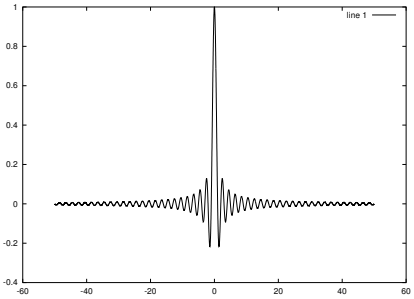
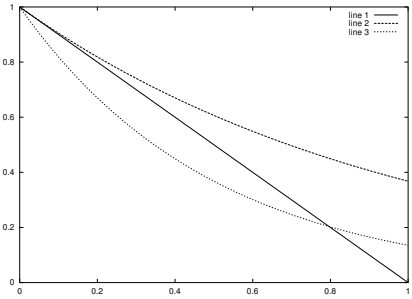




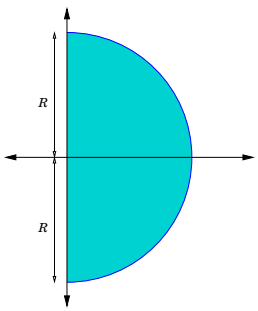
Kettenbrüche

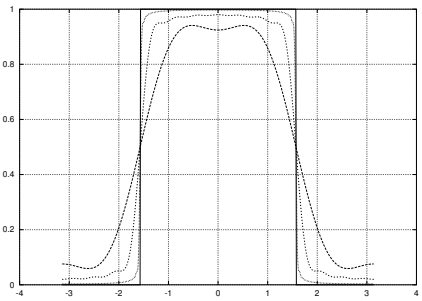
Vorlesung, zuerst gehalten im Wintersemester 2004/5

Tomas Sauer

Lehrstuhl für Numerische Mathematik
Justus–Liebig–Universität Gießen
Heinrich–Buff–Ring 44
D-35392 Gießen

Version 1.0
Version 24.2.2005





Statt einer Leerseite . . . | 0

> The presence of those seeking the truth is infinitely to be preferred to the presence of those who think they've found it.
>
> T. Pratchett, *Monstrous regiment*

> What are the digits that encode beauty, the number-fingers that enclose, transform, transmit, decode, and somehow, in the process, fail to trap or choke the soul of it? Not because of the technology but in spite of it, beauty, that ghost, that treasure, passes undiminished through the new machines.
>
> S. Rushdie, *Fury*

> Although this may seen as a parodox, all exact science is dominated by the idea of approximation.
>
> B. Russell

Und eine Bemerkung zur in diesem Skript verwendeten Orthographie:

> Ich spreche und schreibe Deutsch. Das große, weite und tiefe Deutsch, das die Reformer nicht verstehen. Und nicht ertragen.
>
> R. Menasse

0 Inhaltsverzeichnis

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

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

1 Kettenbrüche und was man mit ihnen anstellen kann

In diesem Abschnitt wollen wir uns erst einmal einen ganz groben Überblick über Kettenbrüche verschaffen und uns so ansehen, mit welchen Objekten sich die Vorlesung befassen wird und was man über diese Objekte sagen kann. Er dient mehr als Motivation denn als systematische oder strukturierte Einführung.

1.1 Die erste Definition

Ein *Kettenbruch*, also das "Studienobjekt" dieser Vorlesung, ist ein Bruch, dessen Nenner wieder als Kettenbruch geschrieben ist. Nun gut, selbstreferenzierende Definitionen sind nicht wirklich gut, also machen wir uns gleich einmal an die formale Definition.

Definition 1.1 *Zu Zahlen $a_0, \ldots, a_n \in \mathbb{Z}$ ist der zugehörige* Kettenbruch *die rationale Zahl*

$$[a_0; a_1, \ldots, a_n] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_{n-1} + \cfrac{1}{a_n}}}}} \tag{1.1}$$

Diese "Pünktchennotation" ist weder wirklich exakt noch ist der Kettenbruch auf diese Weise wohldefiniert, denn es ist ja nicht garantiert, daß der Kettenbruch überhaupt wohldefiniert ist. Wir brauchen uns nur die Fälle anzusehen, in denen beispielsweise $a_n = 0$ oder $a_{n-1} = -1/a_n$ ist – beide Male würden wir durch Null dividieren. Eine einfache rekursive Definition des Kettenbruchs erhalten wir, wenn wir berücksichtigen, daß der Nenner des "großen" Bruchs in (1.1) ja wieder nichts anderes als der Kettenbruch $[a_1; a_2, \ldots, a_n]$ ist und wir somit

$$[a_0; a_1] = a_0 + \frac{1}{a_1}, \qquad [a_0; a_1, \ldots, a_n] = a_0 + \frac{1}{[a_1; a_2, \ldots, a_n]}, \quad n \in \mathbb{N}, \tag{1.2}$$

erhalten. Die rekursive Definition zeigt uns nun auch schön, was in den "degenerierten" Fällen passiert: Ist beispielsweise $[a_k; a_{k+1}, \ldots, a_n] = 0$, dann ist[1]

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

und wenn nicht gerade $a_{k-2} = 0$ ist, dann geht eigentlich alles wieder normal weiter. Man sieht also: Division durch Null ist bei Kettenbrüchen eigentlich nicht so dramatisch. Trotzdem können wir die Schwierigkeiten mit der Division durch Null dadurch vermeiden, daß wir die Parameter in unseren Kettenbrüchen als $a_0 \in \mathbb{Z}$ und $a_j \in \mathbb{N}$ wählen

Natürlich müssen wir Kettenbrüche nicht auf ganzzahlige Koeffizienten einschränken, wir könnten auch $[r_0; r_1, \ldots, r_n]$ mit $r_j \in \mathbb{Q} \setminus \{0\}$ betrachten[2]. Eine einfache Formel, die einem dann sofort in den Schoß fällt ist, daß

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

unter Verwendung des *Restes* $r_k := [a_k; a_{k+1}, \ldots, a_n]$. Solange man $a_j \in \mathbb{Z} \setminus \{0\}$ oder $r_j \in \mathbb{Q} \setminus \{0\}$wählt, ist selbstverständlich der Kettenbruch ebenfalls eine rationale Zahl, was man, wenn man es nicht glaubt, ganz einfach durch Induktion über die Anzahl der Parameter und die Formel (1.2) beweist; alles, was man braucht ist die tiefschürfende Erkenntnis, daß die rationalen Zahlen unter Addition und Bildung des multiplikativen Inversen abgeschlossen sind.

Nun ist ja jede endliche Folge $a_0, \ldots, a_n$ von Zahlen ja Anfangssegment einer "richtigen", also *unendlichen* Folge von Zahlen und wir können somit auch unendliche Kettenbrüche der Form

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

[1] In leicht schlampiger Schreibweise, man möge das bitten nicht als Beweis verstehen

[2] Das soll die Schreibweise mit "r" anstelle von "a" auch immer signalisieren.

betrachten. Als formaler Ausdruck ist das fein, aber was ist der Wert eines solchen unendlichen Kettenbruchs? Eigentlich doch klar: der Grenzwert der endlichen Anfangssegmente, also

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

nur haben wir erst einmal keine Ahnung, wann diese Folge wirklich konvergiert und wann also der unendliche Kettenbruch wirklich konvergiert. Wir werden allerdings sogar ein Kriterium angeben, das außerdem noch sehr handlich ist.

Satz 1.2 *Für $r_j \in \mathbb{Q}$, $r_j > 0$, $j \in \mathbb{N}$, konvergiert der unendliche Kettenbruch $[r_0; r_1, \ldots]$ genau dann, wenn*

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

ist. Insbesondere ist das der Fall für $a_j \in \mathbb{N}$.

Man sieht also – besonders brav werden die unendlichen Kettenbrüche sein, wenn man $a_1, a_2, \ldots$ als natürliche Zahlen wählt. Da dann der Kettenbruch $[a_1; a_2, \ldots]$ positiv ist, erlauben wir uns immerhin $a_0 \in \mathbb{Z}$ um so auch negative Zahlen erzeugen zu können. Und tatschlich bekommen wir auf diese Art und Weise auch tatschlich alles.

Satz 1.3 *Jede reelle Zahl $x \in \mathbb{R}$ lässt sich als Kettenbruch $[a_0; a_1, \ldots]$ darstellen und dieser Kettenbruch ist genau dann endlich, wenn x rational ist.*

Tatschlich haben die Griechen nach der Entdeckung der Irrationalität[3] nicht etwa "normale" Brüche sondern Kettenbrüche für die erste "Definition" reller Zahlen verwendet, zumindest behauptet das [18, S. 359], der seinerseits wieder auf [1] verweist. Und das war auch wirklich eine sehr gute Wahl, denn wir werden sehen, daß Kettenbrüche relle Zahlen bei gleichem Aufwand (also gleicher Zahl von Ziffern) wesentlich genauer approximieren – wir betreiben hier also Approximation von reellen Zahlen durch rationale Zahlen. Und in der Tat wird sich herausstellen, daß die "beste Approximation" an eine irrationale Zahl unter allen rationalen Zahlen mit einem bestimmten Höchstnenner im wesentlichen der zugehörige Kettenbruch ist.

Und es wird sogar möglich sein, die am schlechtesten zu approximierende irrationale Zahl anzugeben, nämlich den *goldenen Schnitt* $\frac{1+\sqrt{5}}{2}$; dabei ist die schlechte Approximierbarkeit der Preis für die besonders einfache Kettenbruchdarstellung $[1; 1, 1, \ldots]$.

1.2 Kettenbrüche von Polynomen

Kettenbrüche lassen sich für die verschiedensten Objekte definieren: wir haben schon $\mathbb{Z} \setminus \{0\}$ und $\mathbb{Q}$ gesehen, aber man erkennt relativ leicht, daß man Kettenbrüche eigentlich über jedem *Ring*[4] betrachten kann, für die Existenz von Kettenbrüchen aber *euklidische Ringe*[5] vorzuziehen

[3] Und wer die Geschichte mit den Pythagoräern immer noch nicht kennt findet sie beispielsweise in [14, 24] in sehr populärwissenschaftlicher Form.

[4] Das ist, grob gesagt, eine Struktur, in der man "vernünftig" addieren und multiplizieren kann.

[5] Division mit Rest spielt bei der Kettenbruchentwicklung eine wichtige Rolle.

sind. Daher werden wir auch Kettenbrüche von *Polynomen* untersuchen, also Ausdrücke der Form

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

für einen geeigneten Körper $\mathbb{K}$. Das führt dann natürlich zu rationalen Funktionen, also zu einem Ausdruck der Form

$$[p_0(x); p_1(x), \ldots, p_n(x)] = \frac{f(x)}{g(x)}, \qquad f, g \in \Pi. \tag{1.3}$$

"Normalerweise" ist dabei jedes der p_j ein lineares oder konstantes Polynom, also ein Polynom vom Höchstgrad 1. Auch hier gibt es eine Approximationstheorie, also den Versuch, eine vorgegebene Funktion[6], dargestellt durch eine Potenzreihe[7], möglichst gut durch ein rationales Objekt[8] anzunähern, und zwar so, daß möglichst viele Terme in der Reihenentwicklung der rationalen Näherungsfunktion mit denen der zu approximierenden Reihe übereinstimmen.

Kettenbrüche mit besonders einfachen Koeffizienten in (1.3) sind natürlich die, bei denen jedes p_j ein *lineares* Polynom ist, also $p_j(x) = \alpha_j x + \beta_j$ gilt. Diese Kettenbrüche haben nun wieder einen engen Bezug zu orthogonalen Polynomen, also Polynomfolgen $f_j \in \Pi$, $j \in \mathbb{N}_0$, mit der Eigenschaft, daß

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

Orthogonale Polynome zu inneren Produkten lassen sich nun sogar durch Kettenbrüche charakterisieren und parametrisieren.

Satz 1.4 *Für jede Folge f_j, $j \in \mathbb{N}_0$, von orthogonalen Polynomen gibt es Koeffizienten $\alpha_j < 0$ und β_j, $j \in \mathbb{N}_0$, so daß*

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

und umgekehrt.

