{p}{q}, \qquad a \in \mathbb{Z},\ p, q \in \mathbb{N}, \quad p < q,$$

[26] In the sense of „faster“.
[27] Namely infinitely many and more.
[28] Even with modern methods like additive manufacturing and nanotechnology, noone has managed so far to produce a cogwheel with a noninteger number of teeth. The same actually holds true for negative numbers.
[29] Not directly for a rational number, as, once again, any rational number can be written in many ways as a fraction. But taking the (more or less) unique irreducible fraction, one could also define the complexity of a rational number.

then the amount of INFORMATION we need to store x is of the order of magnitude $\log a + 2\log q$. This is simply the number of DIGITS of the integer part and in numerator and denominator. Whether we choose these digits decimally or binary is only a constant and not really relevant. Ignoring the integer part[30], the fundamental complexity quantity for a fraction is therefore the size of its denominator and the complexity of a rational number is the size of its denominator in the irreducible representation. This more than justifies the following definition.

Definition 2.4.2. A fraction a/b is called BEST APPROXIMANT to $x \in \mathbb{R}$ if

$$\left|x - \frac{a}{b}\right| \le \left|x - \frac{c}{d}\right|, \qquad d \le b.$$

Here we always consider fractions of the form $\mathbb{Z}/\mathbb{N}$ with positive denominators.

What we will show now is that the convergents of the continued fraction expansions are essentially the best approximants.

Theorem 2.4.3. *Any BEST APPROXIMANT to a realy number $x \in \mathbb{R}$ is either a CONVERGENT of the associated continued fraction expansion or an INTERMEDIATE FRACTION.*

Proof: Let a/b be a[31] best approximant[32] to $x = [a_0; a_1, \dots]$, $a_0 \in \mathbb{Z}$, $a_j \in \mathbb{N}$, $j \in \mathbb{N}$. Then $a/b > a_0$ as otherwise $a/b < a_0 = \lfloor x \rfloor \le x$ and $a_0/1$ were already a better approximant than a/b which were a contractiction. Exactly the same type of argument also shows that $\frac{a}{b} < a_0 + 1$ as then $a_0 + 1$ were a better approximant due to $x < a_0 + 1$. Hence,

$$a_0 \le \frac{a}{b} \le a_0 + 1$$

and with equality in one of the two cases the claim is proved: the best approximant is then either the converget a_0 or the intermediate fraction

$$\frac{a_0 + 1}{1} = \frac{p_1 + p_0}{q_1 + q_0}, \qquad \text{da} \qquad q_0 = 0,\ q_1 = p_0 = 1\, p_1 = a_0.$$

Let us suppose that $a_0 < \frac{a}{b} < a_0 + 1$ and that a/b is neither convergent nor intermediate fraction. We will show that then there exists an intermediate fraction[33] with a smaller denominator that is even closer to x. By Proposition 2.3.6, a/b lies between two intermediate fractions[34] so that there exist n and k such that either

$$\frac{k\,p_n + p_{n-1}}{k\,q_n + q_{n-1}} < \frac{a}{b} < \frac{(k+1)\,p_n + p_{n-1}}{(k+1)\,q_n + q_{n-1}} \qquad \text{or} \qquad \frac{k\,p_n + p_{n-1}}{k\,q_n + q_{n-1}} > \frac{a}{b} > \frac{(k+1)\,p_n + p_{n-1}}{(k+1)\,q_n + q_{n-1}},$$

[30] We can simply restrict x to $[0,1]$, the integer part is fairly easy to approximate; alternatively, we shift the number by multiplying with an approximate power of the basis which is also what happens in FLOATING POINT NUMBERS.

[31] We never claimed that best approximants are unique.

[32] Sometimes called ELEMENT OF BEST APPROXIMATION, cf. [45].

[33] Which may even be a convergent.

[34] Recall that the intermediate fractions for convergents form a sequence that converges monontonically to x, monotonically increasing if the order of the convergent is even, decreasing if it is odd.

and therefore

$$\left|\frac{a}{b}-\frac{k\,p_n+p_{n-1}}{k\,q_n+q_{n-1}}\right| < \left|\frac{(k+1)\,p_n+p_{n-1}}{(k+1)\,q_n+q_{n-1}}-\frac{k\,p_n+p_{n-1}}{k\,q_n+q_{n-1}}\right| = \frac{1}{((k+1)q_n+q_{n-1})\,(kq_n+q_{n-1})}$$

On the other hand, expanding the difference of fractions yields that there exists $c \in \mathbb{N}$ such that[35]

$$0 \neq \left|\frac{a}{b}-\frac{k\,p_n+p_{n-1}}{k\,q_n+q_{n-1}}\right| = \frac{c}{b\,(kq_n+q_{n-1})} \geq \frac{1}{b\,(kq_n+q_{n-1})},$$

which yields

$$\frac{1}{b\,(kq_n+q_{n-1})} < \frac{1}{((k+1)q_n+q_{n-1})\,(kq_n+q_{n-1})} \qquad \Rightarrow \qquad b > (k+1)\,q_n+q_{n-1}.$$

This shows that the $(k+1)$st intermediate fraction that, by construction, is closer to x than a/b, has a smaller denominator than a/b and therefore is a better approximant which is a contradiction. Therefore a/b must be either convergent or approximant. □

Indeed, convergents are even ***unique*** best approximants if the notion of BEST APPROXIMATION is formulated in a slightly sharper way. To motivate this concept, we recall what the expression a/b really means: it is the (rational) number that, when multiplied with b gives the value a. In this respect, x is a good approximation to that number if the difference $|bx-a|$ is as small as possible.

Definition 2.4.4. A fraction a/b is called BEST APPROXIMATION OF THE SECOND KIND[36] to $x \in \mathbb{R}$ provided that

$$\frac{c}{d} \neq \frac{a}{b}, \quad 0 < d \leq b \qquad \Longrightarrow \qquad |bx-a| \leq |dx-c|\,. \tag{2.4.1}$$

Best approximants of the second kind are also best approximant of the first kind in the sense of Definition 2.4.2., as otherwise there would exist a fraction c/d, $d \leq b$, such that

$$\left|x-\frac{a}{b}\right| > \left|x-\frac{c}{d}\right| \qquad \Rightarrow \qquad |bx-a| = b\left|x-\frac{a}{b}\right| > b\left|x-\frac{c}{d}\right| = \frac{b}{d}\,|dx-c| \geq |dx-c|\,,$$

which contradicts the assumption that a/b is best approximant of the second kind. But the converse is not true, not any best approximant of the first kind is also one of the second kind[37]. The simplest example is und damit wäre a/b auch kein Bestapproximant zweiter Art. $x=\frac{1}{5}$ and $\frac{a}{b}=\frac{1}{3}$; it is easy to veryfiy that $\frac{1}{3}$ is closer to $\frac{1}{5}$ than its competitors$\left\{0,\frac{1}{2},\frac{2}{3},1\right\}$ of fractions with numerator ≤ 3, but that

$$\left|3\,\frac{1}{5}-1\right| = \frac{2}{5} > \frac{1}{5} = \left|1\,\frac{1}{5}-0\right|$$

holds. Best approximants of the second kind play an important role as these indeed are convergents and only convergents.

[35]The expression is not zero, hence the numerator is not zero and, since it is an integer, it must be ≥ 1.

[36]There will be no encounter with approximations of the third kind.

[37]Otherwise the distinction would be pointless.

Theorem 2.4.5. *Any* BEST APPROXIMANT OF THE SECOND KIND *to $x \in \mathbb{R}$ is a convergent and any convergent of the continued fraction expansion of x is a best approximant of the second kind.*

*Except the special case $x = a_0 + \frac{1}{2}$ and convergents of first order, all best approximants of the second kind are **unique**.*

Proof: Let us suppose that $\frac{a}{b}$ is a best approximant of the second kind to $x = [a_0; a_1, \dots]$. If $a/b < a_0 = \lfloor x \rfloor < x$, then $b \geq 1$ yields that

$$|1 \cdot x - a_0| = x - a_0 < x - \frac{a}{b} = \frac{1}{b}|bx - a| < |bx - a|$$

and $a_0 = a_0/1$ would be a better approximation of the second kind. Hence, the first convergent of order 0 satiesfies $\frac{p_0}{q_0} = a_0 \leq a/b$. If a/b is no convergent[38], then it either satisfies $\frac{a}{b} > \frac{p_1}{q_1}$ or is enclosed between two convergents $\frac{p_{k-1}}{q_{k-1}}$ und $\frac{p_{k+1}}{q_{k+1}}$ due to the montonic convergence of the convergents, cf. Corollary 2.1.8. In the first case we have that $x < \frac{p_1}{q_1} < \frac{a}{b}$ and the monotnoicity of the denominators q_k yields

$$\left| x - \frac{a}{b} \right| > \left| \frac{p_1}{q_1} - \frac{a}{b} \right| = \frac{|b\,p_1 - a\,q_1|}{b\,q_1} \geq \frac{1}{b\,q_1},$$

that is,

$$|bx - a| > \frac{1}{q_1} = \frac{1}{a_1} = \frac{1}{\lfloor x - a_0 \rfloor^{-1}} \geq |1x - a_0|,$$

contradicting the assumption that a/b is a best approximant of the second kind. If, on the other hand, a/b is enclosed between two convergents, we first have

$$\left| \frac{a}{b} - \frac{p_{k-1}}{q_{k-1}} \right| = \frac{|a\,q_{k-1} - b\,p_{k-1}|}{b\,q_{k-1}} \geq \frac{1}{b\,q_{k-1}} \tag{2.4.2}$$

as well as[39], by Corollary 2.1.6,

$$\left| \frac{a}{b} - \frac{p_{k-1}}{q_{k-1}} \right| < \left| \frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}} \right| = \frac{1}{q_k\,q_{k-1}}, \tag{2.4.3}$$

which allows us to combine (2.4.2) and (2.4.3) to $q_k < b$. Moreover,

$$\left| x - \frac{a}{b} \right| > \left| \frac{p_{k+1}}{q_{k+1}} - \frac{a}{b} \right| \geq \frac{1}{b\,q_{k+1}},$$

which yields, together with (2.3.11), the estimate

$$|bx - a| > \frac{1}{q_{k+1}} = q_k \frac{1}{q_k\,q_{k+1}} > q_k \left| x - \frac{p_k}{q_k} \right| = |q_k\,x - p_k|$$

which would make the kth convergent a better approximant, another contradiction. Hence, all that is left is that a/b is indeed a convergent.

For the converse we fix the order k of the convergent, consider the numbers

$$\min_{a \in \mathbb{Z}} |b\,x - a|, \qquad b \in \{1, \dots, q_k\} \tag{2.4.4}$$

[38] In this case the theorem would be trivially true.

[39] The appearance of the kth convegent is no typo, we make use of the fact that it lies on the „other" side of x.

and denote by b^* the value of b for which this becomes minimal. If b^* is not unique, we take the smallest of these values which makes b^* unique and well–defined. The respective minimizing value for a is denoted by

$$a^* = \operatorname*{argmin}_{a\in\mathbb{Z}} |b^*\, x - a|\,. \tag{2.4.5}$$

We first show that a^* in (2.4.5) is unique. To that suppose that there exists $a' \neq a^*$ which also satisfies (2.4.5), and note that

$$\left|x - \frac{a^*}{b^*}\right| = \left|x - \frac{a'}{b^*}\right| \qquad \Rightarrow \qquad x = \frac{a^* + a'}{2\,b^*}. \tag{2.4.6}$$

The fraction on the right hand side of (2.4.6) has to be irreducible as otherwise there exist an irreducible representation $x = p/q$ with $q \leq b^*$ and thus $|q\,x - p| = 0$, which yields an unbeatable minimal value of (2.4.4) that is assumed exactly for $a = p$ and $b = q \leq b^* \leq q_k$. Developing the rational number x as a continued fraction, $x = [a_0; a_1, \ldots, a_n]$, and[40] writing it as its final convergent

$$x = \frac{p_n}{q_n},$$

irreducibitlity of the convergents and the fraction in (2.4.6) yield that

$$\begin{aligned} p_n &= a^* + a' \\ q_n &= 2b^* = a_n\, q_{n-1} + q_{n-2}, \end{aligned} \qquad a_n \geq 2,$$

so that $q_{j-1} < b^*$ for any $1 \leq j \leq n$. The situation is special for $n = 1$ as in this case we can obtain $q_1 = b^*$ via $a_1 = 2$ and thus $b^* = 1$ due to $q_0 = 1$. This is precisely the special case $x = a_0 + \frac{1}{2}$ for which we have

$$|x - (a_0 + 1)| = \frac{1}{2} = |x - a_0|\,,$$

so that the best approximant of the second kind is not unique.

If, on the other hand $n > 1$, we always have $1 \leq q_{n-1} < b^*$ and thus the assumption $a^* \neq a'$ yields, together with (2.4.6), that

$$\begin{aligned} |q_{n-1}\, x - p_{n-1}| &= \left|q_{n-1}\,\frac{p_n}{q_n} - p_{n-1}\right| = \frac{|q_{n-1}\, p_n - p_{n-1}\, q_n|}{q_n} = \frac{1}{q_n} = \frac{1}{2b^*} \\ &< \frac{1}{2} \leq \frac{|a^* - a'|}{2} = b^*\left|x - \frac{a^*}{b^*}\right| = |b^*\, x - a^*|\,, \end{aligned}$$

once more contradicting the assumption that a^*/b^* is best approximant of the second kind. This eventually proves that a^* is ***unique*** and therefore a^*/b^* is a unique best approximant of the second kind to x with minimal denominator. As we have shown in the first half of the proof, the best approximant of the second kind must be a convergent, hence $a^*/b^* = p_m/q_m$ for some $m \leq k$, where k is the order that we fixed in the beginning. If $m = k$ we are done, otherwise two applications of (2.3.10) yield that

$$\frac{1}{q_{k-1} + q_k} \leq \frac{1}{q_m + q_{m+1}} < |q_m\, x - p_m| < |q_k\, x - p_k| \leq \frac{1}{q_{k+1}}$$

[40] Keep in mind the convention of Theorem 2.3.10: the last component of a *finite* continued fraction expansion is not allowed to have the value 1. Hence, $a_n \geq 2$.

hence, replacing k by $k-1$ in the above and using the recursion once more,

$$q_{k-1} + q_{k-2} > q_k = a_k\, q_{k-1} + q_{k-2} \qquad \Rightarrow \qquad a_k < 1,$$

which is a contradiction to the assumption that we consider only continued fractions with positive components. This finally shows that p_k/q_k is a *strict* best approximant of the second kind which automatically makes it unique, except in the aforementioned special case. □

2.5 Approximation order, quantitative statements

Having identified congruents or intermediate fractions as best approximants, depending on the kind of approximation, we next address *quantitative* issues, i.e., the question *how fast* continued fractions converge to a given real number. Of course, this question is only nontrivial for irrational numbers.

In the proof of Theorem 2.3.10, more precisely, in (2.3.14), we already had an upper estimate of the RATE OF APPROXIMATION of the convergents, namely

$$\left| x - \frac{p_n}{q_n} \right| < \frac{1}{q_n^2}.$$

On the other hand, the rational number $a = [0; n, 1, n]$, for which we have the explicit recursion elements

$$\begin{array}{lllll} p_{-1} = 1, & p_0 = 0, & p_1 = 1, & p_2 = 1, & p_3 = n+1 \\ q_{-1} = 0, & q_0 = 1, & q_1 = n, & q_2 = n+1, & q_3 = n(n+2) \end{array}$$

and thus $a = \frac{n+1}{n(n+2)}$, shows that

$$\left| a - \frac{p_1}{q_1} \right| = \left| \frac{p_3}{q_3} - \frac{p_1}{q_1} \right| = \frac{1}{n} - \frac{n+1}{n(n+2)} = \frac{1}{n(n+2)} = \frac{1}{q_1^2\,(1+2/n)},$$

from which we we can already conclude, even if this only a first convergent, that in general an approximation rate better than q_n^{-2} cannot be expected as for any $\varepsilon > 0$ there exists some $n \in \mathbb{N}$ such that $(1+2/n)^{-1} < 1 - \varepsilon$. Nevertheless we should not overestimate the relevance of such worst-case estimates as the next result shows that tells us that at least half of the convergents improve the rate by a factor of 2.

Proposition 2.5.1. *If the number $x \in \mathbb{R}$ has a kth convergent*[41] *then at least one of the following two inequalities holds:*

$$\left| x - \frac{p_{k-1}}{q_{k-1}} \right| < \frac{1}{2\,q_{k-1}^2}, \qquad \left| x - \frac{p_k}{q_k} \right| < \frac{1}{2\,q_k^2}. \tag{2.5.1}$$

Proof: Since x is enclosed by the two convergents, we can once more use to (2.1.8) to conclude that

$$\left| x - \frac{p_{k-1}}{q_{k-1}} \right| + \left| x - \frac{p_k}{q_k} \right| = \left| \frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}} \right| = \frac{1}{q_k\, q_{k-1}}$$

and the inequality between the ARITHMETIC MEAN and the GEOMETRIC MEAN yields that

$$\frac{1}{q_k\, q_{k-1}} = \sqrt{\frac{1}{q_{k-1}^2}\,\frac{1}{q_k^2}} \le \frac{1}{2} \left(\frac{1}{q_{k-1}^2} + \frac{1}{q_k^2} \right)$$

[41] In other words, if x cannot be written as $x = [x_0; x_1, \ldots, x_m]$ for some $m < k$

and thus

$$\left|x - \frac{p_{k-1}}{q_{k-1}}\right| + \left|x - \frac{p_k}{q_k}\right| \leq \frac{1}{2\,q_{k-1}^2} + \frac{1}{2\,q_k^2}$$

so that the inequalities in (2.5.1) cannot be violated simultaneously. □

Therefore at least one of two successive convergents has an approximation rate not only of $1/q_k^2$, but even of $1/\left(2q_k^2\right)$, and this statement even has a converse.

Theorem 2.5.2. *If for $x \in \mathbb{R}$ there exist $a \in \mathbb{Z}$ and $b \in \mathbb{N}$ such that*

$$\left|x - \frac{a}{b}\right| < \frac{1}{2b^2},$$

then a/b is a convergent of the continued fraction expansion of x.

Proof: According to Theorem 2.4.5 it suffices to show that a/b is a best approximant of the second kind. If there would exist $c \in \mathbb{Z}$ and $d \in \mathbb{N}$ such that $|dx - c| < |bx - a| < 1/2b$, then also

$$\left|x - \frac{c}{d}\right| < \frac{1}{2bd},$$

and, since by assumption $a/b \neq c/d$,

$$\frac{1}{bd} \leq \left|\frac{a}{b} - \frac{c}{d}\right| \leq \left|x - \frac{a}{b}\right| + \left|x - \frac{c}{d}\right| < \frac{1}{2b^2} + \frac{1}{2bd} = \frac{b+d}{2b^2 d}.$$

This means that

$$2b < b + d \qquad \Rightarrow \qquad b < d$$

so that a/b is indeed a BEST APPROXIMANT OF SECOND KIND. □

Proposition 2.5.1 can even be improved by considering *three* successive convergents among which one provides an even better rate of approximation.

Theorem 2.5.3. *If $x \in \mathbb{R}$ has a convergent of order $k > 1$ then at least one of the following three inequalities is satisfied:*

$$\left|x - \frac{p_{k-2}}{q_{k-2}}\right| < \frac{1}{\sqrt{5}\,q_{k-2}^2}, \qquad \left|x - \frac{p_{k-1}}{q_{k-1}}\right| < \frac{1}{\sqrt{5}\,q_{k-1}^2}, \qquad \left|x - \frac{p_k}{q_k}\right| < \frac{1}{\sqrt{5}\,q_k^2}. \tag{2.5.2}$$

We now could try to hope for an extension of this process: maybe among four successive convergents we find an even better rate, then consider five and so on. Unfortunately or fortunately, this is not the case and the counterexample is once more the GOLDEN RATIO

$$x = \frac{1+\sqrt{5}}{2} = [1; 1, \dots], \qquad x = 1 + \frac{1}{x},$$

from Example 2.3.14. Since

$$x = [1; 1, \dots, 1, r_k], \qquad r_k = [1; 1, \dots] = x,$$

we also have that

$$x = \frac{x\,p_k + p_{k-1}}{x\,q_k + q_{k-1}} \qquad \Rightarrow \qquad \left|x - \frac{p_k}{q_k}\right| = \frac{1}{(x\,q_k + q_{k-1})\,q_k} = \frac{1}{q_k^2\,(x + q_{k-1}/q_k)}.$$

Now formula (2.1.12) from Proposition 2.1.10 tells us that

$$\frac{q_k}{q_{k-1}} = [a_k; a_{k-1}, \dots, a_1] = [1; 1, \dots, 1] \to x \quad \text{für } k \to \infty;$$

even if the finite continued fraction in the limit is of „forbidden“ form, keep in mind that it is well–defined. Hence,

$$\frac{q_{k-1}}{q_k} = \frac{1}{x} + \varepsilon_k = x - 1 + \varepsilon_k, \qquad \lim_{k\to\infty} \varepsilon_k = 0,$$

and therefore

$$\left| x - \frac{p_k}{q_k} \right| = \frac{1}{q_k^2 (2x - 1 + \varepsilon_k)} = \frac{1}{q_k^2 \left(\sqrt{5} + \varepsilon_k \right)},$$

due to which there cannot be an approximation rate better than $1/\sqrt{5} q_k^2$, regardless of how many successive convergents we consider.

Proof of Theorem 2.5.3: We set

$$\varphi_k := \frac{q_{k-2}}{q_{k-1}}, \qquad \psi_k := r_k + \varphi_k, \qquad k \geq 2,$$

and first prove that

$$k \geq 2,\ \psi_k \leq \sqrt{5},\ \psi_{k-1} \leq \sqrt{5} \qquad \Rightarrow \qquad \varphi_k > \frac{\sqrt{5}-1}{2}. \tag{2.5.3}$$

Since

$$\frac{1}{\varphi_{k+1}} = \frac{q_k}{q_{k-1}} = \frac{a_k\, q_{k-1} + q_{k-2}}{q_{k-1}} = a_k + \frac{q_{k-2}}{q_{k-1}} = a_k + \varphi_k$$

and

$$r_k = [a_k; a_{k+1}, \dots] = a_k + \frac{1}{[a_{k+1}; a_{k+2}, \dots]} = a_k + \frac{1}{r_{k+1}}$$

we obtain that

$$\frac{1}{\varphi_{k+1}} - \varphi_k = a_k = r_k - \frac{1}{r_{k+1}} \qquad \Rightarrow \qquad \frac{1}{\varphi_{k+1}} + \frac{1}{r_{k+1}} = r_k + \varphi_k = \psi_k,$$

so that the assumptions in (2.5.3) yield the inequalities

$$0 \leq r_k + \varphi_k \leq \sqrt{5}, \qquad 0 \leq \frac{1}{\varphi_k} + \frac{1}{r_k} \leq \sqrt{5},$$

which in turn imply

$$5 - \sqrt{5} \left(\varphi_k + \frac{1}{\varphi_k} \right) = \left(\sqrt{5} - \varphi_k \right) \left(\sqrt{5} - \frac{1}{\varphi_k} \right) - 1 \geq \frac{r_k}{r_k} - 1 = 0.$$

Since φ_k is a rational number, equality[42] cannot be assumed in the above estimate and the inequality is a strict one. Multiplying by $\varphi_k/\sqrt{5} > 0$ then yields that

$$0 < \sqrt{5}\, \varphi_k - \varphi_k^2 + 1 = -\left(\frac{\sqrt{5}}{2} - \varphi_k \right)^2 + \frac{1}{4} \qquad \Rightarrow \qquad -\frac{1}{2} < \frac{\sqrt{5}}{2} - \varphi_k < \frac{1}{2}$$

[42] As then $\sqrt{5}$ would be a rational number. Extending the proof of the famous „$\sqrt{2}$ is irrational“ to that case is a nice exercise here.

and therefore

$$\varphi_k > -\frac{1}{2} + \frac{\sqrt{5}}{2} = \frac{\sqrt{5}-1}{2},$$

as claimed in (2.5.3).

After these preliminaries, we can turn to the proof itself. To that end, we assume that

$$\left| x - \frac{p_n}{q_n} \right| \ge \frac{1}{\sqrt{5}\, q_n^2}, \qquad n \in \{k-2, k-1, k\},$$

which implies, together with

$$\begin{aligned} \left| x - \frac{p_n}{q_n} \right| &= \left| \frac{r_{n+1}\, p_n + p_{n-1}}{r_{n+1}\, q_n + q_{n-1}} - \frac{p_n}{q_n} \right| = \frac{1}{q_n\,(r_{n+1}\, q_n + q_{n-1})} = \frac{1}{q_n^2\,(r_{n+1} + q_{n-1}/q_n)} \\ &= \frac{1}{q_n^2\,(r_{n+1} + \varphi_{n+1})} = \frac{1}{q_n^2\, \psi_{n+1}}, \end{aligned}$$

that

$$\psi_n \le \sqrt{5}, \quad n = k-1, k, k+1 \qquad \Rightarrow \qquad \varphi_n > \frac{\sqrt{5}-1}{2}, \quad n = k, k+1,$$

and, eventually,

$$a_k = \frac{1}{\varphi_{k+1}} - \varphi_k < \frac{2}{\sqrt{5}-1} - \frac{\sqrt{5}-1}{2} = \frac{4 - 5 + 2\sqrt{5} - 1}{2\left(\sqrt{5}-1\right)} = 1,$$

which is impossible since $a_k \in \mathbb{N}$. Hence, we obtained a contradiction and the claim must be true. □

Let us summarize: for *arbitrary* real numbers the approximation order of convergents of the continued fraction expansion is bounded, essentially by $1/\sqrt{5}q_n^2$. This worst approximation rate occurs for the golden ratio which makes it the most irrational number in the sense that its approximation order by convergents is worst.

On the other hand, however, there are irrational numbers that can even be approximated arbitrarily well by convergents.

Theorem 2.5.4. *For any function $\varphi : \mathbb{N} \to \mathbb{R}_+$ there exist $x \in \mathbb{R}$, such that for* infinitely many *values $q \in \mathbb{N}$ the inequality*

$$\left| x - \frac{p}{q} \right| < \varphi(q)$$

holds.

Proof: We construct x by means of its continued fraction expansion. To that end, we choose $a_0 \in \mathbb{Z}$ arbitrarily and, in addition,

$$a_{k+1} > \frac{1}{q_k^2 \varphi\,(q_k)}, \qquad k \in \mathbb{N}_0, \tag{2.5.4}$$

which can be done in a lot of ways. Then $x = [a_0; a_1, \dots] \in \mathbb{R}$, and, once again using (2.2.3) from Theorem 2.2.4,

$$\left| x - \frac{p_k}{q_k} \right| < \frac{1}{q_k\, q_{k+1}} = \frac{1}{q_k\,(a_{k+1} q_k + q_{k-1})} < \frac{1}{a_{k+1}\, q_k^2} < \frac{q_k^2\, \varphi\,(q_k)}{q_k^2} = \varphi\,(q_k),$$

which even hold for *any* $k \in \mathbb{N}_0$, so that all convergents converge with rate φ. $\square$

The estimate (2.5.4) that determines a_k already tells us what we have to do in order to obtain a number x such that the covergents approximate quickly, i.e., with a rapidly decaying φ: the components a_k in the continued fraction expansion of x have to grow. This can be derived from the estimate (2.3.10), from which we obtain

$$\begin{aligned} \frac{1}{a_{k+1}\, q_k^2} &> \left| x - \frac{p_k}{q_k} \right| > \frac{1}{q_k\,(q_{k+1}+q_k)} = \frac{1}{q_k\,(a_{k+1}q_k + q_{k-1} + q_k)} \\ &= \frac{1}{q_k^2\,(a_{k+1} + 1 + q_{k-1}/q_k)} > \frac{1}{(a_{k+1}+2)\, q_k^2} \end{aligned} \tag{2.5.5}$$

which implies an approximation order of $\varphi(q_k) \sim 1/a_{k+1}\, q_k^2$. This suggest the conjecture that good approximation order, i.e., fast approximation has to do with some growth of the coefficients. And this is indeed the case since the next result shows that growth is also necessary for a convergence rate better than the worst case[43].

Theorem 2.5.5. *Let $x \in \mathbb{R} \setminus \mathbb{Q}$ be an irrational number. If the coefficients in the continued fraction expansion of x are bounded then there exists $c > 0$ such that*

$$\left| x - \frac{p}{q} \right| < \frac{c}{q^2}, \qquad p \in \mathbb{Z},\ q \in \mathbb{N}, \tag{2.5.6}$$

*has **no** solution. Conversely, if the coefficients are unbounded, then there exist, for any $c > 0$, **infinitely many** solutions of* (2.5.6).

Proof: If $\sup\{a_k : k \in \mathbb{N}_0\} =: M < \infty$, the lower estimate in (2.5.5) yields that

$$\left| x - \frac{p_k}{q_k} \right| > \frac{1}{(M+2)\, q_k^2}, \qquad k \in \mathbb{N}.$$

For an arbitrary irreducible[44] fraction p/q we now choose k such that $q_{k-1} < q \le q_k$ and since all convergents are best approximants of first and second kind to x, it follows that

$$\begin{aligned} \left| x - \frac{p}{q} \right| &\ge \left| x - \frac{p_k}{q_k} \right| > \frac{1}{(M+2)\, q_k^2} = \frac{1}{(M+2)\, q^2} \left(\frac{q}{q_k} \right)^2 > \frac{1}{(M+2)\, q^2} \left(\frac{q_{k-1}}{q_k} \right)^2 \\ &= \frac{1}{(M+2)\, q^2} \left(\frac{q_{k-1}}{a_k q_{k-1} + q_{k-2}} \right)^2 > \frac{1}{(M+2)\, q^2} \left(\frac{1}{a_k + 1} \right)^2 \\ &> \frac{1}{(M+2)(M+1)^2\, q^2} > \frac{c}{q^2}, \end{aligned}$$

where the constant c satisfies

$$c < \frac{1}{(M+2)(M+1)^2},$$

an estimate that depends only on the bound M of the components but not on the denominator q.

If, on the other hand, $\sup\{a_k : k \in \mathbb{N}\} = \infty$, then there exist, for any $c > 0$ infinitely many indices k with $a_{k+1} > 1/c$ and we can apply the upper estimate (2.5.5) directly for

$$\left| x - \frac{p_k}{q_k} \right| < \frac{1}{a_{k+1}\, q_k^2} < \frac{c}{q_k^2}$$

which yields us an infinity of solutions of (2.5.6). $\square$

[43]Which is still pretty good and much better than digit expansions in whatever basis.

[44]This is, of course, no restriction.

2.6 Algebraic numbers

An algebraic number is a ZERO or ROOT of a POLYNOMIAL with rational or integer coefficients. Since we can always multiply a polynomial with rational coefficients by the LEAST COMMON MULTIPLE (LCM) of the denominators, these two concepts are the same — the zeros are not affected if the polynomial is multiplied by a constant.

Definition 2.6.1. A real number $a \in \mathbb{R}$ is called an ALGEBRAIC NUMBER of order n if there exists a polynomial f of degree at most[45] n,

$$f \in \mathbb{Z}[x], \qquad f(x) = \sum_{k=0}^{n} f_k \, x^k, \qquad f_k \in \mathbb{Z}, \ k = 0, \dots, n,$$

such that $f(a) = 0$ and there exists no polynomial[46] g of degree $< n$ with $g(a) = 0$. A real number that is not algebraic is called TRANSCENDENTAL.

Exercise 2.6.1 Show that every rational number is algebraic. **Hint**: this is very easy. ◊

Classical examples for transcendental numbers are e and π, algebraic numbers are $\sqrt{2}$ and the GOLDEN RATIO. Algebraic numbers are computable in the sense that they allow for symbolic computations by adjoining the polynomial to the base field, cf. [12, 43]. In the end, this may lead to expressions containing `RootOf` in the symbolic solution of systems of polynomial equations that are hard to interpret, but at least correct. They are known to all users of COMPUTER ALGEBRA SYSTMES like `Maple` or `Mathematica`.

What is of interest to us here is the fact that algebraic number admit *slow* approximation by continued fraction, a theorem due to Liouville[47] that even relates the order of approximation to the order of the algebraic number.

Theorem 2.6.2 (Liouville). *For any algebraic number $a \in \mathbb{R} \setminus \mathbb{Q}$ of order n there exists a constant $C > 0$*

$$\left| a - \frac{p}{q} \right| > \frac{C}{q^n}, \qquad p \in \mathbb{Z}, \ q \in \mathbb{N}. \tag{2.6.1}$$

Remark 2.6.3. It seems as if the order n of the algebraic number has to be defined uniquely as the minimal degree in order to make the theorem correct. This is not the case, but the most significant result, that is, the „sharpest“ and thus most relevant lower bound, is obtained by taking the *minimal* degree for a polynomial with $f(a) = 0$.

[45] Sometimes „polynomial of degree n“ means degree *at most* n, sometimes it means degree *exactly* n, the latter requiring the additional condition $f_n \neq 0$. This terminology is not really unique in the literature, so be careful.

[46] In other words, we choose f of MINIMAL DEGREE.

[47] Joseph Liouville, 1809–1882, influenced by Ampère, Cauchy and Poisson at the École Polytechnique. Besides number theory he also contributed to differential theory and differential equations, showing a separation between algebra and analysis ist quite artificial. He is also known for the (transcendental) LIOUVILLE NUMER

$$0.110001000000000000000000010000\dots$$

with „1“ exactly a the positions $1!, 2!, 3!, 4!, \dots$. Source for historical information is [31].

Proof: The algebraic number a of order n is a zero of a degree n polynomial $f \in \mathbb{Z}[x]$, and, choosing the degree minimally, we can write f as

$$f(x) = (x-a)\, g(x), \qquad g \in \mathbb{R}[x],\ g(a) \neq 0. \tag{2.6.2}$$

Indeed, if $g(a) = 0$, we can also divide g by $x - a$ to get $f(x) = (x-a)^2\, h(x)$, hence

$$f'(x) = (x-a)\,(2h(x) + (x-a)h'(x)) \qquad \Rightarrow \qquad f'(a) = 0,$$

and since $f' \in \mathbb{Z}[x]$, the number a would be of order (at most) $n-1$. But $g(a) \neq 0$ implies that, by the continuity of polynomials[48], there exists some $\delta > 0$ such that

$$g(x) \neq 0, \qquad x \in [a-\delta, a+\delta]\,. \tag{2.6.3}$$

Let $p \in \mathbb{Z}$ and $q \in \mathbb{N}$ form a fraction close to a, i.e., they are chosen such that

$$|a - p/q| < \delta. \tag{2.6.4}$$

Since δ depends only a, at least if we choose f as the unique MONIC polynomial of minimal degree with $f(a) = 0$, i.e., $f(x) = x^n + \dots$, all sufficiently good approximants to a must satisfy (2.6.4). According to (2.6.3) this implies that $f\,(p/q) \neq 0$ and substituting $x = p/q$ into $x - a = f(x)/g(x)$, see (2.6.2), we obtain

$$\frac{p}{q} - a \;=\; \frac{f\,(p/q)}{g\,(p/q)} = \frac{f_0 + f_1 \frac{p}{q} + \cdots + f_n \left(\frac{p}{q}\right)^n}{g\,(p/q)} = \frac{f_0\, q^n + f_1\, p\, q^{n-1} + \cdots + f_n\, p^n}{q^n\, g(p/q)}.$$

The numerator of this fraction is different from zero since we assumed that a is irrational and thus $a \neq p/q$. Being an integer, the numerator must be ≥ 1 in absolute value[49] and we can conclude that

$$\left| a - \frac{p}{q} \right| \geq \frac{1}{M\, q^n}, \qquad M = \max_{x \in [a-\delta, a+\delta]} |g(x)|, \tag{2.6.5}$$

whenever $|a - p/q| \leq \delta$. If, on the other hand, $|a - p/q| > \delta$, then trivially[50] we also have $|a - p/q| > \delta/q^n$ and for any contant C with

$$C < \min\left\{\delta, \frac{1}{M}\right\},$$

(2.6.1) is satisfied. □

This theorem gives us a simple recipe for the construction of transcendental numbers: use rapidly growing continued fraction expansions. For example, we could use

$$a_{k+1} > q_k^{k-1}, \qquad [a_0; a_1, \dots, a_k] = \frac{p_k}{q_k}$$

as then $a = [a_0; a_1, \dots]$ satisfies, according to (2.5.5),

$$\left| a - \frac{p_k}{q_k} \right| < \frac{1}{a_{k+1}\, q_k^2} < \frac{1}{q_k^{k+1}}$$

[48] We use analysis to prove a statement in algebraic number theory … Which is in fact not uncommon at all.

[49] Integers are very discrete numbers, they are not giving away secrets easily.

[50] Since $q \geq 1$.

which becomes smaller than C/q_k^n for any numbers C and n.

Exercise 2.6.2 Give an explicit continued fraction expansion of a transcendental number. $\diamond$

This is not the fully story about the approximation order for algebraic numbers. Liouville's theorem, Theorem 2.6.2, says that the order of approximation is *at most* q^{-n} for an algebraic number of order n. But this is just a lower bound that decreases faster if the order of the algebraic number is larger. This raises the question whether the decay rate really depends on the order of the algebraic number[51], which can be rephrased as: is there also an *upper* estimate similar to (2.6.1)? To that end, the question was raised whether, given an ALGEBRAIC NUMBER $x \in \mathbb{R} \setminus \mathbb{Q}$,

$$\left| x - \frac{p}{q} \right| < \frac{1}{q^\alpha}, \qquad \alpha > 0, \tag{2.6.6}$$

can occur for *infinitely* many fractions p/q. The constant 1 in (2.6.6) is no real restriction. Indeed, if (2.6.6) is satisfied by infinitely many fractions for some constant $C > 0$, then it is satisfied for $C = 1$ for any $\alpha' < \alpha$. First results were given by Thue in 1908 who showed that if (2.6.6) holds for infinitely many p/q, then $\alpha \le \frac{1}{2}n + 1$, where again n is the order of the algebraic number. In [9] this was even improved to $\alpha \le \sqrt{2n}$, and Siegel conjectured that α were even independent of n. This was finally verified in [40] is the following famous theorem.

Theorem 2.6.4 (Thue-Siegel-Roth). *Let $x \in \mathbb{R} \setminus \mathbb{Q}$ be an irrational algebraic number and $\alpha > 0$. If*

$$\left| x - \frac{p}{q} \right| < \frac{1}{q^\alpha}$$

holds for infinitely many fractions p/q, then $\alpha \le 2$.

A proof of the Thue-Siegel-Roth is beyond what we can do here[52], but there is a simple consequence of it that shows that practically all algebraic numbers have rational approximation of the „worst possible“ sort.

Corollary 2.6.5. *If $x \in \mathbb{R} \setminus \mathbb{Q}$ is an irrational algebraic number and $\varepsilon > 0$, then there exists a constant $C(\varepsilon)$ such that*

$$\left| x - \frac{p}{q} \right| > \frac{C(\varepsilon)}{q^{2+\varepsilon}} \tag{2.6.7}$$

holds for any fraction p/q.

Proof: Theorem 2.6.4 implies that

$$\left| x - \frac{p}{q} \right| < \frac{1}{q^{2+\varepsilon}}$$

only holds for finitely many fractions $p_1/q_1, \dots, p_N/q_N$ and we can simply set

$$C(\varepsilon) := \min_{j=1,\dots,N} q_j^{2+\varepsilon} \left| x - \frac{p_j}{q_j} \right| > 0,$$

[51] Saying, for example that $\sqrt[500]{5}$ can be approximated faster that $\sqrt{5}$.

[52] Even if the paper has only 20 pages and does not appear to rely on too heavy theory. But it is extremely tricky.

to obtain (2.6.7). □

In summary, this shows that any algebraic number can be approximated like $1/q^2$ by rational numbers, independent of its order. They are all equally bad.

Before we leave the world of [28], not considering the measure theoretic aspects of continued fractions[53] given there, we give a final theorem that shows that any PERIODIC CONTINUED FRACTION can be identified with a SQUARE ROOT, i.e., an algebraic number of order 2. To that end, we a consider periodicity in a slightly more generous way, namely as periodicity after a certain index.

Definition 2.6.6. An infinite continued fraction expansion $[a_0; a_1, a_2, \dots]$ is called PERIODIC if there exists an index $k_0 \in \mathbb{N}_0$ and a PERIOD $\ell \in \mathbb{N}$ such that $a_{k+\ell} = a_k$ for all $k \ge k_0$.

Theorem 2.6.7. *Any periodic continued fraction represents an algebraic number of second order and any algebraic number of second order has a periodic continued fraction expansion.*

Proof: If x has a periodic expansion, then also

$$r_{k+\ell} = [a_{k+\ell}; a_{k+\ell+1}, a_{k+\ell+2}, \dots] = [a_k; a_{k+1}, a_{k+2}, \dots] = r_k, \qquad k \ge k_0,$$

holds for some $k_0 \in \mathbb{N}_0$ and some period length $\ell \in \mathbb{N}$. Therefore,

$$x = [a_0; a_1, \dots] = \frac{r_k\, p_{k-1} + p_{k-2}}{r_k\, q_{k-1} + q_{k-2}} = \frac{r_{k+\ell}\, p_{k+\ell-1} + p_{k+\ell-2}}{r_{k+\ell}\, q_{k+\ell-1} + q_{k+\ell-2}} = \frac{r_k\, p_{k+\ell-1} + p_{k+\ell-2}}{r_k\, q_{k+\ell-1} + q_{k+\ell-2}},$$

and thus

$$(r_k\, p_{k-1} + p_{k-2})(r_k\, q_{k+\ell-1} + q_{k+\ell-2}) - (r_k\, q_{k-1} + q_{k-2})(r_k\, p_{k+\ell-1} + p_{k+\ell-2}) = 0,$$

which is a quadratic equation in r_k with integer coefficients. Therefore r_k and consequently also x is an algebraic number of order 2.

The converse is a bit more work. If $x = [a_0; a_1, \dots]$ satisfies

$$ax^2 + bx + c = 0,$$

we again write x as

$$x = \frac{r_k\, p_{k-1} + p_{k-2}}{r_k\, q_{k-1} + q_{k-2}}$$

and obtain that

$$\begin{aligned} 0 &= a\,(r_k\, p_{k-1} + p_{k-2})^2 + b\,(r_k\, p_{k-1} + p_{k-2})\,(r_k\, q_{k-1} + q_{k-2}) + c\,(r_k\, q_{k-1} + q_{k-2})^2 \\ &= A_k\, r_k^2 + B_k r_k + C_k, \end{aligned}$$

where

$$\begin{aligned} A_k &:= a\, p_{k-1}^2 + b\, p_{k-1}\, q_{k-1} + c\, q_{k-1}^2, && (2.6.8) \\ B_k &:= 2a\, p_{k-1}\, p_{k-2} + b\,(p_{k-1}\, q_{k-2} + p_{k-2}\, q_{k-1}) + 2c\, q_{k-1}\, q_{k-2}, && (2.6.9) \\ C_k &:= a\, p_{k-2}^2 + b\, p_{k-2}\, q_{k-2} + c\, q_{k-2}^2 = A_{k-1}. && (2.6.10) \end{aligned}$$

[53] They are interesting and seem to be Khinchin's genuine contribution to the field, but there are other stories to be told. But the booklet is still highly recommendable.

The DISCRIMINANT $D_k = B_k^2 - 4A_kC_k$ has the value

$$D_k = \left(b^2 - 4ac\right) \underbrace{(p_{k-1}\, q_{k-2} - q_{k-1}\, p_{k-2})^2}_{=1} = b^2 - 4ac =: d,$$

indepently of k. Since the discriminant describes the „square root“ part of the number, this is already a good sign. Next, we record that

$$\left| x - \frac{p_{k-1}}{q_{k-1}} \right| < \frac{1}{q_{k-1}^2} \qquad \Rightarrow \qquad p_{k-1} = q_{k-1}\, x + \frac{\delta_{k-1}}{q_{k-1}}, \quad |\delta_{k-1}| < 1,$$

which we can substitute into (2.6.8) to obtain

$$\begin{aligned} A_k &= a \left(q_{k-1}\, x + \frac{\delta_{k-1}}{q_{k-1}} \right)^2 + b\, q_{k-1} \left(q_{k-1}\, x + \frac{\delta_{k-1}}{q_{k-1}} \right) + c\, q_{k-1}^2 \\ &= \underbrace{\left(ax^2 + bx + c \right)}_{=0} q_{k-1}^2 + (2ax + b)\, \delta_{k-1} + a \frac{\delta_{k-1}^2}{q_{k-1}^2}, \\ |A_k| &\le 2|a|\,|x| + |b| + |a| = (2|x| + 1)\; |a| + |b|. \end{aligned}$$

According to (2.6.10) the numbers A_k and $C_k = A_{k-1}$, but also

$$B_k^2 \le D_k + 4\,|A_k|\,|C_k| \le b^2 + 4|a|\,|c| + [(2|x| + 1)\; |a| + |b|]^2$$

are bounded from above, independently of k. Hence, there are only finitely many combinations of (A_k, B_k, C_k) and at least one of them has to repeat after a while. Thus, there exist k, ℓ satisfying $A_{k+\ell} = A_k$, $B_{k+\ell} = B_k$ and $C_{k+\ell} = C_k$, hence also $r_{k+\ell} = r_k$ and by the construction rule for continued fractions, see the proof of Theorem 2.3.10, it also follows that $r_{k+n\ell} = r_k$, $k \in \mathbb{N}$. □

Exercise 2.6.3 Show that if x is an algebraic number of order, then so is $1/x$. ◇

2.7 Continued fractions and music

The last chapter on number theoretic aspects of continued fractions is concerned with a seemingly unrelated topic: MUSIC and the concept of HARMONY in the sence of CONSONANCE.The connections we present here can be found for example in the books [2, 39]. We will see that continued fractions give an answer to the quesion why there are pentatonic scales in „simple“ music, why the OCTACVE consist of 12 semitones[54] and what would be the next partition of an octave into semitones. Let us begin with the fundamental atom of music analysis.

Definition 2.7.1. A TONE with AMPLITUDE FUNCTION $a : \mathbb{R} \to \mathbb{R}$ is a periodic event, i.e., there exist some $T > 0$ such that $a(\cdot + T) = a$.

This model works with an infinite model of a constant tone which excludes melodies so far. If we would consider melodies[55], we would have to involve concepts of TIME-FREQUENCY-ANALYSIS like a GABOR TRANSFORM or a WAVELET TRANSFORM or an INSTANTANEOUS FREQUENCY. These can be found in [33] and we will not dwell with it here. To be percieved in

[54]This is strange since $8 \neq 12 \times \frac{1}{2}$. Nevertheless few people care about this apparent contradiction.
[55]We which will not do here.

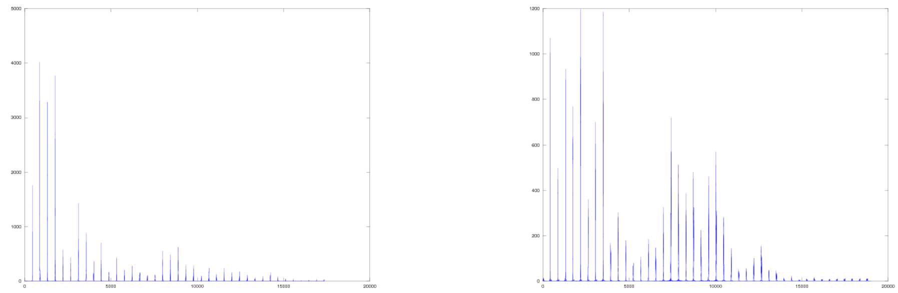


Abbildung 2.7.2: Spectral „fingerprint“ with $|a_k|$ of two bagpipe chanters. Reason which one is louder and sounds more „sharp“.

a melody, a tone would have to be long enough to perform several periods of oscillation which can be seen as the musical version of the HEISENBERG UNCERTAINTY PRINCIPLE.

Since a is a periodic function, which implies that $a\left(\frac{T}{2\pi}\cdot\right)$ is 2π–periodic function that can be considered on the TORUS $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ and has a FOURIER SERIES

$$a\left(\frac{T}{2\pi}\cdot\right) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(k\cdot) + b_k \sin(k\cdot), \qquad \left\{ \begin{array}{c} a_k \\ b_k \end{array} \right\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} a(t) \left\{ \begin{array}{c} \cos kt \\ \sin kt \end{array} \right\} dt.$$

Since the sine is only a phase shift of the cosine and thus physiologically more or less irrelevant, one usually assumes that $b_k = 0$, $k \in \mathbb{N}$, as well as $a_0 = 0$ since a permanent air pressure can be compensated by the environment. Defining the FREQUENCY $\omega = \frac{2\pi}{T}$, our tone can thus be written as

$$a(t) = \sum_{k=1}^{\infty} a_k \cos(k\omega t), \qquad t \in \mathbb{R}. \tag{2.7.1}$$

The $a_k \cos(k\omega\cdot)$ are called PARTIAL TONES of a and their absolute values define the TIMBRE of the tone which depends on and characterizes the instrument, see Fig. 2.7.2.

The second important concept in musical physiology are the BEATS which are an audible version of addition theorem

$$\cos\omega\cdot + \cos\omega'\cdot = 2\cos\frac{\omega+\omega'}{2}\cdot\ \cos\frac{\omega-\omega'}{2}\cdot$$

which says that the sum of two simple tones can be seen[56] as a tone of average frequency $\cos\left(\frac{\omega+\omega'}{2}\cdot\right)$, equipped with an amplitude modulation $\cos\left(\frac{\omega-\omega'}{2}\cdot\right)$. If the two frequencies are close and the difference is small, then these BEATS can very well be perceived, which was actually the way how musical instruments were are are tuned without electronical devices.

This now leads to the concept of consonances and dissonances introduced by Helmholtz [21] which is in fact a property of the partial tones. The maximal consonance is obtained for an OCTAVE which is the simultaneous sound of a and $a(2\cdot)$ as then we get that

$$a(t) + a(2t) = \sum_{k=1}^{\infty} a_k \cos(k\omega t) + \sum_{k=1}^{\infty} a_k \cos(2k\omega t) = \sum_{k=1}^{\infty} \tilde a_k \cos(k\omega t),$$

[56] „Heard“ would be more accurate.

where

$$\tilde{a}_{2k+\epsilon} = \begin{cases} a_{2k+\epsilon}, & \epsilon = 1, \\ a_{2k} + a_k, & \epsilon = 0, \end{cases} \qquad k \in \mathbb{N}_0,\ \epsilon \in \{0,1\},$$

so that we get the same tone, just with a different timbre. The first real consonance is the FIFTH

$$a(t) + a\left(\frac{3}{2}t\right) = \sum_{k=1}^{n} a_k \cos(k\omega t) + \sum_{k=1}^{n} a_{2k} \cos(3k\omega t) + \sum_{k=1}^{n} a_{2k-1} \cos\left(\left(3k - \frac{1}{2}\right)\omega t\right)$$

where half of the partials merge with the fundamental tone and just change the timbre, while the other half of the partials create new tones in the middle between the original partials.

Now, there are complex explanations, see [21] again, to define the following notion of DISSONANCE:

> *Two tones are dissonant if some partials get close to each other and generate perceptible beats.*

Even if they did not have a scientific explanation, the fact itself was already knwon to the Pythagoreans who gave and used the following definition of harmony.

Definition 2.7.2. Two tones with frequencies $\omega < \omega'$ are in HARMONY if $\frac{\omega}{\omega'}$ is a fraction with a small denominator[57].

Example 2.7.3. The octave corresponds to the fraction $\frac{1}{2}$, the fifth to $\frac{2}{3}$. Note that all fractions of the form $\frac{1}{n}$ mean that $\omega' = n\omega$ and therefore all partials merge. In other words, „real“ harmonies have a numerator > 1. We can now predict the next best nonrtivial harmony which has to have denominator 4 and since $\frac{2}{4} = \frac{1}{2}$ the only new choice is $\frac{3}{4}$, the fourth.

The fact that fifth are best possible nontrivial harmonies is the basis for the construction of a SCALE, i.e., a sequence of tones, according to two construction principles:

1. With every tone, it's harmonic relative should be included, i.e., the fifth to the tone.
2. Since octaves are only timbre, we can always go up and down by an octave without really changing the tone.

This construction principle leads to the PYTHAGOREAN SPIRAL of the tones with frequencies $\omega_n := \left(\frac{3}{2}\right)^n \omega$, $n \in \mathbb{Z}$, i.e., to

$$\omega_0 = \omega$$

$$\begin{array}{lcl@{\qquad\qquad}lcl}
\omega_1 & = & \frac{3}{2}\omega & \omega_{-1} & = & \frac{2}{3}\omega \to \frac{4}{3}\omega \\
\omega_2 & = & \frac{9}{4}\omega \to \frac{9}{8}\omega & \omega_{-2} & = & \frac{8}{9}\omega \to \frac{16}{9}\omega \\
\omega_3 & = & \frac{27}{16}\omega \to \frac{27}{32}\omega & \omega_{-3} & = & \frac{32}{27}\omega \\
\omega_4 & = & \frac{81}{64}\omega & \omega_{-4} & = & \frac{64}{81}\omega \to \frac{128}{81}\omega \\
 & \vdots & & & \vdots &
\end{array}$$

[57] And since the fraction is < 1, it also has a small numerator.

where the „→“ indicates that we shifted all tones into the proper octave by normalizing the fractions into the interval $[1,2]$. Using negative steps as well as positive steps has the harmonic advantage that the scale not only considers the fifth but also the fourth.

The name *Pythagorean spiral* reflects the fact that this sequence of tones is infinite and never closes to a circle since $\omega_k = \omega_{k'}$ *modulo octave*[58] would be equivalent to

$$\left(\frac{3}{2}\right)^k \omega = 2^n \left(\frac{3}{2}\right)^{k'} \omega \qquad \Leftrightarrow \qquad 3^{k-k'} = 2^{n+k-k'} \qquad \Leftrightarrow \qquad (k-k')(\log_2 3 - 1) = n \tag{2.7.2}$$

which is impossible except for $k = k'$ and $n = 0$ since 2 and 3 are coprime. But, writing $m := k - k'$ for the „width“ of the scale spanned[59] between ω_k and $\omega_{k'}$, we can replace the right hand condition in (2.7.2) by

$$\min_{m\le M} \min_{n} \left|\log_2\left(\frac{3}{2}\right) - \frac{n}{m}\right| \qquad \text{or} \qquad \min_{m\le M} \min_{n} \left|m \log_2\left(\frac{3}{2}\right) - n\right| \tag{2.7.3}$$

to get the best scale with at most M tones. And the solution of this problem is a best approximant of the first and second kind, respectively, hence a CONVERGENT of the irrational number $\log_2\left(\frac{3}{2}\right)$. Hence, all we have to do is to compute the convergents of this number

```
%%
%% CFconvergent
%% Compute first n components and convergents, return last
%%
function y=CFconvergent( x,n )
  p1 = q0 = 1; q1 = 0; an = floor(x); p0 = an; xx = x-p0;
  printf( "n=0 \t[%d]\t %d / %d \t%f\n",an,p0,q0,abs( x-p0/q0 )*q0^2 );
  y = an;

  for k=1:n
    xx = 1/xx ;
    an = floor( xx );
    xx = xx - an;
    A = [ an 1; 1 0 ] * [ p0,q0 ; p1,q1 ];
    p0 = A(1,1); p1 = A(2,1); q0 = A(1,2); q1 = A(2,2);
    y = [ y,an ];
    printf( "n=%d \t[%d]\t %d / %d \t%f\n",k,an,p0,q0,abs( x-p0/q0 )*(q0^2) );
    if ( xx == 0 ) % Continued fraction computed
      break;
    end
  end
end
```

Abbildung 2.7.3: `CFconvergent.m`: Simple program to compute the first n components and convergents for an arbitrary number.

for which we use a simple `octave` routine `CFconvergent`. This gives us

[58]Do not forget the equivalence relation underlying the construction!

[59]It is only this width that counts, the concrete first tone ω_k only corresponds to a TRANSPOSITION of the scale like C major to G major.

```
>> CFconvergent( log2( 1.5),10 );
n=0     [0]      0 / 1  0.584963
n=1     [1]      1 / 1  0.415037
n=2     [1]      1 / 2  0.339850
n=3     [2]      3 / 5  0.375937
n=4     [2]      7 / 12         0.234600
n=5     [3]      24 / 41        0.678036
n=6     [1]      31 / 53        0.159665
n=7     [5]      179 / 306      0.451282
n=8     [2]      389 / 665      0.041881
n=9     [23]     9126 / 15601   0.409514
n=10    [2]      18641 / 31867  0.334001
```

where the second column shows the components in the continued fraction expansion, the third the convergent and the fourth the error $q_n^2\,|x - p_n/q_n|$ which should be less than $\frac{1}{2}$ for a good and less than $\frac{1}{\sqrt{5}} \approx .44\ldots$ for an exceptional convergent, see Proposition 2.5.1 and Theorem 2.5.3. We thus conclude that the convergents $n = 3$ with a scale of 5 tones, the one for $n = 4$ with a scale of 12 tones and next the one for $n = 6$ with 53 tones are exceptional ones. They correspond to the pentatonic scale, the classical 12 halftone scale and Bosanquet's enharmonic harmonium whose pictures can be found in [2] and various sources over the internet. These three scales can be found by quite simple trial and error, but we also see that the next good one already comprises 665 tones in the scale. This is at least hard for woodwind instruments.

In summary, continued fractions tell us which scales, built on a Pythagorean spiral, hence a sequence of fifth, are almost complete.

Rational functions as continued fractions of polynomials

3

The equations narrowed [...] until they became just a few expressions that appeared to move and sparkle with a life of their own. This was maths without numbers, pure as lightning.

(T. Pratchett, *Men at arms*)

Now it is time to leave continued fractions with integer entries and their role in the representation of real numbers[1] and look at more general situations, in particular rational functions. A RATIONAL FUNCTION is a function of the form

$$f(x) = \frac{p(x)}{q(x)}, \qquad p, q \in \mathbb{K}[x], \tag{3.0.1}$$

i.e., the quotients of polynomials. Note that rational functions are closed under addition, multiplication and division, hence form a FIELD like the rational numbers. To consider rational functions, it is convenient to consider the slightly more general situation of continued fractions over rings. However, we will see that the structure of a EUCLIDEAN RING will be necessary to obtain some desired properties and thus, in the long term, restricts continued fractions to *univariate* Polynomials.

3.1 A beginning with some new notation ...

Finite continued fractions with polynomial components will initially be of the simplified form

$$f(x) = [p; m_1, m_2, \dots, m_n] = p(x) + \cfrac{1}{m_1(x) + \cfrac{1}{m_2(x) + \cfrac{1}{\ddots + \cfrac{1}{m_{n-1}(x) + \cfrac{1}{m_n(x)}}}}},$$

where each component $m_j(x) = a_j\, x^{k_j}$, $a_j \in \mathbb{R}$, $k_j \in \mathbb{N}$, is aMONOMIAL. Such monomial continued fraction are called C–CONTINUED FRACTIONS in [37]. Note that the „1“ appearing

[1]This is not due to lack of interesting questions, for example in solving quadratic diophantine equations of the form $x^2 - Dy^2 = 1$, the so–called PELL EQUATION. The solutions, by the way, are numerators and denominators of the convergents of the continued fraction expansion of $\sqrt{D}$.

in the numerators of the continued fraction is no restriction: a „general“ continued fraction of the form[2]

$$\begin{aligned}
f(x) &= p(x) + \cfrac{b_1}{m_1(x) + \cfrac{b_2}{m_2(x) + \cfrac{b_3}{\ddots + \cfrac{b_{n-1}}{m_{n-1}(x) + \cfrac{b_n}{m_n(x)}}}}} \\
&=: p(x) + \frac{b_1|}{|m_1(x)} + \frac{b_2|}{|m_2(x)} + \cdots + \frac{b_n|}{|m_n(x)}
\end{aligned}$$

can also be written in the form

$$f(x) = [p; \widetilde{m}_1, \ldots, \widetilde{m}_n] = p(x) + \frac{1|}{|\widetilde{m}_1(x)} + \cdots + \frac{1|}{|\widetilde{m}_n(x)},$$

where

$$\widetilde{m}_j(x) = m_j(x) \begin{cases} \prod_{\ell=0}^{k} \dfrac{b_{2\ell}}{b_{2\ell+1}}, & j = 2k+1, \\ \prod_{\ell=0}^{k} \dfrac{b_{2\ell+1}}{b_{2\ell+2}}, & j = 2k+2, \end{cases} \qquad b_0 = 1. \tag{3.1.1}$$

The simplified form (3.1.1) is easily obtained by normalizing the fractions successively which yields

$$\begin{aligned}
f(x) - p(x) &= \frac{b_1|}{|m_1(x)} + \frac{b_2|}{|m_2(x)} + \cdots + \frac{b_n|}{|m_n(x)} \\
&= \frac{1|}{\left|m_1(x)\,\dfrac{1}{b_1}\right.} + \frac{\left.\dfrac{b_2}{b_1}\right|}{|m_2(x)} + \frac{b_3|}{|m_3(x)} + \cdots + \frac{b_n|}{|m_n(x)} \\
&= \frac{1|}{\left|m_1(x)\,\dfrac{1}{b_1}\right.} + \frac{1|}{\left|m_2(x)\,\dfrac{b_1}{b_2}\right.} + \frac{\left.\dfrac{b_1\,b_3}{b_2}\right|}{|m_3(x)} + \cdots + \frac{b_n|}{|m_n(x)} \\
&= \frac{1|}{\left|m_1(x)\,\dfrac{1}{b_1}\right.} + \frac{1|}{\left|m_2(x)\,\dfrac{b_1}{b_2}\right.} + \frac{1|}{\left|m_3(x)\,\dfrac{b_2}{b_1\,b_3}\right.} + \frac{\left.\dfrac{b_2\,b_4}{b_1\,b_3}\right|}{|m_4(x)} + \cdots + \frac{b_n|}{|m_n(x)},
\end{aligned}$$

and so on.

Any continued fraction of the form $[p; m_1, \ldots, m_n]$ is a rational function and, at least for univariate polynomials, any rational function can be expanded into a *finite* continued fraction. We will see this in a more general context soon.

[2]This is a good occasion to introduce another notation for continued fractions, cf. [36, 37].

3.2 Euclidean rings and continued fractions

Let us recall: a RING is a structure in which addition, subtraction and multiplication are well–defined[3] and the structure is closed under these operations. Since we need a little bit more, we have to introduce some more terminology.

Definition 3.2.1 (Euklidean ring). A ring R is called

1. INTEGRAL DOMAIN[4] if there exist no elements $a, b \in R \setminus \{0\}$ such that $ab = 0$. Elements that satisfy this property are called ZERO DIVISOR[5]

2. EUCLIDEAN RING if R is an integral domain and there exists a EUCLIDEAN FUNCTION $d : R \to \mathbb{N} \cup \{-\infty\}$ such that for any $p, q \in R$, $q \neq 0$, there exist a factor $s \in R$ and a REMAINDER $r \in R$ such that we have a DIVISION WITH REMAINDER

$$p = sq + r, \qquad d(r) < d(q). \tag{3.2.1}$$

   We then write $s =: p/q$ and $r =: (p)_q$.

Remark 3.2.2 (Properties of the euclidean functions).

1. Every euclidean function satisfies $d(0) < d(a)$ for all $a \in R \setminus \{0\}$. Assuming that there exists[6] some $a \in R \setminus \{0\}$ with $d(a) \leq d(R)$, then setting $p = q = a$, we get a representation of the form (3.2.1) for a, i.e.,

$$p = sq + r, \quad s \in R, \qquad \Rightarrow \qquad r = p - sq = (1 - s)a.$$

   And regardless of how we choose s each of these remainders would satisfy $d\left((1 - s)r\right) \geq d(a)$ which contradicts the fact that the ring is euclidean.

2. Not any euclidean function has the very natural property

$$d\left(a \cdot b\right) \geq d(a), \qquad a, b \in R \setminus \{0\}, \tag{3.2.2}$$

   that we know from the classical euclidean functions „absolute value“ for $\mathbb{Z}$ and „degree“ for $\mathbb{K}[x]$, but for any integral domain there exists a special euclidean function, called MINIMAL EUCLIDEAN FUNCTION that satisfies (3.2.2). It is defined as the elementwise minimum of all possible euclidean functions, cf. [12, Exercise 3.5]. Thus we could and will always assume that we use the minimal euclidean function and therefore that the euclidean function satisfies (3.2.2).

3. The value $d(a) = -\infty$ can only occur for $a = 0$, but need not be assumed, i.e., $\{a \in R : d(a) = -\infty\} = \emptyset$ is **not** excluded. Indeed, for $R = \mathbb{Z}$ we have $d(0) = 0$ while for $R = \mathbb{K}[x]$ we have $d(0) = -\infty$.

Example 3.2.3.

[3] Including associativity, commutativity and a distributive law that relates the two.

[4] In German „nullteilerfrei“ or „Integritätsring“. The google translation „integrity ring“ of the latter may only earn raised eyebrows among mathematicians.

[5] Hence a ring is an integral domain and if and only if it has no zero divisors.

[6] The function d maps R to $\mathbb{N} \cup \{-\infty\}$, and thus has to have a (possibly nonunique) minimum, i.e., some $r \in R$ such that $d(r) \leq d(R)$, also $d(r) \leq d(q)$, $q \in R$.

1. The integers $\mathbb{Z}$ are a euclidean ring with $d = |\cdot|$.
2. The univariate polynomials $\mathbb{K}[x]$ form a euclidean ring with $d = \deg$, where $\deg 0 = -\infty$.
3. Any field $\mathbb{K}$ is a euclidean ring with $d = (1 - \delta_0)$, however not a very interesting one.
4. A somewhat obscure euclidean function on $\mathbb{Z}$ is $d(3) = 2$ and $d = |\cdot|$ otherwise. This euclidean function is made euclidean by choosing the remainder in $\{-1, 0, 1\}$ when dividing by 3. This euclidean function does **not** satisfy (3.2.2) since $d(-1 \cdot 3) = d(-3) = 3 > 2 = d(3)$. Nevertheless, $d(0)$ is still minimal among all values $d(R)$.

Euclidean rings are useful for an obvious reason: the concept allows us to do division with remainder and the remainders that we obtain this way, are smaller (in the sense of the euclidean function) or „simpler“ than the divisor. And if we recall that division with remainder was one of the fundamental tricks when computing the continued fraction expansions with integer components, it is clear why we insist on euclidean rings: they allow us to transfer the trick almost literally.

Theorem 3.2.4. *Let R be a euclidean ring with one[7]. Than any finite continued fraction $[r_0; r_1, \dots, r_n]$, $r_j \in R$, is rational over R and any rational element over R can be expanded into a continued fraction.*

Definition 3.2.5. The set of all RATIONAL ELEMENTS or FRACTIONS over the commutative ring R with the usal operations for addition, subtraction, multiplication and division will be denoted by

$$R^\star := \left\{ \frac{p}{q} \; : \; p \in R, \; q \in R \setminus \{0\} \right\}.$$

In this notation, $\mathbb{Q} = \mathbb{Z}^\star$, and $R^\star$ is a FIELD if R is an integral domain with one, see [17].

Proof: That finite continued fractions are rational over R can be obtained by expanding the definition or by inductively using the recurrence

$$[r_0; r_1, \dots, r_n] = r_0 + \frac{1}{[r_1; r_2, \dots, r_n]},$$

so this part is quite obvious.

For the converse, let $f = p/q$, $p, q \in R$, $q \neq 0$. Wir set $s_0 = p$, $s_1 = q$ and run the EUCLIDEAN ALGORITHM. To that end, we determine r_0 such that $s_0 = r_0 \, s_1 + s_2$, $d\,(s_2) < d\,(s_1)$, which is possible since we are working in a euclidean ring. For $j = 1, 2, \dots$ we proceed the same way and form

$$s_j = r_j \, s_{j+1} + s_{j+2}, \qquad d\left(s_{j+2}\right) < d\left(s_{j+1}\right),$$

to conclude by induction on k that

$$\frac{p}{q} = \left[r_0; r_1, \dots, r_k, \frac{s_{k+1}}{s_{k+2}} \right], \qquad k \in \mathbb{N}. \tag{3.2.3}$$

Indeed,

$$\left[r_0; \frac{s_1}{s_2} \right] = r_0 + \frac{s_2}{s_1} = \frac{r_0 \, s_1 + s_2}{s_1} = \frac{s_0}{s_1} = \frac{p}{q}$$

[7] This means that there exists a unique neutral element of multiplication, written as „1“.

and because of
$$r_k + \frac{s_{k+2}}{s_{k+1}} = \frac{r_k\, s_{k+1} + s_{k+2}}{s_{k+1}} = \frac{s_k}{s_{k+1}}$$
we also get
$$\left[r_0; r_1, \dots, r_k, \frac{s_{k+1}}{s_{k+2}}\right] = \left[r_0; r_1, \dots, r_k + \frac{s_{k+2}}{s_{k+1}}\right] = \left[r_0; r_1, \dots, r_{k-1}, \frac{s_k}{s_{k+1}}\right] = \frac{p}{q},$$
which proves (3.2.3). Since $d(s_k)$ is a strictly decreasing sequence in $\mathbb{N}_0 \cup \{-\infty\}$, this procedure has to terminate after finitely many steps any give us a finite continued fraction. □

This, of course, was not extremely surprising so far since already the name indicates that *euclidean* ring and *euclidean* algorithm may have something in common and should fit together. But it is getting even better if we assume that R is a COMMUTATIVE RING with (multiplicative) IDENTITY 1. Then the recurrence relation of Theorem 2.1.4 can simply be copied, leading to a lot of interesting formulas for CONVERGENTS or, as the are called NÄHERUNGSBRÜCHE[8] in [36]. The proofs of the preceding chapter can now be transferred literally to the setting of rational elements over arbitrary euclidean rings and can be summarized as follows.

Theorem 3.2.6. *The convergents $\kappa_k := p_k/q_k$, $k \le n$, of the finite continued fraction $[a_0; a_1, \dots, a_n]$, $a_j \in R$, fulfill the recurrence relations*[9]
$$\begin{array}{rcl} p_k &=& a_k p_{k-1} + p_{k-2} \\ q_k &=& a_k q_{k-1} + q_{k-2} \end{array}, \qquad \begin{array}{ll} p_{-1} = 1, & p_0 = r_0, \\ q_{-1} = 0, & q_0 = 1, \end{array} \tag{3.2.4}$$
as well as
$$\frac{p_{k-1}}{q_{k-1}} - \frac{p_k}{q_k} = \frac{(-1)^k}{q_{k-1}\, q_k}, \qquad \frac{p_k}{q_k} - \frac{p_{k-2}}{q_{k-2}} = \frac{(-1)^k\, a_k}{q_{k-2}\, q_k}, \tag{3.2.5}$$
and thus are COPRIME[10].