Schließlich wird es uns diese Theorie erlauben, orthogonale Polynome und Quadraturformeln über die Kettenbrüche zu konstruieren – und so hat Gauß tatschlich auch seine "Original Gaußquadratur" bestimmt, siehe [9] – und den Gaußschen Ansatz und seine Idee wirklich zu verstehen. Die besteht darin, daß die komponentenweise Grenzfunktion der Kettenbrüche zu einem Integral, das heißt, einem inneren Produkt mit der Eigenschaft

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

einfach zu beschreiben ist: Es ist die Laurentreihe[9] zur Momentenfolge, also

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

[6] Entspricht der "rellen Zahl".
[7] Genauer, eine Potenzreihe in z^{-1}, also eine *Laurentreihe*.
[8] Entspricht dem Kettenbruch.
[9] Oder "generating function".

1.3 Digitale Signalverarbeitung

Aber noch ein weiteres Problem soll uns interessieren, bei dem Kettenbrüche eine besondere Rolle in der *digitalen Signalverarbeitung*, genauer bei der Konstruktion von Digitalfiltern spielen. Für den Moment soll es reichen, daß ein Digitalfilter mit einer rationalen Funktion $f = \frac{p}{q}$, $p, q \in \mathbb{C}[z]$ identifiziert[10] werden kann, daß man ihn aber nur dann vernünftig (rekursiv) realisieren kann, wenn f keine Pole im Einheitskreis hat. Das ist der Begriff der *Stabilität* eines rationalen Filters. Mit anderen Worten: Damit ein rationaler Filter sinnvoll ist, muß gewährleistet sein, daß $q(z)$ keine Nullstellen im Inneren des Einheitskreises hat. Mit der gebrochen linearen Transformation $z = \frac{w+1}{w-1}$ ist das dasselbe wie die Forderung, daß $q(w)$ alle seine Nullstellen in der linken Halbebene haben muß. Solch ein Polynom nennt man *Hurwitz–Polynom* und genau diese Polynome werden wir im Satz von Stieltjes mit Hilfe von Kettenbrüchen charakterisieren, und zwar mit Kettenbrüchen der Form

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

in denen sich skalare und lineare Faktoren abwechseln. Zähler und Nenner der zugehörigen rationalen Funktion ergeben dann *zusammen*, wenn man sie richtig mischt, ein Hurwitz–Polynom und jedes Hurwitz–Polynom läßt sich umgekehrt auch so zerlegen.

1.4 Und was gibt's noch?

Natürlich sind die Punkte in dieser Vorlesung nur Teilaspekte der Kettenbruchtheorie. So findet sich beispielsweise in [17] ein Maßtheorie der Kettenbrüche (wie verteilen sich Kettenrbrüche auf der reellen Achse) und das zweibändige Werk von Perron [22, 23] enthält noch einiges, was hier nicht einmal ansatzweise erwähnt wurde, beispielsweise die Frage, wann ein Kettenbruch auch im Sinne einer Potenzreihe gegen eine analytische Funktion konvergiert. Aber anstatt zu jammern, was wir alles nicht machen werden, sollten wir besser loslegen!

[10] Hier brauchen wir wirklich *komplexe* Polynome, auch wenn ihre Koeffizienten zumeist reell sein werden.

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*

2 Kettenbrüche und reelle Zahlen

In diesem Abschnitt betrachten wir die Approximation von reellen Zahlen durch Kettenbrüche, deren Koeffizienten natürliche Zahlen sind. Das meiste Material ist aus dem Buch von Khinchin [17], denn besser kann man es kaum machen.

2.1 Konvergenten und Kontinuanten

Unser erster Schritt in Richtung Verständnis der Kettenbrüche besteht darin, uns den Ausdruck $[r_0; r_1, \ldots, r_n]$ einmal genauer anzusehen. Dazu definieren wir **den** Begriff der Kettenbruchtheorie.

Definition 2.1 *Die n–te* Konvergente *des unendlichen Kettenbruchs $[r_0; r_1, \ldots]$ ist definiert als der endliche Kettenbruch $[r_0; r_1, \ldots, r_n]$.*

Die n–te Konvergente eines Kettenbruchs kann man immer als Quotient zweier Polynome in $r_0, \ldots, r_n$ darstellen:

$$[r_0; r_1, \ldots, r_n] = \frac{p_n(r_0, \ldots, r_n)}{q_n(r_0, \ldots, r_n)}. \tag{2.1}$$

Das ist trivialerweise richtig für $n = 0$[11] und folgt induktiv aus (1.2):

$$[r_0; r_1, \ldots, r_{n+1}] = r_0 + \frac{1}{[r_1; r_2, \ldots, r_{n+1}]} = r_0 + \frac{q_n(r_1, \ldots, r_{n+1})}{p_n(r_1, \ldots, r_{n+1})},$$

was uns sogar eine rekursive Definition von p_{n+1} und q_{n+1} liefert, nämlich

$$\begin{aligned} p_{n+1}(r_0, \ldots, r_{n+1}) &= r_0\, p_n(r_1, \ldots, r_{n+1}) + q_n(r_1, \ldots, r_{n+1}), \\ q_{n+1}(r_0, \ldots, r_{n+1}) &= p_n(r_1, \ldots, r_{n+1}), \end{aligned} \tag{2.2}$$

und somit ist

$$[r_0; r_1, \ldots, r_n] = \frac{p_n(r_0, \ldots, r_n)}{p_{n-1}(r_1, \ldots, r_n)} \tag{2.3}$$

[11] Denn dann hat man das konstante Polynom r_0.

mit der Rekursionsformel

$$p_{n+1}(r_0,\ldots,r_{n+1}) = r_0\, p_n(r_1,\ldots,r_{n+1}) + p_{n-1}(r_2,\ldots,r_{n+1}), \qquad p_{-2}=0,\ p_{-1}=1.$$

Definition 2.2 *Die Polynome $p_n(x_0,\ldots,x_n)$ heißen* Kontinuanten *und wurden von Euler wenn nicht eingeführt so doch untersucht.*

Sehen wir uns doch einmal die ersten Beispiele an:

$$\begin{aligned}
[r_0;] &= r_0\\
[r_0;r_1] &= r_0+\frac{1}{r_1}=\frac{r_0r_1+1}{r_1}\\
[r_0;r_1,r_2] &= r_0+\frac{1}{[r_1;r_2]}=r_0+\frac{r_2}{r_1r_2+1}=\frac{r_0r_1r_2+r_0+r_2}{r_1r_2+1},
\end{aligned}$$

was ja nun wirklich symmetrisch aussieht.

Übung 2.1 Beweisen Sie die Symmetrieeigenschaft

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

der Kontinuante, siehe [18, S. 357]. ♢

Nachdem man ja Brüche bekanntlich auf vielerlei Arten schreiben kann[12], legen wir eine Darstellung über die Kontinuanten fest.

Definition 2.3 *Die Werte*

$$p_k := p_k(r_0,\ldots,r_k), \qquad q_k := p_{k-1}(r_1,\ldots,r_k)$$

heißen kanonische Darstellung *der k–ten Konvergente $[r_0;r_1,\ldots,r_k]=\frac{p_k}{q_k}$.*

Satz 2.4 *Für $k\geq 1$ erfüllen die Faktoren der kanonischen Darstellung die Rekursionsformel*[13]

$$\begin{array}{rcl} p_k &=& r_kp_{k-1}+p_{k-2}\\ q_k &=& r_kq_{k-1}+q_{k-2}\end{array}, \qquad \begin{array}{ll} p_{-1}=1, & p_0=r_0\\ q_{-1}=0, & q_0=1.\end{array} \tag{2.4}$$

Beweis: Den Fall $k=1$ haben wir oben schon nachgerechnet. Um von k nach $k+1$ zu kommen benutzen wir die kanonische Darstellung

$$[r_1;r_2,\ldots,r_{k+1}]=\frac{p'_k}{q'_k}$$

und erhalten, daß

$$\frac{p_{k+1}}{q_{k+1}} = r_0+\frac{1}{[r_1;r_2,\ldots,r_{k+1}]} = r_0+\frac{q'_k}{p'_k}=\frac{p'_k\,r_0+q'_k}{p'_k},$$

[12] Eindeutig ist nur die gekürzte Darstellung.

[13] Wer schon mal orthogonale Polynome gesehen hat, weiß, was uns später erwarten wird.

was uns zusammen mit der Induktionshypothese[14] (2.4) für p'_k und q'_k die Rekursionen

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

was die Induktion vervollstndigt. □

Korollar 2.5 *Für $k\geq 0$ ist*

$$q_k\,p_{k-1}-p_k\,q_{k-1}=(-1)^k, \tag{2.5}$$

bzw.

$$\frac{p_{k-1}}{q_{k-1}}-\frac{p_k}{q_k}=\frac{(-1)^k}{q_{k-1}\,q_k}. \tag{2.6}$$

Beweis: Wir multiplizieren die erste Zeile der Rekursionsformel (2.4) mit $-q_{k-1}$, die zweite mit p_{k-1} und erhalten

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

also (2.5), was wir nur noch durch $q_{k-1}\,q_k$ teilen müssen, um auch (2.6) zu bekommen. □

Und eine Formel haben wir noch.

Satz 2.6 *Für $k\geq 2$ gilt*

$$p_k\,q_{k-2}-q_k\,p_{k-2}=(-1)^k\,r_k \qquad \textit{bzw.} \qquad \frac{p_k}{q_k}-\frac{p_{k-2}}{q_{k-2}}=\frac{(-1)^k\,r_k}{q_{k-2}\,q_k}. \tag{2.7}$$

Beweis: Der Beweis ist nicht sonderlich überraschend: Wir multiplizieren die beiden Zeilen von (2.4) mit q_{k-2} bzw $-p_{k-2}$, addieren das Ganze und landen bei

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

gemäß (2.5). □

Dieser so unschuldig erscheinende Satz gibt uns bereits Information über die Konvergenz unendlicher Kettenbrüche, zumindest wenn $r_j\in\mathbb{Q}_+$, $j\in\mathbb{N}$, wobei $\mathbb{Q}_+$ die Menge der positiven rationalen Zahlen bezeichnet.

Korollar 2.7 *Ist $r_j\in\mathbb{Q}_+$, $j\in\mathbb{N}$, dann ist die Folge der Konvergenten gerader Ordnung, $[r_0;r_1,\ldots,r_{2k}]$, monoton steigend, die der Konvergenten ungerader Ordnung, $[r_0;r_1,\ldots,r_{2k+1}]$ hingegen monoton fallend. Außerdem ist*

$$\inf_{k\in\mathbb{N}}\left[r_0;r_1,\ldots,r_{2k-1}\right]\geq\sup_{k\in\mathbb{N}}\left[r_0;r_1,\ldots,r_{2k}\right]. \tag{2.8}$$

[14] Und unter Berücksichtigung der Indexverschiebung.

Beweis: Ein Blick auf die Rekursionsformel (2.4) zeigt uns, daß $q_k > 0$ ist, $k \in \mathbb{N}$, solange nur alle r_j strikt positiv sind[15]. Dann liefert aber (2.7), daß

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

bzw.

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

Als nächstes zeigen wir, daß jede Konvergente gerader Ordnung kleiner ist als eine beliebige Konvergente ungerader Ordnung. Dazu seien $m, m' \in \mathbb{N}$ gegeben und $\ell \geq \max\{m, m'\}$. Aus (2.6) mit $k = 2\ell + 1$ folgt, daß

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

und die Monotonieeigenschaft der Konvergenten liefert uns

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

wie behauptet – und damit natürlich auch (2.8). □

Halten wir es fest: Wir haben es bei den geraden Konvergenten mit einer monoton steigenden, nach oben beschränkten, bei den ungeraden Konvergenten hingegen mit einer monoton fallenden, nach unten beschränkten Folge zu tun. Diese beiden Folgen müssen aber konvergieren und damit hat die Folge der Konvergenten $[r_0; r_1, \ldots, r_k]$, $k \in \mathbb{N}$, maximal zwei Häufungspunkte und der unendliche Kettenbruch ist genau dann konvergent, wenn in (2.8) Gleichheit herrscht. Diese Form der einschließenden Konvergenz ist durchaus begrüßenswert, gibt sie uns doch immer eine obere und eine untere Abschätzung für den Grenzwert – vorausgesetzt natürlich, daß es überhaupt einen Grenzwert gibt.

Zum Abschluß dieses Abschnitts noch zwei nette Formeln für Kettenbrüche und deren Konvergenten.

Proposition 2.8 *Für $1 \leq k \leq n$ gilt*

$$[a_0; a_1, \ldots, a_n] = \frac{p_{k-1}\, r_k + p_{k-2}}{q_{k-1}\, r_k + q_{k-2}}, \qquad r_k := [a_k; a_{k+1}, \ldots, a_n]\,, \tag{2.9}$$

sowie

$$\frac{q_k}{q_{k-1}} = [a_k; a_{k-1}, \ldots, a_1]\,. \tag{2.10}$$

[15] Auch ein paar Nullen würden nicht stören, wenn wir nur einmal einen positiven Wert erreicht haben.

Beweis: Aus der Rekursionsformel (1.2) zur Definition der Kettenbrüche folgt[16]

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

Sind nun p_{k-1}, q_{k-1} Zähler und Nenner der $(k-1)$–ten Konvergente und p_k, q_k die Bestandteile der k–ten Konvergente von $[a_0; a_1, \ldots, a_{k-1}, r_k]$, dann ist nach (2.4)

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

was nichts anderes als (2.9) ist.

Formel (2.10) beweisen wir durch Induktion über k; für $k = 1$ erhalten wir dabei die korrekte Identität[17]

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

Haben wir (2.10) einmal für $k \geq 1$ bewiesen, dann setzen wir die Induktionshypothese (2.10) in (2.4) ein und erhalten, daß

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

ist – genau das, was wir wollen. □

2.2 Unendliche Kettenbrüche und deren Konvergenz

In diesem Abschnitt befassen wir uns mit unendlichen Kettenbrüchen der Form $[a_0; a_1, \ldots]$ und deren Konvergenz. Dazu werden wir die Annahme machen, daß

$$a_j > 0, \qquad j \in \mathbb{N}. \tag{2.11}$$

Daß wir jetzt a_j verwenden, um die Bestandteile des Kettenbruchs von den Resten $r_k = [a_k; a_{k+1}, \ldots]$ zu unterscheiden, würde von Haus aus noch nicht bedeuten, daß sie natürliche Zahlen sein müssen. Tatsächlich werden die Beweise deutlich zeigen, daß sie auch für $a_1, a_2, \ldots \in \mathbb{Q}_+$ funktionieren. Nachdem wir aber im nächsten Abschnitt sowieso zeigen werden, daß Kettenbrüche mit positive ganzzahligen Einträgen "ausreichend" sind, müssen wir die Sache ja nicht bis ins Letzte ausreizen. Unser Ziel in diesem Abschnitt ist es, Information über die Konvergenz unendlicher Kettenbrüche zu sammeln und insbesondere Satz 1.2 zu beweisen. Dazu aber erst einmal ein paar Vorbemerkungen, vor allem die Klarstellung, was Konvergenz eigentlich bedeutet.

[16] Wie wir ja bereits in der Einleitung auf Seite 3 gesehen haben.

[17] Achtung, der Kettenbruch beginnt ja erst bei a_1!

Definition 2.9 *Der unendliche Kettenbruch* $[a_0; a_1, \ldots]$ *heißt* konvergent*, wenn der Grenzwert*

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

existiert und endlich ist[18]*. Andernfalls heißt der Kettenbruch* divergent[19].

Proposition 2.10 *Konvergiert der unendliche Kettenbruch* $a = [a_0; a_1, \ldots]$*, dann konvergieren auch alle seine Reste* $r_k = [a_k; a_{k+1}, \ldots]$*. Konvergiert umgekehrt mindestens ein Rest* r_k*, dann konvergiert auch* a*.*

Beweis: Wir wählen $k, n \in \mathbb{N}$ und betrachten die n–te Konvergente

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

von r_k. Unter Verwendung von (2.9) erhalten wir dann

$$\frac{p_{k+n}}{q_{k+n}} = [a_0; a_1, \ldots, a_{k+n}] = [a_0; a_1, \ldots, a_{k-1}, r_n] = \frac{p_{k-1}\, r'_n + p_{k-2}}{q_{k-1}\, r'_n + q_{k-2}} \tag{2.12}$$

Lösen wir das nach r'_n auf, dann erhalten wir, daß

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

so daß

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

Wäre nun der Grenzwert des Nenners $= 0$, also die Folge r'_n divergent, dann müssen wir uns nur die Einträge r'_{2n+1} ansehen, um zu sehen, daß was nicht stimmen kann: Nach Korollar 2.7 bilden sie eine *monoton fallende* Folge endlicher Werte, die gegen $+\infty$ divergiert!

Existiert nun umgekehrt der Grenzwert $r'_n \to r_k$ für $n \to \infty$, dann ist

$$a = \lim_{n\to\infty} [a_0; a_1, \ldots] = \frac{p_{k-1} \lim\limits_{n\to\infty} r'_n + p_{k-2}}{q_{k-1} \lim\limits_{n\to\infty} r'_n + q_{k-2}} = \frac{p_{k-1}\, r_k + p_{k-2}}{q_{k-1}\, r_k + q_{k-2}}$$

und der Kettenbruch konvergiert. □

Als nächstes gönnen wir uns eine *quantitative* Aussage über die Konvergenz. Dieser Satz wird das "Herzstück"[20] der Theorie werden, die besagt, daß man relle Zahlen besonders gut durch Kettenbrüche approximieren kann.

[18] Es sollen ja schon Leute vom "Grenzwert ∞" gesprochen haben ...

[19] Divergenz bedeutet also nicht automatisch, daß die Folge gegen $\pm\infty$ divergiert!

[20] Am Tag bevor dieser Eintrag entstand, ging durch die Presse, daß das Wort "Habseligkeiten" zum schönsten Wort der deutschen Sprache gewählt wurde. Nicht, daß dies was mit Kettenbrüchen zu tun hätte, aber die Abstimmung über die schwachsinngste und unnötigste Abstimmung steht bis heute noch aus. Ganz zu schweigen von überflüssigen Fußnoten.

Satz 2.11 *Existiert der Wert $a = [a_0; a_1, \ldots]$, so gilt für jedes $k \geq 0$ die Abschätzung*

$$\left| a - \frac{p_k}{q_k} \right| < \frac{1}{q_k\, q_{k+1}}. \tag{2.13}$$

Beweis: Der Beweis beruht auf der monotonen Konvergenz von Konvergenten gerader und ungerader Ordnung: Ist k gerade, dann ist nach Korollar 2.7

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

und (2.6) liefert uns, daß

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

wohingegen wir für ungerade k die Abschätzung

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

erhalten. Zusammen liefert das (2.13). □

So, jetzt haben wir alles beisammen, um uns an den Beweis des Konvergenzkriteriums zu machen, den wir aber der Vollständigkeit halber nochmal formulieren wollen.

Satz 2.12 (Satz 1.2) *Für $a_0 \in \mathbb{Q}$, $a_j \in \mathbb{Q}_+$, $j \in \mathbb{N}$, konvergiert der unendliche Kettenbruch $[a_0; a_1, \ldots]$ dann und nur dann, wenn*

$$\sum_{j=0}^{\infty} a_j = \infty. \tag{2.14}$$

Korollar 2.13 *Jeder unendliche Kettenbruch $[a_0; a_1, \ldots]$ mit $a_0 \in \mathbb{Z}$ und $a_j \in \mathbb{N}$, $j \in \mathbb{N}$, konvergiert.*

Beweis: Korollar 2.13 folgt unmittelbar aus Satz 2.12, also machen wir uns an dessen Beweis. Nach Korollar 2.7 bedeutet das, daß wir zeigen müssen, daß die Folge der geraden und der ungeraden Konvergenten *denselben* Grenzwert haben, denn individuell konvergieren sie ja. Konvergieren nun alle Konvergenten[21], dann muß wegen (2.6) natürlich $(q_k\, q_{k-1})^{-1}$ gegen Null konvergieren, was nach (2.13) aber auch notwending für die Konvergenz ist. Mit anderen Worten: Der Kettenbruch konvergiert genau dann, wenn

$$\lim_{k\to\infty} q_k\, q_{k+1} = \infty. \tag{2.15}$$

[21] Machen sie also ihrem Namen alle Ehre.

Nehmen wir nun an, daß die Reihe in (2.14) *konvergieren* würde; damit ist a_k eine Nullfolge und es gibt ein $k_0 \in \mathbb{N}$, so daß $a_k < 1$ für $k \geq k_0$ ist. Die Rekursionsformel (2.4) für die q_k sagt uns, daß diese Werte für $k \geq 1$ alle positiv sein und daß deswegen auch

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

ist. Damit ist dann entweder $q_{k-1} \leq q_{k-2}$ und somit $q_{k-1} < q_k$ oder aber $q_{k-1} > q_{k-2}$. Im ersten Fall erhalten wir, wieder mit (2.4), daß

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

im anderen Fall

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

also gibt es $\ell \in \{k-1, k-2\}$, sp daß

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

Ist $\ell \geq k_0$, dann können wir das Argument wiederholen und erhalten, daß

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

für ein $m \in \{k-2, k-3, k-4\}$ ist und allgemein[22]

$$q_k < \frac{q_{\ell_m}}{(1-a_k)\,(1-a_{\ell_1})\cdots\left(1-a_{\ell_{m-1}}\right)}, \qquad \ell_m < k_0. \tag{2.16}$$

wobei $\ell_j \in \{k-j, \ldots, k-2j\}$. Da die Reihe in (2.14) konvergiert gilt dasselbe auch für das unendliche Produkt[23]

$$0 < \lambda := \prod_{j=k_0}^{\infty} (1-a_j) \leq \prod_{j=0}^{m-1} \left(1-a_{\ell_j}\right), \qquad \ell_0 = k. \tag{2.17}$$

Mit $Q := \max\{q_j \ : \ j < k_0\}$ erhalten wir dann aus (2.16), daß $q_k < Q/\lambda$ für $k \geq k_0$ sein muß und die Folge $q_k\, q_{k+1}$ ist durch

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

am Divergieren gehindert. Damit ist (2.14) notwendig für die Konvergenz.

[22] Wenn wir bis zum "bitteren Ende" iterieren.

[23] Um Khinchin [17] wörtlich zu zitiieren: *"... the infinite product [...], as we know, converges: that is, it has positive value ..."* Nachdem es aber nicht ganz so einfach war, dafür eine Referenz zu finden, machen wir den Beweis halt in Lemma 2.14 schnell selbst.

Nehmen wir nun an, daß (2.14) erfüllt ist, die Reihe also divergiert. Nachdem immer noch $q_k > q_{k-2}$ ist, $k \geq 2$, setzen wir $q := \min\{q_0, q_1\}$ und erhalten, daß $q_k > q$ ist, zumindest für alle $k \geq 2$. Und wieder einmal mißbrauchen wir die Rekursionsformel, diesmal für die Abschätzung

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

was uns

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

also

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

liefert. Damit ist aber

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

und da der andere Werte zumindest $> q$ ist, kommen wir schließlich zu

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

was uns die Konvergenz garantiert. □

Lemma 2.14 *Für $a_j \in [0,1)$, $j \in \mathbb{N}$, hat das unendliche Produkt*

$$\prod_{j=0}^{\infty} (1 - a_j)$$

genau dann einen positiven Grenzwert, wenn die unendliche Reihe

$$\sum_{j=0}^{\infty} a_j$$

konvergiert.

Beweis: Da die endlichen Teilprodukte $(1-a_1)\cdots(1-a_n)$, $n \in \mathbb{N}$, eine monoton fallende Folge von positiven Zahlen bilden, ist es klar, daß der Grenzwert

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

existiert – die einzige Frage ist, ob er positiv oder Null ist. Außerdem sieht man sofort, daß $\lambda = 0$ ist, wenn a_j keine Nullfolge ist. Also müssen wir uns beim Beweis von Lemma 2.14 nur für Nullfolgen wirklich "anstrengen".