But continued fractions give us even more[11]. Whenever the recursion in the proof of Theorem 3.2.4 stops, which means $s_{n+2} = 0$, then we have computed a greatest common divisor, cf. [12, 43]. In ohter words, $r_n = \gcd(p, q)$ and since the components $p_n = p/r_n$, $q_n = q/r_n$ of the convergent are coprime[12], we have that
$$\frac{p}{q} = [r_0; r_1, \dots, r_n] = \frac{p_n}{q_n} = \frac{r_n\, p_{n-1} + p_{n-2}}{r_n\, q_{n-1} + q_{n-2}},$$
hence, using (3.2.5),
$$q_{n-1} p - p_{n-1} q \;=\; r_n \left(q_{n-1}\, p_n - p_{n-1} q_n\right) = (-1)^{n+1}\, r_n = (-1)^{n+1} \gcd(p, q).$$

[8] Probably best translated as *approximating fractions*.

[9] Conveniently initialized by $\kappa_{-1} = 0/1 = 0$, i.e., $p_{-1} = 0$ and $q_{-1} = 1$.

[10] Two elements $p, q \in R$ of a commutative ring R with identity are called COPRIME if $p \in q\, R^\times$ where $R^\times = \left\{ r \in R \;:\; r^{-1} \in R \right\}$ denotes the UNITS in R. The units of $\mathbb{Z}$ are $\mathbb{Z}^\times = \{\pm 1\}$, the units among the polynomials $\mathbb{K}[x]$ are $\mathbb{K}[x]^\times = \mathbb{K}^\times = \mathbb{K} \setminus \{0\}$ which means that not all units are identities, not even in absolute value.

[11] I am grateful to H. M. Möller who pointed out that fact to me.

[12] Everything that can be divided off is found in the gcd.

In other words, numerator and denominator of the PENULTIMATE CONVERGENT which is the last „real“ convergent[13] are the solutions of the BÉZOUT IDENTITY

$$a\,p + b\,q = \gcd(p, q) \qquad \Leftrightarrow \qquad a = (-1)^{n+1} q_{n-1}, \quad b = (-1)^n p_{n-1}. \tag{3.2.6}$$

This is no new observation as the EXTENDED EUCLIDEAN ALGORITHM is well-known to compute such a solution. But still it is a very nice and also useful connection.

3.3 One result of one Bernoulli

It is a quite natural question for continued fractions on arbitrary euclidean rings like polynomials which rational objects can be convergents; of course, we consider here the full sequence of convergents since for any $p < q$ the first convergent of

$$[0; a, b] = \frac{1}{a + \frac{1}{b}} = \frac{b}{a+b}$$

equals $\frac{p}{q}$ as soon as $a = q - p$ and $b = p$. Hence, any rational number is a convergent of some continued fraction. So the question is:

> For which sequences $c_n \in R^\star$ does there exist a continued fraction which has this sequence as sequence of convergents?

According to [37] this question was already answered in 1775 by D. Bernoulli[14] in [3], and this for continued fractions of the quite general form

$$r_0 + \frac{s_1|}{|r_1} + \frac{s_2|}{|r_2} + \cdots + \frac{s_n|}{|r_n}, \qquad r_j, s_j \in R^\star \setminus \{0\}. \tag{3.3.1}$$

Theorem 3.3.1 (D. BERNOULLI). *A sequence $c_n \in R^\star$ has a continued fraction expansion as*

$$c_n = r_0 + \frac{s_1|}{|r_1} + \frac{s_2|}{|r_2} + \cdots + \frac{s_n|}{|r_n}, \qquad r_j, s_j \in R^\star \setminus \{0\},$$

if and only if $c_{n+1} \neq c_n$ ist, $n \in \mathbb{N}_0$. In this case, the coefficients can be given explicitly as

$$r_n = \frac{1}{q_{n-1}} \frac{c_n - c_{n-2}}{c_{n-2} - c_{n-1}}, \qquad s_n = \frac{1}{q_{n-2}} \frac{c_{n-1} - c_n}{c_{n-2} - c_{n-1}}. \tag{3.3.2}$$

Proof: The proof is based on a RECURRENCE RELATION for the convergents

$$\frac{p_k}{q_k} = r_0 + \frac{s_1|}{|r_1} + \frac{s_2|}{|r_2} + \cdots + \frac{s_k|}{|r_k}, \qquad k \in \mathbb{N}_0,$$

of continued fractions of the form (3.3.1). This recurrence,

$$\begin{array}{rcl} p_k &=& r_k\, p_{k-1} + s_k\, p_{k-2} \\ q_k &=& r_k\, q_{k-1} + s_k\, q_{k-2} \end{array}, \qquad \begin{array}{ll} p_{-1} = 1, & p_0 = r_0 \\ q_{-1} = 0, & q_0 = 1, \end{array} \tag{3.3.3}$$

[13] The last convergent is the fraction itself and thus not really an approximation.

[14] DANIEL BERNOULLI, 1700-1782, son of JOHANN BERNOULLI, brother of NICOLAUS II BERNOULLI and nephew of JACOB BERNOULLI, thus right in the middle of the famous Bernoulli clan. Althought his father originally wanted him to become a merchant, he obtained a doctoral degree in *medicine* on the mechanics of breathing. Besides mathematics and physics he also worked on applications of these sciences in medicine.

is obtained in the same as (2.1.6) in Theorem 2.1.4 by induction on k; the case $k = 0$ is simply the definition of p_0 and q_0 while $k = 1$ is obtained by a straightforward computation:

$$r_0 + \frac{s_1|}{|r_1} = r_0 + \frac{s_1}{r_1} = \frac{r_0\, r_1 + s_1}{r_1} = \frac{r_1\, p_0 + s_1\, p_{-1}}{r_1\, q_0 + s_1\, q_{-1}}.$$

For the inductive step $k \to k+1$ we again set

$$\frac{p_k'}{q_k'} = r_1 + \frac{s_2|}{|r_2} + \cdots + \frac{s_{k+1}|}{|r_{k+1}},$$

which immediately yields

$$\frac{p_{k+1}}{q_{k+1}} = r_0 + \cfrac{s_1}{r_1 + \frac{s_2|}{|r_2} + \cdots + \frac{s_{k+1}|}{|r_{k+1}}} = r_0 + \frac{s_1\, q_k'}{p_k'} = \frac{r_0\, p_k' + s_1\, q_k'}{p_k'}$$

and the shifted induction hypothesis then gives

$$\begin{aligned} p_{k+1} &= r_0 \left(r_{k+1} p_{k-1}' + s_{k+1}\, p_{k-2}' \right) + s_1 \left(r_{k+1} q_{k-1}' + s_{k+1}\, q_{k-2}' \right) \\ &= r_{k+1} \left(r_0\, p_{k-1}' + s_1\, q_{k-1}' \right) + s_{k+1} \left(r_0\, p_{k-2}' + s_1\, q_{k-2}' \right) = r_{k+1}\, p_k + s_{k+1}\, p_{k-1} \\ q_{k+1} &= p_k' = r_{k+1}\, p_{k-1}' + s_{k+1}\, p_{k-2}' = r_{k+1}\, q_k + s_{k+1}\, q_{k-1}, \end{aligned}$$

which proves (3.3.3). Multiplying the first line by $-q_{k-1}$, the second one by p_{k-1} and adding everything, we get that

$$\begin{aligned} p_{k-1}\, q_k - p_k\, q_{k-1} &= r_k \left(-p_{k-1}\, q_{k-1} + p_{k-1}\, q_{k-1} \right) - s_k \left(p_{k-2}\, q_{k-1} - p_{k-1}\, q_{k-2} \right) \\ &= -s_k \left(p_{k-2}\, q_{k-1} - p_{k-1}\, q_{k-2} \right) = s_k\, s_{k-1} \left(p_{k-3}\, q_{k-2} - p_{k-2}\, q_{k-3} \right) \\ &= \cdots = (-1)^k \prod_{j=1}^k s_j \left(p_{-1}\, q_0 - p_0\, q_{-1} \right), \end{aligned}$$

hence

$$p_{k-1}\, q_k - p_k\, q_{k-1} = (-1)^k \prod_{j=1}^k s_j. \tag{3.3.4}$$

This already gives one direction of our theorem: if c_n, $n \in \mathbb{N}$, is a sequence of convergents, then

$$c_n - c_{n-1} = \frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = \frac{(-1)^{n+1}\, s_1 \cdots s_n}{q_n\, q_{n-1}} \neq 0,$$

since $s_j \neq 0$ for all j was assumed[15]

For the converse we use the recurrence (3.3.3) to obtain

$$c_n = \frac{p_n}{q_n} = \frac{r_n\, p_{n-1} + s_n\, p_{n-2}}{r_n\, q_{n-1} + s_n\, q_{n-2}} \qquad \Leftrightarrow \qquad \begin{bmatrix} p_n \\ q_n \end{bmatrix} = \begin{bmatrix} p_{n-1} & p_{n-2} \\ q_{n-1} & q_{n-2} \end{bmatrix} \begin{bmatrix} r_n \\ s_n \end{bmatrix}$$

which can be solved ***uniquely*** for r_n, s_n since

$$\begin{aligned} \det \begin{bmatrix} p_{n-1} & p_{n-2} \\ q_{n-1} & q_{n-2} \end{bmatrix} &= p_{n-1}\, q_{n-2} - p_{n-2} q_{n-1} = q_{n-1}\, q_{n-2} \left(\frac{p_{n-1}}{q_{n-1}} - \frac{p_{n-2}}{q_{n-2}} \right) \\ &= q_{n-1}\, q_{n-2} \left(c_{n-1} - c_{n-2} \right) \neq 0, \end{aligned}$$

[15] It is immediate from Definition (3.3.1), that any continued fraction with $s_k = 0$, $k \leq n$, is a continued fraction of length $k - 1 < n$ for which all convergents beyond the kth coincide.

dure to our assumption on the c_k and by induction on q_k, $k = n-1, n-2$, respectively.
CRAMER'S RULE now implies that

$$r_n = \frac{\det\begin{bmatrix} p_n & p_{n-2} \\ q_n & q_{n-2} \end{bmatrix}}{\det\begin{bmatrix} p_{n-1} & p_{n-2} \\ q_{n-1} & q_{n-2} \end{bmatrix}} = \frac{q_n\, q_{n-2}\,(c_n - c_{n-2})}{q_{n-1}\, q_{n-2}\,(c_{n-1} - c_{n-2})} = \frac{q_n}{q_{n-1}} \frac{c_n - c_{n-2}}{c_{n-1} - c_{n-2}}$$

$$s_n = \frac{\det\begin{bmatrix} p_{n-1} & p_n \\ q_{n-1} & q_n \end{bmatrix}}{\det\begin{bmatrix} p_{n-1} & p_{n-2} \\ q_{n-1} & q_{n-2} \end{bmatrix}} = \frac{q_n\, q_{n-1}\,(c_{n-1} - c_n)}{q_{n-1}\, q_{n-2}\,(c_{n-1} - c_{n-2})} = \frac{q_n}{q_{n-2}} \frac{c_{n-1} - c_n}{c_{n-1} - c_{n-2}}.$$

Replacing r_n, s_n by $r_n' = a\, r_n$, $s_n' = a\, s_n$ for an arbitrary $a \in R \setminus \{0\}$, we still have

$$\frac{p_n'}{q_n'} = \frac{a\, p_n}{a\, q_n} = \frac{p_n}{q_n} = c_n,$$

where we only have to set $a = 1/q_n$ to end up with(3.3.2). ☐

The last remark in the proof returns us to teh normalized continued fractions $[r_0; r_1, \dots, r_n]$ where $s_1 = \cdots = s_n = 1$. Indeed, settin $a = 1/s_n$ in the above division argument, we obtain

$$r_n' = \frac{q_{n-2}}{q_{n-1}} \frac{c_n - c_{n-2}}{c_{n-1} - c_n}, \qquad s_n' = 1$$

and thus an expansion in the „old", slightly more restrictive form $[a_0; a_1, \dots]$ of a continued fraction.

Corollary 3.3.2 (Normalized BERNOULLI). *If the sequence $c_n \in R^\star$, $n \in \mathbb{N}_0$, satisfies $c_n \neq c_{n-1}$, then*

$$c_n = [r_0; r_1, \dots, r_n], \qquad n \in \mathbb{N}_0,$$

where

$$r_n = \frac{q_{n-2}}{q_{n-1}} \frac{c_n - c_{n-2}}{c_{n-1} - c_n}, \qquad n \geq 2, \qquad r_{-1} = 0,\ r_0 = c_0,\ r_1 = \frac{1}{c_1 - c_0}. \tag{3.3.5}$$

Proof: We can obtain (3.3.5) directly from (3.2.5) if we solve for proper terms taking into account the assumption $c_n = \frac{p_n}{q_n}$:

$$c_{n-1} - c_n = \frac{(-1)^n}{q_{n-1}\, q_n} \qquad \Rightarrow \qquad q_n = \frac{(-1)^n}{q_{n-1}\,(c_{n-1} - c_n)}; \tag{3.3.6}$$

and

$$c_n - c_{n-2} = \frac{(-1)^n\, r_n}{q_{n-2}\, q_n}. \tag{3.3.7}$$

Solving (3.3.7) for r_n and substituting (3.3.6), finally get

$$r_n = (-1)^n\, q_{n-2} q_n\, (c_n - c_{n-2}) = \frac{q_{n-2}}{q_{n-1}} \frac{c_n - c_{n-2}}{c_{n-1} - c_n},$$

which is (3.3.5). ☐

Remark 3.3.3 (Continued fraction expansions).

1. The above observation shows that in $R^\star$ the continued fraction expansion (3.3.1) is *not* unique in general, mainly because R can have too many units. Recall that, for exmple, in the polynomial ring $\mathbb{K}[x]$ the units consist of $\mathbb{K} \setminus \{0\}$. This leads to the notion of EQUIVALENT CONTINUED FRACTIONS: two continued fractions are called EQUIVALENT if all their convergents coincide.

2. The continued fraction expansion from Corollary 3.3.2, that is, the one with $s_n = 1$, $n \in \mathbb{N}$, plays a particular role in its equivalent family of continued fractions[16]: they are those continued fraction expansion where the components of the convergent, formed by the RECURRENCE RELATION, are IRREDUCIBLE, i.e., those where the convergent is in normalized form. This follows immediately from (3.2.5), the argument is exactly the same as in Theorem 2.3.4.

3. In general continued fraction expansions, common divisors of numerator and denominator cannot be excluded any more, see (3.3.4).

With the help of Bernoulli's theorem, we now can compute continued fraction expansions of a POWER SERIES which is the counterpiece to a real number in the world of rational functions. Let us study this by means of an example.

Example 3.3.4. The EXPONENTIAL FUNCTION $f(x) = e^x$ has the power series expansion

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots = \sum_{j=0}^{\infty} \frac{x^j}{j!},$$

and we can determine the continued fraction expansions of the PARTIAL SUM

$$\sum_{j=0}^{n} \frac{x^j}{j!} =: c_n = [r_0; r_1, \ldots, r_n], \qquad r_0, \ldots, r_n \in \mathbb{K}[x], \qquad n \in \mathbb{N}.$$

According to Corollar 3.3.2 this is possible since $c_n - c_{n-1} = \frac{x^n}{n!} \neq 0$, where for a polynomial $p \neq 0$ means that the polynomial is not the neutral element of addition in the ring which is the ZERO POLYNOMIAL. The first two values $r_0 = 1$, $r_1 = 1/x$ and therefore also[17] $q_0 = 1$, $q_1 = 1/x$ yield together with

$$\frac{c_n - c_{n-2}}{c_{n-1} - c_n} = -\left(1 + \frac{n}{x}\right)$$

[16] That is, the equivalence class modulo the above equivalence relation of coinciding convergents, to be formal.

[17] Keep in mind that whenever the r_j are RATIONAL, the same holds true for *numerator* and *denominator* of the convergents.

the values

$$
\begin{aligned}
r_2 &= -\frac{1}{1/x}\left(1+\frac{2}{x}\right) = -(x+2) & q_2 &= r_2\, q_1 + q_0 = -\frac{x+2}{x} + 1 = 2x^{-1} \\
r_3 &= \frac{1}{2} + \frac{3}{2}x^{-1} & q_3 &= -3x^{-2} \\
r_4 &= -\frac{2}{3}x - \frac{8}{3} & q_4 &= 8x^{-2} \\
r_5 &= \frac{3}{8} + \frac{15}{8}x^{-1} & q_5 &= 15x^{-3} \\
r_6 &= -\frac{8}{15}x - \frac{48}{15} & q_6 &= -48x^{-3} \\
r_7 &= \frac{5}{16} + \frac{35}{16}x^{-1} & q_7 &= -105x^{-4} \\
r_8 &= -\frac{16}{35}x - \frac{128}{35} & q_9 &= 384x^{-4}
\end{aligned}
$$

and so on. It would be a bit nicer for $f(x) = e^{1/x}$ when x is replaced by x^{-1}.

The example already shows that the „natural environment“ for continued fractions might be the ring of LAURENT POLYNOMIALS, i.e., all finite sums

$$
f(x) = \sum_{k\in\mathbb{Z}} f_k\, x^k, \qquad \#\{k : f_k \neq 0\} < \infty.
$$

But note that although any Laurent polynomial can be written as $f(x) = x^{-k}\, p(x)$, $k \in \mathbb{N}_0$, $p \in \mathbb{K}[x]$, the ring has a completely different structure: all nonzero multiples of monomial are know units,

$$
\left(cx^k\right)^{-1} = c^{-1}\, x^{-k}, \qquad c \in \mathbb{K} \setminus \{0\}, \quad k \in \mathbb{Z},
$$

and therefore the ring is generated by units as a vectors space wich already implies that the notion of degree is impossible here.

The method of Example 3.3.4 can be generalized into a general equivalence between continued fractions and series over $R^\star$. More precisely, we use the following concept, which is due to Seidel [53].

Definition 3.3.5. A series $c_0 + c_1 + \cdots$, $c_j \in R^\star \setminus \{0\}$, and a continued fraction $r_0 + \frac{s_1|}{|r_1} + \cdots$, $r_j, s_j \in R^\star \setminus \{0\}$ are called EQUIVALENT if

$$
\sum_{j=0}^{n} c_j = \frac{p_n}{q_n} = r_0 + \frac{s_1|}{|r_1} + \cdots + \frac{s_n|}{|r_n}, \qquad n \in \mathbb{N}. \tag{3.3.8}
$$

Then any series has an equivalent continued fraction expansion and vice versa and the conversion is explicit.

Theorem 3.3.6 (Euler). *The continued fraction* $r_0 + \frac{s_1|}{|r_1} + \cdots$ *and the series*

$$
\sum_{n=0}^{\infty} \frac{(-1)^{n+1}}{q_{n-1}\, q_n} \prod_{j=1}^{n} s_j \tag{3.3.9}
$$

and the series $c_0 + c_1 + \cdots$ and the continued fraction

$$c_0 + \frac{c_1|}{|1} - \frac{\frac{c_2}{c_1}\Big|}{\Big|1 + \frac{c_2}{c_1}} - \cdots - \frac{\frac{c_j}{c_{j-1}}\Big|}{\Big|1 + \frac{c_j}{c_{j-1}}} - \cdots \tag{3.3.10}$$

are equivalent.

Proof: Equivalence is equivalent[18] to $c_0 = r_0$ and

$$c_n = \sum_{j=0}^{n} c_j - \sum_{j=0}^{n-1} c_j = \frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}} = \frac{p_n\, q_{n-1} - p_{n-1}\, q_n}{q_{n-1}\, q_n} = \frac{(-1)^{n+1}}{q_{n-1}\, q_n} \prod_{j=1}^{n} s_j, \qquad n \ge 1,$$

due to (3.3.4), from which (3.3.9) follows. For the converse, we apply Theorem 3.3.1 to the sequence

$$a_n = \sum_{j=0}^{n} c_j, \qquad n \in \mathbb{N}_0,$$

which satisfies the conditions of the theorem since $c_j \neq 0$. Then,

$$r_n = \frac{1}{q_{n-1}} \frac{a_n - a_{n-2}}{a_{n-2} - a_{n-1}} = -\frac{1}{q_{n-1}} \frac{c_n + c_{n-1}}{c_{n-1}} = -\frac{1}{q_{n-1}} \left(1 + \frac{c_n}{c_{n-1}}\right) \tag{3.3.11}$$

and

$$s_n = \frac{1}{q_{n-2}} \frac{a_{n-1} - a_n}{a_{n-2} - a_{n-1}} = \frac{1}{q_{n-2}} \frac{c_n}{c_{n-1}}, \tag{3.3.12}$$

hence, by the recurrence (3.3.3),

$$q_n = r_n\, q_{n-1} + s_n\, q_{n-2} = -\frac{q_{n-1}}{q_{n-1}} \left(1 + \frac{c_n}{c_{n-1}}\right) + \frac{q_{n-2}}{q_{n-2}} \frac{c_n}{c_{n-1}} = -\left(1 + \frac{c_n}{c_{n-1}}\right) + \frac{c_n}{c_{n-1}} = -1$$

for any $n \in \mathbb{N}$. Resubstituting this into (3.3.11) and (3.3.12), respectively, gives

$$r_n = 1 + \frac{c_n}{c_{n-1}}, \qquad s_n = -\frac{c_n}{c_{n-1}}, \qquad n \in \mathbb{N}, \tag{3.3.13}$$

and verifies the equivalence to (3.3.10). □

Exercise 3.3.1 Compute the (non-normalized) continued fraction expansion for an arbitrary power series and especially for $f(x) = e^x$. ◊

Remark 3.3.7. Computing the equivalent representation for a power series is nice, but in the next section we will see that, at least in some cases, we can do better and determine continued fractions whose convergents cover more coefficients of a given Laurent series.

[18]This nice statement of course abbreviates *„equivalence of series and continued fraction is equivalent to"*.

3.4 Orthogonal polynomials, continued fractions and Gauß

In this chapter we will have a look at the close connection between continued fractions and orthogonal polynomials which, is essentially a consequence of the THREE TERM RECURRENCE (3.2.4) common to both concepts. This relationship was used by GAUSS in his original development of the so called GAUSS QUADRATURE which is a fundamental concept in numerics, more precisely is numerical integration, see [14, 26, 44]. The second tool used by Gauß was to expand a certain series in terms of continued fraction and by means of Bernoulli's theorem, hence following precisely the way of the preceding chapter.

We now get more specific than in the preceding chapters and explicitly consider the ring $R = \Pi = \mathbb{R}[x]$ of univariate polynomials with *real* coefficients[19] as well as the VECTOR SPACE

$$\Pi_n = \mathrm{span}\,\{1, x, \dots, x^n\} = \{f \in \Pi \ : \ \deg f \le n\}$$

of all polynomials of DEGREE at most n, $n \in \mathbb{N}$. What we also need is an INNER PRODUCT that induces the notion of orthogonality.

Definition 3.4.1. A bilinear form

$$\langle \cdot, \cdot \rangle \ : \ \Pi \times \Pi \to \mathbb{R},$$

on Π is called INNER PRODUCT if it is SYMMETRIC, $\langle f, g\rangle = \langle g, f\rangle$, and DEFINITE, i.e.

$$\langle f, f\rangle > 0, \qquad f \ne 0.$$

We want this inner product to be induced by a SQUARE POSITIVE LINEAR FUNCTIONAL, i.e., $\langle f, g\rangle = L(fg)$, where

$$L \ : \ \Pi \to \mathbb{R}, \qquad L\left(f^2\right) > 0, \quad f \in \Pi. \tag{3.4.1}$$

Exercise 3.4.1 Show that any square positive functional defines an inner product. Yes, this is easy. ◊

Remark 3.4.2. The most popular and standard case of a square positive linear functional is of course the INTEGRAL

$$L(f) = \int_0^1 f(x)\,dx.$$

Definition 3.4.3 (Moments).

1. The nth MOMENT of the inner product $\langle \cdot, \cdot\rangle$ is defined as

$$\mu_n = L\left((\cdot)^n\right) = \langle 1, (\cdot)^n\rangle, \qquad n \in \mathbb{N}; \tag{3.4.2}$$

together, the moments define the MOMENT SEQUENCE $(\mu_n : n \in \mathbb{N})$.

2. The MOMENT MATRIX is the biinfinite matrix

$$M = \left[\left\langle (\cdot)^j, (\cdot)^k \right\rangle \ : \ j,k \in \mathbb{N}_0\right] = \left[\mu_{j+k} \ : \ j,k \in \mathbb{N}_0\right]. \tag{3.4.3}$$

which represents an operator acting on real valued sequences.

[19] „Real“ actually makes the problem a little more complex than one might originally think. But there's a lot of magic in Gauß quadrature, probably because continued fractions are hidden somewhere.

3. A matrix A of the form $a_{j,k} = a_{j+k}$ is called a HANKEL MATRIX or HANKEL OPERATOR for the sequence $a = (a_n \,:\, n \in \mathbb{N})$.

Of course, the simplest way of obtaining square positive functionals is to chose $a, b \in \mathbb{R}$, $a \le b$ and $w : [a, b] \to \mathbb{R}$ as a nonzero, nonnegative (continous[20]) function, and to set

$$L(f) := \int_a^b f(x)\, w(x)\, dx, \qquad f \in C[a, b]. \tag{3.4.4}$$

However, in order to emphasize the algebraic approach here, we will avoid such explicit representations of the square positive linear functional and focus on moment sequences only.

Exercise 3.4.2 Show that L from (3.4.4) is square positive. Easy again. ◊

Remark 3.4.4. A natural question is which sequences μ_n can be moment sequences and how to recover L or maybe even a, b and w from the moment sequence. Questions of this type are known as MOMENT PROBLEM and there is a substantial literature on it, cf. [11].

On Π inner products induced by square positive functionals and moment matrices are easily seen to be equivalent. Of course, any inner product defines a MOMENT MATRIX and conversely, for any two polynomials

$$f(x) = \sum_{j=0}^n f_k\, x^k, \qquad g(x) = \sum_{j=0}^n g_k\, x^k, \qquad n = \max\{\deg f, \deg g\},$$

we simple get

$$\langle f, g\rangle = f^T M_n g = [f_0, \dots, f_n] \begin{bmatrix} \mu_0 & \mu_1 & \dots & \mu_n \\ \mu_1 & \mu_2 & \iddots & \vdots \\ \vdots & \vdots & \iddots & \mu_{2n-1} \\ \mu_n & \mu_{n-1} & \dots & \mu_{2n} \end{bmatrix} \begin{bmatrix} g_0 \\ \vdots \\ g_n \end{bmatrix},$$

where the Hankel structure ensures that $\langle f, g\rangle = L(fg)$.

Exercise 3.4.3 Show that whenever L is square positive, then M_n is a symmetric, strictly positive definite matrix for any $n \in \mathbb{N}_0$. ◊

Definition 3.4.5. A sequence $f_n \in \Pi_n \setminus \{0\}$, $n \in \mathbb{N}$, of nonzero polynomials is called SEQUENCE OF ORTHOGONAL POLYNOMIALS if

$$\langle f_n, \Pi_{n-1}\rangle = 0, \qquad \text{i.e.,} \qquad \langle f_n, f\rangle = 0, \quad f \in \Pi_{n-1}. \tag{3.4.5}$$

The polynomial f_n is called ORTHOGONAL POLYNOMIAL of degree n.

The orthogonal polynomials are of degree *exactly* n unique up to normalization and can be easily determined from the moment matrix. To that end note that for any $g \in \Pi_{n-1}$ we have

$$0 = \langle g, f_n\rangle = [g_0, \dots, g_{n-1}] \begin{bmatrix} \mu_0 & \dots & \mu_n \\ \vdots & \ddots & \vdots \\ \mu_{n-1} & \dots & \mu_{2n-1} \end{bmatrix} \begin{bmatrix} f_0 \\ \vdots \\ f_n \end{bmatrix},$$

[20]That makes the question of integrability much easier as the a Riemann integral works.

and since this has to hold for *any* $g \in \Pi_{n-1}$, it follows that

$$0 = \begin{bmatrix} \mu_0 & \dots & \mu_n \\ \vdots & \ddots & \vdots \\ \mu_{n-1} & \dots & \mu_{2n-1} \end{bmatrix} \begin{bmatrix} f_0 \\ \vdots \\ f_n \end{bmatrix} = \begin{bmatrix} M_{n-1} & \begin{bmatrix} \mu_n \\ \vdots \\ \mu_{2n-1} \end{bmatrix} \end{bmatrix} \begin{bmatrix} f_0 \\ \vdots \\ f_n \end{bmatrix}$$

and since M_{n-1} is positive definite, we get a unique nonzero solution of

$$M_{n-1} \begin{bmatrix} f_0 \\ \vdots \\ f_{n-1} \end{bmatrix} = - \begin{bmatrix} \mu_n \\ \vdots \\ \mu_{2n-1} \end{bmatrix} f_n$$

for any $f_n \neq 0$. This could also be expressed in terms of a SCHUR COMPLEMENT of M_{n-1} in M_n, cf. [25].

Theorem 3.4.6. *A sequence f_n, $n \in \mathbb{N}$, is a sequence of orthogonal polynomials with positive leading coefficients for an inner product if and only if there exist real coefficients $\alpha_n > 0$, $\beta_n \in \mathbb{R}$ and $\gamma_n > 0$, $n \in \mathbb{N}$, such that*

$$f_n = (\alpha_n x + \beta_n) f_{n-1} - \gamma_n f_{n-2}, \qquad n \in \mathbb{N}, \qquad f_0 = 1, \qquad f_{-1} = 0. \tag{3.4.6}$$

Remark 3.4.7. The request $\alpha_n, \gamma_n > 0$ in Theorem 3.4.6 could be weakened into $\alpha_n \gamma_n > 0$ as this would only result in alternatingly changing the sign of the leading terms of f_n.

Proof: Let f_n, $n \in \mathbb{N}$, be a sequence of orthogonal polynomials. We will show by induction on n that the polynomial

$$g_{n+1}(x) = x\, f_n(x) - \underbrace{\frac{\langle x f_n, f_n \rangle}{\langle f_n, f_n \rangle}}_{=:\beta'_n} f_n - \underbrace{\frac{\sqrt{\langle g_n, g_n \rangle \langle f_n, f_n \rangle}}{\langle f_{n-1}, f_{n-1} \rangle}}_{=:\gamma'_n > 0} f_{n-1}(x), \qquad x \in \mathbb{R}, \tag{3.4.7}$$

in nonzero and orthogonal to Π_n, hence must be a positive multiple of f_n. Indeed, for $n = 0$ we obtain that

$$g_1(x) = x\, f_0(x) - \langle x, 1 \rangle f_0 \qquad \Rightarrow \qquad \langle g_1, f_0 \rangle = \langle g_1, 1 \rangle = \langle x, 1 \rangle - \langle x, 1 \rangle = 0,$$

while for the induction step we first note that for $n \in \mathbb{N}_0$ and any $f \in \Pi_{n-2}$

$$\langle g_{n+1}, f \rangle = \langle f_n, x f \rangle - \beta'_n \langle f_n, f \rangle - \gamma'_n \langle f_{n-1}, f \rangle = 0$$

holds. Using the induction hypothesis we also get that $g_n = \lambda_n f_n$ with[21]

$$\langle g_n, g_n \rangle = \lambda_n^2 \langle f_n, f_n \rangle \qquad \Rightarrow \qquad \lambda_n = \sqrt{\frac{\langle g_n, g_n \rangle}{\langle f_n, f_n \rangle}}$$

and end up with

$$\begin{aligned} \langle g_{n+1}, f_{n-1} \rangle &= \langle x\, f_n, f_{n-1} \rangle - \beta'_n \ \langle f_n, f_{n-1} \rangle - \gamma'_n \langle f_{n-1}, f_{n-1} \rangle \\ &= \ \langle f_n, x\, f_{n-1} \rangle - \gamma'_n \langle f_{n-1}, f_{n-1} \rangle = \left\langle f_n, g_n + \beta'_{n-1} f_{n-1} + \gamma'_{n-1} f_{n-2} \right\rangle - \gamma'_n \langle f_{n-1}, f_{n-1} \rangle \\ &= \ \sqrt{\frac{\langle g_n, g_n \rangle}{\langle f_n, f_n \rangle}} \langle f_n, f_n \rangle - \gamma'_n \langle f_{n-1}, f_{n-1} \rangle = 0, \end{aligned}$$

[21]We choose the positive solution of the quadratic equation $-\lambda_n$ would work equally well, cf. Remark 3.4.7.

as well as

$$\langle g_{n+1}, f_n\rangle = \langle x f_n, f_n\rangle - \beta_n' \langle f_n, f_n\rangle - \gamma_n' \langle f_n, f_{n-1}\rangle = 0.$$

This proves (3.4.7) and we can even explicitly give the coefficients as

$$\alpha_n \in \mathbb{R}_+, \qquad \beta_n = -\alpha_n \beta_n', \qquad \gamma_n = \alpha_n \gamma_n',$$

where $\alpha_n > 0$ is a free normalization paramter.

Suppose conversely that f_n is a sequence of polynomials that satisfies (3.4.6) and let us choose, for simplicity, $\alpha_n = 1$, so that we obtain a sequence of MONIC polynomials $f_n(x) = x^n + \widetilde{f}_n(x)$. We also assume inductively that we already determined the inner product on $\Pi_{n-1} \times \Pi_{n-1}$, and know the moments $\mu_0, \dots, \mu_{2n-3}$. Now we consider the polynomials

$$f_n(x) = x\, f_{n-1}(x) + \beta_n\, f_{n-1}(x) - \gamma_n\, f_{n-2}(x)$$

and remark that for $f \in \Pi_{n-3}$ the inner product with f_n is already defined, since

$$\langle f_n, f\rangle := \langle f_{n-1}, x\, f\rangle + \beta_n \langle f_{n-1}, f\rangle - \gamma_n \langle f_{n-2}, f\rangle$$

only contains monomials up to degree $2n - 3$. On the other hand, the additional orthogonality conditions and the recurrence relation (3.4.6) yield

$$\begin{aligned}
0 &= \left\langle f_n, x^{n-2}\right\rangle = \left\langle x f_{n-1} + \beta_n f_{n-1} - \gamma_n f_{n-2}, x^{n-2}\right\rangle \\
&= \left\langle f_{n-1}, x^{n-1}\right\rangle + \beta_n \underbrace{\left\langle f_{n-1}, x^{n-2}\right\rangle}_{=0} - \gamma_n \left\langle f_{n-2}, x^{n-2}\right\rangle \\
&= \left\langle f_{n-1}, x^{n-1}\right\rangle - \gamma_n \left\langle f_{n-2}, x^{n-2}\right\rangle = \left\langle x^{n-1} + \tilde{f}_{n-1}, x^{n-1}\right\rangle - \gamma_n \left\langle f_{n-2}, x^{n-2}\right\rangle \\
&= \mu_{2n-2} + \left\langle \widetilde{f}_{n-1}, x^{n-1}\right\rangle - \gamma_n \left\langle f_{n-2}, x^{n-2}\right\rangle \qquad (3.4.8) \\
&= \mu_{2n-2} + \sum_{j=0}^{2n-3} a_{n,j}\, \mu_j \qquad (3.4.9)
\end{aligned}$$

for some coefficients $a_{n,0}, \dots, a_{n,2n-3}$, and

$$\begin{aligned}
0 &= \left\langle f_n, x^{n-1}\right\rangle = \langle f_{n-1}, x^n\rangle + \beta_n \left\langle f_{n-1}, x^{n-1}\right\rangle - \gamma_n \left\langle f_{n-2}, x^{n-1}\right\rangle \\
&= \mu_{2n-1} + \sum_{j=0}^{2n-2} b_{n,j}\, \mu_j, \qquad (3.4.10)
\end{aligned}$$

for some $b_{n,0}, \dots, b_{n,2n-2}$. Now (3.4.9) defines μ_{2n-2} uniquely in terms of its predecessors and then (3.4.10) does the same for μ_{2n-1}. In summary, this process defines the moments up to the choice of the normalization $\mu_0 > 0$:

$$\begin{aligned}
\mu_1 &= -\beta_1 \mu_0 \\
\mu_2 &= -a_{2,0}\, \mu_0 - a_{2,1}\, \mu_1 \\
\mu_3 &= -\beta_2 \mu_2 - b_{2,0}\, \mu_0 - b_{2,1}\, \mu_1 \\
&\vdots \\
\mu_{2n-2} &= -\sum_{j=0}^{2n-3} a_{n,j}\, \mu_j \\
\mu_{2n-1} &= -\sum_{j=0}^{2n-2} b_{n,j}\, \mu_j.
\end{aligned}$$

It remains to show that the inner product is definite, that is, that $\langle f_n, f_n\rangle > 0$ for $n \in \mathbb{N}_0$ which we will prove, once more, by induction[22] on n, where the easy case $n = 0$ is the assumption $\mu_0 > 0$. Next, we consider

$$\langle f_n, f_n\rangle = \langle f_n, x f_{n-1}\rangle = \langle f_n, x^n\rangle = \mu_{2n} + \left\langle \widetilde{f}_n, x^n \right\rangle \tag{3.4.11}$$

and replacing n in (3.4.8) by $n+1$, we can use

$$\mu_{2n} + \left\langle \tilde{f}_n, x^n \right\rangle = \gamma_{n+1}\left\langle f_{n-2}, x^{n-2}\right\rangle,$$

together with the induction hypothesis to obtain

$$\langle f_n, f_n\rangle = \langle f_n, x^n\rangle = \mu_{2n} + \left\langle \widetilde{f}_n, x^n \right\rangle = \gamma_n \left\langle f_{n-1}, x^{n-1}\right\rangle = \gamma_n \left\langle f_{n-1}, f_{n-1}\right\rangle > 0, \tag{3.4.12}$$

hence the symmetric bilinear form is positive and therefore an inner product. □

Remark 3.4.8. A closer inspection of (3.4.12) even yields an explicit formula for $\langle f_n, f_n\rangle$, namely,

$$\langle f_n, f_n\rangle = \gamma_n \left\langle f_{n-1}, f_{n-1}\right\rangle = \gamma_n\gamma_{n-1} \left\langle f_{n-2}, f_{n-2}\right\rangle = \cdots = \left(\prod_{j=1}^{n} \gamma_j\right) \langle f_0, f_0\rangle = \mu_0 \prod_{j=1}^{n} \gamma_j.$$

Therefore, if we divide (3.4.6) by γ_n, we get a sequence of ORTHONORMAL POLYNOMIALS.

This way we can always get orthogonal polynomials as convergents of continued fractions. And there is not even something to prove any more, we just have to compare the respective three term recurrences.

Corollary 3.4.9. *The orthogonal polynomials with parameters $\alpha_n, \beta_n, \gamma_n$ in the recurrence* (3.4.6) *are obtained as DENOMINATOR of the convergents of the continued fractions*

$$\frac{-\gamma_1|}{|(\alpha_1 x + \beta_1)} - \frac{\gamma_2|}{|(\alpha_2 x + \beta_2)} - \frac{\gamma_3|}{|(\alpha_3 x + \beta_3)} + \cdots$$

or

$$\left[0; -\frac{\alpha_1\, x + \beta_1}{\gamma_1}, -\frac{\alpha_2\, x + \beta_2}{\gamma_2}, \dots\right],$$

respectively. Conversely, the denominators of all continued fractions of the form

$$[0; -\alpha_1 x + \beta_1, -\alpha_2 x + \beta_2, \dots], \qquad \alpha_j > 0,\ \beta \in \mathbb{R},$$

are a system of orthogonal polynomials for an appropriate inner product $\langle\cdot,\cdot\rangle$.

Remark 3.4.10. Orthogonal polynomials can also be defined in several variables, but the geometric and algebraic issues are significantly more intricate [8]. Recurrence relations can be defined, but are based on matrices of increasing block size [59] and by far not all properties that we will list here can be recovered. In addition, the study of multivariate moment problems is also quite recent [50]. Since polynomials in several variables are *not* a euclidean ring, hence there exist no multivariate continued fractions to speak of, we will not touch the issue here.

[22] This is not lack of mathematical proving techniques, but properties defined by recurrence usually ask for induction.

We found out that any sequence of orthogonal polynomials for a strictly square positive linear functional can be written as denominators of convergents of an infinite continued fraction. But what does this continued fraction mean or represent? In other words, what is the analogy for the real number represented by an infinite continued fraction with positive integer coefficients? To answer these questions, we will consider LAURENT SERIES which are usually more popular in complex analysis [23, 56, 57].

Definition 3.4.11 (Laurent series and convergence).

1. The LAURENT SERIES $\lambda(x)$ associated to a sequence $\left(\lambda_j \;:\; j \in \mathbb{N}_0\right)$ is defined as

$$\lambda(x) = \sum_{j=0}^{\infty} \lambda_j \, x^{-j}. \tag{3.4.13}$$

2. A sequence $\lambda_n(x)$, $n \in \mathbb{N}$, of Laurent series is CONVERGENT to a Laurent series $\lambda^*(x)$, if for any $k \in \mathbb{N}_0$ there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ one has

$$\lambda_n(x) - \lambda^* x = x^{-k-1} \, \widetilde{\lambda}_n(x), \qquad \text{i.e.,} \qquad \lambda_{n,j} = \lambda_j^*, \quad j = 0, \dots, k-1. \tag{3.4.14}$$

Remark 3.4.12. Note that Definition 3.4.11 deals with *formal* Laurent series only. We are not interested so far in the convergence radius in (3.4.13) and (3.4.14) is a purely formal comparison of coefficients in the sequence of Laurent series which could as well be used entirely in the context of sequences. The advantage of Laurent series to sequences will become evident soon when we will multiply them.

To make it clear: CONVERGENCE in Definition 3.4.11 means that after a certain index the first k terms of any Laurent series in the sequence coincide with the first k terms of the limit, and that occurs for any $k \in \mathbb{N}_0$. Whether λ^* or some λ_n are analytic functions, we do not care so far.

A first, very simple but surprisingly fundamental, observation is that any reciprocal of a polynomial can be expanded into a power series with a lot of zero initial coefficients.

Lemma 3.4.13. *For $p \in \Pi_n$ with $p_n \neq 0$ one has*

$$\frac{1}{p(x)} = \sum_{j=n}^{\infty} \lambda_j \, x^{-j} =: \lambda(x).$$

Proof: We write $p(x) = p_0 + p_1 \, x + \cdots + p_n \, x^n$ and set $1/p(x) = \lambda(x)$ which yields

$$\begin{aligned} 1 &= p(x) \, \lambda(x) = \left(\sum_{j=0}^{n} p_j \, x^j \right) \left(\sum_{k=0}^{\infty} \lambda_j \, x^{-k} \right) = \sum_{j=0}^{n} \sum_{k=0}^{\infty} p_j \, \lambda_k \, x^{j-k} \\ &= \sum_{j=-\infty}^{n} x^j \sum_{k-\ell=j} p_k \, \lambda_\ell = \sum_{j=-\infty}^{n} x^j \sum_{\ell=-j}^{n-j} p_{j+\ell} \lambda_\ell, \end{aligned}$$

where $\lambda_{-n} = \cdots = \lambda_{-1} = 0$. Comparison of coefficients gives

$$\sum_{k=-j}^{n-j} p_{j+k} \lambda_k = \delta_{j,0} = \begin{cases} 0, & j \neq 0, \\ 1, & j = 0, \end{cases}$$

in particular

$$\begin{aligned} 0 &= p_n\,\lambda_0 \\ 0 &= p_{n-1}\,\lambda_0 + p_n\,\lambda_1 \\ &\vdots \\ 0 &= p_1\,\lambda_0 + \cdots + p_n\,\lambda_{n-1}, \end{aligned}$$

which we can write in matrix form and make use of $p_n \neq 0$ to see that

$$0 = \begin{bmatrix} p_n & & \\ \vdots & \ddots & \\ p_1 & \dots & p_n \end{bmatrix} \begin{bmatrix} \lambda_0 \\ \vdots \\ \lambda_{n-1} \end{bmatrix} \qquad \Rightarrow \qquad \lambda_0 = \cdots = \lambda_{n-1} = 0.$$

The other coefficients are obtained by successively solving the systems

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \end{bmatrix} = \begin{bmatrix} p_n & & & \\ \vdots & \ddots & & \\ p_0 & \dots & p_n & \\ & \ddots & & \ddots \end{bmatrix} \begin{bmatrix} \lambda_n \\ \lambda_{n+1} \\ \vdots \end{bmatrix},$$

that determine $\lambda_n, \lambda_{n+1}, \dots$ *uniquely.* □

Now we get our polynomial analogue for real numbers.

Definition 3.4.14. An infinite continued fraction $[0; a_1, a_2, \dots]$, $a_j \in \Pi \setminus \Pi_0$ is called CONVERGENT, if there exists a LAURENT SERIES $\lambda(x)$ such that

$$\lim_{n\to\infty} \frac{p_n(x)}{q_n(x)} = \lambda(x)$$

in the sense of Definition 3.4.11.

Remark 3.4.15 (Convergence of continued fractions).

1. Definition 3.4.14 still lives entirely in the context of *formal* Laurent series.
2. Definition 3.4.14 makes sense. Since $p_0 = 0$ and $p_1 = 1$, it follows that $\deg q_n > \deg p_n$ and thus, by Lemma 3.4.13,

$$\frac{p_n(x)}{q_n(x)} = p_n(x) \sum_{j=\deg q_n}^{\infty} \lambda_j\, x^{-j} = \sum_{j=\deg q_n - \deg p_n}^{\infty} \widetilde{\lambda}_j\, x^{-j}$$

   any convergent can be represented as a Laurent series.
3. One could also expand the rations functions with respect to positive powers of x which would give the TAYLOR SERIES. However one would then need a slightly different notion of continued fractions, see [37].
4. We can illustrate the idea behind convergence of continued fractions of polynomials by recalling how the objects are generated: we expand a finite segment into a rational function, transfer that into a Laurent series and consider the limit of this sequence of Laurent series in the sense of Definition3.4.11:

$$[0; a_1, \dots] \to [0; a_1, \dots, a_n] = \frac{p_n}{q_n} = \lambda_n \to \lambda, \qquad n \to \infty.$$

Indeed, there are plenty of convergent continued fractions in the sense of Definition 3.4.14, in particular those that we already know from three term recurrences with at least *linear* components.

Theorem 3.4.16. *Any continued fraction of the form* $[0; r_1, \dots]$, $r_j \in \Pi$, $\deg r_j \ge 1$, $j \in \mathbb{N}$, *converges to a Laurent series* $\lambda(x)$ *in such a way that*

$$\lambda(x) - \frac{p_n(x)}{q_n(x)} = O\left(x^{-d_{n+1}-d_n}\right), \tag{3.4.15}$$

that is,

$$\frac{p_n(x)}{q_n(x)} = \lambda_0 + \cdots + \lambda_{d_{n+1}+d_n-1}\, x^{-d_{n+1}-d_n+1} + \cdots, \tag{3.4.16}$$

where $d_n := \deg q_n$, $n \in \mathbb{N}_0$.

Proof: In the formal Laurent series

$$\lambda(x) - \frac{p_n(x)}{q_n(x)} = \sum_{j=n}^{\infty} \left(\frac{p_{j+1}(x)}{q_{j+1}(x)} - \frac{p_j(x)}{q_j(x)} \right) = \sum_{j=n}^{\infty} \frac{(-1)^j}{q_{j+1}(x)\, q_j(x)} = \sum_{j=d_{n+1}+d_n}^{\infty} \gamma_j\, x^{-j} =: \gamma(x)$$

all coefficients γ_j are well–defined, since γ_j depends only on finitely many values q_k. Then convergence follows since

$$\frac{p_{n+k}(x)}{q_{n+k}(x)} - \frac{p_n(x)}{q_n(x)} = O\left(x^{-d_n-d_{n+1}}\right), \qquad k \in \mathbb{N},$$

and thus we have an analogy to a Cauchy sequence. This carries over to the limit series $\lambda(x)$ and gives (3.4.15). □

Returning to orthogonal polynomials this particularly implies that continued fractions with *affine coefficients*[23] always converge and that even of a very simple order.

Corollary 3.4.17. *Any continued fraction of the form* $[0; r_1, \dots]$, $r_j \in \Pi_1 \setminus \Pi_0$, $j \in \mathbb{N}$, *converges to a Laurent series* $\lambda(x)$ *in such a way that*

$$\lambda(x) - \frac{p_n(x)}{q_n(x)} = O\left(x^{-2n-1}\right). \tag{3.4.17}$$

These continued fractions converge rapidly in the sense that the number of coefficients captured is twice the degree of the denominator and thus fit particularly well with the Laurent series λ, due to which we should have a closer look at them. The theory could even be developed in a more general framework of continued fractions with $r_j \in \Pi \setminus \Pi_0$, but we will restrict ourselves to continued fractions with factors of degree 1, i.e., $r_j(x) = \alpha_j x + \beta_j$, $\alpha_j \neq 0$, for which we have $\deg q_n = \deg p_n + 1 = n$. And the good representations of that type for a given Laurent series get a special name.

Definition 3.4.18. The infinite continued fraction $[0; r_1, \dots]$, $r_j \in \Pi_1 \setminus \Pi_0$ is called ASSOCIATED to the LAURENT SERIES $\lambda(x)$ if

$$\lambda(x) - \frac{p_n(x)}{q_n(x)} = O\left(x^{-2n-1}\right), \qquad n \in \mathbb{N},$$

that is,

$$\frac{p_n(x)}{q_n(x)} = \sum_{j=0}^{2n} \lambda_j\, x^{-j} + \sum_{j=2n+1}^{\infty} \gamma_{n,j}\, x^{-j}, \qquad n \in \mathbb{N}. \tag{3.4.18}$$

[23]That is, polynomials of degree ≤ 1.

It would be too optimistic to assume that *all* Laurent series have associated continued fractions[24], but it will actually turn out that a description of Laurent series for which there exists an associated continued fraction is even more interesting and will involve the concept of a HANKEL MATRIX which we already know from Definition 3.4.3, (3.4.3).

Theorem 3.4.19. *A Laurent series $\lambda(x)$ has an* ASSOCIATED *continued fraction* $[0; r_1, \dots]$, $r_j \in \Pi_1 \setminus \Pi_0$, *if and only if* $\lambda_0 = 0$ *and*

$$\det \Lambda_n \neq 0, \qquad \Lambda_n = \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_n \\ \lambda_2 & \lambda_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_{2n-2} \\ \lambda_n & \dots & \lambda_{2n-2} & \lambda_{2n-1} \end{bmatrix}, \qquad n \in \mathbb{N}. \tag{3.4.19}$$

Proof: The continued fraction is associated if and only if for any $n \in \mathbb{N}$ we have

$$\begin{aligned} \frac{p_n(x)}{q_n(x)} &= \lambda_0 + \dots + \lambda_{2n} x^{-2n} + \gamma_{n,2n+1} x^{-2n-1} + \gamma_{n,2n+2} x^{-2n-2} + \cdots \\ \frac{p_{n+1}(x)}{q_{n+1}(x)} &= \lambda_0 + \dots + \lambda_{2n} x^{-2n} + \lambda_{2n+1} x^{-2n-1} + \lambda_{2n+2} x^{-2n-2} + \cdots \end{aligned} \tag{3.4.20}$$

Subtracting the second equation in (3.4.20) from the first one and keeping in mind that this equals $(-1)^n/(q_{n+1}\, q_n)$, we find that

$$\frac{(-1)^{n+1}}{q_{n+1}(x)\, q_n(x)} = \frac{p_n(x)}{q_n(x)} - \frac{p_{n+1}(x)}{q_{n+1}(x)} = \left(\gamma_{n,2n+1} - \lambda_{2n+1}\right) x^{-2n-1} + \left(\gamma_{n,2n+2} - \lambda_{2n+2}\right) x^{-2n-2} + \cdots$$

Next, we note that the recursion formula (3.2.4) takes the form

$$q_n(x) = r_n(x)\, q_{n-1}(x) + q_{n-2}(x) = (\alpha_n x + \beta_n)\; q_{n-1}(x) + q_{n-2}(x),$$

from which it follows by induction that

$$q_n(x) = \left(\prod_{j=1}^n \alpha_j \right) \left(x^n + x^{n-1} \sum_{j=1}^n \frac{\beta_j}{\alpha_j} \right) + \cdots. \tag{3.4.21}$$

Indeed, we have $q_0 = 1$, $q_1 = \alpha_1 x + \beta_1$, and, in general the two highest degree terms are given as

$$\begin{aligned} &(\alpha_n x + \beta_n) \left(\prod_{j=1}^{n-1} \alpha_j \right) \left(x^{n-1} + x^{n-2} \sum_{j=1}^{n-1} \frac{\beta_j}{\alpha_j} \right) \\ &= \left(\prod_{j=1}^{n} \alpha_j \right) \left(x^n + x^{n-1} \sum_{j=1}^{n-1} \frac{\beta_j}{\alpha_j} \right) + \beta_n \left(\prod_{j=1}^{n-1} \alpha_j \right) x^{n-1} + O\left(x^{n-2}\right) \\ &= \left(\prod_{j=1}^{n} \alpha_j \right) \left(x^n + x^{n-1} \sum_{j=1}^{n-1} \frac{\beta_j}{\alpha_j} + \frac{\beta_n}{\alpha_n} x^{n-1} \right) + O\left(x^{n-2}\right) = \left(\prod_{j=1}^{n} \alpha_j \right) \left(x^n + x^{n-1} \sum_{j=1}^{n} \frac{\beta_j}{\alpha_j} \right) + O\left(x^{n-2}\right). \end{aligned}$$

Thus,

$$q_{n+1}(x)\, q_n(x) = \alpha_{n+1} \left(\prod_{j=1}^n \alpha_j \right)^2 x^{2n+1} + \left(\prod_{j=1}^n \alpha_j \right)^2 \left(\beta_{n+1} + \alpha_{n+1} \sum_{j=1}^n \frac{\beta_j}{\alpha_j} \right) x^{2n} + \cdots \tag{3.4.22}$$

[24]So finally here is a difference to real numbers and their continued fraction expansions.

By Lemma 3.4.13 we have that

$$\frac{1}{q_{n+1}(x)\,q_n(x)} = \alpha_{n+1}^{-1}\left(\prod_{j=1}^{n}\alpha_j\right)^{-2} x^{-2n-1} + \cdots$$

and comparing coefficients implies

$$(-1)^{n+1}\left(\gamma_{n,2n+1} - \lambda_{2n+1}\right) = \alpha_{n+1}^{-1}\left(\prod_{j=1}^{n}\alpha_j\right)^{-2},$$

and thus the equivalent identities

$$\alpha_{n+1} = \frac{(-1)^{n+1}}{\left(\gamma_{n,2n+1} - \lambda_{2n+1}\right)(\alpha_1\cdots\alpha_n)^2}, \qquad \gamma_{n,2n+1} - \lambda_{2n+1} = \frac{(-1)^{n+1}}{\alpha_{n+1}\,(\alpha_1\cdots\alpha_n)^2}. \tag{3.4.23}$$

Let us summarize what we obtained so far: The existence of an associated continued fraction with $r_j \in \Pi_1 \setminus \Pi_0$ is equivalent to the validity of (3.4.23) with all $\alpha_j \neq 0$ which is in turn equivalent to $\gamma_{n,2n+1} \neq \lambda_{2n+1}$.

To see what this means, we multiply the first line of (3.4.20) by $q_n(x)$, which[25] leads to

$$\begin{aligned} p_n(x) &= \left(\sum_{j=0}^{2n}\lambda_j\,x^{-j} + \sum_{j=2n+1}^{\infty}\gamma_{n,j}\,x^{-j}\right)\left(\sum_{k=0}^{n} q_{n,k}\,x^k\right) \\ &= \sum_{j=0}^{2n}\sum_{k=0}^{n}\lambda_j\,q_{n,k}\,x^{k-j} + \sum_{j=2n+1}^{\infty}\sum_{k=0}^{n}\gamma_{n,j}\,q_{n,k}\,x^{k-j} = \sum_{k=0}^{n}\sum_{j=k-2n}^{k}\lambda_{k-j}\,q_{n,k}x^j + O\left(x^{-n-1}\right) \\ &= \sum_{-n\le k-2n\le j\le k\le n}\lambda_{k-j}\,q_{n,k}x^j + O\left(x^{-n-1}\right) = \sum_{-n\le j\le n}\sum_{j\le k\le j+2n}\lambda_{k-j}\,q_{n,k}x^j + O\left(x^{-n-1}\right) \\ &= \sum_{j=-n}^{n} x^j \sum_{k=j}^{j+2n}\lambda_{k-j}\,q_{n,k} + O\left(x^{-n-1}\right) = \sum_{j=-n}^{n} x^{-j}\sum_{k=0}^{n}\lambda_{j+k}q_k + O\left(x^{-n-1}\right), \end{aligned}$$

hence

$$p_n(x) = \sum_{j=0}^{n}\eta_{-j}x^j + \sum_{j=1}^{n}\eta_j x^{-j} + \eta_{n+1}x^{-n-1} + O\left(x^{-n-2}\right), \tag{3.4.24}$$

where

$$\begin{bmatrix}\eta_{-n}\\ \vdots \\ \eta_0\end{bmatrix} = \begin{bmatrix}\lambda_0 & & \\ \vdots & \ddots & \\ \lambda_n & \dots & \lambda_0\end{bmatrix}\begin{bmatrix} q_{n,n}\\ \vdots \\ q_{n,0}\end{bmatrix} \tag{3.4.25}$$

and[26]

$$\begin{bmatrix}\eta_1\\ \vdots \\ \eta_n \\ \eta_{n+1}\end{bmatrix} = \begin{bmatrix}\lambda_1 & \lambda_2 & \dots & \lambda_{n+1}\\ \lambda_2 & \lambda_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_{2n-1} \\ \lambda_n & \dots & \lambda_{2n-1} & \lambda_{2n} \\ \lambda_{n+1} & \dots & \lambda_{2n} & \gamma_{n,2n+1}\end{bmatrix}\begin{bmatrix} q_{n,0}\\ \vdots \\ q_{n,n}\end{bmatrix}. \tag{3.4.26}$$

[25]Using the *convention* that $0 = \lambda_j = p_k$, $j,k < 0$ or $k > n$, respectively.

[26]The rule for η_{n+1} is obvious once one understands how $\eta_1,\dots,\eta_n$ are formed.

Since the left hand side of (3.4.24) is a polynomial, comparison of coefficients yields that $\eta_1 = \cdots = \eta_{n+1} = 0$, hence, since $q \neq 0$, the determinant of the matrix in (3.4.26) is 0. From (3.4.23) we now determine

$$\gamma_{n,2n+1} = \lambda_{2n+1} + \frac{(-1)^{n+1}}{\alpha_{n+1}\,(\alpha_1 \cdots \alpha_n)^2}$$

and substitute this into (3.4.26), which gives

$$\begin{aligned}
0 &= \det \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_{n+1} \\ \lambda_2 & \lambda_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_{2n-1} \\ \lambda_n & \dots & \lambda_{2n-1} & \lambda_{2n} \\ \lambda_{n+1} & \dots & \lambda_{2n} & \gamma_{n,2n+1} \end{bmatrix} \\
&= \det \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_{n+1} \\ \lambda_2 & \lambda_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_{2n-1} \\ \lambda_{n+1} & \dots & \lambda_{2n} & \lambda_{2n+1} \end{bmatrix} + \left(\alpha_{n+1} \prod_{j=1}^{n} \alpha_j^2 \right)^{-1} \det \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & 0 \\ \lambda_2 & \lambda_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \lambda_n & \dots & \lambda_{2n-1} & 0 \\ \lambda_{n+1} & \dots & \lambda_{2n} & (-1)^{n+1} \end{bmatrix} \\
&= \det \Lambda_{n+1} + \left(\alpha_{n+1} \prod_{j=1}^{n} \alpha_j^2 \right)^{-1} \det \Lambda_n,
\end{aligned}$$

that is,

$$\det \Lambda_{n+1} = -\frac{\det \Lambda_n}{\alpha_{n+1}\,(\alpha_1 \cdots \alpha_n)^2} \qquad \text{and} \qquad \alpha_{n+1} = -\left(\prod_{j=1}^{n} \alpha_j^2 \right)^{-1} \frac{\det \Lambda_n}{\det \Lambda_{n+1}}. \tag{3.4.27}$$

Let us summarize: if the continued fraction with coefficients in $\Pi_0 \backslash \Pi_0$ is associated to a Laurent series $\lambda(x)$, then the left hand side of (3.4.27) yields inductively on n that $\det \Lambda_n \neq 0$, while, conversely, the right hand side of (3.4.27) shows that all components $r_j = \alpha_j x + \beta_j$ are nonconstant polynomials as long as the determinant condition is valid. The coefficients β_j are then determined from looking at the second nonzero term in the Laurent expansion of $1/(q_{n+1} q_n)$ and solving for β_{n+1} which will depend on $\alpha_1, \dots, \alpha_{n+1}$, $\beta_1, \dots, \beta_n$ and $\gamma_{2n,2n+2} - \lambda_{2n+2}$.

The condition on λ_0 is sinmpler: we only observe in (3.4.25) that $\deg p_n = n - 1$, hence $0 = \eta_{-n} = \lambda_0\, q_{n,n}$, whereas $\deg q_n = n$ which implies that $q_{n,n} \neq 0$. □

Exercise 3.4.4 Give an explicit method to compute β_{n+1}.
Hint: Use Lemma 3.4.13 to compute the term θ_{n+2} of

$$\frac{1}{q_{n+1}(x)\, q_n(x)} = \sum_{j=2n+1}^{\infty} \theta_j\, x^{-j}$$

and solve that for β_{n+1}. ◊

Theorem 3.4.19 is already quite nice with a cute proof, but the real beauty of this observation is only contained in the next result that really connects orthogonal polynomials and continued fractions – and also provides Gauß' implicit definition of orthogonal polynomials.

Theorem 3.4.20. *Let μ be the* MOMENT SEQUENCE *for a square positive linear functional. Then the orthogonal polynomials for this functional are the numerators q_n, $n \in \mathbb{N}$, of the continued fraction for the associated* LAURENT SERIES

$$\mu(x) = \sum_{j=1}^{\infty} \mu_{j-1} x^{-j}.$$

Proof: The matrices $\Lambda_n = M_{n-1}$, $n \in \mathbb{N}$, are *strictly* positive definite and thus all have positive determinants. Thus there exists an associated continued fraction. Due to (3.4.26) and the comparison of coefficients in (3.4.24) we moreover have that

$$\begin{aligned} 0 &= \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_{n+1} \\ \lambda_2 & \lambda_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_{2n-1} \\ \lambda_n & \dots & \lambda_{2n-1} & \lambda_{2n} \end{bmatrix} \begin{bmatrix} q_{n,0} \\ \vdots \\ q_{n,n} \end{bmatrix} = \begin{bmatrix} \mu_0 & \mu_1 & \dots & \mu_n \\ \mu_1 & \mu_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mu_{2n-2} \\ \mu_{n-1} & \dots & \mu_{2n-2} & \mu_{2n-1} \end{bmatrix} \begin{bmatrix} q_{n,0} \\ \vdots \\ q_{n,n} \end{bmatrix} \\ &= \begin{bmatrix} \langle 1, q \rangle \\ \vdots \\ \left\langle (\cdot)^{n-1} q \right\rangle \end{bmatrix}, \end{aligned}$$

which implies orthogonality of the polynomials. And according to (3.4.27) the coefficients α_j of the recurrence even have the proper sign for an orthogonal polynomial. □

Another remark: we now even have a way to find the parameters in the recurrence relation by determining α_n via (3.4.27) and then β_n by using the coefficient vector[27] $q_n = \left[q_{n,j} \ : \ j = 0, \dots, n \right]$ for q_n and the identity

$$\begin{aligned} 0 &= \langle q_{n+1}, q_n \rangle = \alpha_{n+1} \langle (\cdot) q_n, q_n \rangle + \beta_{n+1} \langle q_n, q_n \rangle + \langle q_{n-1}, q_n \rangle \\ &= \alpha_{n+1} \, q_n^T \begin{bmatrix} \mu_1 & \dots & \mu_{n+1} \\ \vdots & \ddots & \vdots \\ \mu_{n+1} & \dots & \mu_{2n+1} \end{bmatrix} q_n + \beta_{n+1} \, q_n^T M_n q_n \end{aligned}$$

as

$$\beta_{n+1} = -\alpha_{n+1} \frac{q_n^T \widetilde{M}_n q_n}{q_n^T M_n q_n}, \qquad \widetilde{M}_n = \begin{bmatrix} \mu_1 & \dots & \mu_{n+1} \\ \vdots & \ddots & \vdots \\ \mu_{n+1} & \dots & \mu_{2n+1} \end{bmatrix}. \tag{3.4.28}$$

What does all that have to do with Gauß? The connection is that continued fractions were the key in the original method to determine the so–called GAUSS QUADRATURE formula in [13]. Such a quadrature formula consists of WEIGHTS ω_j and KNOTS x_j, $j = 0, \dots, n$, such that

$$0 = L(f) - \Omega(f) = L(f) - \sum_{j=0}^{n} \omega_j \, f\left(x_j\right), \qquad f \in \Pi_{2n+1}, \tag{3.4.29}$$

where, once more, L denotes a square positive linear functional. The QUADRATURE FORMULA Ω in (3.4.29) has the maximal EXACTNESS $2n+1$. *Maximal* means that there cannot

[27] Using the same symbol for the polynomial and its coefficient vector is quite reasonable and, after all, it is also the way how polynomials are usually stored on a computer: by means of their coefficients

be a quadrature formula with $n+1$ weights and knots that is EXACT on Π_{2n+2}, at least if the functional comes from an integral

$$L(f) = \int_a^b f(x)\, w(x)\, dx, \qquad w(x) > 0, \quad x \in (a,b),$$

with strictly positive WEIGHT FUNCTION w. This is seen by considering the polynomial $f(x) = (x-x_0)^2 \cdots (x-x_n)^2 \in \Pi_{2n+2}$ which satisfies

$$L\left(f^2\right) > 0 = \sum_{j=0}^n \omega_j\, f\left(x_j\right),$$

so that (3.4.29) fails for this f. To given points $x_0, \dots, x_n$ or a given polynomial $w(x) = (x-x_0)\cdots(x-x_n)$ the weights ω_j, $j = 0, \dots,$ are determined as

$$\omega_j = L\left(\ell_j\right), \qquad \ell_j = \prod_{k \neq j} \frac{\cdot - x_k}{x_j - x_k} = \frac{w}{w'\left(x_j\right)\,\left(\cdot - x_j\right)}, \qquad j = 0, \dots, n. \tag{3.4.30}$$

Exercise 3.4.5 Prove formula (3.4.30) for the polynomials ℓ_j. ◇

Writing $w(x) = w_0 + w_1\, x + \cdots + w_n x^n + x^{n+1}$, we get that[28]

$$\begin{aligned}
w'\left(x_j\right)\, \ell_j(x) &= \frac{w(x)}{x - x_j} = \frac{w(x) - w\left(x_j\right)}{x - x_j} \\
&= \frac{w_1\left(x - x_j\right) + \cdots + w_n\left(x^n - x_j^n\right) + \left(x^{n+1} - x_j^{n+1}\right)}{x - x_j} \\
&= \sum_{k=1}^{n+1} w_k \frac{x^k - x_j^k}{x - x_j} = \sum_{k=1}^{n+1} w_k \sum_{m=0}^{k-1} x^m\, x_j^{k-1-m} \\
&= \begin{array}{ccccccccc}
x^n & + & x_j\, x^{n-1} & + & x_j^2\, x^{n-2} & + & \cdots & + & x_j^n \\
 & + & w_n\, x^{n-1} & + & w_n\, x_j\, x^{n-2} & + & \cdots & + & w_n\, x_j^{n-1} \\
 & & & + & w_{n-1}\, x^{n-2} & + & \cdots & + & w_{n-1}\, x_j^{n-2} \\
 & & & & & & \ddots & & \vdots \\
 & & & & & & & + & w_1
\end{array} \\
&= \sum_{k=0}^n x^k \frac{w\left(x_j\right)}{x_j^{k+1}} + O\left(x_j^{-1}\right),
\end{aligned}$$

hence

$$\begin{aligned}
w_j'\left(x_j\right)\, L\left(\ell_j\right) &= \mu_n + \mu_{n-1}\left(x_j + w_n\right) + \cdots + \mu_0\left(x_j^n + w_n\, x_j^{n-1} + \cdots + w_1\right) \\
&= \sum_{k=0}^n \mu_k \left(x_j^{n-k} + \sum_{m=k+1}^n w_k\, x_j^{n-k}\right) =: \widetilde{w}\left(x_j\right), \qquad \widetilde{w} \in \Pi_n,
\end{aligned}$$

which yields

$$\omega_j = \frac{\widetilde{w}\left(x_j\right)}{w'\left(x_j\right)}, \qquad j = 0, \dots, n. \tag{3.4.31}$$

[28]What follows now is taken almost literally from the original paper by Gauß, only the notation is slightly modernized.

This formula allows us to determine the WEIGHTS of the QUADRATURE FORMULA *directly* from the MOMENTS once we fix the KNOTS and hence the polynomial w and its coefficients.