Die einfache Idee hierbei ist die Abschätzung[24]

$$e^{-2x} \leq 1 - x \leq e^{-x}, \qquad 0 \leq x \leq \frac{1}{2}\log 2, \tag{2.18}$$

denn an $x = 0$ haben alle drei Ausdrücke den Wert 1 und für die Ableitungen gilt

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

Ist also nun a_j eine Nullfolge, dann ist $a_j < \frac{1}{2}\log 2$ für $j \geq n_0$ und wir haben, daß

$$\prod_{j=n_0}^{\infty}(1-a_j) \geq \prod_{j=n_0} e^{-2a_j} = \exp\left(-2\sum_{j=n_0}^{\infty} a_j\right) \tag{2.19}$$

sowie

$$\prod_{j=n_0}^{\infty}(1-a_j) \leq \prod_{j=n_0} e^{-a_j} = \exp\left(-\sum_{j=n_0}^{\infty} a_j\right). \tag{2.20}$$

Konvergiert nun die Reihe, so konvergiert auch die Teilreihe, die bei n_0 beginnt, sagen wir gegen a und (2.19) liefert uns, daß

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

divergiert die Reihe hingegen, so liefert uns (2.20), daß

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

was genau die Behauptung war. □

2.3 Kettenbrüche mit natürlichen Koeffizienten

Nachdem wir also gesehen haben, daß Kettenbrüche mit natürlichen Koeffizienten[25] so schön und handlich sind – schließlich ist ja die Konvergenzfrage immer geklärt – sollten wir jetzt zuerst einmal nachweisen, daß sie auch "ausreichend" sind, daß sich also zumindest alle Brüche als Kettenbrüche mit natürlichen Elementen schreiben lassen.

Satz 2.15 *Jede rationale Zahl $x = \frac{p}{q}$ kann durch einen endlichen Kettenbruch mit ganzzahligen positiven Koeffizienten dargestellt werden.*

[24] Die, wie Abb 2.1 zeigt, sogar auf einem viel größeren Bereich erfüllt ist.

[25] Natürlich außer a_0 natürlich.

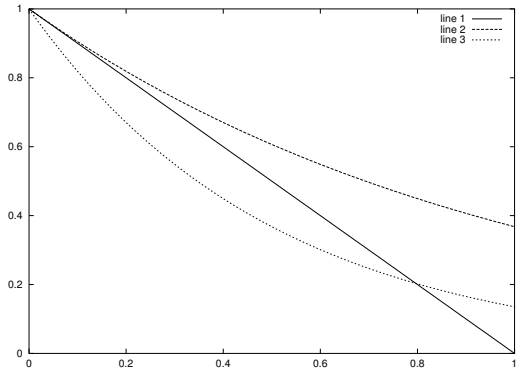


Abbildung 2.1: Die drei Funktionen aus der Abschätzung (2.18) – wie man sieht, funktioniert das fast bis 0.8.

Beweis: Wir nehmen an, daß p/q bereits gekürzt sind, das heißt, daß $\mathrm{ggT}\,(p,q) = 1$ ist, ansonsten kann man ja (mit Euklid) den größten gemeinsamen Teiler ermitteln und dann kürzen. Dann definieren wir a_0 und r durch $p = a_0 q + r$, $0 \leq r < q$; ist $r = 0$, dann ist $x = \frac{p}{q} = a_0 = [a_0;\,]$, andernfalls haben wir

$$\frac{p}{q} = \frac{a_0\,q + r}{q} = a_0 + \frac{r}{q} = a_0 + \frac{1}{\dfrac{q}{r}} \tag{2.21}$$

Nun führen wir Induktion über den Nenner q durch[26] und erhalten aus der Induktionshypothese die endliche Kettenbruchdarstellung

$$\frac{q}{r} = [a_1; a_2, \ldots, a_k]\,,$$

was wir nur noch in (2.21) einsetzen müssen, um

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

zu bekommen. □

Der Beweis suggeriert natürlich die folgende Abschätzung für die Länge des Kettenbruchs.

Korollar 2.16 *Ist* $\frac{p}{q} = [a_0; a_1, \ldots, a_k]$*, dann ist* $k \leq q$*.*

Übung 2.2 Kann in Korollar 2.16 $k = q$ sein? ♢

[26] Der Fall $q = 0$ ist Unsinn under Fall $q = 1$ trivial!

Die Rekursionsformel (2.4) für die kanonische Darstellung der k–ten Konvergenten zeigen uns sofort, daß der Zähler p_k eine ganze Zahl[27], der Nenner q_k sogar eine natürliche Zahl ist. Damit stellt sich natürlich die Frage nach der Kürzbarkeit dieser Darstellung, also ob p_k/q_k bereits "optimal" ist. Die Antwort ist ja und der Beweis ist einfach.

Satz 2.17 *Die kanonischen Darstellungen der $\frac{p_k}{q_k}$ der k–ten Konvergenten sind irreduzibel.*

Beweis: Ein gemeinsamer Teiler von p_k und q_k würde auch den Ausdruck

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

aus (2.5) teilen und kann daher nur 1 sein. □

Die Nenner in der kanonischen Darstellung der k–ten Konvergenten wachsen, wie man sofort aus der Rekursionsformel sieht:

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

Aber sie wachsen auch noch ziemlich schnell, nämlich

$$q_k \geq 2^{(k-1)/2}, \qquad k \geq 1. \tag{2.22}$$

Auch ist eine Konsequenz aus der Rekursionsformel, die zusammen mit dem monotonen Wachstum

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

liefert, woraus wir mit $q_0 = 1$ und $q_1 = a_0 \geq 1$ auf (2.22) kommen[28].

Definition 2.18 *Für $k \geq 2$ bezeichnet man die Brüche*

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

als Zwischenbrüche *zwischen der $(k-2)$–ten und der k–ten Konvergenten des Kettebruchs.*

Der Name "Zwischenbrüche" ist schnell erklärt: für $j = 0$ erhalten wir die $(k-2)$–te Konvergente[29], für $j = a_k$ hingegen die k–te Konvergente – wieder einmal nichts anderes als eine Konsequenz aus der Rekursionsformel (2.4).

Proposition 2.19 *Für gerades k bilden die Zwischenbrüche eine monoton steigende, für ungerades k eine monoton fallende Folge.*

[27] Je nachdem, ob a_0 positiv oder negativ ist, denn irgendwo muß das Vorzeichen ja landen.
[28] Wer will, kann gerne eine formale und vollständige Induktion durchführen.
[29] Genauer gesagt, ihre kanonische Darstellung, aber das ist ja jetzt sowieso **der** gekürzte Bruch.

Beweis: Für $j \geq 0$ betrachten wir die Differenz

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

und die ist für gerade k positiv, für ungerade k hingegen negativ. □

Die Zwischenbrüche sind *Medianten* zwischen der $(k-1)$–ten und der k–ten Konvergente; generell ist der *Mediant* zweier Brüche a/b und c/d als

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

definiert[30]. Wie man aus Proposition 2.19 hängt der Wert des Medianten sehr wohl von der Darstellung des Bruches ab: Die Zwischenbrüche sind Medianten von

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

haben aber für unterschiedliches j unterschiedliche, entweder monoton steigende oder monoton fallende, Werte. Wir können es aber auch noch anders sehen:

> *Der j–te Zwischenbruch ist der Mediant aus dem $(j-1)$–ten Zwischenbruch und der $(k-1)$–ten Konvergente:*
>
> $$\frac{j\,p_{k-1}+p_{k-2}}{j\,q_{k-1}+q_{k-2}} = \frac{(j-1)\,p_{k-1}+p_{k-2}}{(j-1)\,q_{k-1}+q_{k-2}} \oplus \frac{p_{k-1}}{q_{k-1}}$$

Generell liegt der Wert eines Medianten immer zwischen den beiden Brüchen, genauer,

$$b,d>0, \quad \frac{a}{b} < \frac{c}{d} \qquad \Rightarrow \qquad \frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d}, \tag{2.23}$$

Die Annahme $a/b < c/d$ oder äquivalent $bc-ad>0$ ist keine Einschränkung, denn wenn die beiden Brüche nicht gerade gleich sind, dann muß einer von beiden der kleinere sein, warum also nicht a/b? Die Ungleichungen in (2.23) folgen nun einfach aus der Beobachtung, daß

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

und

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

[30] Der Mediant ist also das, was herauskommt, wenn man Brüche so "addiert", wie man in der Schule immer wollte und nie durfte.

ist. Also liegt insbesondere jeder Zwischenbruch zwischen zwei aufeinanderfolgenden Konvergenten. Dazu betrachtet man die Folge der potentiellen Zwischenbrüche b_j definiert durch

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

Übung 2.3 Zeigen Sie, daß die Darstellung

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

irreduzibel ist. ♢

Als anständiger Mediant liegt dann b_j zwischen b_{j-1} und p_{k-1}/q_{k-1}. Da die k–te Konvergente gerade b_{a_k} ist und da der Grenzwert $a = [a_0; a_1, \ldots]$ eines unendlichen Kettenbruchs zwischen der $(k-1)$–ten und der k–ten Konvergente liegt, finden wir den Grenzwert also immer zwischen b_1 und p_{k-1}. Andererseits ist aber b_1 wieder der Mediant der $(k-2)$–ten und der $(k-1)$–ten Konvergente. Und bevor das nun zu verwirrend wird, stellen wir die Situation einmal für gerades k dar:

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

für ungerade k drehen sich einfach alle Ungleichungszeichen um. Ersetzen wir nun in dieser Ungleichungskette k durch $k+2$, dann erhalten wir insbesondere, daß für gerade k die Beziehung

$$\frac{p_k}{q_k} < \frac{p_k}{q_k} \oplus \frac{p_{k+1}}{q_{k+1}} < a < \frac{p_{k+1}}{q_{k+1}}, \tag{2.24}$$

für ungerade k dasselbe mit umgedrehten Ungleichungszeichen, gilt. Diese Beobachtung hat eine interessante Konsequenz für die Approximationsgüte der Kettenbrüche.

Satz 2.20 *Für $a = [a_0; a_1, \ldots]$ und $k \geq 0$ ist*

$$\frac{1}{q_k\,(q_{k+1} + q_k)} < \left| a - \frac{p_k}{q_k} \right| < \frac{1}{q_k\,q_{k+1}}. \tag{2.25}$$

Dieser Satz sagt uns also, daß die obere Abschätzung für die Konvergenzordnung der Kettenbrüche praktisch optimal ist! Da die Nenner q_k monoton steigend sind, ist $q_{k+1} > q_k$, also $q_{k+1} + q_k < 2q_{k+1}$ und somit

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

was uns die etwas gröbere, aber instruktivere Einschließung

$$\frac{1}{2\,q_k\,q_{k+1}} < \left| a - \frac{p_k}{q_k} \right| < \frac{1}{q_k\,q_{k+1}} \tag{2.26}$$

liefert. Da die q_k ja wie $2^{k/2}$ wachsen, ist der Faktor 2 in (2.26) eher irrelevant und wir können sagen, daß die k–ten Konvergenten etwas wie 2^{-k} konvergieren. Mit jeder Konvergente bekommen wir also eine Binärziffer des Bruchs "sicher".

Beweis von Satz 2.20: Die obere Abschätzung in (2.25) ist ja genau Satz 2.11, für die untere Abschätzung schauen wir uns nochmals die Medianten etwas genauer an[31]; tatsächlich sagt uns (2.24), daß der Mediant der k–ten und $(k+1)$–ten Konvergente immer näher bei der k–ten Konvergente liegt als die k–te Konvergente, daß als

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

genau wie behauptet. □

Satz 2.21 *Jede reelle Zahl $x\in\mathbb{R}$ kann auf genau eine Art durch einen Kettenbruch dargestellt werden. Der Kettenbruch ist endlich, wenn die Zahl rational ist und unendlich, wenn sie irrational ist.*

So wie er momentan da steht, ist Satz 2.21 allerdings nicht richtig! Endliche Kettenbrüche können nämlich gar nicht eindeutig sein. Das zeigt das ganz einfache Beispiel, daß

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

ist. Daraus folgt sofort, daß *immer* $[a_0;a_1,\ldots,a_n]=[a_0;a_1,\ldots,a_n-1,1]$ bzw. $[a_0;a_1,\ldots,a_n,1]=[a_0;a_1,\ldots,a_n+1]$ ist. Endliche Kettenbrüche, die auf "1" enden, haben also die Mehrdeutigkeit bereits eingebaut. Daher ist es am besten, wir einigen uns auf die folgende Konvention: *Kein endlicher Kettenbruch darf 1 als letzte Komponente haben.*

Beweis von Satz 2.21: Daß rationale Zahlen durch endliche Kettenbrüche dargestellt werden können wissen wir ja schon aus Satz 2.15. Was wir noch nicht gezeigt haben sind die Kettenbruchentwicklung für irrationale Zahlen und vor allem die Eindeutigkeit der Kettenbruchdarstellung.