Now let $\lambda_k = \Omega\left((\cdot)^k\right)$ denote the moments of the quadrature formula, write $\theta_k = \mu_k - \lambda_k$ for the moments of the error $E = L - \Omega$ and let $\lambda(x)$ and $\theta(x)$ be the associated LAURENT SERIES. Using the (formal) identity[29]

$$\frac{1}{x-\xi} = \sum_{j=1}^{\infty} \frac{\xi^{j-1}}{x^j} \tag{3.4.32}$$

we note that

$$\lambda(x) = \sum_{k=1}^{\infty} \frac{\Omega\left((\cdot)^{k-1}\right)}{x^k} = \sum_{k=1}^{\infty} x^{-k} \sum_{j=0}^{n} \omega_j\, x_j^{k-1} = \sum_{j=0}^{n} \omega_j \sum_{k=1}^{\infty} x_j^{k-1} x^{-k} = \sum_{j=0}^{n} \frac{\omega_j}{x - x_j},$$

from which we conclude that

$$\theta(x) = \mu(x) - \lambda(x) = \mu(x) - \sum_{j=0}^{n} \frac{\omega_j}{x - x_j} \tag{3.4.33}$$

has to hold. By construction, the quadrature formula is INTERPOLATORY[30], yielding $\theta_0 = \cdots = \theta_n = 0$ and thus

$$O\left(x^{-1}\right) = w(x)\,\theta(x) = w(x)\,\mu(x) - \sum_{j=0}^{n} \omega_j \frac{w(x)}{x - x_j} = w(x)\,\mu(x) - \underbrace{\sum_{j=0}^{n} \omega_j\, w'\left(x_j\right)\, \ell_j(x)}_{\in \Pi_n}.$$

And now we are at the point where Gauß uses the magic of continued fractions in [13]: if we choose specifically $w(x) = q_{n+1}(x)$ as the denominator of the $n+1$st convergent of $\mu(x)$, which exists by Theorem 3.4.20, then

$$\mu(x) = \frac{p_{n+1}(x)}{q_{n+1}(x)} + O\left(x^{-2n-3}\right) \qquad \Rightarrow \qquad q_{n+1}(x)\,\mu(x) = p_{n+1}(x) + O\left(x^{-n-2}\right),$$

and therefore

$$w(x)\,\theta(x) = \underbrace{p_{n+1}(x) - \sum_{j=0}^{n} \omega_j\, w'\left(x_j\right)\, \ell_j(x)}_{=:p(x)} + O\left(x^{-n-2}\right) = O\left(x^{-1}\right),$$

hence the polynomial p must satisfy $p = 0$ which yields

$$w(x)\,\theta(x) = O\left(x^{-n-2}\right) \qquad \Rightarrow \qquad \theta(x) = \frac{O\left(x^{-n-2}\right)}{w(x)} = O\left(x^{-2n-3}\right),$$

and, consequently,

$$0 = \theta_0 = \cdots = \theta_{2n+1}. \tag{3.4.34}$$

In other words, the quadrature formula with the zeros of q_{n+1} provides the desired EXACTNESS.

[29]Which is proved by multiplying and comparing coefficients

[30]It is formed by integrating the interpolation polynomial at $x_0, \ldots, x_n$.

Remark 3.4.21. This way to determine the quadrature knots makes **no** use of any sort of integral, is purely algebraic and only applies formal manipulations to the formal Laurent series associated to the moment sequence.

In our enthusiasm about this really beautiful construction[31], we forgot one important point: q_{n+1} has to have *real* and *simple* zeros, otherwise the whole approach makes no sense and we'd contradict our implicit assumption of simple zeros at the end. But fortunately, the zeros are real and simple as is ensured by the following proposition that again relates directly to continued fractions, more precisely, to their associated recurrence.

Proposition 3.4.22. *If a sequence f_n, $n \in \mathbb{N}$, of polynomials satisfies a recurrence as in* (3.4.6)*, then each f_n has simple and real zeros.*

In standard numerical analysis, one would first refer to Theorem 3.4.6 and then rely on the well–known fact that orthogonal polynomials have only real and simple zeros, cf. [14]. This proof, however, usually relies on an integral representation of the functional, which we do not have when we start only with a moment sequence. This can be somewhat compensated by using the fact that any POSITIVE POLYNOMIAL, i.e., any polynomial $p \neq 0$ with $p(x) \geq 0$, $x \in \mathbb{R}$, can be decomposed into a sum of squares, thus relating positive and square positive functions.

Here we will follow a direct approach and since we will need Sturm chains later on anyway, we immediately present a proof based on those.

3.5 Sturm chains

Sturm chains give a method to *count* the zeros or sign changes of a polynomial within a given interval without having to *determine* them. This is done by counting the sign changes of a certain sequence of numbers which makes them a useful and fairly popular tool in the numerics for univariate polynomials, due to which they can be found in various places of the literature. Here, we follow the terminology and notation from [11].

Definition 3.5.1. A finite sequence $f_0, \dots, f_n \in \Pi$ of polynomials is called a STURM CHAIN for an interval[32] I if

1. at each ZERO of f_k the polynomials f_{k+1} and f_{k-1} have opposite SIGN:
$$f_k(x) = 0 \qquad \Rightarrow \qquad f_{k-1}(x)\, f_{k+1}(x) < 0, \qquad k = 1, \dots, n-1. \tag{3.5.1}$$
2. the polynomial f_0 has no zero in I, i.e., $0 \notin f(I)$.

Remark 3.5.2. The second condition in Definition 3.5.1 means that the continuous function f_0 has to be either strictly positive or strictly negative on I. Since $f_0, \dots, f_n$ is a Sturm sequence if and only if $-f_0, \dots, -f_n$ is a Sturm sequence, we could replace this requirement by $f_0(I) > 0$ without an essential loss of generality.

What this concept has to do with zeros becomes clear if for some $x \in \mathbb{R}$ we consider the number $V(x)$ of true or proper SIGN CHANGES in the vector $(f_0(x), \dots, f_n(x))$; proper sign change means that zero values in the vector are ignored or erased from the vector so that we

[31] After all, it is due to Gauß, so what else should we expect?
[32] Open or closed, bounded or unbounded.

only count strict sign changes from + to − or from − to +. Then, we let x vary and consider $V(x)$ as a function in x. As long as $f_j(x) \neq 0$, $j = 0, \dots, n$, the value of $V([x-\varepsilon, x+\varepsilon))$ is constant for a sufficiently small $\varepsilon > 0$, again due to the continuity of polynomials. If, however, f_k, $1 < k < n$, has a zero at x, i.e., $f_k(x) = 0$, then, because of (3.5.1), either f_{k+1} or f_{k-1} has the same sign as f_k restricted to $[x-\varepsilon, x)$ and the same holds for the other half interval $(x, x+\varepsilon]$. But this means that $V(x)$ remains unchanged:

$$V(x-\varepsilon) = V(x+\varepsilon) = V(x) = V(y), \qquad y \in [x-\varepsilon, x+\varepsilon].$$

In other words: $V(x)$ changes only if f_n changes its sign relative to f_{n-1}. If f_{n-1} and f_n have a joint sign change at x, then V is again constant on $[x-\varepsilon, x+\varepsilon]$, otherwise the number of sign changes increases or decreases depending on whether f_{n-1} and f_n had the same or opposite sign at $x-\varepsilon$, respectively. This is depicted in the following table:

	$x-\varepsilon$	x	$x+\varepsilon$
f_n	$\pm$	0	$\mp$
f_{n-1}	$\pm$	$\pm$	$\pm$
V	k	k	$k+1$

	$x-\varepsilon$	x	$x+\varepsilon$
f_n	$\pm$	0	$\mp$
f_{n-1}	$\mp$	$\mp$	$\mp$
V	k	$k-1$	$k-1$

If we now track this along an interval and take into account that changes become active on the right of the zero, we get the following result.

Theorem 3.5.3 (Zero counting). *Define*[33]

$$\sigma_+(f, I) := \#Z_+(f, I) := \#\{x \in I \ :\ f(x-\varepsilon) > f(x) = 0 > f(x+\varepsilon)\},$$

and

$$\sigma_-(f, I) := \#Z_-(f, I) := \#\{x \in I \ :\ f(x-\varepsilon) < f(x) = 0 < f(x+\varepsilon)\},$$

then we get, for $I = [a, b)$, that

$$\sigma_+\left(\frac{f_n}{f_{n-1}}, I\right) - \sigma_-\left(\frac{f_n}{f_{n-1}}, I\right) = V(b) - V(a). \tag{3.5.2}$$

Proof: If $f(a) = 0$, then $V(a+\varepsilon) = V(a) \pm 1$ depending on whether a belongs to Z_+ or to Z_-. Then, $V(x)$ is piecewise constant and increases by 1 on Z_+ and decreases by 1 on Z_-. Thus, eventually

$$V(b) = V(a) + \sigma_+\left(\frac{f_n}{f_{n-1}}, I\right) - \sigma_-\left(\frac{f_n}{f_{n-1}}, I\right),$$

from which (3.5.2) follows immediately. □

And this is all we need to show that polynomials which obey a three term recurrence always have simple real zeros.

Proposition 3.5.4. *For any polynomial sequence f_n, $n \in \mathbb{N}_0$, defined by a three term recurrence relation*

$$f_0 = 1, \qquad f_{n+1}(x) = (x+\beta_n) f_n(x) - \gamma_n f_{n-1}(x), \qquad \gamma_n > 0, \qquad n \in \mathbb{N}_0, \tag{3.5.3}$$

the following holds:

[33] The notation is slightly sloppy, but here $x-\varepsilon$ always includes „for all sufficiently small $\varepsilon > 0$".

1. *Each finite sequence* $f_0, \dots, f_n$ *is a* STURM CHAIN *for* any *interval* $I \subseteq \mathbb{R}$.
2. *The polynomial* f_n *has exactly* n *simple real zeros, that is*

$$\#Z_{\mathbb{R}}(f_n) = n, \qquad Z_I(f) = \{x \in I \ : \ f(x) = 0\}.$$

Hence

$$f_n(x) = \prod_{j=1}^{n} (x - \xi_j), \qquad \xi_1 < \cdots < \xi_n. \tag{3.5.4}$$

Remark 3.5.5. According to Theorem 3.4.6 the recurrence relations from (3.5.3) are *exactly* the recurrence for MONIC ORTHOGONAL POLYNOMIALS with respect to a square positive linear functional. The proof, hoewever, is purely algebraic and does not use any underlying functionals or measures.

Proof: That $f_0 = 1$ has no zeros is obvious. If, for some $n \in \mathbb{N}$, the point x is such that $f_n(x) = 0$, then the recurrence (3.5.3) yields

$$f_{n+1}(x) = -\gamma_n \, f_{n-1}(x),$$

hence f_{n+1} and f_{n-1} have the opposite sign at x so that either $f_{n+1}(x)\, f_{n-1}(x) < 0$ or[34] $f_{n+1}(x) = f_n(x) = f_{n-1}(x) = 0$. In the latter case we would also have that

$$f_{n-2}(x) = \frac{f_n(x) - (x + \beta_{n-1})\, f_{n-1}(x)}{\gamma_{n-1}} = 0,$$

and, repeating the argument, eventually $0 = f_{n-3}(x) = \cdots = f_0(x)$, contradicting $f_0 = 1$. Therefore, since n was arbitrary, and finite sequence $f_0, \dots, f_n$ is a Sturm chain.

This allows us to apply Theorem 3.5.3. Since σ_+ and σ_- may only capture a part[35] of the zeros of, we have for $I = [a, b)$, $a < b \in \mathbb{R}$, that

$$\left| \sigma_+ \left(\frac{f_n}{f_{n-1}}, I \right) - \sigma_- \left(\frac{f_n}{f_{n-1}}, I \right) \right| \le \sigma_+ \left(\frac{f_n}{f_{n-1}}, I \right) + \sigma_- \left(\frac{f_n}{f_{n-1}}, I \right) \le \#Z_{\mathbb{R}}(f_n) \le n. \tag{3.5.5}$$

Since all the polynomials are MONIC, i.e., $f_k(x) = x^k + \cdots$, it follows that

$$\lim_{x \to -\infty} f_k(x) = (-1)^k \infty, \qquad \lim_{x \to +\infty} f_k(x) = \infty,$$

hence,

$$\lim_{x \to -\infty} \operatorname{sgn} \begin{bmatrix} f_n(x) \\ f_{n-1}(x) \\ \vdots \\ f_0(x) \end{bmatrix} = (-1)^n \begin{bmatrix} 1 \\ -1 \\ \vdots \\ (-1)^n \end{bmatrix}, \qquad \lim_{x \to +\infty} \operatorname{sgn} \begin{bmatrix} f_n(x) \\ f_{n-1}(x) \\ \vdots \\ f_0(x) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

and we can conclude that

$$\lim_{a \to -\infty} V(a) = n, \qquad \lim_{b \to +\infty} V(b) = 0.$$

Thus, for sufficiently small a and sufficiently large b,

$$n = |V(b) - V(a)| = \left| \sigma_+ \left(\frac{f_n}{f_{n-1}}, I \right) - \sigma_- \left(\frac{f_n}{f_{n-1}}, I \right) \right|.$$

[34] This has not been excluded so far.

[35] We have not yet excluded double zeros or complex, nonreal ones.

Substituting this into (3.5.5) we get that

$$n \le \#Z_{\mathbb{R}}(f_n) \le n \qquad \Rightarrow \qquad \#Z_{\mathbb{R}}(f_n) = n,$$

as claimed. □

Actually, the proof tells us even more. Since σ_+ and σ_- are nonnegative numbers, the identity

$$-n = V(b) - V(a) = \sigma_+\left(\frac{f_n}{f_{n-1}}, \mathbb{R}\right) - \sigma_-\left(\frac{f_n}{f_{n-1}}, \mathbb{R}\right)$$

can only be obtained if

$$\sigma_+\left(\frac{f_n}{f_{n-1}}, \mathbb{R}\right) = 0 \qquad \text{und} \qquad \sigma_-\left(\frac{f_n}{f_{n-1}}, \mathbb{R}\right) = n$$

Hence, *all* sign changes of f_n/f_{n-1} are sign changes from $-$ to $+$. But this can only be obtained if f_{n-1} changes its sign between two sign changes of f_n. With this insight we can summarize the findings of this section in the following way.

Theorem 3.5.6. *If a polynomial sequence f_n, $n \in \mathbb{N}_0$, is defined by the recurrence* (3.5.3)*, then f_n has n simple real zeros, $n \in \mathbb{N}$, and the zeros of f_n and f_{n-1} are* NESTED.

This is a well–known property of orthogonal polynomials, cf. [7, 14], but we now know that actually it is a property of polynomials that satisfy a certain recursion, hence also a property of the convergents of certain continued fractions. That these continued fractions produce orthogonal polynomials, is again stated in Theorem 3.4.6.

3.6 Prony's problem

Finally, we relate continued fractions to yet another, seemingly unrelated problem which was considered and solved by Prony in [38]. It consists, in modern language, of recovering a function of a certain type, namely an EPONENTIAL SUM,

$$f(x) = \sum_{j=1}^{n} f_j\, e^{\omega_j x}, \qquad \omega_j \in \mathbb{R} + i\mathbb{T}, \qquad f_j \neq 0, \tag{3.6.1}$$

from samples, i.e., from finitely many function values which we assume to be equally distributed, and hence as $f(0), \dots, f(N)$, $N \in \mathbb{N}$. Of course, N will depend on n, at least if we want to obtain a reconstruction of f.

Remark 3.6.1 (Normalizations).

1. We normalize the FREQUENCIES ω_j to

$$\mathbb{R} + i\mathbb{T} = \mathbb{R} + i\,(\mathbb{R}/2\pi\mathbb{Z}) \simeq \mathbb{R} + i[-\pi, \pi],$$

   to avoid ambiguities in the representation (3.6.1) that may make the problem unsolvable, for example generating functions like $\sin(\pi\cdot) = \frac{1}{2i}\left(e^{i\pi\cdot} - e^{-i\pi\cdot}\right)$ that cannot be recovered from any subset of $\mathbb{Z}$.

2. The request that the COEFFICIENTS f_j are nonzero makes the representation SPARSE, that is, it contains no „phantom" frequencies.

3. Sampling on $0, \dots, N$ is no restriction since

$$f(ax+b) = \sum_{j=1}^{n} f_j \, e^{\omega_j \, (ax+b)} = \sum_{j=1}^{n} \left(e^{\omega_j \, b} f_j \right) \, e^{(a\omega_j) x} =: \sum_{j=1}^{n} \tilde{f}_j \, e^{\tilde{\omega}_j \, x}$$

shows that any AFFINE TRANSFORMATION only changes the coefficients and the frequencies but does not affect structure or solvability of the problem. In other words, sampling on $x_0 + kh$, $k = 0, \dots, N$, $x_0 \in \mathbb{R}$, $h > 0$, can be easily reduced to sampling at integers $0, \dots, N$.

The interesting part of Prony's problem consists of recovering the frequencies. Once these are known, one obtains the LINEAR SYSTEM

$$f(k) = \sum_{j=1}^{n} f_j \, e^{\omega_j \, k}, \qquad k = 0, \dots, N,$$

that can be written in the standard matrix form

$$\begin{bmatrix} f(0) \\ \vdots \\ f(N) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ e^{\omega_1} & \dots & e^{\omega_n} \\ \vdots & \ddots & \vdots \\ e^{N\,\omega_1} & \dots & e^{N\,\omega_n} \end{bmatrix} \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix}, \tag{3.6.2}$$

or

$$[f(j) : j = 0, \dots, N] = V \left[f_j : j = 1, \dots, n \right], \tag{3.6.3}$$

respectively. It is well–known that the VANDERMONDE MATRIX V has rank n whenever the ω_j are all distinct and $N \ge n-1$, so that the coefficients are uniquely determined already from n samples as soon as the frequencies are known.

Exercise 3.6.1 Show that for any distinct $\omega_1, \dots, \omega_n \in \mathbb{C}$ the matrix

$$\left[e^{j\omega_k} : \begin{array}{c} j = 0, \dots, n-1 \\ k = 1, \dots, n \end{array} \right] := \begin{bmatrix} 1 & \dots & 1 \\ e^{\omega_1} & \dots & e^{\omega_n} \\ \vdots & \ddots & \vdots \\ e^{(n-1)\,\omega_1} & \dots & e^{(n-1)\,\omega_n} \end{bmatrix} \in \mathbb{C}^{n \times n}$$

is invertible. **Hint**: polynomial interpolation ... ◊

Prony's ingenious idea to solve the problem consists of the following simple idea: let $p(x) = p_0 + p_1\, x + \cdots + p_m x^m$ be a polynomial of degree $m \ge n$ and consider, for fixed $0 \le j \le N-m$,

$$\begin{aligned} \sum_{k=0}^{m} f(j+k)\, p_k &= \sum_{k=0}^{m} \sum_{\ell=1}^{n} f_\ell \, e^{\omega_\ell \, (j+k)} p_k = \sum_{\ell=1}^{n} f_\ell e^{\omega_\ell \, j} \sum_{k=0}^{m} p_k \, (e^{\omega_\ell})^k \\ &= \sum_{\ell=1}^{n} f_\ell e^{\omega_\ell \, j} \, p\,(e^{\omega_\ell}) \,. \end{aligned}$$

In matrix notation this is

$$\begin{bmatrix} f(0) & \dots & f(m) \\ f(1) & \dots & f(m+1) \\ \vdots & \ddots & \vdots \\ f(N-m) & \dots & f(N) \end{bmatrix} \begin{bmatrix} p_0 \\ \vdots \\ p_m \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \dots & 1 \\ e^{\omega_1} & \dots & e^{\omega_n} \\ \vdots & \ddots & \vdots \\ e^{(N-m)\omega_1} & \dots & e^{(N-m)\omega_n} \end{bmatrix} \begin{bmatrix} f_1 & & \\ & \ddots & \\ & & f_n \end{bmatrix} \begin{bmatrix} p\left(e^{\omega_1}\right) \\ \vdots \\ p\left(e^{\omega_n}\right) \end{bmatrix} \tag{3.6.4}$$

and shows the appearance of yet another Vandermonde matrix. Taking into account the above remark, we get the following result that is at the heart of Prony's method.

Lemma 3.6.2. *If f is of the form* (3.6.1) *and $N \geq 2m-1$ then*

$$\sum_{k=0}^{m} f(j+k)\, p_k = 0 \qquad j = 0, \dots, N-m$$

if and only if

$$p\left(e^{\omega_j}\right) = 0, \qquad j = 0, \dots, n,$$

where $p = p_0 + p_1 x + \cdots + p_m x^m$.

Definition 3.6.3. The least degree polynomial p with $p(e^{\omega_j}) = 0$ is called the PRONY POLYNOMIAL for the function f from (3.6.1).

Lemma 3.6.2 already gives us a way to solve Prony's problem, i.e., to recover (3.6.1), provided that the number n of exponentials is known: determine the KERNEL of the HANKEL MATRIX

$$F_n := \begin{bmatrix} f(0) & \dots & f(n) \\ \vdots & \ddots & \vdots \\ f(n) & \dots & f(2n) \end{bmatrix} \in \mathbb{C}^{n+1 \times n+1},$$

identify the solution $p \in \mathbb{C}^{n+1}$ such that $F_n p = 0$ but $p \neq 0$ with a polynomial $p(x)$ and compute its zeros, these are e^{ω_j}, $j = 1, \dots, n$. This was already proposed by Prony in his original paper [38], see also [48], and much later refined into the algorithms MUSIC [49] and ESPRIT [41], both in the context of multisource radar.

There is, however, also an interpretation by means of continued fractions. To that end, we first note that $f(k)$ can be interpreted as a MOMENT SEQUENCE itself.

Definition 3.6.4. The DIRAC DISTRIBUTION δ_x for $x \in \mathbb{R}$ is defined as

$$\int_{\mathbb{R}} f(t)\, \delta_x(t)\, dt = f(x), \qquad f \in C_{00}(\mathbb{R}),$$

where $C_{00}(\mathbb{R})$ denotes the (real or complex valued) functions on $\mathbb{R}$ with COMPACT SUPPORT. Alternatively, one could use the POINT MEASURE

$$\int_{\mathbb{R}} f(t)\, d\mu_x(t) = f(x)$$

for all MEASURABLE f.

If we now define the measure

$$\mu := \sum_{j=1}^{n} f_j \, \mu_{e_j^\omega},$$

then we obtain the moments

$$\mu_k = \int_{\mathbb{R}} x^k d \left(\sum_{j=1}^{n} f_j \, \mu_{e_j^\omega} \right)(x) = \sum_{j=1}^{n} f_j \int_{\mathbb{R}} x^k \, d\mu_{e_j^\omega}(x) = \sum_{j=1}^{n} f_j \, (e^{\omega_j})^k = \sum_{j=1}^{n} f_j e^{\omega_j \, k} = f(k),$$

hence $f(k)$ is indeed a moment sequence for the (possible signed) point measure μ and we can consider the Laurent series

$$\mu(x) := \sum_{j=0}^{\infty} \mu_j \, x^{-j}$$

it defines, or even better

$$\lambda(x) := x^{-1}\mu(x) = \sum_{j=1}^{\infty} \mu_{j-1} \, x^{-j}, \qquad \text{i.e.} \qquad \lambda_j := \mu_{j-1}, \; \lambda_0 = 0. \tag{3.6.5}$$

The square Hankel matrices

$$M_n := \begin{bmatrix} \mu_0 & \cdots & \mu_n \\ \vdots & \ddots & \vdots \\ \mu_n & \cdots & \mu_{2n} \end{bmatrix} \in \mathbb{C}^{n+1 \times n+1} \tag{3.6.6}$$

can be considered as finite segments of the Hankel operator

$$M = \begin{bmatrix} \mu_0 & \mu_1 & \cdots \\ \mu_1 & \ddots & \ddots \\ \vdots & \ddots & \ddots \end{bmatrix}$$

that maps the sequence space

$$\ell(\mathbb{N}_0) := \{c = (c_k : k \in \mathbb{N}_0) : c_k \in \mathbb{C}\} \tag{3.6.7}$$

to itself by means of the correlation

$$(Mc)_k = \mu \star c = \sum_{j=0}^{\infty} \mu_{k+j} \, c_j, \qquad k \in \mathbb{Z}.$$

Definition 3.6.5. The rank of the Hankel operator M is defined as

$$\operatorname{rank} M := \sup_{n \in \mathbb{N}_0} \operatorname{rank} M_n = \sup_{n \in \mathbb{N}_0} \operatorname{rank} \begin{bmatrix} \mu_0 & \cdots & \mu_n \\ \vdots & \ddots & \vdots \\ \mu_n & \cdots & \mu_{2n} \end{bmatrix}. \tag{3.6.8}$$

The sequence μ is called nondegenerate if, for $n = \operatorname{rank} M$

$$1 = \operatorname{rank} M_0 < \operatorname{rank} M_1 < \cdots < \operatorname{rank} M_{n-1} = \operatorname{rank} M_n = \cdots = \operatorname{rank} M. \tag{3.6.9}$$

We already know Hankel operators of finite rank. Indeed, if we set

$$\mu_k = \sum_{j=1}^{n} f_j e^{\omega_j k}, \qquad k \in \mathbb{N}_0,$$

as in Prony's problem or moments of finite sums of point measures, then we continue (3.6.4) to get for $k \in \mathbb{N}_0$ and $p \in \mathbb{C}^{k+1}$ the identity

$$\begin{aligned} M_k\, p &= \begin{bmatrix} 1 & \dots & 1 \\ e^{\omega_1} & \dots & e^{\omega_n} \\ \vdots & \ddots & \vdots \\ e^{k\omega_1} & \dots & e^{k\omega_n} \end{bmatrix} \begin{bmatrix} f_1 & & \\ & \ddots & \\ & & f_n \end{bmatrix} \begin{bmatrix} p\,(e^{\omega_1}) \\ \vdots \\ p\,(e^{\omega_n}) \end{bmatrix} \\ &= \begin{bmatrix} 1 & \dots & 1 \\ e^{\omega_1} & \dots & e^{\omega_n} \\ \vdots & \ddots & \vdots \\ e^{k\omega_1} & \dots & e^{k\omega_n} \end{bmatrix} \begin{bmatrix} f_1 & & \\ & \ddots & \\ & & f_n \end{bmatrix} \begin{bmatrix} 1 & e^{\omega_1} & \vdots & e^{k\omega_1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{\omega_n} & \vdots & e^{k\omega_n} \end{bmatrix} \begin{bmatrix} p_0 \\ \vdots \\ p_k \end{bmatrix} \end{aligned}$$

which can be summarized as

$$M_k = \underbrace{\begin{bmatrix} 1 & \dots & 1 \\ e^{\omega_1} & \dots & e^{\omega_n} \\ \vdots & \ddots & \vdots \\ e^{k\,\omega_1} & \dots & e^{k\,\omega_n} \end{bmatrix}}_{k+1\times n} \underbrace{\begin{bmatrix} f_1 & & \\ & \ddots & \\ & & f_n \end{bmatrix}}_{n\times n} \underbrace{\begin{bmatrix} 1 & e^{\omega_1} & \dots & e^{k\,\omega_1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{\omega_n} & \dots & e^{k\,\omega_n} \end{bmatrix}}_{n\times k+1} =: V_{k,\Omega}\, F\, V_{k,\Omega}^T. \tag{3.6.10}$$

with the VANDERMONDE MATRIX $V_{k,\Omega}$ and the nonsingular diagonal matrix $F := \operatorname{diag}(f_1, \dots, f_n)$. Noting that the rank of $V_{k,\Omega}$ is $\min(k+1, n)$, we can record that

$$\operatorname{rank} M = n. \tag{3.6.11}$$

Remark 3.6.6. It is easy to construct degenerate measures by means of exponential functions (3.6.1). The rank of the associated Hankel operator will be n. Let us only choose arbitrary frequencies ω_j as well as $0 \le k < k' \le n$ and a polynomial $p \in \Pi_k$ with $p\,(e^{\omega_j}) \neq 0$, $j = 1, \dots, n$. Now we let $f \in \mathbb{R}^n$ be any solution of the undetermined system

$$\begin{aligned} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} &= \begin{bmatrix} (p\,(e^{\omega_1}))^2 & \dots & (p\,(e^{\omega_n}))^2 \\ e^{\omega_1}\,(p\,(e^{\omega_1}))^2 & \dots & e^{\omega_n}\,(p\,(e^{\omega_n}))^2 \\ \vdots & \ddots & \vdots \\ (e^{\omega_1})^{k'-k}\,(p\,(e^{\omega_1}))^2 & \dots & (e^{\omega_n})^{k'-k}\,(p\,(e^{\omega_n}))^2 \end{bmatrix} f \\ &= \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ (e^{\omega_1})^{k'-k} & \dots & (e^{\omega_n})^{k'-k} \end{bmatrix} \begin{bmatrix} (p\,(e^{\omega_1}))^2 & & \\ & \ddots & \\ & & (p\,(e^{\omega_n}))^2 \end{bmatrix} f. \end{aligned} \tag{3.6.12}$$

Such a solution exists since the matrix in (3.6.12) has maximal rank $k' - k + 1$. Since, for

$0 \le \ell \le k' - k$

$$
\begin{aligned}
&\left[0_{1\times\ell} p^T \right] M_{k+\ell} \begin{bmatrix} 0_{\ell\times 1} \\ p \end{bmatrix} \\
&= \sum_{r,s=0}^{k} f(\ell + r + s)\, p_r p_s = \sum_{r,s=0}^{k} \sum_{j=1}^{n} f_j\, e^{\omega_j (r+s+\ell)}\, p_r p_s \\
&= \sum_{j=1}^{n} f_j e^{\omega_j \ell} \sum_{r,s=0}^{k} p_r \left(e^{\omega_j}\right)^r\, p_r \left(e^{\omega_j}\right)^s\, p_s = \sum_{j=1}^{n} f_j e^{\omega_j \ell} \left(p\left(e^{\omega_j}\right)\right)^2 = \begin{cases} 1, & \ell = 0, \\ 0, & \ell = 1, \dots, k' - k - 1, \\ 1, & \ell = k - k', \end{cases}
\end{aligned}
$$

it follows immediately that

$$
\operatorname{rank} M_{k-1} < \operatorname{rank} M_k = \cdots = \operatorname{rank} M_{k'-1} < \operatorname{rank} M_{k'}
$$

and all the Hankel matrices M_j, $j = k+1, \dots, k'$, are automatically *singular.*

Hankel operators of finite rank can be characterized in many equivalent ways, one of which we will give next, cf. [15, 35].

Theorem 3.6.7 (KRONECKER'S THEOREM[36]). *The Hankel operator M has finite rank if and only if $\mu(x)$ is a rational funtion, i.e.,*

$$
\mu(x) = \frac{p(x)}{q(x)}, \qquad p, q \in \Pi. \tag{3.6.13}
$$

Remark 3.6.8. That the Hankel operator is of finite rank does *not* mean that the associated sequence μ is finitely supported, quite the contrary. It can be shown that any finitely supported sequence μ always defines a Hankel operator of infinite rank - at least as long as it is nonzero.

To prove the theorem, we introduce the bilinear form

$$
(\cdot, \cdot) : \ell(\mathbb{Z}) \times \Pi \to \ell(\mathbb{Z}), \qquad (\mu, p) := \mu \star p, \tag{3.6.14}
$$

and note that, for the SHIFT OPERATOR τ, $(\tau c)_k = c_{k+1}$ we have

$$
(\tau\mu, p)_j = (\tau\,(\mu, p))_j = \sum_{k=0}^{\infty} \mu_{j+1+k}\, p_k = \sum_{k=1}^{\infty} \mu_{j+k} p_{k-1} = (\mu, (\cdot) p)_j\,. \tag{3.6.15}
$$

Though (3.6.15) is almost trivial to prove[37], it has a fundamental consequence.

Lemma 3.6.9. *The set*

$$
\ker(\mu, \cdot) = \{ p \in \Pi : (\mu, p) = 0 \} \tag{3.6.16}
$$

is an ideal, *i.e., it is closed under addition and multiplication with arbitrary polynomials.*

[36] There are several, quite different results known as *Kronecker's theorem*, for example also a number theoretic one on lattices generated by real numbers that are linearly independent over $\mathbb{Q}$, see [20], which, by the way, also contains a nice chapter on continued fractions. So the lesson is that the name alone is not always helpful, one should look for the *meaning* of a result.

[37] It is just an index shift.

Proof: The shift invariance of the zero sequence gives

$$0 = \tau 0 = \tau(\mu, p) = (\mu, (\cdot)\, p), \qquad p \in \ker(\mu, \cdot),$$

and closure under addition is trivial because of bilinearity. □

Proof of Theorem 3.6.7: If M is of finite rank, then

$$0 \in \left\{ M \begin{bmatrix} p \\ 0 \end{bmatrix} : p \in \Pi \setminus \{0\} \right\},$$

as otherwise the rank would be infinite. Thus, there exists $0 \neq q \in \Pi$ of minimal degree such that $0 = (\mu, q) = (\mu, \Pi\, q)$, where $\Pi\, q$ denotes the PRINCIPAL IDEAL generated by q. Thus,

$$\begin{aligned} 0 &= (\mu, q)(x) = \sum_{j=0}^{\infty} (\mu, q)_j \; x^{-j} = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \mu_{j+k}\, q_k\, x^{-j} = \sum_{j,k=0}^{\infty} \mu_{j+k} x^{-j-k}\, q_k x^k \\ &= \sum_{k=0}^{n} q_k x^k \sum_{j=k}^{\infty} \mu_j\, x^{-j} = \sum_{k=0}^{n} q_k x^k \left(\mu(x) - \sum_{j=0}^{k-1} \mu_j\, x^{-j} \right) \\ &= q(x)\, \mu(x) - \sum_{k=0}^{n} q_k \sum_{j=0}^{k-1} \mu_j x^{k-j} \end{aligned}$$

that is,

$$\mu(x) = \frac{p(x)}{q(x)}$$

with

$$p(x) = \sum_{k=0}^{n} q_k \sum_{j=0}^{k-1} \mu_j x^{k-j} = \sum_{k=0}^{n} q_k \sum_{j=0}^{k-1} \mu_{k-1-j} x^{j+1} = x \sum_{j=0}^{n-1} x^j \sum_{k=j+1}^{n} q_k\, \mu_{k-(j+1)} \tag{3.6.17}$$

as claimed.

For the converse, we note that[38]

$$\mu(x)\, q(x) = p(x), \qquad p \in \Pi_m,\ q \in \Pi_n, \qquad q_n \neq 0,$$

implies[39], setting $q_k = 0$ for $k < 0$,

$$\begin{aligned} \sum_{j=0}^{m} p_j x^j &= \left(\sum_{j=0}^{\infty} \mu_j\, x^{-j} \right) \left(\sum_{k=0}^{n} q_k x^k \right) = \sum_{j=0}^{\infty} \sum_{k=0}^{n} \mu_j\, q_k x^{k-j} = \sum_{k=0}^{n} \sum_{j=-k}^{\infty} x^{-j}\, q_k \mu_{j+k} \\ &= \sum_{j=-n}^{\infty} x^{-j} \sum_{k=-j}^{n} \mu_{j+k}\, q_k = \sum_{j=-n}^{\infty} x^{-j} \sum_{k=\max(0,-j)}^{n} \mu_{j+k}\, q_k. \end{aligned}$$

Since the left hand side of this equation is a polynomial, it follows that all coefficients with negative power have to have a zero coefficient, i.e.,

$$0 = \sum_{k=0}^{n} \mu_{j+k}\, q_k = (\mu \star q)_j = \left(M \begin{bmatrix} q \\ 0 \end{bmatrix} \right)_j, \qquad j \in \mathbb{N}_0, \tag{3.6.18}$$

[38] We request that the denominator polynomial has degree *exactly* n.