Dazu betrachten wir zuerst einmal die allgemeine Entwicklungsregel für Kettenbrüche, ausgehend von einer Zahl $x\in\mathbb{R}\setminus\mathbb{Q}$. Dann bleibt uns gar nicht viel mehr übrig, als

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

zu setzen. Das liefert uns dann entweder $a=a_0$ oder wir setzen

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

da $0<x-a_0<1$ ist. Und jetzt machen wir so weiter, indem wir

$$a_j=\lfloor r_j\rfloor,\qquad r_{j+1}=\frac{1}{r_j-a_j},\qquad j=1,2,\ldots \tag{2.27}$$

[31] Irgendeinen Grund muss es ja gehabt haben, daß wir sie eingeführt haben.

setzen – die so erhaltene Folge erfüllt schon einmal $a_0 \in \mathbb{Z}$, $a_j \in \mathbb{N}$, $j \in \mathbb{N}$, und definiert daher einen konvergenten Kettenbruch. Abbrechen würde diese Folge nur, wenn $a_j = r_j$ wäre, aber dann hätten wir einen *endlichen* Kettenbruch erhalten und somit wäre x eine rationale Zahl. Nun gilt aber, nach Konstruktion und unter Verwendung der "unendlichen Variante" von (2.9), die Indentität

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

Dann ist aber[32]

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

und somit

$$\left| x - \frac{p_n}{q_n} \right| < \frac{1}{q_n^2} \tag{2.28}$$

und die Kettenbrüche konvergieren tatsächlich gegen a, und zwar mit der gewohnten Geschwindigkeit.

Bleibt noch die Eindeutigkeit. Die macht Gebrauch von der Tatsache, daß wir, um eine Zahl $x \in \mathbb{R}$ durch einen Kettenbruch mit *natürlichen* Einträgen darzustellen, den Wert a_0 als $a_0 = \lfloor x \rfloor$ wählen *müssen*[33], denn der Rest, also ein Kettenbruch der Form $[0; a_1, \ldots]$ liegt ja immer zwischen 0 und 1. Sind also nun $[a_0; a_1, \ldots]$ und $[a_0'; a_1', \ldots]$ zwei Kettenbruchentwicklungen derselben Zahl $a \in \mathbb{R}$, dann erhalten wir sofort, daß $a_0 = a_0'$. Haben wir, per Induktion über $k \geq 0$, bereits gezeigt, daß

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

dann liefert uns

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

daß

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

und somit muß nach unserem obige Argument auch $a_{k+1}' = a_{k+1}$ sein. □

Beispiel 2.22 *Die Konstruktion der Kettenbruchentwicklung erlaubt es uns auch, diejenigen Zahlen anzugeben, die eine Kettenbruchentwicklung der Form*

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

[32] Muß man noch dazusagen, daß hier wieder einmal die Rekursionsformel ins Spiel kommt?

[33] Es sei denn, x wäre eine ganze Zahl, dann hätten wir wieder das Problem mit der Eindeutigkeit!

haben, denn bei diesen ist $r_1 = x$*, also*

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

Die negative Nullstelle der quadratischen Gleichung macht natürlich hier keinen Sinn. Insbesondere erhalten wir also, daß

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

und das ist der goldene Schnitt.

Wir können das Spiel noch ein bißchen weitertreiben und uns mal ansehen, was wir mit 2*–periodischen Kettenbrüchen der Form* $x = [k_1; k_2, k_1, k_2, \ldots]$ *bekommen. Die Fixpunktgleichung lautet nun, daß* $r_2 = x$ *sein muß, also*

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

und wir suchen jetzt also die Nullstellen von

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

Übung 2.4 Zeigen Sie: Jeder periodische Kettenbruch gehört zu $\mathbb{Q} + \sqrt{\mathbb{Q}}$, kann also als $q + r$, $q, r^2 \in \mathbb{Q}$, geschrieben werden.
Hinweis: Beweisen Sie zuerst, daß jedes x, das als periodischer Kettenbruch geschrieben werden kann, eine Gleichung der Form

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

erfüllt. ♢

2.4 Konvergenten als beste Approximanten

Nachdem wir die Frage der *Darstellbarkeit* reeller Zahlen durch Kettenbrüche geklärt haben, können wir jetzt die Benutzung der Kettenbrüche dadurch rechtfertigen, daß wir zeigen, daß die Kettenbrüche reelle Zahlen besser approximieren als alle anderen Brüche. Denn da $\mathbb{Q}$ dicht in $\mathbb{R}$ liegt gibt es natürlich viele Folgen von Brüchen, die gegen eine vorgegebene Zahl $x \in \mathbb{R}$ konvergieren.

Ein gutes Maß dafür, wie "kompliziert" ein Bruch, also eine rationale Zahl $x \in \mathbb{Q}$ ist, ist die Größe des Nenners: Schreiben wir nmlich x als

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

dann ist die Information, die wir brauchen, um den Bruch zu speichern, maximal von der Größenordnung[34] $\log a + 2\log q$. Da der Aufwand für den ganzzahligen Anteil einer reellen Zahl sowieso unabhängig davon ist, wie *genau* wir die Zahl als Bruch darstellen können, ist also die Länge des Nenners die für uns relevante Größe. Das ist mehr als genug Rechtfertigung für das folgende Qualitätskriterium.

Definition 2.23 *Ein Bruch a/b heißt* bester Approximant *an $x \in \mathbb{R}$, wenn*

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

Dabei sind Brüche immer in der Form $\mathbb{Z}/\mathbb{N}$ zu sehen, Nenner sind also immer nichtnegativ.

Und nun kommt die Stärke der Kettenbrüche: Sie sind beste Approximanten!

Satz 2.24 *Jeder beste Approximant an eine reelle Zahl x ist entweder eine Konvergente der zugehörigen Kettenbruchentwicklung oder ein Zwischenbruch.*

Beweis: Sei a/b ein[35] bester Approximant[36] an $x = [a_0; a_1, \ldots]$. Dann ist $a/b > a_0$, denn ansonsten wäre $a/b < a_0 = \lfloor x \rfloor < x$ und $a_0/1$ wäre bereits ein besserer Approximant als a/b. Mit exakt derselben Argumentation sehen wir auch, daß $\frac{a}{b} < a_0 + 1$ sein muß, denn sonst wäre ja $a_0 + 1$ ein besserer Approximant, da $x < a_0 + 1$ ist. Wir haben also $a_0 \leq \frac{a}{b} \leq a_0 + 1$, und würde bei einer der beiden Ungleichungen Gleichheit eintreten, dann ist die Behauptung ja bewiesen – der beste Approximant wäre entweder die Konvergente a_0 oder der Zwischenbruch

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

Nehmen wir jetzt einmal an, daß $a_0 < \frac{a}{b} < a_0 + 1$ ist und daß a/b weder Konvergente noch Zwischenbruch ist. Wir werden zeigen, daß es dann einen Zwischenbruch[37] gibt, der kleineren Nenner hat, aber näher an x liegt. Nach dem, was wir vorher über Zwischenbrüche in Erfahrung gebracht haben[38], insbesondere nach Proposition 2.19, liegt also a/b zwischen zwei Zwischenbrüchen[39], das heißt, es gibt n und k, so daß entweder

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

[34] Dezimalzahlen mit k Stellen decken beispielsweise den Bereich von 0 bis $10^k - 1$ ab, das heißt, der Speicheraufwand für eine Dezimalzahl d beträgt $\log_{10} d$. Und wenn uns die Basis nicht interessiert, dann stecken wir das in eine Konstante und reden nur von Größenordnungen.

[35] Wir haben ja nie behauptet, beste Approximanten wären eindeutig!

[36] Manche Leute sagen auch "Element bester Approximation".

[37] Der natürlich auch eine Konvergente sein kann.

[38] Für irgendwas muß das ja gut sein.

[39] Zur Erinnerung: Die Zwischenbrüche zu Konvergenten gerader Ordnung bilden eine (verfeinerte) Folge, die aufsteigend gegen x konvergiert, die zu Konvergenten ungerader Ordnung hingegen eine absteigen konvergente Folge.

und damit ist

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

Andererseits gibt es $c\in\mathbb{N}$, so daß

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

was uns

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

liefert. Damit hat aber der $(k+1)$–te Zwischenbruch, der nach Konstruktion näher beim Grenzwert x liegt als a/b einen *kleineren* Nenner als a/b und ist damit ein besserer Approximant. Also muß also der beste Approximant Konvergente oder Zwischenbruch sein. □

Tatsächlich sind die Konvergenten aber sogar eindeutige beste Approximanten, nämlich dann, wenn man den Begriff der Bestapproximation ein klein wenig schärfer formuliert. Dabei erinnert man sich am besten erst einmal daran, was ein Bruch a/b eigentlich ist, nämlich diejenige Zahl, die mit b multipliziert den Wert a ergibt. Dann ist ein Bruch aber auch eine gute Näherung an x, oder, anders gesagt, x eine gute Näherung an den Bruch, wenn die Differenz $|bx-a|$ möglichst klein ist. Man könnte auch sagen[40], daß in so einem Bruch a viele gültige Stellen hat, wenn der Fehler klein ist.

Definition 2.25 *Ein Bruch a/b heißt* bester Approximant der zweiten Art *an $x\in\mathbb{R}$, wenn*

$$\frac{c}{d}\neq\frac{a}{b},\quad 0<d\le b\qquad\Longrightarrow\qquad |bx-a|\le|dx-c|\,. \tag{2.29}$$

Beste Approximanten zweiter Art sind auch beste Approximanten erster Art[41], denn wären sie das nicht, gäbe es einen Bruch c/d, $d\le b$, so daß

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

und damit wäre a/b auch kein Bestapproximant zweiter Art. Aber nicht jeder Bestapproximant erster Art ist auch ein Bestapproximant zweiter Art! Das einfachste Beispiel ist $x=\frac{1}{5}$ und $\frac{a}{b}=\frac{1}{3}$; man sieht sofort, daß $\frac{1}{3}$ näher bei $\frac{1}{5}$ liegt als die "Kokurrenz" $\left\{0,\frac{1}{2},\frac{2}{3},1\right\}$, aber leider ist

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

Trotzdem spielen Bestapproximanten zweiter Art eine wichtige Rolle, denn das sind nun wirklich die Konvergenten!

[40] Man denke an $b=10^k$ für $k\in\mathbb{N}$.

[41] Also im Sinne von Definition 2.23.

Satz 2.26 *Jeder beste Approximant der zweiten Art an $x \in \mathbb{R}$ ist eine Konvergente und jede Konvergente ist auch ein bester Approximant der zweiten Art. Bis auf den Spezialfall $x = a_0 + \frac{1}{2}$ und die Konvergente erste Ordnung sind auch alle besten Approximanten zweiter Art* eindeutig.

Beweis: Nehmen wir an, $\frac{a}{b}$ wäre ein bester Approximant zweiter Art an $x = [a_0; a_1, \ldots]$. Wäre $a/b < a_0 = \lfloor x \rfloor < x$, dann wäre, da $b \geq 1$ ist,

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

und $a_0/1$ würde bereits die Bestapproximation zweiter Art zunichte machen. Also ist $a_0 \leq a/b$, liegt also rechts von der Konvergente $\frac{p_0}{q_0}$. Damit liegt a/b, angenommen der Bruch wäre keine Konvergente, entweder rechts vom Maximalwert, also $\frac{a}{b} > \frac{p_1}{q_1}$, oder aber zwischen zwei aufeinanderfolgenden Konvergenten $\frac{p_{k-1}}{q_{k-1}}$ und $\frac{p_{k+1}}{q_{k+1}}$. Im ersten Fall ist $x < \frac{p_1}{q_1} < \frac{a}{b}$ und daher liefert uns das strikte Ansteigen der Nenner q_k

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

also

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

und schon kann a/b kein Bestapproximant zweiter Art mehr sein. Liegt a/b zwischen zwei Konvergenten, dann ist zuerst einmal wieder

$$\left| \frac{a}{b} - \frac{p_{k-1}}{q_{k-1}} \right| = \frac{|a\,q_{k-1} - b\,p_{k-1}|}{b\,q_{k-1}} \geq \frac{1}{b\,q_{k-1}} \tag{2.30}$$

und außerdem[42], nach Korollar 2.5,

$$\left| \frac{a}{b} - \frac{p_{k-1}}{q_{k-1}} \right| < \left| \frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}} \right| = \frac{1}{q_k\,q_{k-1}}, \tag{2.31}$$

so daß wir (2.30) und (2.31) zu $q_k < b$ kombinieren können. Andererseits ist aber auch

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

was zusammen mit der quantitativen Approximationsaussage aus (2.26) die Abschätzung

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

liefert, weswegen die k–te Konvergente ein besserer Approximant zweiter Art wäre.