[39] For „advaced tricks“ to manipulate double sums, see [16].

so that, by Lemma 3.6.9,

$$\ker(\mu,\cdot) \supseteq q\,\Pi \qquad \Rightarrow \qquad 0 = M \begin{bmatrix} (\cdot)^k\, q \\ 0 \end{bmatrix} = M \begin{bmatrix} 0_k \\ q \\ 0 \end{bmatrix}, \quad k \in \mathbb{N}_0,$$

hence $\operatorname{rank} M \le \deg q = n$. □

Exercise 3.6.2 Show that the infinite vectors $\begin{bmatrix} 0_k \\ q \\ 0 \end{bmatrix}$, $k \in \mathbb{N}_0$, are *linearly independent.* ◇

Remark 3.6.10. The explicit formula (3.6.17) shows that the numerator polynomial $p(x)$ in (3.6.13) is always of the form $p(x) = x\,\tilde p(x)$, hence

$$\mu(x) = x\,\frac{\tilde p(x)}{q(x)}, \qquad \text{i.e.,} \qquad \lambda(x) = \frac{\tilde p(x)}{q(x)},$$

which indicates that the shifted sequence λ from (3.6.5) may be more appropriate to consider later.

Definition 3.6.11. A Hankel operator will be called SIMPLE if it has finite rank and the denominator in the normalized representation (3.6.13) has only *simple zeros.*

Remark 3.6.12. It is not really difficult to extend the theory to the case of multiple zeros of q. The functions to consider are still of the type (3.6.1), but now the coefficients are *polynomials* whose degree is one less than the multiplicity of the respective zero. Indeeed, this extension even works in several variables, cf. [34, 46, 47].

An inspection of the proof of Theorem 3.6.7 leads to the following observation.

Corollary 3.6.13 (Hankel & Prony).

1. *The polynomial q in the normalized representation $\mu(x) = \frac{p(x)}{q(x)}$ is the Prony polynomial for μ.*
2. *Any simple Hankel operator is generated by exponential functions, i.e., $\mu_j = f(j)$ for some f of the form* (3.6.1).
3. *Any simple Hankel operator factorizes as*

$$M = \underbrace{\begin{bmatrix} 1 & \dots & 1 \\ e^{\omega_1} & \dots & e^{\omega_n} \\ e^{2\omega_1} & \dots & e^{2\omega_n} \\ \vdots & \ddots & \vdots \end{bmatrix}}_{=:V_\Omega} \begin{bmatrix} f_1 & & \\ & \ddots & \\ & & f_n \end{bmatrix} \underbrace{\begin{bmatrix} 1 & e^{\omega_1} & e^{2\omega_1} & \dots \\ \vdots & \vdots & \vdots & \ddots \\ 1 & e^{\omega_n} & e^{2\omega_n} & \dots \end{bmatrix}}_{=V_\Omega^T}. \tag{3.6.19}$$

Proof: For 1), we note that q was defined by the property $(\mu, q) = 0$, which is in turn the definition of the Prony polynomial.

To verify 2), we divide q by any factor of the form $(\cdot)^k$, $k \in \mathbb{N}$, if necessary[40], normalize it into a monic polynomial and let $e^{\omega_1}, \dots, e^{\omega_n}$ be the (remaining) zeros of q, i.e.

$$q = (\cdot - e^{\omega_1}) \cdots (\cdot - e^{\omega_n}).$$

Then the proof of Theorem 3.6.7 shows that μ is a solution of the homogeneous DIFFERENCE EQUATION

$$0 = \sum_{k=0}^{n} \mu_{j+k}\, q_k, \qquad j \in \mathbb{N}_0.$$

This solution space has dimension n, cf. [27], and since

$$\sum_{k=0}^{n} e^{\omega(j+k)} q_k = e^{\omega j} \sum_{k=0}^{n} q_k\, e^{\omega k} = e^{\omega j}\, q(e^{\omega}), \qquad j \in \mathbb{N}_0,$$

we see that the sequences $k \mapsto e^{\omega_j k}$, $j = 1, \dots, n$, form a basis for this space. Consequently, μ must be a linear combination of theses sequences, hence of the form (3.6.1).

For 3) we first record that, according to 2), we can write

$$\mu_k = \sum_{j=1}^{n} f_j e^{\omega_j k}, \qquad k \in \mathbb{N}_0,$$

hence, for $k, \ell \in \mathbb{N}_0$,

$$\begin{aligned}
e_k^T M e_\ell &= \mu_{k+\ell} = \sum_{j=1}^{n} f_j e^{\omega_j (k+\ell)} = \sum_{j=1}^{n} f_j e^{\omega_j k} e^{\omega_j \ell} \\
&= \left[e^{\omega_1 k}, \dots, e^{\omega_n k}\right] \begin{bmatrix} f_1 & & \\ & \ddots & \\ & & f_n \end{bmatrix} \begin{bmatrix} e^{\omega_1 j} \\ \vdots \\ e^{\omega_n j} \end{bmatrix} \\
&= e_k^T \begin{bmatrix} 1 & \dots & 1 \\ e^{\omega_1} & \dots & e^{\omega_n} \\ e^{2\omega_1} & \dots & e^{2\omega_n} \\ \vdots & \ddots & \vdots \end{bmatrix} \begin{bmatrix} f_1 & & \\ & \ddots & \\ & & f_n \end{bmatrix} \begin{bmatrix} 1 & e^{\omega_1} & e^{2\omega_1} & \dots \\ \vdots & \vdots & \vdots & \ddots \\ 1 & e^{\omega_n} & e^{2\omega_n} & \dots \end{bmatrix} e_\ell,
\end{aligned}$$

which is (3.6.19). Note that this is the „infinite version" of the argument that lead to the finite factorization (3.6.10). □

Now we can combine our findings with Theorem 3.4.19. The shifted moment sequence λ from (3.6.5) has an associated continued fraction expansion if and only if $\det \Lambda_n \neq 0$ which is in turn equivalent to μ being NONDEGENERATE. If this is satisfied, we can apply the full machinery of continued fractions to Prony's problem.

Corollary 3.6.14. *If, for an exponential f of the form* (3.6.1), *the sequence $\lambda = (f(j-1) : j \in \mathbb{N})$ and $\lambda_0 = 0$ is nondegenerate, then the continued fraction expansion of $\lambda(x)$ terminates after n steps and the denominator of the convergent is the Prony polynomial.*

Exercise 3.6.3 Derive a recurrence relation that eventually computes the Prony polynomial. ◇

[40]Without mentioning it explicitly, we use here the more convenient approach of considering rational functions of LAURENT POLYNOMIALS.

3.7 Flat extensions of moment sequences

Finally, let us briefly touch the issue of TRUNCATED MOMENT SEQUENCES, i.e., the question how moment sequences can be extended. Here one usually restricts oneself to the case that the initial segment $\mu_0, \dots, \mu_{2n}$ of a moment sequence μ is known and assumes that

$$M_n := \begin{bmatrix} \mu_0 & \cdots & \mu_n \\ \vdots & \ddots & \vdots \\ \mu_n & \cdots & \mu_{2n} \end{bmatrix}$$

is SYMMETRIC and POSITIVE DEFINITE which implies the same for the Hankel submatrices M_k, $k = 0, \dots, n-1$. This can be seen as the moments coming from a linear functional that is at least SQUARE POSITIVE on Π_{2n}. Based on that knowledge, we define a particular type of extension of the MOMENT SEQUENCE μ.

Definition 3.7.1. A sequence $\hat\mu \in \ell(\mathbb{N}_0)$ is called a FLAT EXTENSION of the moment sequence $\mu = (\mu_0, \dots, \mu_{2n}, \dots)$ if

1. $\hat\mu_j = \mu_j$, $j = 0, \dots, 2n$,
2. $\operatorname{rank} \hat M_k = \operatorname{rank} M_n = n+1$, $k \ge n$.

In other words, a flat extension leads to a moment sequence whose associated HANKEL OPERATOR has rank $n+1$; „flatness“ means that the dimension is all defined by means of M_n from where on the rank is constant:

$$1 = \operatorname{rank} \hat M_0 < \operatorname{rank} \hat M_1 < \cdots < \operatorname{rank} \hat M_n = \operatorname{rank} \hat M_{n+1} = \cdots = n+1. \tag{3.7.1}$$

Continued fractions help us to construct flat extensions.

Theorem 3.7.2. *Any moment sequence μ such that M_n is positive definite has a flat extension $\hat\mu$.*

Proof: We construct the sequence of convergents for $\lambda(x)$ with $\Lambda_k = M_{k-1}$, $k = 1, \dots, n+1$. Since

$$0 < \det M_j, \qquad j = 0, \dots, n,$$

Theorem 3.4.19 implies that

$$\lambda(x) = [0; r_1, \dots, r_{n+1}, \dots] \qquad \text{and} \qquad \frac{p_j(x)}{q_j(x)} = \sum_{k=1}^{2j-2} \mu_k x^{-k} + \cdots, \qquad k = 1, \dots, n+1,$$

and at least the first $n+1$ convergents are well–defined. Setting

$$\hat\mu(x) = \frac{p_{n+1}(x)}{q_{n+1}(x)}$$

then already gives the desired flat extension. □

By Proposition 3.5.4, the zeros of q_{n+1} are real and simple, hence can be written as e^{ω_j}, $j = 1, \dots, n+1$, as long as[41] $q_{n+1}(0) \neq 0$. In other words,

$$q_{n+1}(x) = q_{n+1,n+1} \prod_{j=1}^{n+1} (x - e^{\omega_j}), \qquad q_{n+1,n+1} \in \mathbb{R} \setminus \{0\}. \tag{3.7.2}$$

[41] A zero at the origin is a „spurious“ zero when passing to Laurent polynomials and must be excluded in this theory.

Then Corollary 3.6.13 implies that, defining a finite rank Hankel operator, the flat extension $\hat{\mu}$ must be samples of an exponential function

$$f(x) = \sum_{j=1}^{n+1} f_j \, e^{\omega_j x}, \qquad f_j \in \mathbb{R}, \quad j = 1, \dots, n+1.$$

Finally, define

$$\Pi_n \ni \ell_j(x) = \prod_{k \neq j} \frac{x - e^{\omega_k}}{e^{\omega_j} - e^{\omega_k}} = C \, \frac{q_{n+1}(x)}{x - e^{\omega_j}}, \qquad C \in \mathbb{R} \setminus \{0\},$$

note that $\ell_j(e^{\omega_k}) = \delta_{j,k}$, $j, k = 1, \dots, n+1$, and apply (3.6.4) to obtain that

$$M_n \ell_j = \begin{bmatrix} 1 & \dots & 1 \\ e^{\omega_1} & \dots & e^{\omega_{n+1}} \\ \vdots & \ddots & \vdots \\ e^{n\,\omega_1} & \dots & e^{n\,\omega_{n+1}} \end{bmatrix} \begin{bmatrix} f_1 & & \\ & \ddots & \\ & & f_{n+1} \end{bmatrix} \begin{bmatrix} \ell_j(e^{\omega_1}) \\ \vdots \\ \ell_j(e^{\omega_{n+1}}) \end{bmatrix} = f_j \begin{bmatrix} 1 \\ e^{\omega_j} \\ \vdots \\ e^{n\,\omega_j} \end{bmatrix},$$

i.e., $f_j = \left(M_n \ell_j \right)_1$, which even gives a direct way to obtain the coefficients $f_j > 0$. Summarizing all that, we get the final small piece of insight.

Corollary 3.7.3. *A flat extension of a moment sequence is equivalent to a Gaussian quadrature formula.*

Signal processing, Hurwitz and Stieltjes

4

When the epoch of analogue (which was to say also the richness of language, of analogy*) was giving way to the digital era, the final victory of the numerate over the literate.*

(S. Rushdie, *Fury*)

Even in SIGNAL PROCESSING continued fractions are unavoidable, this time by means of a classical theorem due to STIELTJES from [11]. The context will be HURWIRTZ POLYNOMIALS which are, in turn, closely related to the stability of an IIR FILTER. To understand what this really means, we first need some additional terminology.

4.1 Signals and filters

A *time discrete* SIGNAL is a doubly infinite sequence of the form

$$\sigma = \left(\sigma_j \ : \ j \in \mathbb{Z}\right) \in \ell(\mathbb{Z}).$$

Of course, realistic signals have a beginning and an end, hence a FINITE SUPPORT[1] and at least FINITE ENERGY, i.e.,

$$\|\sigma\|_2 = \left(\sum_{j\in\mathbb{Z}} \left|\sigma_j\right|^2\right)^{1/2}.$$

Anyway, it is much more convenient to work with bi–infinite signals as we do not have to worry about any boundary issues which are very inconvenient to track.

Definition 4.1.1 (Filter). A FILTER $F : \ell(\mathbb{Z}) \to \ell(\mathbb{Z})$ is an operator on the discrete signal space. It is called an LTI FILTER[2] if F is a linear operator that is TIME INVARIANT, i.e.

$$\sigma_j' = \sigma_{j+k}, \quad j \in \mathbb{Z} \qquad \Rightarrow \qquad \left(F\sigma'\right)_j = \left(F\sigma\right)_{j+k}, \quad j \in \mathbb{Z}. \tag{4.1.1}$$

Remark 4.1.2. It is common practice in signal processing to use „filter“ or „digital filter“ synonomously for „LTI filter“, cf. [19].

With the SHIFT OPERATOR $\tau : \ell(\mathbb{Z}) \to \ell(\mathbb{Z})$, defined as $(\tau\sigma)_j = \sigma_{j+1}$ there is a nice and simple way to describe LTI filters.

Lemma 4.1.3. *A filter F is an LTI filter if and only if it commutes with τ, i.e.,*

$$\tau F = F\tau. \tag{4.1.2}$$

[1]One might even say COMPACT SUPPORT which is the same for discrete signals.
[2]**L**inear **T**ime **I**nvariant

Proof: Writing σ' in (4.1.1) as $\sigma' = \tau^k \sigma$, the LTI property is equivalent to

$$F\tau^k \sigma = F\sigma' = \tau^k F\sigma, \qquad k \in \mathbb{Z},$$

hence (4.1.2) follows for any LTI filter by setting $k = 1$. Conversely, we simply observe that (4.1.2) implies for $k > 0$ that

$$\tau^k F = \tau^{k-1} F\tau = \cdots = F\tau^k,$$

and the argument for $k < 0$ is similar. □

Any linear filter F can be written as a bi-infinite matrix $F = \left[F_{jk} \; : \; j,k \in \mathbb{Z}\right]$, such that

$$(Ff)_j = \sum_{k\in\mathbb{Z}} F_{jk}\, f_k, \qquad j \in \mathbb{Z}.$$

If F is an LTI filter, then

$$\left[F_{j+1,k} \; : \; j,k \in \mathbb{Z}\right] = \tau F = F\tau = \left[F_{j,k-1} \; : \; j,k \in \mathbb{Z}\right]. \tag{4.1.3}$$

To verify (4.1.3), recall that for $j \in \mathbb{Z}$

$$((F\tau)\, f)_j = (F\,(\tau f))_j = \sum_{k\in\mathbb{Z}} F_{j,k} (\tau f)_k = \sum_{k\in\mathbb{Z}} F_{j,k} f_{k+1} = \sum_{k\in\mathbb{Z}} F_{j,k-1} f_k.$$

Since the two biinfinite matrices in (4.1.3) define the same operator, they must coincide in all components, hence $F_{j+1,k} = F_{j,k-1}$, $j,k \in \mathbb{Z}$, or, after iteration thereof,

$$F_{j+\ell,k} = F_{j,k-\ell}, \qquad \ell \in \mathbb{Z}.$$

This holds true whenever $F_{jk} = f_{j-k}$ for some $f \in \ell(\mathbb{Z})$, but conversely we also have that $j - k = \ell - m$ implies $j - \ell = k - m$ and thus

$$F_{jk} = F_{\ell+(j-\ell),k} = F_{\ell,k-(k-m)} = F_{\ell,m}$$

so that F_{jk} depends only on $j - k$, which can be summarized as follows.

Proposition 4.1.4. *A filter $F \; : \; \ell(\mathbb{Z}) \to \ell(\mathbb{Z})$ is an LTI filter if there exists $f \in \ell(\mathbb{Z})$ such that $F_{jk} = f_{j-k}$, $j,k \in \mathbb{Z}$. In that case,*

$$(F\sigma)_j = \sum_{k\in\mathbb{Z}} f_{j-k}\, \sigma_k, \qquad j \in \mathbb{Z}. \tag{4.1.4}$$

Definition 4.1.5. The sum in (4.1.4) is called the CONVOLUTION $f * \sigma$ between f and σ. F is then called a TOEPLITZ OPERATOR.

Remark 4.1.6. A Toeplitz operator is almost the same as a Hankel operator, just with the difference that the matrix elements are formed as f_{j-k} in the first and f_{j+k} in the second case, respectively.

Next, some terminology.

Definition 4.1.7 (Pulse, filter types and z transform).

1. The PULSE $\delta \in \ell(\mathbb{Z})$ is defined as $\delta_j = \delta_{j0}$.

2. The *impulse response* of a filter F is the signal $F\delta$.

3. The SUPPORT of a signal $\sigma \in \ell(\mathbb{Z})$ is defined as

$$\operatorname{supp}\sigma = \left\{ j \in \mathbb{Z} \ : \ \sigma_j \neq 0 \right\},$$

and the ZERO NORM[3] is $\|\sigma\|_0 := \# \operatorname{supp}\sigma$.

4. A filter is called FIR FILTER[4] if it is an LTI filter with finitely supported impulse response:

$$F\delta \in \ell_0(\mathbb{Z}) := \{\sigma \in \ell(\mathbb{Z}) \ : \ \|\sigma\|_0 < \infty\} = \left\{\sigma \in \ell(\mathbb{Z}) \ : \ \# \operatorname{supp}\sigma < \infty\right\}.$$

Otherwise the filter is called an IIR FILTER[5]

5. The z TRANSFORM of a signal $f \in \ell(\mathbb{Z})$ is the formal bi-infinite LAURENT SERIES

$$f(z) = \sum_{k \in \mathbb{Z}} f_k \, z^{-k}, \qquad z \in \mathbb{C}^\times = \mathbb{C} \setminus \{0\}.$$

The reason for the introduction of the z transform is easily seen: for arbitrary signals $f, g \in \ell(\mathbb{Z})$ one has

$$(f * g)(z) = \sum_{j \in \mathbb{Z}} \left(\sum_{k \in \mathbb{Z}} f_{j-k} \, g_k \right) z^{-j} = \sum_{j \in \mathbb{Z}} \sum_{k \in \mathbb{Z}} f_j \, g_k \, z^{-j-k} = \left(\sum_{j \in \mathbb{Z}} f_j z^{-j} \right) \left(\sum_{k \in \mathbb{Z}} g_k z^{-k} \right)$$

and hence

$$(f * g)(z) = f(z) \, g(z), \tag{4.1.5}$$

so that the z transform turns convolutions into products. In particular any LTI filer F can be expressed as

$$(F\sigma)(z) = f(z) \, \sigma(z). \tag{4.1.6}$$

That we turn convolutions into multiplications is, however, only one part of the story. More relevantly, we can implement *fast* filterings by setting $z = e^{-i\omega}$ in (4.1.6), discretizing the whole thing and applying the FAST FOURIER TRANSFORM (FFT), see for example [32, 52]. `Matlab` und `octave` have a special routine, `fftfilt`, for that purpose. Roughly speaking, the COMPUTATIONAL COMPLEXITY of filtering with a filter of length[6] N can be reduced from $O\left(N^2\right)$ to the significantly better and probably optimal[7] $O(N \log N)$.

If F is an FIR filter, its z transform is of the form

$$f(z) = \sum_{j=n_0}^{n_1} f_j z^j, \qquad n_0 \leq n_1 \in \mathbb{Z},$$

[3]Which is *not* a norm!

[4]**F**inite **I**mpulse **R**esponse

[5]**Exercise**: discover the meaning of „I“, perhaps by exhaustive literature research.

[6]The length is the difference between the largest and smallest index of a nonzero filter coefficient, hence the size of the support interval.

[7]To my knowledge there exists no proof that the FFT is really optimal for its job. However, since it re-invention [6], see [4, 5] for some historical remarks, noone found anything better. And meanwhile FFT is also used to multiply certain matrices or even large integers, see [12, 51].

i.e., a LAURENT POLYNOMIAL. In case $n_0 \geq 0$, hence $\operatorname{supp} f \subseteq \mathbb{N}_0$, the filter is called a CAUSAL FILTER as for any $j \in \mathbb{Z}$ one has

$$(F\sigma)_j = \sum_{k \in \mathbb{Z}} f_{j-k}\, \sigma_k = \sum_{k \in \mathbb{Z}} f_k\, \sigma_{j-k} = \sum_{k \in \mathbb{N}_0} f_k\, \sigma_{j-k},$$

and the filtered signal at time j depends only on σ_k, $k \leq j$, that is, on knowledge from the past. This is what real time filters can realize, predicting the future is usually nontrivial to implement.

4.2 Rational filters and stability

FIR filters are a nice thing since they can be realized physically at least when a certain LATENCY, i.e., a delay of the output, is accepted. Indeed, any FIR FILTER can be built by

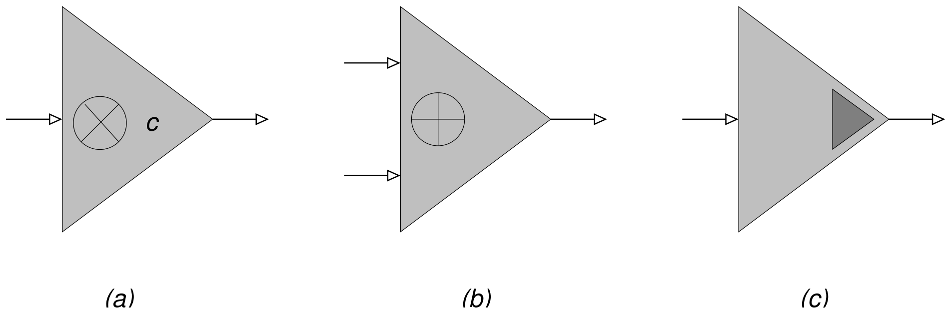


Abbildung 4.2.1: Symbolic representation of the three components: multiplier (a), adder (b) and delay element (c).

cascading the three components from Fig. 4.2.1. Such a cascade for a causal FIR filter with coefficients $f_0, \dots, f_N$ is shown in Fig. 4.2.2. On the other hand, FIR filters are somewhat

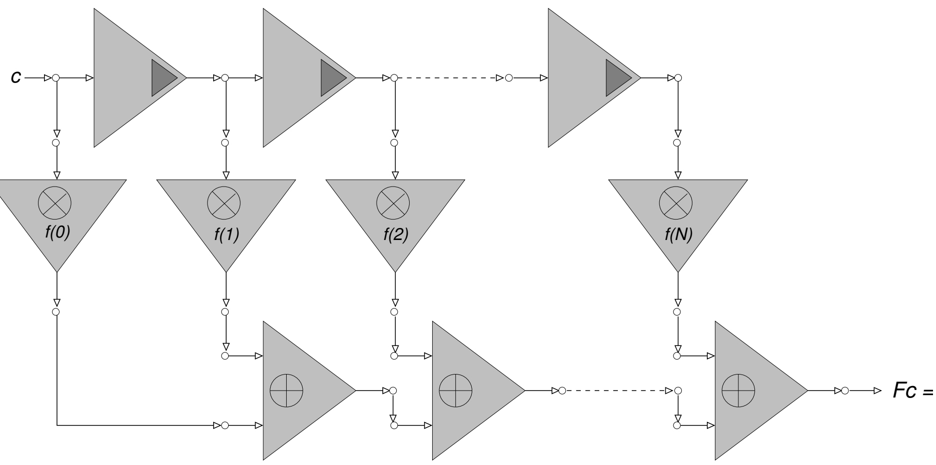
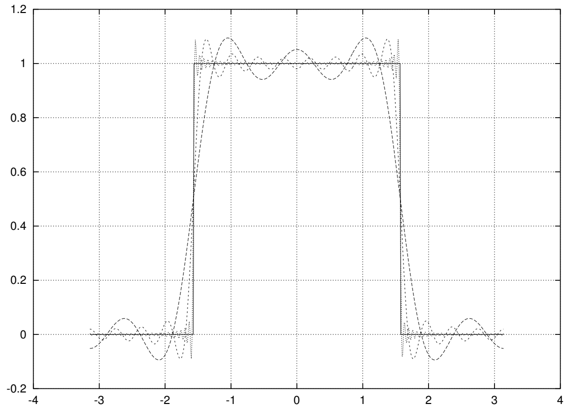


Abbildung 4.2.2: Realization of an FIR FILTER by means of the components from Fig. 4.2.1. The delay elements take care of the translations and the latency of the system is N clock tics.

limited in their flexibility, in particular when one wants to realized band pass filters with precise localization. A BAND PASS FILTER is a filter that blocks everything except a certain frequency band. Its Fourier transform or transfer function would ideally be a characteristic function, which is impossible since the Fourier transform of an FIR filter is a trigonometric polynomial. Even worse, the best approximation[8] to a band pass filter by means of FIR filters *always* shows an oscillation behavior, known as the GIBBS PHENOMENON, see Fig.4.2.3. This can be repaired by different approximation methods, but the price to pay is a loss in accuracy and localization.



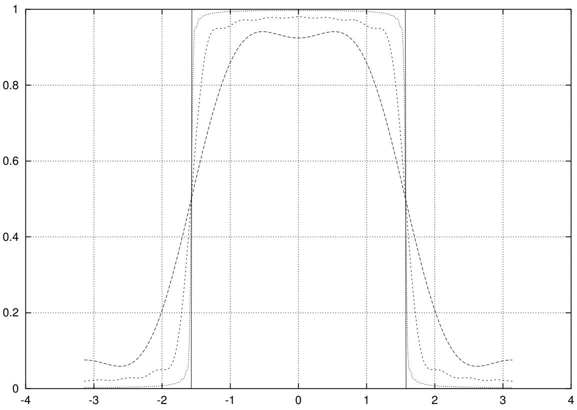


Abbildung 4.2.3: *Left:* (Best) approximation of a bandpass by partial sums of the associated Fourier series for $n = 5, 15, 100$ to illustrate the Gibbs phenonemon. Observe that the overshooting effects only get more narrow, not smaller. *Right:* A SHAPE PRESERVING APPROXIMATION by so called FEJÉR MEANS. The overall quality is not so good, but the oscillations are gone.

Another approach is to extend the class of admissible filters by choosing rational instead of polynomial functions.

Definition 4.2.1. A RATIONAL FILTER F is a filter that has a rational function as its z transform,

$$(F\sigma)(z) = f(z)\,\sigma(z) = \frac{p(z)}{q(z)}\,\sigma(z), \qquad p(z) = \sum_{j\in\mathbb{N}_0} p_j\, z^{-j}, \quad q(z) = \sum_{j\in\mathbb{N}_0} q_j\, z^{-j}. \tag{4.2.1}$$

Keep in mind that it makes no difference whether we define numerator and denominator as Laurent polyomials or as polynomials since we can always expand the fraction by an arbitrary power of z and a constant. Thus, we can always assume that $q(z) = 1 + q_1 z^{-1} + \cdots + q_n z^{-n}$, $q_n \neq 0$, for some $n \in \mathbb{N}_0$, hence $q(z) = z^{-n}\widehat{q}(z)$, where $\widehat{q}(z) = q_n + q_{n-1}\, z + \cdots + z^n$ is a polynomial. By Lemma 3.4.13,

$$\frac{1}{q(z)} = z^n\,\frac{1}{\widehat{q}(z)} = z^n \sum_{j=n}^{\infty} \lambda_j\, z^{-j} = \sum_{j=0}^{\infty} \lambda_j\, z^{-j}, \qquad \lambda \in \ell(\mathbb{Z}),$$

so that

$$f(z) = \sum_{j=0}^{\infty} f_j\, z^{-j} \qquad \Rightarrow \qquad f \in \ell(\mathbb{Z}), \quad \operatorname{supp} f \subseteq \mathbb{N}_0.$$

[8]In the L_2 norm, the best approximation usually depends on the underlying norm.

We should therefore neither hope nor expect that f is still an FIR filter. Nevertheless, this filter can still be implemented effectively. To see that, we rephrase the definition of $F\sigma$ as

$$p(z)\sigma(z) = (F\sigma)(z)\, q(z) = (F\sigma)(z) + z^{-1}\, \widetilde{q}(z)\, (F\sigma)(z), \qquad \widetilde{q}(z) = q_1 + \cdot + q_n z^{-n},$$

that is,

$$(F\sigma)(z) = p(z)\sigma(z) - \left[z^{-1}\,(F\sigma)(z)\right] \widetilde{q}(z) = p(z)\sigma(z) - q(z)\left(\tau^{-1}\, F\sigma\right)(z) \tag{4.2.2}$$

da

$$z^{-1}\,(F\sigma)(z) = \sum_{j\in\mathbb{Z}} (F\sigma)_j\, z^{-j-1} = \sum_{j\in\mathbb{Z}} (F\sigma)_{j-1}\, z^{-j} = \left(\tau^{-1}\, F\sigma\right)(z).$$

By definition, $\widetilde{q}$ is a CAUSAL FIR filter and therefore is determined at time step j only by the values of $\tau^{-1}\, F\sigma$ until time step j, that is, the values of $F\sigma$ until time step $j-1$, and those are known. In other words, we compute $F\sigma$ by filtering σ with p and feedback using $\widetilde{q}$. This is shown in Fig. 4.2.4, for details see [18, 19]. What is interesting for us at this point

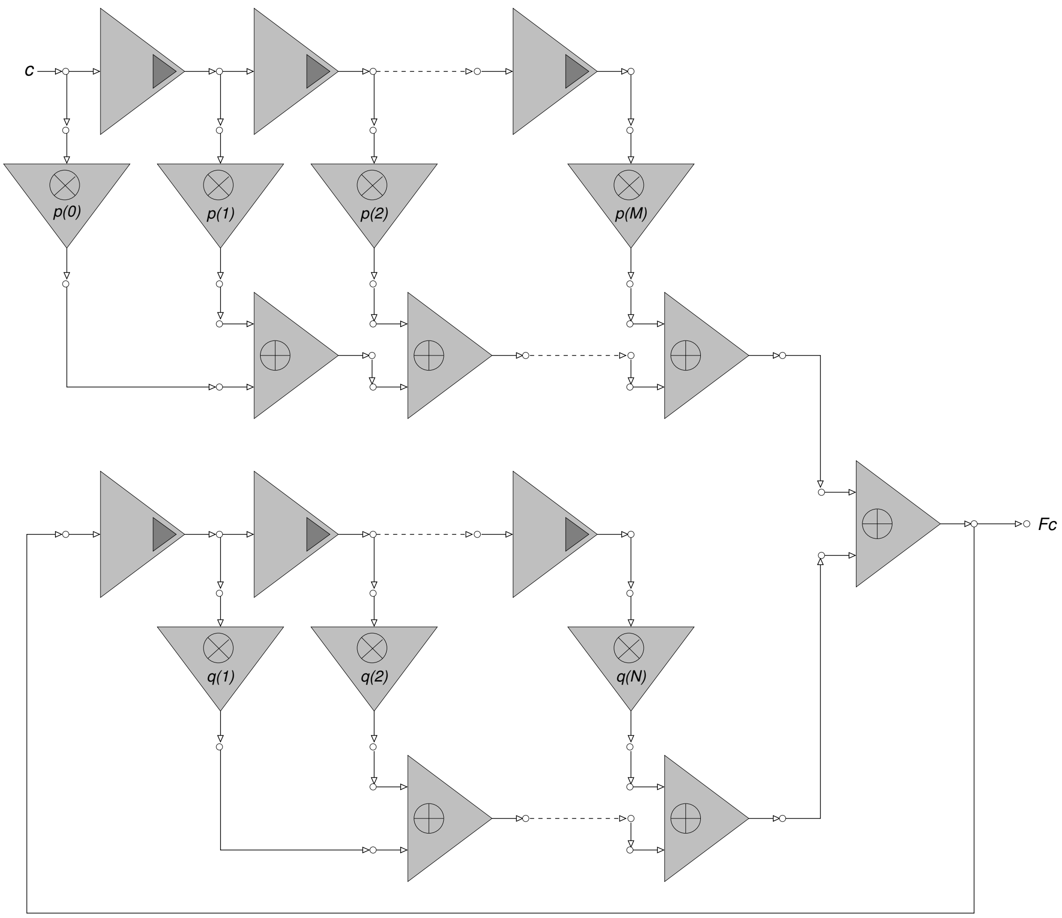


Abbildung 4.2.4: A rational filter, realized by means fo DELAYED FEEDBACK: The signal filterd by p (filter on top) is sent into the filter q (filter below) and the results are added.

is the fact that rational filters are of real practical relevance since they can be implemented physically.

Unfortunately, such a feedback system can also show quite an unwanted behavior. To understand this, we expand $1/q$ as the Laurent series

$$\frac{1}{q(z)} = \sum_{j=0}^{\infty} \lambda_j \, z^{-j},$$

obtain under the assumption[9] that $\operatorname{supp} p \subseteq [0,m]$ the identity

$$f(z) = \sum_{j=0}^{\infty} \sum_{k=0}^{m} p_k \, \lambda_j \, z^{-j-k} = \sum_{j=0}^{\infty} \left[\sum_{k=0}^{m} p_k \lambda_{j-k} \right] z^{-j} = (\lambda * p)(z)$$

and have a look at the behavior of λ_j and therefore also f_j for $j \to \infty$. Indeed, q can show a *damping* behavior if $\lambda_j \to 0$, $j \to \infty$ or it can be *exciting*[10], in case $|\lambda_j| \to \infty$, $j \to \infty$. Since $f_j = (\lambda * p)_j$ this convergence or divergence behavior carries over to the impulse response f. A „good“ filter better should have a decaying impulse response as otherwise it would, after a certain time, not even react to its input any more.

Definition 4.2.2. The LTI[11] filter F is called STABLE if

$$\lim_{j \to -\infty} f_j = \lim_{j \to \infty} f_j = 0.$$

What does stability mean for the denominator polynomial q? Let us look at the simplest nontrivial case, namely $q(z) = 1 - \zeta \, z^{-1} = z^{-1} (z - \zeta)$, $\zeta \in \mathbb{C}^\times$. Recalling (3.4.32), it follows that

$$\frac{1}{q(z)} = z \, \frac{1}{z - \zeta} = \sum_{j=0}^{\infty} \frac{\zeta^j}{z^j} \qquad \Rightarrow \qquad \lambda_j = \zeta^j,$$

and thus stability is equivalent to $|\zeta| < 1$, the ZERO ζ of $q(z)$ has to be *inside* the UNIT DISC

$$z \in \mathbb{D}^0 = \{ z \in \mathbb{C} \ : \ |z| < 1 \} = \mathbb{D} \setminus \partial \mathbb{D}, \qquad \mathbb{D} := \{ z \in \mathbb{C} \ : \ |z| \le 1 \}. \tag{4.2.3}$$

If, on the other hand, $|\zeta| > 1$, the filter will „explode“, if the zero lies on the UNIT CIRCLE $\partial \mathbb{D}$, i.e., $|\zeta| = 1$, we cannot make general statements about the impulse response. For an arbitrary rational filter we factorize the denominator q into

$$q(z) = z^{-n} (z - \zeta_1) \cdots (z - \zeta_n)$$

and use the PARTIAL FRACTION DECOMPOSITION

$$f(z) = \frac{p(z)}{q(z)} = \sum_{j=1}^{k} \frac{p_j(z)}{(z - \zeta_j)^{\alpha_j}}, \qquad \alpha_1 + \cdots + \alpha_k = n,$$

where α_j denotes the MULTIPLICITY of the zero ζ_j, $j = 1, \dots, k$. Now the converge or divergence are decided by the the zero z_j of maximal modulus: if it is inside the unit circle, we have convergence, if it is outside the (closed) unit circle, we have to face divergence. And this is the main result about the stabilily of rational filters.