[42] Daß hier die k–te Konvergente auftaucht, ist kein Fehler. Sie liegt auf alle Fälle sogar auf der "anderen" Seite von x!

Dann also ran an die Umkehrung! Hierbei fixieren wir erst einmal k und betrachten wir die Werte

$$|b\,x - a|\,, \qquad a \in \mathbb{Z},\, b \in \{1, \ldots, q_k\} \tag{2.32}$$

und bezeichnen mit b^* denjenigen Wert von b, für den der Ausdruck (für passendes a) minimal wird. Gibt es mehrere b, für die der Minimalwert angenommen wird, dann nehmen wir als b^* einfach den kleinsten dieser Werte. Der zugehörige a–Wert sei dann a^*. Wir zeigen zuerst, daß a^* eindeutig ist. Gäbe es auch noch ein $a' \neq a^*$ für das ebenfalls der Minimalwert angenommen wird, dann wäre

$$\left|x - \frac{a^*}{b^*}\right| = \left|x - \frac{a'}{b^*}\right| \qquad \Rightarrow \qquad x = \frac{a^* + a'}{2\,b^*}. \tag{2.33}$$

Der Bruch auf der rechten Seite von (2.33) muß aber irreduzibel sein, denn sonst wäre[43] $x = p/q$ mit $q \leq b^*$ eine irreduzible Darstellung und damit $|q\,x - p| = 0$, ein Minimum in (2.32), das nicht unterboten werden kann und *genau* für $a = p$ und $b = q \leq b^* \leq q_k$ angenommen wird. Entwickeln wir die rationale Zahl x nun als Kettenbruch, dann ist $x = [a_0; a_1, \ldots, a_n]$ und[44]

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

so daß $q_{n-1} < b^*$ für $n > 1$ ist. Im Spezialfall $n = 1$ können wir $q_1 = b^*$ genau dadurch erreichen, daß $a_n = 2$ und damit $b^* = 2$ ist[45]. Das ist genau der Spezialfall $x = a_0 + \frac{1}{2}$ und hier ist gerade

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

das heißt, der beste Approximant zweiter Art ist nicht eindeutig. Ansonsten ist immer $1 \leq q_{n-1} < b^*$ und damit ist, wegen der Annahme $a^* \neq a'$ und mit (2.33)

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

im Widerspruch zur Annahme, daß a^*/b^* bester Approximant zweiter Art ist. Damit haben wir also gezeigt, daß a^* *eindeutig* ist, und folglich ist a^*/b^* *eindeutiger* Bestapproximant zweiter Art an x. Aber jeder Bestapproximant zweiter Art muß, das haben wir in der ersten Hälfte des Beweises ja schon gezeigt, eine Konvergente p_m/q_m mit $m \leq k$ sein. Ist $m = k$, dann sind wie fertig, andernfalls erhalten wir mit zwei Anwendungen von (2.25), daß

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

[43] Einen Faktor 2 müssen wir ja mindestens kürzen.

[44] Nicht vergessen: die letzte Stelle in einem Kettenbruch darf nie den Wert 1 haben, siehe Satz 2.21.

[45] Schließlich ist ja immer $q_0 = 1$.

also, indem wir oben k druch $k-1$ ersetzen und wieder einmal mit der Rekursionsformel arbeiten,

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

was ein Widerspruch ist. Damit haben wir also endlich gezeigt, daß p_k/q_k ein *strikter* Bestapproximant zweiter Art ist. Damit sind diese Bestapproximanten natürlich auch eindeutig – bis auf den einen Sonderfall natürlich. □

Die Bestapproximationseigenschaft der Kettenrbüche zeichnet auch für eine ihrer ersten Anwendungen verantwortlich, zumindet laut [17]: Als Huygens[46] ein Modell des Sonnensystems konstruierte, brauchte er Zahnräder, mit denen er den Umlaufzeiten der Planeten möglichst nahe kommen konnte – aber die Verhltnisse der Anzahl der Zähne ist eine rationale Zahl, aus technischen Gründen nach oben begrenzt und somit erhält man die beste Näherung durch einen Kettenbruch!

2.5 Approximationsaussagen

Nachdem wir also jetzt die bestapproximierenden Brüche[47] als Konvergenten oder Zwischenbrüche identifiziert haben, je nachdem, ob wir uns für Bestapproximanten erster oder zweiter Art interessieren, wollen wir uns jetzt mit der Frage herumschlagen, wie gut allgemein relle Zahlen durch rationale Zahlen mit einem bestimmten Maximalnenner approximiert werden können. Im Beweis von Satz 2.21, genauer, in (2.28), haben wir bereits eine obere Abschtzung für die Approximationsgüte der Konvergenten kennengelernt, nämlich

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

Andererseits zeigt uns die Zahl $a = [0; n, 1, n]$, für die wir

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

und somit $a = \frac{n+1}{n(n+2)}$ erhalten, daß

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

und es daher im allgemeinen auch nicht besser geht als wie q_n^{-2}, denn für jedes $\varepsilon > 0$ gibt es ein $n \in \mathbb{N}$, so daß $(1+2/n)^{-1} < 1 - \varepsilon$ ist. Allerdings sind solche "worst–case"–Aussagen so viel auch wieder nicht wert, wie die nchste Aussage zeigt.

[46] Christiaan Huygens, 1629–1695, studierte Jura (zu dieser Zeit konnten auch und sogar aus Juristen noch etwas vernünftiges werden) und Mathematik in Leiden; Zeitgenosse von und bekannt mit Descartes, Leibniz, Newton und anderen

[47] Relativ zur Nennergröße.

Proposition 2.27 *Wenn eine Zahl $x \in \mathbb{R}$ eine k–te Konvergente besitzt*[48]*, dann gilt mindestens eine der beiden folgenden Ungleichungen:*

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

Beweis: Da x zwischen den beiden Konvergenten liegt ist, wieder einmal mit (2.6),

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

und die Ungleichung zwischen dem geometrischen und arithmetischen Mittel liefert

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

und somit erhalten wir, daß

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

ist, was sicherlich nicht erfüllt wäre, wenn die Aussage der Proposition nicht gelten würde. □

Also hat mindestens jede zweite Konvergente eine Approximationsordnung nicht nur von $1/q_k^2$, sondern sogar von $1/\left(2q_k^2\right)$, und zu dieser Aussage gibt es sogar eine Umkehrung.

Satz 2.28 *Ist für $x \in \mathbb{R}$*

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

dann ist a/b eine Konvergente der Kettenbruchentwicklung von x.

Beweis: Nach Satz 2.26 brauchen wir nur zu zeigen, daß a/b ein Bestapproximant zweiter Art an x ist. Wäre nun $|dx - c| < |bx - a| < 1/2b$, dann wäre auch

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

und somit, unter der Annahme, daß $a/b \neq c/d$,

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

Doch das bedeutet, daß

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

ist, also ist a/b wirklich Bestapproximant zweiter Art. □

Proposition 2.27 hat eine Verschärfung in dem Sinne, daß man unter *drei* aufeinanderfolgenden sogar noch besser approximieren kann als $1/2q_k^2$, und zwar wie folgt.

[48] Wenn also x nicht als ganzzahliger Kettenbruch der Ordnung $k-1$ dargestellt werden kann.

Satz 2.29 *Wenn $x \in \mathbb{R}$ eine Konvergente der Ordnung $k > 1$ besitzt, dann gilt mindestens eine der folgenden drei Ungleichungen:*

$$\left| x - \frac{p_{k-2}}{q_{k-2}} \right| < \frac{1}{\sqrt{5}\, q_{k-2}^2}, \qquad \left| x - \frac{p_{k-1}}{q_{k-1}} \right| < \frac{1}{\sqrt{5}\, q_{k-1}^2}, \qquad \left| x - \frac{p_k}{q_k} \right| < \frac{1}{\sqrt{5}\, q_k^2}. \tag{2.34}$$

Diese Aussage gibt natürlich Hoffnung, daß man das vielleicht noch weiter verallgemeinern könnte: unter vier aufeinanderfolgenden Konvergenten findet sich mindestens eine, die noch ein bißchen besser konvergiert, unter fünf Konvergenten eine noch viel bessere, und so weiter. Dem ist aber leider[49] **nicht** so, und unser Standardbeispiel ist wieder der gute alte goldene Schnitt

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

siehe Beispiel 2.22. Da

$$x = [1; 1, ..., 1, r_k], \qquad r_k = [1; 1, \ldots] = x$$

ist, haben wir also auch

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

Die fast vergessene Formel (2.10) aus Proposition 2.8 sagt uns dann aber, daß[50]

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

Also ist

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

und daher

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

weswegen es keine bessere Approximationsordnung als $1/\sqrt{5}q_k^2$ geben kann, ganz egal wieviele aufeinanderfolgende Konvergenten wir betrachten.

Beweis von Satz 2.29: Wir setzen

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

und beweisen zuerst einmal, daß

$$k \geq 2,\ \psi_k \leq \sqrt{5},\ \psi_{k-1} \leq \sqrt{5} \qquad \Rightarrow \qquad \varphi_k > \frac{\sqrt{5}-1}{2}. \tag{2.35}$$

[49] Oder vielleicht auch glücklicherweise?

[50] Hier taucht übrigens ein endlicher Kettenbruch mit 1 am Ende auf – macht aber nix, denn einen Wert hat dieser Ausdruck ja!

Da

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

und

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

ist, erhalten wir, daß

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

so daß uns die Annahmen auf der linke Seite von (2.35) die Ungleichungen

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

liefern, weswegen

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

sein muß – und da φ_k eine rationale Zahl ist, also keine Gleichheit auftreten kann, gilt sogar die *strikte* Ungleichung. Durchmultiplizieren mit $\varphi_k/\sqrt{5} > 0$ ergibt dann, daß

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

wie in (2.35) behauptet.

Jetzt können wir uns aber endlich an unseren eigentlichen Beweis machen! Dazu nehmen wir an, es wäre

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

dann folgt aus dieser Annahme zusammen mit[51]

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

daß

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

[51] Das alte Spiel ...

und damit wäre[52]

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

was ja wohl nicht ganz sein kann. Somit haben wir den gewünschten Widerspruch. □

Also: für *beliebige* relle Zahlen ist die Approximationsordnung der Konvergenten von Kettenbrüchen beschränkt, und zwar im wesentlichen durch $1/\sqrt{5}q_n^2$. Auf der anderen Seite gibt es aber irrationale Zahlen, die beliebig gut, gut im Sinne einer Approximationsordnung, durch Konvergenten angenähert werden können.

Satz 2.30 *Für jede Funktion $\varphi : \mathbb{N} \to \mathbb{R}_+$ gibt es eine Zahl $x \in \mathbb{R}$, so daß für* unendlich viele *Werte $q \in \mathbb{N}$ die Ungleichung*

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

erfüllt ist.

Beweis: Wir konstruieren x über seine Kettenbruchentwicklung, indem wir einfach $a_0 \in \mathbb{Z}$ beliebig und

$$a_{k+1} > \frac{1}{q_k^2 \varphi\left(q_k\right)}, \qquad k \in \mathbb{N}_0, \tag{2.36}$$

wählen – das können wir natürlich auf vielerlei Art tun. Dann ist für $x = [a_0; a_1, \ldots] \in \mathbb{R}$, wieder einmal unter Verwendung von (2.13) aus Satz 2.11,

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

und zwar sogar *für alle* $k \in \mathbb{N}_0$. □

Die Formel (2.36) zur Bestimmung von a_k sagt uns schon, was wir tun müssen, um eine Zahl x zu bekommen, die durch die Konvergenten schnell, also mit sehr schnell abfallendem φ, approximiert werden kann: Die Komponenten a_k in der Kettenbruchentwicklung von x müssen wachsen. Daß das so sein muß zeigt uns die untere Abschätzung aus (2.25), mit deren Hilfe wir auf

$$\begin{aligned} \frac{1}{a_{k+1}\, q_k^2} \;&>\; \left| x - \frac{p_k}{q_k} \right| > \frac{1}{q_k\left(q_{k+1} + q_k\right)} = \frac{1}{q_k\left(a_{k+1} q_k + q_{k-1} + q_k\right)} \\ &=\; \frac{1}{q_k^2\left(a_{k+1} + 1 + q_{k-1}/q_k\right)} > \frac{1}{\left(a_{k+1} + 2\right) q_k^2} \end{aligned} \tag{2.37}$$

kommen, so daß wir bei einer Approximationsgüte in der Größenordnung $\varphi\left(q_k\right) \sim 1/a_{k+1}\, q_k^2$ landen. Das legt natürlich die Vermutung nahe, daß "schnelle" Approximierbarkeit etwas mit dem Wachstum der Koeffizienten zu tun haben könnte. Und das ist auch der Fall.

[52] Hurrah, wir dürfen zweimal nach unten abschätzen, einmal wegen der Division, einmal wegen des Vorzeichens.

Satz 2.31 *Sei $x \in \mathbb{R} \setminus \mathbb{Q}$ eine irrationale Zahl. Sind die Koeffizienten in dern Kettenbruchentwicklung von x beschränkt, dann gibt es eine Zahl $c > 0$, so daß die Ungleichung*

$$\left| x - \frac{p}{q} \right| < \frac{c}{q^2}, \qquad p \in \mathbb{Z},\, q \in \mathbb{N}, \tag{2.38}$$

keine Lösung hat. Sind umgekehrt die Koeffizienten der Kettenbruchentwicklung unbeschränkt, dann gibt es für jedes $c > 0$ unendlich viele *Lösungen von (2.38).*

Beweis: Ist $\sup\{a_k \,:\, k \in \mathbb{N}_0\} = M < \infty$, dann folgt aus der unteren Abschätzung in (2.37), daß

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

Für einen beliebigen Bruch p/q wählen wir k so, daß $q_{k-1} < q \le q_k$ ist, und da alle Konvergenten beste Approximanten zweiter und somit auch erster Art an x sind, ergibt sich, daß

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

wobei die Konstante c nur von der Schranke M, nicht aber vom Nenner q abhängt.

Ist andererseits $\sup\{a_k \,:\, k \in \mathbb{N}\} = \infty$, dann gibt es für jedes feste $c > 0$ unendlich viele Indizes k, so daß $a_{k+1} > 1/c$, und damit können wir die obere Abschätzung aus (2.37) direkt für

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

verwenden, was natürlich unendlich viele "Lösungen" liefert. □

2.6 Algebraische Zahlen

Algebraische Zahlen sind die Nullstellen oder *Wurzeln* von Polynomen mit rationalen oder ganzzahligen Koeffizienten – für eine Nullstelle ist es nicht relevant, ob man mit dem Hauptnenner der Koeffizienten durchmultipliziert oder nicht. Formal ist $a \in \mathbb{R}$ eine *algebraische Zahl der Ordnung* n, wenn es

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

gibt, so daß $f(a) = 0$ ist. Dabei können und werden wir annehmen, daß n minimal gewählt ist, daß es also kein Polynom g vom Grad $< n$ mit ganzzahligen Koeffizienten gibt, so daß $g(a) = 0$ ist. Eine reelle Zahl, die nicht algebraisch ist, nennt man bekanntlich *transzendent*, die

klassischen Beispiele sind e und π. Mit algebraischen Zahlen kann man noch halbwegs rechnen, indem man diese Zahlen, genauer gesagt, die Polynome, die sie definieren, zum Grundkörper adjungiert. Das passiert auch in der Computeralgebra, siehe [8, 26], und führt zu den Ausdrücken der Form `RootOf`, mit denen Computeralgebrasysteme wie `Maple`, `Mathematica` oder `MuPAD`[53] ihre Benutzer erfreuen.

Was uns hier allerdings interessieren soll, ist die Tatsache, daß algebraische Zahlen durch Kettenbrüche wieder eher langsam approximierbar sind, ein Satz, der auf Liouville[54] zurückgeht.

Satz 2.32 (Liouville) *Für jede algebraische Zahl $a \in \mathbb{R} \setminus \mathbb{Q}$ der Ordnung n gibt es eine Konstante $C > 0$, so daß*

$$\left| a - \frac{p}{q} \right| > \frac{C}{q^n}, \qquad p \in \mathbb{Z},\, q \in \mathbb{N}. \tag{2.39}$$

Beweis: Die algebraische Zahl a der Ordnung n ist eine Nullstelle eines Polynoms $f \in \mathbb{Z}[x]$, und wir können f als

$$f(x) = (x-a)\; g(x), \qquad g \in \mathbb{R}[x],\; g(a) \neq 0, \tag{2.40}$$

schreiben. Wäre nämlich $g(a) = 0$, dann könnten wir g ebenfalls durch $x - a$ teilen und es wäre $f(x) = (x-a)^2\, h(x)$, also

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

und da f' ebenfalls zu $\mathbb{Z}[x]$ gehört, wäre a nur eine algebraische Zahl der Ordnung $n-1$. Ist aber $g(a) \neq 0$, dann gibt es, wegen der Stetigkeit der Polynome, ein $\delta > 0$, so daß

$$g(x) \neq 0, \qquad x \in [a-\delta, a+\delta]\,. \tag{2.41}$$

Seien nun $p \in \mathbb{Z}$ und $q \in \mathbb{N}$ so nahe bei a, daß $|a - p/q| < \delta$; da δ nur von a abhängt[55], müssen alle hinreichend guten Approximationen an a diese Eigenschaft haben. Nach (2.41) ist dann $f\,(p/q) \neq 0$, und durch Einsetzen von $x = p/q$ in $x - a = f(x)/g(x)$, siehe (2.40), erhalten wir, daß

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

[53] Siehe `www.mupad.org`

[54] Joseph Liouville, 1809–1882, beeinflusst von Ampère, Cauchy und Poisson an der École Polytechnique, beschäftigte sich neben Zahlentheorie auch mit Differentialgeometrie und Differentialgleichungen, ist also ein gutes Beispiel, daß die Trennung zwischen Analysis und Algebra künstlich ist. Von ihm stammt auch die "Liouville–Zahl"

$$0.11000100000000000000000010000\ldots$$

mit "1" genau an den Nachkommastellen $1!, 2!, 3!, 4!, \ldots$, die ebenfalls transzendent ist. Quelle für historische Informationen ist wie immer [20].

[55] Die Polynome f und g gehören ja zur Definition der algebraischen Zahl a.

Der Zähler dieses Bruchs ist eine von Null verschiedene ganze Zahl, denn schließlich haben wir ja gefordert, daß a irrational und nicht p/q sein soll. Damit ist der Zähler aber ≥ 1 im Absolutbetrag und somit gilt

$$\left|a - \frac{p}{q}\right| \geq \frac{1}{M\, q^n}, \qquad M = \max_{x\in[a-\delta,a+\delta]} |g(x)|\,, \tag{2.42}$$

wann immer $|a - p/q| \leq \delta$. Ist andererseits aber $|a - p/q| > \delta$, dann ist trivialerweise[56] auch $|a - p/q| > \delta/q^n$ und für jede Konstante

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

ist (2.39) erfüllt. □

Dieser Satz liefert uns einen einfachen "Baukasten" für transzendente Zahlen, nämlich, indem wir sehr schnell wachsende Kettenbruchentwicklungen verwenden. Beispielsweise könnte man

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

verwenden, denn dann ist für $a = [a_0; a_1, \ldots]$ nach (2.37)

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

was natürlich irgendwann kleiner als C/q_k^n wird, wenn C und n von vornherein festgelegt werden.

Übung 2.5 Geben Sie eine explizite Kettenbruchentwicklung einer transzendenten Zahl an. ♢

Bevor wir [17] verlassen[57], einen letzten Satz, der die periodischen Kettenbrüche mit den Quadratwurzeln, also den algebraischen Zahlen zweiter Ordnung, identifiziert.

Satz 2.33 *Jeder periodische Kettenbruch stellt eine algebraische Zahl zweiter Ordnung dar und jede algebraische Zahl zweiter Ordnung hat eine periodische Kettenbruchentwicklung.*

Beweis: Hat x eine periodische Kettenbruchentwicklung, dann bedeutet das ja, daß es eine Periodenlänge ℓ und einen Anfangsindex k_0 gibt, so daß $a_{k+\ell} = a_k$ für alle $k \geq k_0$, und somit auch $r_{k+\ell} = r_k$, $k \geq k_0$, gilt. Dann ist

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

also

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

[56] Da $q \geq 1$ ist.

[57] Die maßtheoretischen Aspekte von Kettenbrüchen sind zwar ganz amüsant, aber wir wollen uns jetzt mal langsam mit Polynomen befassen.

was eine quadratische Gleichung in r_k mit ganzzahligen Koeffizienten ist. Daher ist r_k und somit auch x eine algebraische Zahl zweiter Ordnung.

Die Umkehrung ist etwas aufwendiger. Wenn $x = [a_0; a_1, \ldots]$ die Bedingung

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

erfüllt, dann schreiben wir wieder x als

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

und erhalten, daß

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

wobei

$$A_k = a\, p_{k-1}^2 + b\, p_{k-1}\, q_{k-1} + c\, q_{k-1}^2, \tag{2.43}$$

$$B_k = 2a\, p_{k-1}\, p_{k-2} + b\,(p_{k-1}\, q_{k-2} + p_{k-2}\, q_{k-1}) + 2c\, q_{k-1}\, q_{k-2}, \tag{2.44}$$

$$C_k = a\, p_{k-2}^2 + b\, p_{k-2}\, q_{k-2} + c\, q_{k-2}^2 = A_{k-1}. \tag{2.45}$$

Die *Diskriminante* $D_k = B_k^2 - 4A_kC_k$ hat dann den Wert[58]

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

und zwar unabhngig von k. Das ist schon mal gut, denn schließlich bestimmt ja die Diskriminante den "Wurzelanteil" der Zahl. Als nächstes halten wir fest, daß

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

was wir in (2.43) einsetzen können, so daß wir

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

erhalten. Nach (2.45) sind damit die Zahlen A_k und $C_k = A_{k-1}$ aber auch

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

unabängig von k nach oben beschränkt. Es gibt also nur endlich viele Kombinationen (A_k, B_k, C_k) und mindestens eine davon muß sich irgendwann wiederholen, es gibt also k, ℓ, so daß $A_{k+\ell} = A_k$, $B_{k+\ell} = B_k$ und $C_{k+\ell} = C_k$, also $r_{k+\ell} = r_k$ und nach dem Bildungsgesetz für Kettenbrüche, siehe Beweis von Satz 2.21 ist dann auch $r_{k+n\ell} = r_k$, $k \in \mathbb{N}$. □

Übung 2.6 Zeigen Sie: Ist x eine algebraische Zahl zweiter Ordnung, so auch $1/x$. ♢

[58] Wer's nicht glaubt, kann's gerne nachrechnen.

> *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*

Kettenbrüche und Polynome | 3

Jetzt ist es an der Zeit, die Kettenbrüche (mit ganzzahligen Einträgen) als Mittel zur Darstellung reeller Zahlen zu verlassen und uns andere, allgemeinere Situationen anzusehen. Insbesondere interessieren wir uns für die Darstellung *rationaler Funktionen* $f(x) = p(x)/q(x)$, p, q Polynome, mit Hilfe von Kettenbrüchen. Dazu werden wir auch ein bißchen "Theorie" über Kettenbrüche über Ringen machen, um uns dann aber ganz gezielt mit Polynomen zu befassen – allerdings nur univariate Polynomen, denn wir werden sehen, daß *euklidische Ringe* ein klein wenig unverzichtbar sind.