Theorem 4.2.3. *A rational LTI filter F with z transform $f(z) = p(z)/q(z)$ is stable if and only if all its zeros belong to $\mathbb{D}^\circ$.*

[9] Without any loss of generality, one more normalization issue.

[10] Which is usually not so exciting.

[11] It has to be an LTI filter, otherwise the impulse response f would not be defined.

4.3 Fourier and sampling

The preceding section tells us that it is important to construct polynomials without zeros in the unit circle as those are the denominators of stable rational filters. Before we consider such a construction, we first remark why the UNIT CIRCEL

$$\partial\mathbb{D} = \{z \in \mathbb{C} \ : \ |z| = 1\} = \left\{e^{-i\theta} \ : \ \theta \in [-\pi, \pi]\right\}$$

plays such a fundamental role. Instead the z transform of a $\sigma(z)$ of a signal σ we can also consider the associated TRIGONOMERTRIC SERIES or FOURIER SERIES

$$\widehat{\sigma}(\theta) = \sigma\left(e^{i\theta}\right) = \sum_{k\in\mathbb{Z}} \sigma_k\, e^{-ik\theta} = \sum_{k\in\mathbb{Z}} \sigma_k \,\cos k\theta + i \sum_{k\in\mathbb{Z}} \sigma_k \,\sin k\theta$$

which satisfies

$$(f * \sigma)^\wedge(\theta) = (f * \sigma)\left(e^{i\theta}\right) = f\left(e^{i\theta}\right)\,\sigma\left(e^{i\theta}\right) = \widehat{f}(\theta)\,\widehat{\sigma}(\theta). \tag{4.3.1}$$

The complex valued function $\widehat{f}(\theta)$ is called the TRANSFER FUNCTION of the filter and is usually given in the logarithmic DECIBEL[12] scale with the unit DB. Instead of the value y, the value $10\log_{10} y$ is used and a ***dB*** is added.

Since sine and cosine are odd and even functions, respectively, we have that

$$\widehat{f}(\theta) = f_0 + \sum_{k=1}^{\infty} (f_k + f_{-k}) \,\cos k\theta + i \sum_{k=1}^{\infty} (f_k - f_{-k}) \,\sin k\theta$$

hence this function is real valued if and only if $f_k = f_{-k}$, i.e., if and only if the filter is a SYMMETRIC FILTER. The important point is the fact that by switching from z transform to trigonometric polynomials we have objects that are only defined on the unit circle $\partial\mathbb{D}$ instead of $\mathbb{C}^\times$.

Another advantage of this representation is that frequencies are represented in a much more natural way since now a band pass filter is really of the form $\widehat{f} = \chi_{[\omega_0,\omega_1]}$. But since $\widehat{f}$ is always defined in $\mathbb{T}$, there has to be conversion factor between absolute frequencies and their representation in $\mathbb{T}$. This is done by means of the SAMPLING RATE. Indeed, we always assumed that σ is a *discrete* signal which means that

$$\sigma_k = s\,(t_0 + k\tau), \qquad k \in \mathbb{Z}, \qquad t_0 \in \mathbb{R},\ \tau > 0,$$

is a SAMPLED version of the original signal s, where τ is called the SAMPLING INTERVAL and τ^{-1} the SAMPLING RATE. Intuitively, it is quite clear that the frequency resolution will be related to the sampling rate: the finer the sampling, the higher the sampling rate, the higher are the frequecies that can be detected. This is formalized in the famous SAMPLING THEOREM, called Shannon, Shannon-Whittaker oder Shannon-Whittaker-Kotelnikov sampling theorem[13]. In fact, Whittaker proved the recovery result in the context of infinite cardinal interpolation in 1915 [58], see also citeWhittaker35, but Shannon discovered its meaning in the context of digital signal processing later in [54, 55]. Kotelnikov [30] is in-between the two but was more popular in the Russian literature. The sampling theorem is based on a fundamental concept.

[12]Named after Alexander Graham Bell, despite the missing „l". The „deci" refers to the fact that a decimal logarithm with basis 10 is used.

[13]Despite the variation in the naming, it is the same result.

Definition 4.3.1. A function $f \in L_1(\mathbb{R})$ is called BANDLIMITED with BANDWIDTH T if its FOURIER TRANSFORM

$$\widehat{f}(\xi) = \int_{\mathbb{R}} f(t)\, e^{-i\xi t}\, dt$$

vanished outside $[-T, T]$:

$$\widehat{f}(\xi) = 0, \qquad \xi \notin [-T, T].$$

Bandlimited means that the function f, seen as a signal defined on the continuum $\mathbb{R}$, has only frequency content between $-T$ and T, hence the energy is localized in a compact subset of the spectrum of f. Bandlimited functions can be recovered *exactly* from discrete samples.

Theorem 4.3.2 (Sampling theorem). *If f is a T bandlimited function and $\tau < \tau^* = \frac{\pi}{T}$, then*

$$f(x) = \sum_{k \in \mathbb{Z}} \sigma_k \frac{\sin \pi\,(x/\tau - k)}{\pi\,(x/\tau - k)}, \qquad \sigma_k = f(k\tau), \quad k \in \mathbb{Z}. \tag{4.3.2}$$

The critical sampling rate $1/\tau^* = T/\pi$ or the half it[14] is called the NYQUIST RATE for the signal and describes how finely the signal has to be sampled in order to recover it. The function

$$g(x) = \frac{\sin \pi x}{\pi x} =: \operatorname{sinc} x, \qquad x \in \mathbb{R},$$

is called SINUS CARDINALIS or CARDINAL SINE FUNCTION, where the name is due to its behavior

$$\operatorname{sinc} k = \delta_{0k} = \begin{cases} 1, & k = 0, \\ 0, & \text{sonst}, \end{cases}$$

at the CARDINAL NUMBERS $\mathbb{Z}$, see Fig. 4.3.5. A proof of Theorem 4.3.2 can be found, for

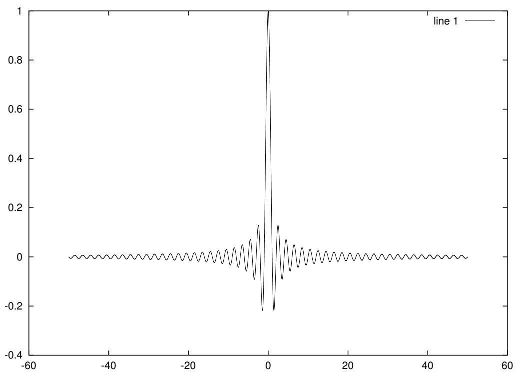


Abbildung 4.3.5: The sinc function. It decays for $|x| \to \infty$, unfortunately only like $|x|^{-1}$ which makes it inconvenient for numerical applications.

example, in [32].

This answers the question of the frequency range of digital filters: the values $\theta \in [0, \pi]$ correspond to the frequencies $\left[0, \tau^{-1}\right]$, hence the frequency range is determined by the sampling rate.

This shows that filter construction is not as simple as it may occur first: the transfer function for a rational filter determines the rational function on $\partial\mathbb{D}$ while, on the other

[14]This is a question of how the sampling rates are normalized

hand, the function should not have any poles *inside* the unit circle in order to define a stable filter. Fortunately, such problems have been discussed in the theory of functions and in systems theory.

4.4 Zeros of polynomials

Having learned that stability of rational filters is connected to the location of the zeros of the denominator, i.e., to the POLES of the z transform. We are interested in good filters which have all their poles inside the unit disc.

Remark 4.4.1. We will see soon that the „good" location of poles[15] can vary under simple transformations. Sometimes the good locations are inside, sometimes outside $\mathbb{D}$ and sometimes they have to lie in a certain half plane. All this depends mainly on which way the result can be proved most easily.

A polynomial q has all its zeros inside[16] $\mathbb{D}$ if

$$q\left(z^{-1}\right) = \sum_{j=0}^{n} q_j\, z^j, \qquad z \in \mathbb{C}^\times,$$

has all its zeros *outside* $\mathbb{D}$. Fortunately, the literature on complex analysis, for example [22], provides some results that study precisely this question: When does a complex polynomial $f \in \mathbb{C}[z]$ have *all* or *no* zeros inside the unit disc. A classic in this respect which can also be found in [10], is the Eneström–Kakeya theorem that provides a *sufficient* condition for a polynomial to have no zeros inside the unit disc.

Theorem 4.4.2 (ENESTRÖM–KAKEYA). *If $p_0 > p_1 > \cdots > p_n > 0$, then the polynomial $p(z) = p_0 + \cdots + p_n z^n$ has no zero in $\mathbb{D}$.*

Proof:[17] For $z \in \mathbb{C}$ we have

$$(1-z)\,p(z) = p_0 + \sum_{j=1}^{n} \left(p_j - p_{j-1}\right) z^j - p_n z^{n+1}$$

and therefore for $|z| \le 1$, by a double triangle inequality downwards,

$$\begin{aligned} |1-z|\;|p(z)| &\ge p_0 - \left| \sum_{j=1}^{n} \left(p_j - p_{j-1}\right) z^j - p_n z^{n+1} \right| \\ &\ge\; p_0 - \sum_{j=1}^{n} \left|p_j - p_{j-1}\right| \left|z^j\right| - |p_n| \left|z^{n+1}\right| \ge p_0 + \sum_{j=1} \left(p_j - p_{j-1}\right) - p_n = 0 \end{aligned}$$

with equality if and only if $|z| = 1$, i.e., $z = e^{i\theta}$ for some $\theta \in [0, 2\pi)$, and if all the powers $z^j = e^{i\theta j}$ have the same argument which is the case if and only if $\theta = 0$ or $z = 1$ ist. Since

[15] The location of poles in theory of functions and signal processing has lead to a lot of mathematical jokes; they are slightly politically incorrect but nevertheless funny.

[16] To clarify the terminology one more time: *„inside"* means *„in the interior"*

[17] The proof is not needed for what follows, but it it nice, short and simple, so let us have a look at it

$p(1) = p_0 + \cdots + p_n > 0$, p ther polynomial p cannot have a zero at $z = 1$, hence $0 \notin p(\mathbb{D})$. $\square$

While the Eneström–Kakeya is indeed a nice and interesting result, it is only a *suffcient* condition. The question is whether it is possible to characterize polynomials without zeros inside the unit circle *without* having to factorize[18] it. To that end, we first modify the problem by means of a fractional linear RATIONAL TRANSFORM of the form

$$w = \frac{z+1}{z-1}, \qquad z = \frac{w+1}{w-1}.$$

These two transforms are inverses of each other, which is easily verified by noting that both can be rewritten as $zw - z - w - 1 = 0$. Writing $w = u + iv$, we then get

$$|z|^2 = \left|\frac{w+1}{w-1}\right|^2 = \frac{(u+1)^2 + v^2}{(u-1)^2 + v^2} \qquad \Rightarrow \qquad \begin{cases} |z| > 1, & u > 0, \\ |z| = 1, & u = 0, \\ |z| < 1, & u < 0. \end{cases}$$

Consequently, the transform $z \to w$ maps the complex plane $\mathbb{C}$ to itself and in such a way that $|z| < 1$ holds if and only if the associated w has negative real part: $\Re w < 0$. If now $p(z)$ is a Laurent polynomial, then

$$\begin{aligned} p(z) &= \sum_{j=0}^{n} p_j \, z^{-j} = \sum_{j=0}^{n} p_j \left(\frac{w+1}{w-1}\right)^{-j} = \left(\frac{1}{w+1}\right)^n \sum_{j=0}^{n} p_j \, (w-1)^j (w+1)^{n-j} \\ &= \left(\frac{1}{w+1}\right)^n \sum_{j=0}^{n} p_j^w \, w^j = (1+w)^{-n} \, p^w(w), \end{aligned}$$

where

$$(1+w)^{-1} = \left(1 + \frac{z+1}{z-1}\right)^{-1} = \left(\frac{2z}{z-1}\right)^{-1} = \frac{z-1}{2z}.$$

If z is a zero of p such that[19] $0 < |z| < 1$, then $w \neq 1$ and therefore $p^w(w) = 0$ where w lies in the left half plane. We record this fact in a formal way.

Theorem 4.4.3. *The Laurent polynomial $p(z)$ has all its zeros inside the unit circle if and only if p^w has all its zeros in the* LEFT HALF PLANE $\mathbb{H}_- := \{z \in \mathbb{C} \; : \; \Re z < 0\}$.

4.5 Hurwitz polynomials and Stieltjes' theorem

Looking at the definition of the transformation, we can easily observe that the coefficients of p^w are real if the coefficients of p are real. This leads us to a class of polynomials which will become the object of investigation for the rest of this chapter. From now on we write polynomials as polynomials in the variable z, to indicate that now we *explicitly* consider polynomials in complex variable over the domain $\mathbb{C}$. And moreover we are not so much interested in the unit circle but in the left half plane.

Definition 4.5.1. A polynomial $f \in \mathbb{C}[z]$ is called a HURWITZ POLYNOMIAL if it has real coefficients[20] and all its zeros have negative real part, i.e.,

$$Z(f) := \{z \in \mathbb{C} : f(z) = 0\} \subset \mathbb{H}_-. \tag{4.5.1}$$

[18] That's to cheap (conceptionally) and to expensive (computationally) at the same time.

[19] Recall that **Laurent** polynomials have no meaning at $z = 0$.

[20] Although formally it is a polynomial with *complex* ceofficients as indicated by the notation $f \in \mathbb{C}[z]$.

Before we will collect further information on Hurwitz polynomials, we first address the question what justifies their appearance in the context of continued fractions. To that end, we first mention a classical way of decomposing polynomials which is actually used a lot in subdivision and wavelet theory. We write $f(z)$ as

$$f(z)=\sum_{j=0}^{n} f_j\, z^j=\sum_{j\le n/2} f_{2j}\, z^{2j}+\sum_{j<n/2} f_{2j+1}\, z^{2j+1}=h\left(z^2\right)+zg\left(z^2\right)$$

where h contains the coefficients of f with even index while g contains those with an odd index. Splitting a polynomial into such a pair can become useful and interesting if this pair has a special property.

Definition 4.5.2. Two *real* polynomials $p(x)$ und $q(x)$ with $\deg p=\deg q=n$ or $\deg p=n$ and $\deg q=n-1$ form a POSITIVE PAIR if their zeros $x_1,\dots,x_n$ und $x_1',\dots,x_n'$ or $x_1',\dots,x_{n-1}'$, respectively, INTERLACE, i.e.,

$$\begin{array}{ll} x_1'<x_1<x_2'<\cdots<x_n'<x_n<0, & q\in\Pi_n, \\ x_1<x_1'<x_2<\cdots<x_{n-1}'<x_n<0, & q\in\Pi_{n-1} \end{array} \tag{4.5.2}$$

and the leading coefficients of p and q have the same sign[21].

The nice thing is that positive pairs characterize Hurwitz polynomials and can in turn be characterized by means of continued fractions.

Theorem 4.5.3 (Stieltjes). *For a polynomial $f(z)=g\left(z^2\right)+zh\left(z^2\right)$ the following statements are equivalent:*

1. *f is a Hurwitz polynomial.*
2. *The polynomials g and h form a positive pair*[22]
3. *There exist $c_0\ge 0$ and positive number $c_j,d_j>0$, $j=1,\dots,m$, such that*

$$\frac{h(x)}{g(x)}=[c_0;d_1x,c_1,d_2x,c_2,\dots,d_mx,c_m], \tag{4.5.3}$$

where $c_0=0$ iff $\deg f\in 2\mathbb{N}_0+1$, i.e., is an odd number.

Besides having positive coefficients, the continued fraction in (4.5.3) can also offer a quite amazing structure: in the partial denominator polynomials of degree 1 and degree 0 take turns, so that the degrees only increase slowly.

To prove Theorem 4.5.3 we have to work a bit harder and learn some more concepts and ideas, but the result is worth it and a highlight. In addition, it allows us to construct and even to „enumerate“ denominators of stable rational filters, hence has a meaning in signal processing as well. But before we attack the steps of the proof of this theorem, we record another simple property of Hurwitz polynomials concerning the sign of their coefficients.

Lemma 4.5.4. *If $f\in\Pi_n$ is a Hurwitz polynomial of degree n with $f_n>0$, then $f_j>0$, $j=0,\dots,n$.*

[21] Which is only a normalization issue since sign and even absolute value of leading coefficients are not relevant for the zeros of a polynomial.

[22] Keep in mind that the degree of h can be smaller than that of g.

Proof: We factorize f as

$$f(z) = f_n \prod_{j=1}^{n} \left(z - \zeta_j\right), \qquad \zeta_j \in \mathbb{H}_-.$$

Since in a real polynomial[23] all zeros have to appear as complex conjugate pairs, f contains factors either of the form $(z+\alpha)$, $\alpha \in \mathbb{R}_+$, if the zero $-\alpha$ is real or of the form

$$(z-\zeta)\left(z-\overline{\zeta}\right) = z^2 - \underbrace{\left(\zeta + \overline{\zeta}\right)}_{=\Re\zeta<0} z + \underbrace{\zeta\,\overline{\zeta}}_{=|\zeta|^2>0} = z^2 + \beta z + \gamma, \qquad \beta,\gamma \in \mathbb{R}_+,$$

if the zero is complex. Hence,

$$f(z) = f_n \left[\prod_{j=0}^{k} \left(z+\alpha_j\right)\right] \left[\prod_{j=0}^{k'} \left(z^2 + \beta_j z + \gamma_j\right)\right]$$

can only have positive coefficients. □

4.6 Cauchy index and the argument of the argument

It is getting time to recall the Sturm chain where, for an interval $I = [a,b]$ we counted the *weighted*[24] sign changes $\Sigma_a^b f = \sigma(f,[a,b])$ of a function f. In the proof of Proposition 3.5.4 we then considered a rational function f defined as the quotient of two successive orthognal polynomials or polynomials that satisfied a three term recurrence. Such a rational function, however, does not only have zeros – which are zeros of the numerator – but also zeros of the denominator, that is poles. Each POLE again provides a sign change, this time from $\pm\infty$ to $\mp\infty$ and nothing can prevent us from counting these sign changes as well.

Definition 4.6.1 (Sign changes across poles & Cauchy index).

1. We say that f has a SINGULAR SIGN CHANGE or SIGN CHANGE ACROSS A POLE at a point x if
$$\lim_{x'\to x_-} = \pm\infty \qquad \text{and} \qquad \lim_{x'\to x_+} = \mp\infty. \tag{4.6.1}$$
2. The CAUCHY INDEX $I_a^b f$ of a function f on the interval $[a,b]$ is the weighted sum of singular sign changes or sign changes across poles where the changes from $-\infty$ to $+\infty$ are counted as positive, those from $+\infty$ to $-\infty$ as negative.

In a slightly more formal way the Cauchy index can be defined by means of „normal“ sign changes as
$$I_a^b f := -\Sigma_a^b f^{-1}. \tag{4.6.2}$$

It does not require much imagination to get the idea that also the Cauchy index will be strongly connected to Sturm chains. But to really follow the proof from [11], we need a little bit of function theory[25], cf. [10, Theorem 2, S. 175].

[23] Which we use synonymously for „polynomial with real coefficients“.
[24] With + and − depending of the direction in which the sign changed.
[25] Or complex analysis.

Definition 4.6.2. The ARGUMENT $\theta =: \arg z$ of a complex number z is defined as

$$\Re z + i\,\Im z = z = |z|\,e^{i\theta} = |z|\,(\cos\theta + i\,\sin\theta)\,. \tag{4.6.3}$$

It follows immediately from (4.6.3) that

$$\cos\theta = \Re z/|z|, \qquad \sin\theta = \Im z/|z|. \tag{4.6.4}$$

Theorem 4.6.3 (Argument principle). *If f is analytic on a domain $D \subset \mathbb{C}$ and γ is a positively oriented piecewise smooth closed curve in D, enclosing a domain $\Omega \subset D$, then*

$$\frac{1}{2\pi}\,\Delta_\gamma \arg f(z) = \#\,\{z \in \Omega \;:\; f(z) = 0\}\,,$$

where Δ_γ stands for the number of changes in the argument modulo 2π along the curve γ.

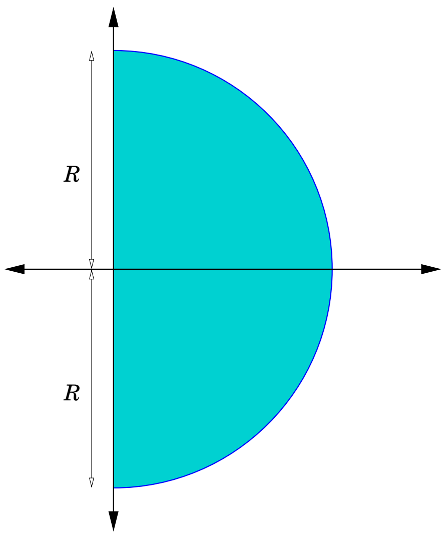


Abbildung 4.6.6: The domain of integration that certainly contains **no** zero of f if f is a Hurwitz polynomial, regardless of how large we choose R.

Now let f be a Hurwitz polynomial and consider, for $R > 0$ the integral along the curve γ that consists of the interval $[-Ri, Ri]$ and the semicircle of radius R in $\mathbb{H}_+$, see Fig. 4.6.6. For this domain we have that

$$0 = \Delta_{-R}^{R} \arg f\,(ix) - \Delta_{-\pi}^{\pi} f\left(R\,e^{ix}\right),$$

and for sufficiently large values of R the change of argument along the semicircle is determined by the leading term of $f_n\,z^n$ of f, $n = \deg f$, and has the value $n\pi$. Thus,

$$\Delta_{-\infty}^{\infty} \arg f\,(ix) = \lim_{R\to\infty} \Delta_{-R}^{R} \arg f\,(ix) = n\pi. \tag{4.6.5}$$

Writing f in the slightly excentric form

$$f(z) = a_0\,z^n + b_0\,z^{n-1} + a_1\,z^{n-2} + b_1 z^{n-3} + \cdots, \qquad a_0 \neq 0,$$

we get for $n = 2m$

$$\begin{aligned} f(ix) &= (-1)^m a_0\, x^n + i\,(-1)^{m-1}\, x^{n-1} + (-1)^{m-1} a_1\, x^{n-2} + \cdots \\ &= (-1)^m \left(a_0\, x^n - a_1\, x^{n-2} + a_2\, x^{n-4} + \cdots\right) + i\,(-1)^{m-1} \left(b_0\, x^{n-1} - b_1\, x^{n-3} + \cdots\right) \end{aligned}$$

and for $n = 2m+1$

$$f(ix) = (-1)^m \left(b_0\, x^{n-1} - b_1\, x^{n-3} + \cdots\right) + i\,(-1)^m \left(a_0\, x^n - a_1\, x^{n-2} + \cdots\right),$$

respectively, which shows that in both cases

$$f(ix) = p(x) + i\, q(x), \qquad x \in \mathbb{R}, \tag{4.6.6}$$

holds, where

$$p(x) = \begin{cases} (-1)^m \left(a_0\, x^n - a_1\, x^{n-2} + \cdots + (-1)^m a_m\right), & n = 2m, \\ (-1)^m \left(b_0\, x^{n-1} - b_1\, x^{n-3} + \cdots + (-1)^m b_m\right), & n = 2m+1, \end{cases} \tag{4.6.7}$$

and

$$q(x) = \begin{cases} (-1)^{m-1} \left(b_0\, x^{n-1} - b_1\, x^{n-3} + \cdots + (-1)^{m-1}\, b_{m-1}\, x\right), & n = 2m, \\ (-1)^m \left(a_0\, x^n - a_1\, x^{n-2} + \cdots + (-1)^m\, a_m\, x\right), & n = 2m+1. \end{cases} \tag{4.6.8}$$

By (4.6.4) we have that

$$\tan\theta = \frac{\mathfrak{I}z}{\mathfrak{R}z}, \quad \cot\theta = \frac{\mathfrak{R}z}{\mathfrak{I}z} \qquad \Rightarrow \qquad \theta = \arctan\frac{\mathfrak{I}z}{\mathfrak{R}z} = \operatorname{arccot}\frac{\mathfrak{R}z}{\mathfrak{I}z}.$$

Applied to (4.6.6) this implies that

$$\arg f(ix) = \arctan\frac{q(x)}{p(x)} = \operatorname{arccot}\frac{p(x)}{q(x)}$$

Now any inkrement of the argument, that is, any winding of $f(ix)$, corresponds to a pole or singularity of the tangent and therefore

$$\frac{1}{\pi}\, \Delta_{-\infty}^{\infty} \arg f(ix) = \begin{cases} I_{-\infty}^{\infty} \dfrac{p(x)}{q(x)}, & n = 2m+1, \\ -I_{-\infty}^{\infty} \dfrac{q(x)}{p(x)}, & n = 2m, \end{cases}$$

so that we obtain for our Hurwitz polynomial by means of (4.6.5) the identity

$$n = I_{-\infty}^{\infty} \frac{b_0\, x^{n-1} - b_1\, x^{n-3} + \cdots}{a_0\, x^n - a_1 x^{n-2} + \cdots} = -\Sigma_{-\infty}^{\infty} \frac{a_0\, x^n - a_1 x^{n-2} + \cdots}{b_0\, x^{n-1} - b_1\, x^{n-3} + \cdots}. \tag{4.6.9}$$

Now it is time to return to the decomposition $f(z) = g\left(z^2\right) + z\, h\left(z^2\right)$. Let us begin with the case $n = 2m$ where

$$g(x) = f_n\, x^m + f_{n-2}\, x^{m-1} + \cdots + f_0, \qquad h(x) = f_{n-1}\, x^{m-1} + f_{n-3}\, x^{m-2} + \cdots + f_1, \tag{4.6.10}$$

hence[26],

$$g\left(-z^2\right) = (-1)^m \left(a_0\, z^n - a_1\, z^{n-2} + \cdots\right), \qquad h\left(-z^2\right) = (-1)^m \left(b_0\, z^{n-2} - b_1\, z^{n-4} + \cdots\right),$$

[26] Recall that $a_j = f_{n-2j}$ and $b_j = f_{n-1-2j}$.

from which we conclude with the help of (4.6.9) that

$$n = -I_{-\infty}^{\infty} \frac{z\,h\left(-z^2\right)}{g\left(-z^2\right)}. \tag{4.6.11}$$

The respective identities for $n = 2m + 1$ are

$$g(x) = f_{n-1}\,x^m + f_{n-3}\,x^{m-1} + \cdots + f_0, \qquad h(x) = f_n\,x^m + f_{n-2}\,x^{m-1} + \cdots + f_1 \tag{4.6.12}$$

and

$$n = -I_{-\infty}^{\infty} \frac{g\left(-z^2\right)}{z\,h\left(-z^2\right)}. \tag{4.6.13}$$

Next, we derive a property of the Cauchy index similar by making use of its similarity to Sturm chains, so that the following lemma is mainly a reformulation of Theorem 3.5.3.

Lemma 4.6.4. *Let $a < c < b$ and $\phi : [a, b] \to \mathbb{R}$. Then*

$$I_a^b \phi = I_a^c \phi + I_c^b \phi + \eta_c \phi,$$

where

$$\eta_c \phi := \begin{cases} 1 \\ -1 \\ 0 \end{cases} \quad \text{if} \quad \lim_{x \to c_-} \phi(x) \begin{cases} = +\infty \\ = -\infty \\ \in \mathbb{R}. \end{cases} \tag{4.6.14}$$

Proof: Since the Cauchy index counts sign changes of ϕ^{-1}, we can proceed like in Theorem 3.5.3 just taking into account the fact that any singular sign change[27] of ϕ is a normal sign change of ϕ^{-1} and vice versa. If, on the other hand, such a sign change happens exactly at c, it is not recognized by the indices for the subintervals and has to be compensated explicitly by the quantity η_c from (4.6.14). □

Taking into account that the factor z in the denominator is irrelevant for the Cauchy index since the denominator polynomial g satisfies $g(0) = f_0 \neq 0$, hence there cannot be an η_0 term, we can expand (4.6.11) for $n = 2m$ in the following way:

$$\begin{aligned} n &= -I_{-\infty}^{\infty} \frac{z\,h\left(-z^2\right)}{g\left(-z^2\right)} = -\left(I_{-\infty}^{0} + I_0^{\infty}\right) \frac{z\,h\left(-z^2\right)}{g\left(-z^2\right)} = -2\,I_{-\infty}^{0} \frac{z\,h\left(-z^2\right)}{g\left(-z^2\right)} \\ &= 2\,I_{-\infty}^{0} \frac{h\left(-z^2\right)}{g\left(-z^2\right)} = 2\,I_{-\infty}^{0} \frac{h\,(x)}{g\,(x)} = I_{-\infty}^{0} \frac{h\,(x)}{g\,(x)} - I_{-\infty}^{0} \frac{x\,h\,(x)}{g\,(x)} \\ &= I_{-\infty}^{0} \frac{h\,(x)}{g\,(x)} - I_{-\infty}^{0} \frac{x\,h\,(x)}{g\,(x)} + \underbrace{I_0^{\infty} \frac{h\,(x)}{g\,(x)} - I_0^{\infty} \frac{x\,h\,(x)}{g\,(x)}}_{=0} = I_{-\infty}^{\infty} \frac{h\,(x)}{g\,(x)} - I_{-\infty}^{\infty} \frac{x\,h\,(x)}{g\,(x)}. \end{aligned}$$

For $n = 2m + 1$ we obtain the analogous

$$n = I_{-\infty}^{\infty} \frac{g\,(x)}{x\,h\,(x)} - I_{-\infty}^{\infty} \frac{g\,(x)}{h\,(x)},$$

and therefore

$$n = \begin{cases} I_{-\infty}^{\infty} \dfrac{h\,(x)}{g\,(x)} - I_{-\infty}^{\infty} \dfrac{x\,h\,(x)}{g\,(x)}, & n = 2m, \\[2ex] I_{-\infty}^{\infty} \dfrac{g\,(x)}{x\,h\,(x)} - I_{-\infty}^{\infty} \dfrac{g\,(x)}{h\,(x)}, & n = 2m + 1. \end{cases} \tag{4.6.15}$$

[27] That is, sign change via a pole of ϕ.

This already allows us to tackle one statement of Theorem 4.5.3 which even has a name of its own[28]

Theorem 4.6.5 (Hermite–Biehler theorem). *A polynomial $f(z) = g\left(z^2\right) + z\, h\left(z^2\right)$ is a Hurwitz polynomial if and only if g and h form a positive pair.*

Proof: We have already shown that f is a Hurwitz polynomial if and only if (4.6.15) is satisfied. For the rest, we once more have to distinguish two cases.
$n = 2m$: the denominator polynomial g has degree m and therefore at most m zeros. Therefore[29], because of

$$2m = I_{-\infty}^{\infty} \frac{h\,(x)}{g\,(x)} - I_{-\infty}^{\infty} \frac{x\,h\,(x)}{g\,(x)} \qquad \Rightarrow \qquad I_{-\infty}^{\infty} \frac{h\,(x)}{g\,(x)} = -I_{-\infty}^{\infty} \frac{x\,h\,(x)}{g\,(x)} = m$$

the quotient $h(x)/g(x)$ can only have singular sign changes or sign changing poles from $-\infty$ to $+\infty$, the quotient $x\,h(x)/g(x)$ on the other hand, only those from $+\infty$ to $-\infty$. This is turn is possible if between any such pair of jumps there is a regular sign change, i.e., a zero of h. Since g has exactly m such zero $x_1, \dots, x_m$ and h has the $m-1$ zeros $x_1', \dots, x_{m-1}'$, these zeros can thus only be arranged as

$$x_1 < x_1' < x_2 < x_2` < \cdots < x_{m-1}' < x_m < 0.$$

According to (4.6.10) and Lemma 4.5.4, f_n and f_{n-1} have to have the same sign and therefore we can assume that g and h have both leading coefficients of the same and even positive sign, which makes g and h a positive pair. Since all arguments were equivalences, the converse is obtained by repeating the proof backwards.
$n = 2m + 1$: now the $n = 2m + 1$ sign changes across poles have to be obtained by $m + 1$ sign changes of $x\,h(x)$ and m sign changes of $h(x)$ with opposite parities which just means that the $m + 1$ sign changes of $x\,h(x)$ occur at the positions $x_1' < \cdots < x_m' < 0$ and at 0; note that $x = 0$ is the only additional zero when passing from $h(x)$ to $xh(x)$ which has exactly one more zero. Between these sign changes there have to be the sign changes of g, that is

$$x_1' < x_1 < x_2' < \cdots < x_m' < x_m < 0,$$

as claimed. □

Now the identity (4.6.15), which is equivalent to f being a Hurwitz polynomial or, equivalently, that g and h form a positive pair, allows us to draw another conclusion.

Proposition 4.6.6. *Two polynomials g and h, $\deg g = m$, form a positive pair if and only if*

$$m = I_{-\infty}^{\infty} \frac{h(x)}{g(x)} = -I_{-\infty}^{\infty} \frac{x\,h(x)}{g(x)} \tag{4.6.16}$$

and if in the case $\deg g = \deg h$ we additionally have

$$\epsilon_\infty = \lim_{x \to +\infty} \operatorname{sgn} \frac{h(x)}{g(x)} = 1. \tag{4.6.17}$$

[28]To be precise: according to [11] this is a *special case* of the Hermite–Biehler theorem.
[29]We already used that argument in the proof of Proposition 3.5.4, when we showed that orthogonal polynomials must have the *maximal* number of zeros, hence those zeros had to be simple.

Proof: We have already seen that (4.6.16) follows for $n = 2m$ directly from (4.6.15), but to also obtain a respective result for $n = 2m + 1$, that is, to get from (4.6.15) to the statement of Proposition 4.6.6, we need a formular for the Cauchy index of a rational function whose numerator degree exceeds that of the denominator, namely

$$I_{-\infty}^{\infty} f(x) + I_{-\infty}^{\infty} f^{-1}(x) = \frac{\epsilon_\infty - \epsilon_{-\infty}}{2}, \qquad \epsilon_{\pm\infty} = \lim_{x \to \pm\infty} \operatorname{sgn} f(x), \tag{4.6.18}$$

Indeed, the expression on the left hand side is exactly the number[30] of all singular and normal sign changes of f and those sum to 1 1 if $\epsilon_\infty = 1$ and $\epsilon_{-\infty} = -1$, to -1, if the limits have signs $-$ and $+$, and to 0 whenever $\epsilon_\infty = \epsilon_{-\infty}$.

Using (4.6.18) we can now rewrite the second line of (4.6.15) into

$$2m + 1 = n = I_{-\infty}^{\infty} \frac{g(x)}{x\,h(x)} - I_{-\infty}^{\infty} \frac{g(x)}{h(x)} = I_{-\infty}^{\infty} \frac{h(x)}{g(x)} - \frac{1-1}{2} - I_{-\infty}^{\infty} \frac{x\,h(x)}{g(x)} + \frac{1+1}{2},$$

which gives (4.6.16) again. And the equal sign of the leading coefficients of g and h, a necessary condition for being a positive pair, follows for $n = 2m$, and thus $\deg h = \deg g - 1$, directly from (4.6.16), for $n = 2m+1$, i.e., $\deg h = \deg g$, the additional assumption (4.6.17) becomes necessary. □

To prove the second equivalance in Theorem 4.5.3, we need the following auxiallary statement.

Lemma 4.6.7. *Suppose that the polynomials g and h, $\deg g = m$ form a positive pair*[31] *and there are constants c, d as well as polynomials $g_1, h_1 \in \Pi_{m-1}$ such that*

$$\frac{h(x)}{g(x)} = c + \cfrac{1}{dx + \cfrac{g_1(x)}{h_1(x)}} = \left[c; dx, \frac{g_1(x)}{h_1(x)} \right]. \tag{4.6.19}$$

Then c, d and g_1, h_1 are determined uniquely by g and h and the following holds true:

1. $c \ge 0$, $d > 0$,
2. $\deg g_1 = \deg h_1 = m - 1$,
3. *g_1 and h_1 form apositive pair.*

If, conversely, the numbers c, d and the polynomials g_1, h_1 satisfy the above three conditions and g, h are defined by (4.6.19), *then g and h form a positive pair.*

Proof: If g, h are a positive pair then g has m real zeros and we obtain by (4.6.19) that[32]

$$m = I_{-\infty}^{\infty} \frac{h(x)}{g(x)} = I_{-\infty}^{\infty} \left[c + \cfrac{1}{dx + \cfrac{g_1(x)}{h_1(x)}} \right] = I_{-\infty}^{\infty} \frac{h_1(x)}{dx\,h_1(x) + g_1(x)}. \tag{4.6.20}$$

[30] And this number is finite since a rational function has only a finite number of zeros and poles.
[31] This implies, in particular, that $\deg h \in \{m-1, m\}$.
[32] Here the Cauchy index is helpful and useful: in contrast to normal sign changes singular sign changes are not affected by sign changes when a constant is added to the function.

This can only hold if the denominator is a polynomial of degree at least m, hence $d \neq 0$ and $\deg h_1 = m-1$, as otherwise the denominator degree could not exceed $m-1$. Without loss of generality we can also assume that the leading term of h_1 is positive[33]. Now (4.6.20) tells us that both rational functions $h(x)/g(x)$, as well as $h_1(x)/(dx\, h_1(x) + g_1(x))$, have a maximal number of singular sign changes from $-$ to $+$ and thus are ***negative*** for sufficiently small x and ***positive*** for sufficiently large x. Thus,

$$-1 = -\operatorname{sgn} d = \lim_{x\to-\infty} \frac{h_1(x)}{dx\, h_1(x) + g_1(x)}, \qquad 1 = \operatorname{sgn} d = \lim_{x\to-\infty} \frac{h_1(x)}{dx\, h_1(x) + g_1(x)},$$

implying $d > 0$. By (4.6.20) the function h/g has precisely m singular sign changes from $-\infty$ to $+\infty$, which interlace with $m-1$ sign changes from $+$ to $-$, so that

$$-I_{-\infty}^{\infty} \left[d\,x + \frac{g_1(x)}{h_1(x)} \right] \geq m - 1. \tag{4.6.21}$$

Since $\deg h_1 = m-1$, this Cauchy index is at most $m-1$ so that equality has to hold int (4.6.21) and thus

$$m - 1 = -I_{-\infty}^{\infty} \left[d\,x + \frac{g_1(x)}{h_1(x)} \right] = -I_{-\infty}^{\infty} \frac{g_1(x)}{h_1(x)}. \tag{4.6.22}$$

From the second identity in (4.6.16) we moreover conclude that

$$\begin{aligned} m &= -I_{-\infty}^{\infty} \frac{x\,h(x)}{g(x)} = -I_{-\infty}^{\infty} \left[c\,x + \frac{x}{d\,x + \dfrac{g_1(x)}{h_1(x)}} \right] = -I_{-\infty}^{\infty} \left[c\,x + \frac{1}{d + \dfrac{g_1(x)}{x\,h_1(x)}} \right] \\ &= -I_{-\infty}^{\infty} \left[\frac{1}{d + \dfrac{g_1(x)}{x\,h_1(x)}} \right] = I_{-\infty}^{\infty} \left[d + \frac{g_1(x)}{x\,h_1(x)} \right] = I_{-\infty}^{\infty} \frac{g_1(x)}{x\,h_1(x)} \end{aligned} \tag{4.6.23}$$

so that also $\deg g = m-1$, since there must be a sign change between any pair of singular sign changes. This completes the proof of 2).

Since the two polynomials g_1, h_1 have the same degree, it follows that

$$\lim_{x\to\pm\infty} \frac{g_1(x)}{h_1(x)} = \mu \neq 0 \quad \Rightarrow \quad \lim_{x\to\pm\infty} d\,x + \frac{g_1(x)}{h_1(x)} = \pm\infty \quad \Rightarrow \quad \lim_{x\to\pm\infty} \frac{1}{d\,x + \dfrac{g_1(x)}{h_1(x)}} = 0$$

and therefore, by (4.6.19)

$$c = \lim_{x\to\infty} \left[\frac{h(x)}{g(x)} - \frac{1}{d\,x + \dfrac{g_1(x)}{h_1(x)}} \right] = \lim_{x\to\infty} \frac{h(x)}{g(x)} \begin{cases} > 0, & \deg g = \deg h, \\ = 0, & \deg g > \deg h, \end{cases}$$

which also verifies the claim 1).

It remains to show that g_1 and h_1 indeed form a positive pair. To that end, we apply (4.6.18) to (4.6.23) to and obtain that

$$I_{-\infty}^{\infty} \frac{x\,h_1(x)}{g_1(x)} = -m + \frac{\epsilon_\infty - \epsilon_{-\infty}}{2} = -m + \epsilon_\infty, \tag{4.6.24}$$

[33]Otherweise we multiply both g_1 and h_1 by -1.

since

$$\lim_{x\to+\infty} \operatorname{sgn} \frac{h_1(x)}{g_1(x)} = \epsilon_\infty := \lim_{x\to+\infty} \operatorname{sgn} \frac{x\, h_1(x)}{g_1(x)} = -\lim_{x\to-\infty} \operatorname{sgn} \frac{x\, h_1(x)}{g_1(x)} = -\epsilon_{-\infty}.$$

If we normalize g_1 and h_1 in such a way that $\epsilon_\infty = 1$, then this identity, togehter with (4.6.22) and (4.6.24) is exactly what need to apply Proposition 4.6.6, hence g_1 and h_1 form a positive pair.

For the converse, we just note that all arguments used here were either identities or equivalences. □

With this lemma at hand, the proof of Theorem 4.5.3 is no magic any more since it shows us that positive pairs are transferred to positive pairs by such a „double step“ of the continued fraction expansion. Indee, Theorem 4.5.3 follows from assembling the Hermite–Biehler theorem, Theorem 4.6.5, and the following result.

Theorem 4.6.8. *Two polynomials g and h, $\deg g = m$, form a positive pair if and only if there exist*

$$c_0 \begin{cases} > 0, & \deg g = \deg h, \\ = 0, & \deg g = \deg h + 1, \end{cases} \qquad c_j, d_j \in \mathbb{R}_+, \quad j = 1, \dots, m,$$

such that

$$\frac{h(x)}{g(x)} = [c_0; d_1\, x, c_1, \dots, d_m\, x, c_m]\,. \tag{4.6.25}$$

Proof: Due to Lemma 4.6.7 we only have to show that to any positive pair g, h there exists a decomposition into g_1, h_1 as in (4.6.19). If $m = \deg g = \deg h$, then we can perform a division of h by g with remainder h_1, that is, $h = c_0\, g + h_1$, where even $c_0 > 0$ since as a positive pair g and h have leading coefficients of the same sign. Hence, $\deg h_1 = m - 1$. Therefore,

$$\frac{h(x)}{g(x)} = \frac{c\, g(x) + h_1(x)}{g(x)} = c_0 + \frac{h_1(x)}{g(x)} = c_0 + \cfrac{1}{\cfrac{g(x)}{h_1(x)}}.$$

On the other hand, $\deg g = m = \deg h_1 + 1$, hence $g(x) = d_1\, x\, h_1(x) + g_1(x)$, $\deg g_1 \le m - 1$, and therefore

$$\frac{h(x)}{g(x)} = c_0 + \cfrac{1}{\cfrac{d_1 x\, h_1(x) + g_1(x)}{h(x)}} = c_0 + \cfrac{1}{d_1\, x + \cfrac{g_1(x)}{h_1(x)}},$$

so that Lemma 4.6.7 implies $d_1 > 0$ and $\deg g_1 = \deg h_1 = m - 1$. For $\deg h = \deg g - 1$ the same holds, only with $c = 0$ and therefore $h_1 = h$. In summary, we have shown that in both cases

$$\frac{h(x)}{g(x)} = c_0 + \cfrac{1}{dx + \cfrac{1}{h_1(x)/g_1(x)}} = \left[c_0; d_1\, x, \frac{h_1(x)}{g_1(x)}\right], \qquad \deg g_1 = \deg h_1 = m - 1, \tag{4.6.26}$$

holds. This allows us to write h_1/g_1 as $\left[c_1; d_2\, x, \frac{h_2(x)}{g_2(x)}\right]$ with $\deg g_2 = \deg h_2 = m-2$. Iterating this decomposition in (4.6.26), we finally obtain that

$$\frac{h(x)}{g(x)} = \left[c_0; d_1\, x, c_1, \dots, d_j\, x, \frac{h_j(x)}{g_j(x)}\right], \qquad \deg g_j = \deg h_j = m - j, \qquad j = 1, \dots, m, \tag{4.6.27}$$

and the case $j = m$ together with the observation that $g_m, h_m \neq 0$ gives $c_m \neq 0$ and thus (4.6.25). The converse follows directly from expanding the continued fraction. □

4.7 The Routh–Hurwitz theorem

The famous theorem by Routh–Hurwitz[34] provides another characterization for a Hurwitz polynomial, this time by means of certain determinants. And since determinants cannot be imagined without (square) matrices, we start with a another peculiar type of matrices.

Definition 4.7.1. Let $p \in \Pi$ be a polynomial of degree n. The HURWITZ MATRIX associated to p is the $n \times n$ matrix

$$H_p := \begin{bmatrix} p_{n-1} & p_{n-3} & p_{n-5} & \dots & 0 \\ p_n & p_{n-2} & p_{n-4} & \dots & 0 \\ 0 & p_{n-1} & p_{n-3} & \dots & 0 \\ 0 & p_n & p_{n-2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & p_0 \end{bmatrix}. \tag{4.7.1}$$

Example 4.7.2. Let us consider some examples of such Hurwitz matrices for small values of n and a generic polynomial $p(x) = p_0 + \cdots + p_n x^n$ of that degree:

$n = 1$: we only have the 1×1 matrix $H_p = [p_0]$.

$n = 2$ the Hurwitz matrix is

$$H_p = \begin{bmatrix} p_1 & 0 \\ p_2 & p_0 \end{bmatrix}$$

and contains a zero for the first time.

$n = 3$: some more structure becomes visible:

$$H_p = \begin{bmatrix} p_2 & p_0 & 0 \\ p_3 & p_1 & 0 \\ 0 & p_2 & p_0 \end{bmatrix}$$

$n = 4$: we see even more structure:

$$H_p = \begin{bmatrix} p_3 & p_1 & 0 & 0 \\ p_4 & p_2 & p_0 & 0 \\ 0 & p_3 & p_1 & 0 \\ 0 & p_4 & p_2 & p_0 \end{bmatrix}$$

Looking carefully at the examples we see that once again we have to distingish between odd and even values of n, namely

$$H_p = \begin{bmatrix} p_{n-1} & \dots & p_3 & p_1 & 0 & 0 & \dots & 0 & 0 \\ p_n & \dots & p_4 & p_2 & p_0 & 0 & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & p_{n-1} & p_{n-3} & p_{n-5} & \dots & p_1 & 0 \\ 0 & \dots & 0 & p_n & p_{n-2} & p_{n-4} & \dots & p_2 & p_0 \end{bmatrix}, \qquad n = 2m, \tag{4.7.2}$$

[34]And this does not refer to the statement *"A PhD dissertation is a paper of the professor written under aggravating circumstances"* which is attributed in [31] to A. Hurwitz, but also to O. Töplitz. Since the matrices bearing their names have a lot in common, this does not really make a difference anyway.

and

$$H_p = \begin{bmatrix} p_{n-1} & \dots & p_2 & p_0 & 0 & \dots & 0 \\ p_n & \dots & p_3 & p_1 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & p_{n-1} & p_{n-3} & \dots & p_0 \end{bmatrix}, \qquad n = 2m+1, \tag{4.7.3}$$

respectively. Next, we need the fundamental concept of the minors of a matrix.

Definition 4.7.3. Let $A \in \mathbb{R}^{n \times n}$ and $I \subset \{1, \dots, n\}$. The I–MINOR of A is defined as

$$m_I(A) = \det A(I,I) = \det \left[a_{jk} \ : \ j,k \in I \right],$$

and the jth PRINCIPAL MINOR as

$$m_j(A) = m_{\{1,\dots,j\}}(A) = \det \left[a_{k\ell} \ : \ k,\ell = 1,\dots,j \right].$$

Theorem 4.7.4 (ROUTH–HURWITZ THEOREM). *A polynomial $f \in \Pi$ with positive leading coefficient is a Hurwitz polynomial if and only if*

$$m_k \left(H_f \right) > 0, \qquad j = 1, \dots, \deg f. \tag{4.7.4}$$

Before we turn to the proof of this theorem, to which the next section will be devoted, we again have a look at the first special cases.

$n = 1$: a polynomial $f(x) = f_1 x + f_0$, $f_1 > 0$, is a Hurwitz polynomial, according to Theorem 4.7.4, iff $0 < m_1 \left(H_f \right) = f_0$, which can be easily verified „manually“:

$$f(x) = 0 \qquad \Leftrightarrow \qquad x = -\frac{f_0}{f_1}$$

and the zero, which is always real in this case, is negative if and only if f_0 and f_1 have the same sign, hence are positive.

$n = 2$: here the positivity of the principal minors of

$$H_f = \begin{bmatrix} f_1 & 0 \\ f_2 & f_0 \end{bmatrix}$$

leads to

$$0 < f_1, \quad 0 < f_0 \, f_1 \qquad \Leftrightarrow \qquad 0 < f_0, f_1.$$

And indeed the zeros of f are the numbers

$$x = \frac{-f_1 \pm \sqrt{f_1^2 - 4 f_0 f_2}}{2 f_0} \qquad \Rightarrow \qquad \Re x < 0 \quad \text{for} \quad 0 < f_0, f_1, f_2,$$

since the root is either imaginary or less than f_1 as long as $f_0 f_2 > 0$, that is, $f_2 > 0$. So we can verify the Routh–Hurwitz criterion directly again.

$n = 3$: Now all principal minors of the matrix

$$M_f = \begin{bmatrix} f_2 & f_0 & 0 \\ f_3 & f_1 & 0 \\ 0 & f_2 & f_0 \end{bmatrix}$$

have to positive which in turn is equivalent to

$$f_0, \, f_2 > 0 \qquad \text{and} \qquad f_1 f_2 > f_0 f_3,$$

where the latter also implies that $f_1 > 0$.

After these special cases it is getting time to return to the general theory.

4.8 The Routh scheme or the return of Sturm's chains

Starting point for the proof of Theorem 4.7.4 is the characterization (4.6.9) of a Hurwitz polynomial by means of the Cauchy index:

$$n = I_{-\infty}^{\infty} \frac{b_0\, x^{n-1} - b_1\, x^{n-3} + \cdots}{a_0\, x^n - a_1 x^{n-2} + \cdots} =: I_{-\infty}^{\infty} \frac{f_1(x)}{f_0(x)}. \tag{4.8.1}$$

The two polynomials f_1 and f_2 cannot have a common zero as otherwise we could divide by the respective linear factor and the nominator would have only degree at most $n-1$ with at most $n-1$ zeros and also a Cauchy index of $n-1$. Therefore we can construct a sequence of polynomials $f_2, \dots, f_m$ by means of division with remainder in the following way:

$$f_j(x) = q_j(x)\, f_{j+1}(x) - f_{j+2}, \qquad \deg f_{j+2} < \deg f_{j-1}. \tag{4.8.2}$$

This is just the euclidean algortihm which has a the following property.

Lemma 4.8.1. *If f_0, f_1 are two polynomials without common zero and $f_m \in \Pi_0 \setminus \{0\}$ is the sequence from* (4.8.2), *then $f_0, \dots, f_m$ form a Sturm chain*[35].

Proof: Since the two polynomials have no common zero, the euclidean algorithm ends with the greatest common divisor $f_m \neq 0$ being a constant function. We have to show that at each zero of f_j the two polynomials[36] f_{j-1} and f_{j+1} have opposite sign. If we replace j by $j-1$ in (4.8.2), then it follows at each zero x of f_j that

$$0 = q_j(x)\, f_j(x) = f_{j-1}(x) + f_{j+1}(x)$$

so that eihter $f_{j-1}(x) = f_{j+1}(x) = 0$ or the two polynomials indeed have opposite sign. If, on the other hand, $f_j(x) = f_{j+1}(x) = 0$, then[37] (4.8.2) implies $f_{j+2}(x) = 0$ and, eventually, $f_m(x) = 0$, which is a contradiction. $\square$

Performing the euclidean algorithm explicitly, we obtain the sequence of polynomials

$$\begin{aligned} f_2(x) &= \frac{a_0}{b_0} x\, f_1(x) - f_0(x) = c_0\, x^{n-2} - c_1\, x^{n-4} + \cdots \\ f_3(x) &= \frac{b_0}{c_0} x\, f_2(x) - f_1(x) = d_0\, x^{n-3} - d_1\, x^{n-5} + \cdots \\ f_j(x) &= a_0^j\, x^{n-j} - a_1^j\, x^{n-j-2} + \cdots = \frac{a_0^{j-2}}{a_0^{j-1}} x\, f_{j-1}(x) - f_{j-2}(x), \end{aligned} \tag{4.8.3}$$

where

$$a_k^0 = a_k, \qquad a_k^1 = b_k, \qquad a_k^j = \frac{a_0^{j-1}\, a_{k+1}^{j-2} - a_0^{j-2}\, a_{k+1}^{j-1}}{a_0^{j-1}}, \tag{4.8.4}$$

[35]Note, however, that the indexing is reversed in comparison to Definition 3.5.1.

[36]These are the „neighboring" ones, so the only effect of idexing (potential) Sturm chain is whether the initial and the zero-free function are the first or last one in this order, respectively.

[37]Yes, this is precisely the argument that we already used in the proof of Proposition 3.5.4.

since[38]

$$
\begin{aligned}
f_j(x) &= \frac{a_0^{j-2}}{a_0^{j-1}} x \left[\sum_{k=0}^{(n-j+1)/2} (-1)^k a_k^{j-1} x^{n-j+1-2k} \right] - \left[\sum_{k=0}^{(n-j)/2+1} (-1)^k a_k^{j-2} x^{n-j+2-2k} \right] \\
&= \sum_{k=1}^{(n-j)/2+1} (-1)^k \frac{a_0^{j-2}\, a_k^{j-1} - a_0^{j-1}\, a_k^{j-2}}{a_0^{j-1}} x^{n-j+2-2k} \\
&= \sum_{k=0}^{(n-j)/2} (-1)^k \underbrace{\frac{a_0^{j-1}\, a_{k+1}^{j-2} - a_0^{j-2}\, a_{k+1}^{j-1}}{a_0^{j-1}}}_{=a_k^j} x^{n-j-2k}.
\end{aligned}
$$

In general it could happen that at some step of this process we run into

$$
0 = a_0^j = \frac{a_0^{j-2}\, a_1^{j-1} - a_0^{j-1}\, a_1^{j-2}}{a_0^{j-1}}, \qquad a_0^{j-1} \neq 0,
$$

so that we would divide by zero in the next step. In that case, we replace a_1^{j-2} by $a_1^{j-2} + \varepsilon$ with a sufficiently small $\varepsilon > 0$. Even if we would have to do that several time, we can eventually pass to the limit $\varepsilon \to 0$. This continuity arguments works as long as f has no zeros on the imaginary axis, for detail see [11].

This allows us to restrict ourselves to the *regular* case that he Routh scheme (4.8.3) produces a Sturm chain of length n. Now all polynomials with even index $f_0, f_2, \dots,$ have the same PARITY, i.e., are either all an ODD FUNCTION or an EVEN FUNCTION, that is $f(-x) = -f(x)$ or $f(-x) = f(x)$, respectively, while those with odd indices, $f_1, f_3, \dots,$ share the oppositite parity[39]. This however implies

$$
\begin{aligned}
V(-x) &= V\left(f_0(-x), f_1(-x), \dots, f_{n-1}(-x), f_n(-x)\right) \\
&= \begin{cases} V\left(f_0(x), -f_1(x), \dots, -f_{n-1}(x), f_n(x)\right), & n = 2m, \\ V\left(-f_0(x), f_1(x), \dots, f_{n-1}(x), -f_n(x)\right), & n = 2m+1. \end{cases}
\end{aligned}
$$

and therefore[40]

$$
n = V(-\infty) + V(\infty) := \lim_{x \to \infty} V(-x) + V(x), \tag{4.8.5}
$$

as there is either a sign change from $f_j(\infty)$ to $f_{j+1}(\infty)$ or from $\pm f_j(\infty)$ to $\pm f_{j+1}(-\infty) = \mp f_{j+1}(\infty)$. On the other hand, (4.8.1), (4.6.2) and Theorem 3.5.3 imply that

$$
n = I_{-\infty}^{\infty} \frac{f_1(x)}{f_0(x)} = -\Sigma_{-\infty}^{\infty} \frac{f_0(x)}{f_1(x)} = V(-\infty) - V(\infty),
$$

hence f is a Hurwitz polynomial if and only if

$$
0 = V(\infty) = V\left(a_0^j \;:\; j = 0, \dots, n\right), \qquad n = V(-\infty). \tag{4.8.6}
$$

All togehter, this proves the following theorem.

[38] Here the summation limits are always meant as the integer part of the respective numbers.

[39] This in an immediate consequence of the fact that each polynomial contains only powers of the same parity.

[40] Note that the limit in (4.8.5) is already assumed at all $x > x_0$ for some x_0.

Theorem 4.8.2 (Routh criterion). *The polynial $f(z)$ is a Hurwitz polynomial if and only if all the numbers a_0^j, $j = 0, \ldots, n$, are either strictly positive or strictly negative.*

Remark 4.8.3. According to (4.8.6) the vector whose sign changes define $V(\infty)$ has to have at least $n+1$ entries for a Hurwitz polynomial – how else could one obtain n sign changes. This means that the euclidean algorithm for a Hurwitz polymial cannot have any degree jumps, all quotient polynomials q_j must be of degree 1 and no more. Or, in other words: if would divide by zero in (4.8.4) then the underlying polynomial cannot be a Hurwitz polynomial.

We can arrange all coefficients of the polynomials $f_0, f_1, \ldots, f_n$ into a table which is called the Routh scheme:

$$\begin{array}{lll} a_0^0 & a_1^0 & \cdots \\ a_0^1 & a_1^1 & \cdots \\ \vdots & & \\ a_0^n & & \end{array}$$

This table can be explicitly computed by (4.8.4). The Routh criterion of Theorem 4.8.2 can now be rephrased as that we can recognize a Hurwitz polynomial from the property that all entries of of the *first column* of the Routh scheme have the same strict sign[41], which is now really easy to check.

Example 4.8.4. Let us try to get an idea what the Routh criterion means.

1. For $n = 2$ and $f(z) = f_0 + f_1\, z + f_2\, z^2$ we get that $a_0^0 = f_2$, $a_1^0 = f_0$ and $a_0^1 = f_1$, hence
$$a_0^2 = \frac{a_0^1\, a_1^0}{a_1^0},$$
and we see that this polynomial is a Hurwitz polynomial if and onyl if f_0, f_1, f_2 have the same strict sign.

2. A slightly more intricate example from [11], where one can also see the "ε–Argument" applied, is the polynomial $f(z) = z^4 + z^3 + 2z^2 + 2z + 1$, leading to the scheme
$$\begin{array}{cccccc} 1 & 2 & 1 & & & \\ 1 & 2 & & & & \\ \varepsilon & 1 & & \leftarrow & 0 & 1 \\ 2 - \dfrac{1}{\varepsilon} & & & & & \\ 1 & & & & & \end{array}$$
of length n. Here f is *no* Hurwitz polynomial as any positive choice $\varepsilon > 0$ leads to a sign distribution $+, +, +, -, +$, while $\varepsilon < 0$ leads to $+, +, -, +, +$ and in both cases $V(\infty) = 2$. This shows, by the way, that f must have two zeros in $\mathbb{H}_+$.

The way from the Routh scheme to Theorem 4.7.4 is now very short: we first observe that the Hurwitz matrix is

$$H_f = \begin{bmatrix} b_0 & -b_1 & b_2 & \cdots \\ a_0 & -a_1 & a_2 & \cdots \\ 0 & b_0 & -b_1 & \cdots \\ 0 & a_0 & -a_1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

[41] Zero is forbidden. Either everything is strictly positive or strictly negative.

Like in Gauß elimination we multiply the first row by a_0/b_0 and subtract that from the third row, then the same with the second and fourth row and so on, leading to a matrix of the form

$$H_f' = \begin{bmatrix} b_0 & -b_1 & b_2 & \dots \\ 0 & c_0 & -c_1 & \dots \\ 0 & b_0 & -b_1 & \dots \\ 0 & 0 & c_0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \qquad c_k = \frac{b_0\, a_{k+1} - a_0\, b_{k+1}}{b_0}.$$

The formula for the c_k is already familiar to us as it is precisely (4.8.4) and, consequently,

$$H_f^{(1)} := H_f' = \begin{bmatrix} a_0^1 & a_1^1 & a_2^1 & \dots \\ 0 & a_0^2 & a_1^2 & \dots \\ 0 & a_0^1 & a_1^1 & \dots \\ 0 & 0 & a_0^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

from where it starts to become fun. Now we multiply the second row by a_0^1/a_0^2, subtract that from the third row and apply similar operations to the fourth and fifth, the sixth and seventh row and so on. Again we encounter the recurrence (4.8.4) and obtain the matrix

$$H_f^{(2)} = \begin{bmatrix} a_0^1 & a_1^1 & a_2^1 & \dots \\ 0 & a_0^2 & a_1^2 & \dots \\ 0 & 0 & a_0^3 & \dots \\ 0 & 0 & a_0^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Assuming that there was no division by zero during this process[42], this iteration ends with the upper triangular matrix

$$H_f^{(n)} = \begin{bmatrix} a_0^1 & \dots & * \\ & \ddots & \vdots \\ & & a_0^n \end{bmatrix}$$

and since we only subtracted multiples of the earlier rows $1, \dots, k-1$ from the kth row, $k = 1, \dots, n$, the principal minors of H_f and H_f^n coincide, that is,

$$m_k\left(H_f\right) = m_k\left(H_f^{(n)}\right) = \prod_{j=1}^{k} a_0^j, \qquad k = 1, \dots, n. \tag{4.8.7}$$

So we can finally complete all the proofs of this chapter.

Proof of Theorem 4.7.4: According to Theorem 4.8.2 the polynomial $f(z)$ with $a_0^0 = f_n > 0$ is a Hurwitz polynomial of and only if $a_0^j > 0$, $j = 1, \dots, n$, which, according to (4.8.7) is equivalent to all principal minors of $H_f^{(n)}$ and thus also all principal minors of H_f being positive. □

[42]Which would request he ε–modification and never happens for Hurwitz polynomials

Literaturverzeichnis

[1] O. Becker, *Quellen und Studien zur Geschichte*, Math., Astron., Physik **B2** (1933), 311–333.

[2] D. J. Benson, *Music. A mathematical offering*, Cambridge University Press, 2007.

[3] D. Bernoulli, *Disquisitiones ulteriores de idole fractionum continuarum*, N. C. Pet. **20** (1775).

[4] J. W. Cooley, *The re–discovery of the Fast Fourier Transform*, Mikrochimica Acta **3** (1987), 33–45.

[5] ———, *How the FFT gained acceptance*, A History of Scientific Computing (S. G. Nash, ed.), ACM–Press and Addison–Wesley, 1990, pp. 133–140.

[6] J. W. Cooley and J. W. Tukey, *An algorithm for machine calculation of complex Fourier series*, Math. Comp. **19** (1965), 297–301.

[7] P. J. Davis, *Interpolation and approximation*, Dover Books on Advanced Mathematics, Dover Publications, 1975.

[8] Ch. Dunkl and Y. Xu, *Orthogonal polynomials in several variables*, Cambridge University Press, 2001.

[9] F. J. Dyson, *The approximation to algebraic numbers by rationals*, Acta Mathematica (1947).

[10] S. D. Fisher, *Complex variables*, Wadsworth & Brooks, 1990, Dover Reprint 1999.

[11] F. R. Gantmacher, *Matrix Theory. Vol. II*, Chelsea Publishing Company, 1959, Reprinted by AMS, 2000.

[12] J. von zur Gathen and J. Gerhard, *Modern computer algebra*, Cambridge University Press, 1999.

[13] C. F. Gauss, *Methodus nova integralium valores per approximationem inveniendi*, Commentationes societate regiae scientiarum Gottingensis recentiores **III** (1816).

[14] W. Gautschi, *Numerical analysis. an introduction*, Birkhäuser, 1997.

[15] W. B. Gragg and L. Reichel, *On singular values of Hankel operators of finite rank*, Linear Algebra Appl. **121** (1989), 53–70.

[16] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete mathematics*, 2nd ed., Addison–Wesley, 1998.

[17] W. Gröbner, *Algebraische Geometrie I*, B.I–Hochschultaschenbücher, no. 273, Bibliographisches Institut Mannheim, 1968.

[18] D. Ch. von Grüningen, *Digitale Signalverarbeitung*, VDE Verlag, AT Verlag, 1993.

[19] R. W. Hamming, *Digital filters*, Prentice–Hall, 1989, Republished by Dover Publications, 1998.

[20] G. H. Hardy and E. M. Wright, *An introduction to the theory of numbers*, 3rd ed., Oxford University Press, 1954.

[21] H. Helmholtz, *On the sensations of tone*, Longmans & Co, 1885, Translated by A. J. Ellis, Dover reprint 1954.

[22] P. Henrici, *Applied and computational complex analysis. volume 2*, Wiley, 1977.

[23] E. Hille, *Analytic function theory*, 2nd ed., Chelsea Publishing Company, 1982.

[24] D. R. Hofstadter, *Gödel, Escher, Bach: ein endloses geflochtenes Band*, Klett–Cotta, 1985.

[25] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

[26] E. Isaacson and H. B. Keller, *Analysis of Numerical Methods*, John Wiley & Sons, 1966.

[27] Ch. Jordan, *Calculus of finite differences*, 3rd ed., Chelsea, 1965.

[28] A. Ya. Khinchin, *Continued fractions*, 3rd ed., University of Chicago Press, 1964, Reprinted by Dover 1997.

[29] D. E. Knuth, *The art of computer programming. seminumerical algorithms*, 3rd ed., Addison–Wesley, 1998.

[30] Kotelnikov, *On the carrying capacity of the "ether" and wire in telecommunications*, First All Union Conference of Communications I, zd. Red. Upr. Svyazi RKKA, Moscov, 1933, In Russian.

[31] MacTutor, *The MacTutor History of Mathematics archive*, `http://www-groups.dcs.st-and.ac.uk/~history`, 2003, University of St. Andrews.

[32] S. Mallat, *A wavelet tour of signal processing*, 2. ed., Academic Press, 1999.

[33] ———, *A wavelet tour of signal processing: The sparse way*, 3rd ed., Academic Press, 2009.

[34] B. Mourrain, *Polynomial-exponential decomposition from moments*, (2016), arXiv:1609.05720v1.

[35] J. R. Partington, *An introduction to Hankel operators*, London Mathematical Society Student Texts, vol. 13, Cambridge University Press, 2010.

[36] O. Perron, *Die Lehre von den Kettenbrüchen I*, 3rd ed., B. G. Teubner, 1954.

[37] ———, *Die Lehre von den Kettenbrüchen II*, 3rd ed., B. G. Teubner, 1954.

[38] C. Prony, *Essai expérimental et analytique sur les lois de la dilabilité des fluides élastiques, et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, à différentes températures*, J. de l'École polytechnique **2** (1795), 24–77.

[39] G. E. Roberts, *From music to mathematics. exploring the connec+ions*, Johns Hopkins University Press, 2016.

[40] K. F. Roth, *Rational approximations to algebraic numbers*, Mathematika **2** (1955), 1–20.

[41] R. Roy and Th. Kailath, *ESPRIT – estimation of signal parameters via rotational invariance techniques*, IEEE Trans. Acoustics, Speech and Signal Processing **37** (1989), 984–995.

[42] C. Sagan, *Unser Kosmos*, Droemersche Verlagsanstalt Th. Knaur Nachf., 1989, Deutsche Taschenbuchausgabe.

[43] T. Sauer, *Computeralgebra*, Vorlesungsskript, Justus–Liebig–Universität Gießen, Universität Passau, 2001.

[44] ———, *Einführung in die Numerische Mathematik*, Vorlesungsskript, Universität Passau, 2013.

[45] ———, *Constructive Approximation*, Lecture notes, University of Passau, 2017.

[46] ———, *Reconstructing sparse exponential polynomials from samples: Difference operators, Stirling numbers and Hermite interpolation*, Mathematical Methods for Curves and Surfaces. 9th International Conference, MMCS 2016, Tønsberg, Norway. Revised Selected Papers (M. Floater, T. Lyche, M.-L. Mazure, K. Moerken, and L.-L. Schumaker, eds.), Lecture Notes in Computer Science, vol. 10521, Springer, 2017, arXiv:1610.02780, pp. 233–251.

[47] ———, *Hankel and Toeplitz operators of finite rank and Prony's problem in several variables*, (2018), submitted for publication, arXiv:1805.08494.

[48] ———, *Prony's method: an old trick for new problems*, Snapshots of modern mathematics from Oberwolfach (2018).

[49] R. Schmidt, *Multiple emitter location and signal parameter estimation*, IEEE Transactions on Antennas and Propagation **34** (1986), 276–280.

[50] K. Schmüdgen, *The moment problem*, Graduate Texts in Mathematics, Springer, 2017.

[51] A. Schönhage and V. Strassen, *Schnelle Multiplikation großer Zahlen*, Computing **7** (1971), 281–292.

[52] H. W. Schüßler, *Digitale Signalverarbeitung*, 3. ed., Springer, 1992.

[53] L. Seidel, *Bemerkungen über den Zusammenhang zwischen dem Bildungsgesetze eines Kettenbruchs und der Art des Fortgangs seiner Näherungsbrüche*, Abh. München **7** (1855).

[54] C. E. Shannon, *A mathematical theory of communication*, Bell System Tech. J. **27** (1948), 379–423.

[55] ———, *Communications in the presence of noise*, Proc. of the IRE **37** (1949), 10–21.

[56] G. E. Shilov, *Elementary real and complex analysis*, MIT Press, 1973, Dover reprint, 1996.

[57] R. A. Silverman, *Introductionary complex analysis*, Prentice Hall, 1967, Dover reprint 1972.

[58] E. T. Whittaker, *On the functions which are represented by the expansions of the interpolation–theory*, Edinb. R. S. Proc. **35** (1915), 181–194.

[59] Yuan Xu, *Common zeros of polynomials in several variables and higher dimensional quadrature*, Pittman Research Monographs, Longman Scientific and Technical, 1994.

Index