3.1 Zum Einstieg . . .

Kettenbrüche mit Polynomen werden in ihrer endlichen Version von der Form

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

sein, wobei $m_j(x) = a_j\, x^{k_j}$, $a_j \in \mathbb{R}$, $k_j \in \mathbb{N}$, ein *Monom* ist. Derartige Kettenbrüche mit Monomen werden in [23] als *C–Kettenbrüche* bezeichnet. Auch hier wieder sind die 1–en in den Zählern des Kettenbruchs wieder einmal *keine* Einschränkung: Ein "allgemeiner" Kettenbruch

der Form[59]

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

läßt sich ja auch immer als

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

schreiben, wobei

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

Wie man auf diese Formel kommt? Ganz einfach, man "kürzt" die Brüche sukzessive und erhält

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

und so weiter.

Jeder Kettenbruch der Form $[p; m_1, \ldots, m_n]$ ist eine rationale Funktion und, zumindest für univariate Polynome, jede rationale Funktion kann in einen endlichen Kettenbruch entwickelt werden. Aber das werden wir gleich in einem allgemeineren Zusammenhang sein.

[59] Das ist eine gute Gelegenheit, auch einmal die andere Notation für Kettenbrüche, siehe z.B. [22, 23], kennenzulernen.

3.2 Euklidische Ringe und Kettenbrüche

Zur Erinnerung: ein Ring ist eine Menge in der Addition, Subtraktion und Multiplikation wohldefiniert sind, in der also all die beliebten Gesetze wie Assoziativität, Kommutativität (bei der Addition) und Distributivität[60] gelten. Ein klein wenig spezieller wird es dann schon mit den folgenden Begriffen.

Definition 3.1 (Euklidischer Ring) *Ein Ring R heißt*

1. nullteilerfrei *oder* Integritätsring*, wenn es kein $a, b \in R \setminus \{0\}$ gibt, so daß $ab = 0$.*
2. euklidischer Ring*, wenn R eine Integritätsring ist und es eine* Bewertungsfunktion *oder* euklidische Funktion $d : R \to \mathbb{N} \cup \{-\infty\}$ *gibt, so daß es für alle $p, q \in R$, $q \neq 0$, einen* Quotienten $s \in R$ *und einen* Rest $r \in R$ *gibt, so daß*

$$p = sq + r, \qquad d(r) < d(q).$$

Wir schreiben dann auch $s =: p/q$ und $r =: (p)_q$.

Bemerkung 3.2 (Eigenschaften der Bewertungsfunktion)

1. *Jede euklidische Funktion hat die Eigenschaft, daß $d(0) < d(a)$ für alle $a \in R\setminus\{0\}$. Gäbe es nämlich*[61] *ein $a \in R \setminus \{0\}$, so daß $d(a) \leq d(R)$, dann erhalten wir mit $p = q = a$, daß eine euklidische Darstellung von a die Form*

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

haben muß, aber ganz egal wie wir s wählen, würde jeder dieser Reste $d((1 - s)r) \geq d(a)$ erfüllen und damit wäre der Ring nicht euklidisch.

2. *Es hat zwar nicht jede euklidische Funktion die Eigenschaft*

$$d(a \cdot b) \geq d(a), \qquad a, b \in R \setminus \{0\}, \tag{3.1}$$

die man von den "normalen" euklidischen Funktionen für $\mathbb{Z}$ und $\mathbb{F}[x]$ kennt und fast für "selbstverständlich" hält, aber für jeden Ring Integritätsring R gibt es so eine Bewertungsfunktion, und zwar die minimale euklidische Funktion *bzw.* minimale Bewertungsfunktion*, die als punktweises Minimum* aller *möglichen euklidischen Funktionen definiert ist, siehe [8, Übung 3.5]. Wir können und werden also immer annehmen, daß wir die* minimale *euklidische Funktion verwenden.*

3. *Der Wert $d(a) = -\infty$ ist nur für $a = 0$ zulässig – muß aber nicht angenommen werden.*

[60] Irgendwie müssen Addition und Multiplikation ja auch zusammenpassen.

[61] Die Funktion d bildet ja R nach $\mathbb{N} \cup \{-\infty\}$ ab und damit muß es zwar nicht notwendigerweiße ein Maximum, aber auf alle Fälle ein Minimum geben, d.h. ein Element $r \in R$, so daß $d(r) \leq d(R)$, also $d(r) \leq d(q)$, $q \in R$, ist.

Beispiel 3.3

1. *Die ganzen Zahlen $\mathbb{Z}$ bilden zusammen mit der Funktion $d = |\cdot|$ einen euklidischen Ring.*
2. *Die Polynome $\mathbb{K}[x]$ bilden einen euklidischen Ring mit der Funktion $d = \deg$, wobei $\deg 0 = -\infty$.*
3. *Ein Körper $\mathbb{K}$ ist ein euklidischer Ring mit $d = (1 - \delta_0)$.*
4. *Eine etwas obskurere euklidische Funktion auf $\mathbb{Z}$ ist $d(3) = 2$ und $d = |\cdot|$ sonst. Diese euklidische Funktion*[62] *erfüllt* nicht *die Bedingung (3.1), da $d(-1 \cdot 3) = d(-3) = 3 > 2 = d(3)$ ist. Trotzdem ist aber immer noch $d(0)$ minimal ...*

Der Nutzen des euklidischen Rings ist klar: Dieses Konzept erlaubt es uns, Division mit Rest durchzuführen und Reste zu erhalten, die, im Sinne der Bewertungsfunktion, kleiner sind als das Objekt, das wir dividiert haben. Und wenn man sich erinnert, daß Division mit Rest eigentlich der entscheidende Trick bei der Kettenbruchentwicklung war, dann brauchen wir eigentlich nicht mehr lange um den heißen Brei herumzureden.

Satz 3.4 *Sei R ein euklidischer Ring mit Einselement. Dann ist jeder endliche Kettenbruch $[r_0; r_1, \ldots, r_n]$, $r_j \in R$, eine rationales Element über R und jedes rationale Element läßt sich in einen Kettenbruch entwickeln.*

Definition 3.5 *Die Menge aller rationalen Elemente über dem kommutativen*[63] *Ring R mit den üblichen Rechenoperationen für Addition, Subtraktion, Multiplikation und Division bezeichnen*[64] *wir mit*

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

Die rationalen Zahlen wären dann also $\mathbb{Q} = \mathbb{Z}^\star$ und $R^\star$ ist ein Körper, wenn R ein Integritätsring mit Einselement ist, siehe [11].

Beweis: Daß endliche Kettenbrüche rational über R sind erhält man wie gehabt durch Ausmultipizieren oder durch induktive Anwendung der Rekursion

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

Sei umgekehrt $f = p/q$, $p, q \in R$, $q \neq 0$. Wir setzen $s_0 = p$, $s_1 = q$ und werfen wieder den euklidischen Algorithmus an; dabei bestimmen wir r_0 so, daß $s_0 = r_0\, s_1 + s_2$, $d\,(s_2) < d\,(s_1)$, ist. Das ist möglich, weil wir es ja mit einem *euklidischen* Ring zu tun haben. Für $j = 1, 2, \ldots$ bilden wir dann wieder

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

[62] Euklidisch ist sie dadurch, daß der Divisionsrest bei Division in $\{-1, 0, 1\}$ gewählt wird.

[63] Mit nichtkommutativen Ringen wollen wir uns hier nicht beschäftigen. Abwegig sind diese aber nicht, man denke nur an Matrizen!

[64] Diese Notation ist in keinster Weise Standard, aber irgendwie müssen wir die rationalen Elemente ja schreiben.

und sehen per Induktion über k, daß

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

In der Tat ist

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

und wegen

$$r_k + \frac{s_{k+2}}{s_{k+1}} = \frac{r_k\, s_{k+1} + s_{k+2}}{s_{k+1}} = \frac{s_k}{s_{k+1}}$$

auch

$$\left[r_0; r_1, \ldots, r_k, \frac{s_{k+1}}{s_{k+2}}\right] = \left[r_0; r_1, \ldots, r_k + \frac{s_{k+2}}{s_{k+1}}\right] = \left[r_0; r_1, \ldots, r_{k-1}, \frac{s_k}{s_{k+1}}\right] = \frac{p}{q}.$$

Da aber $d\,(s_k)$ eine monoton fallende Folge in $\mathbb{N}_0$ ist, muß dieser Vorgang nach endlich vielen Schritten terminieren und uns so den endlichen Kettenbruch liefern. □

Na gut, so überraschend war das jetzt nicht – schließlich wurden ja die Zutaten, euklidischer Ring und euklidischer Algorithmus, zu diesem Eintopf so gewählt, daß sie gut zusammenpassen. Aber es wird noch besser! Wenn wir annehmen, daß R ein kommutativer Ring *mit Einselement* 1 ist[65], dann können wir den Beweis der Rekursionsformel aus Satz 2.4 einfach abschreiben und erhalten sofort eine ganze Menge von interessanten Formeln für die Konvergenten oder, wie sie in [22, 23] heißen, *Näherungsbrüche*.

Satz 3.6 *Die Konvergenten $\kappa_k := p_k/q_k$, $k \le n$, des Kettenbruchs $[r_0; r_1, \ldots, r_n]$, $r_j \in R$, erfüllen die Rekursionsformeln*[66]

$$\begin{array}{rcl} p_k &=& r_k p_{k-1} + p_{k-2} \\ q_k &=& r_k q_{k-1} + q_{k-2} \end{array}, \qquad \begin{array}{ll} p_{-1} = 1, & p_0 = r_0, \\ q_{-1} = 0, & q_0 = 1, \end{array} \tag{3.2}$$

sowie

$$\frac{p_{k-1}}{q_{k-1}} - \frac{p_k}{q_k} = \frac{(-1)^k}{q_{k-1}\, q_k}, \qquad \frac{p_k}{q_k} - \frac{p_{k-2}}{q_{k-2}} = \frac{(-1)^k\, r_k}{q_{k-2}\, q_k}, \tag{3.3}$$

und sind damit teilerfremd[67].

[65] Das sind in gewissem Sinne unsere "Minimalringe", diejenigen, die wenigstens die Elemente 0 und 1 – und zwar in dem Sinn, in dem man sie normalerweise versteht – enthalten.

[66] Daß wir $\kappa_{-1} = 0/1 = 0$ setzen, ist reine Geschmackssache.

[67] Teilerfremd in einem allgemeinen Ring mit Einselement heißt, daß $p \in q \cdot R^*$ ist, wobei $R^\times = \left\{r \in R : r^{-1} \in R\right\}$ die *Einheiten* in R bezeichnet. Bei Polynomen $\mathbb{K}[x]$ ist beispielsweise $R^\times = \mathbb{K}[x]^\times = \mathbb{K}$ und nicht alle Einheiten sind Einsen, noch nicht einmal im Betrag!

Jetzt aber liefern uns die Kettenbrüche sogar ein klein wenig mehr[68]! Wenn im Beweis von Satz 3.4 die Reduktionskette abbricht, also $s_{n+2} = 0$ ist, dann ist ja[69] $r_n = \text{ggT}\,(p, q)$ und, wegen der Teilerfremdheit der Terme im Zwischenbruch, $p_n = p/r_n$, $q_n = q/r_n$, mit

$$\frac{p}{q} = [r_0; r_1, \ldots, r_n] = \frac{p_n}{q_n} = \frac{r_n\, p_{n-1} + p_{n-2}}{r_n\, q_{n-1} + q_{n-2}},$$

und, unter Verwendung von (3.3),

$$q_{n-1}p - p_{n-1}q = r_n\,(q_{n-1}\,p_n - p_{n-1}q_n) = p_{n-2}\,q_n - p_n\,q_{n-2} = (-1)^{n+1}\,r_n = (-1)^{n+1}\text{ggT}\,(p, q)$$

liefert. Zähler und Nenner der vorletzten Konvergente, das heißt, der letzten "echten" Konvergente liefern also automatisch die Lösung der *Bézout–Identität*

$$a\,p + b\,q = \text{ggT}\;(p, q) \qquad \Leftrightarrow \qquad a = (-1)^{n+1}q_{n-1}, \quad b = (-1)^n p_{n-1}. \tag{3.4}$$

3.3 Ein Satz von einem Bernoulli

Eine Frage, die bei Kettenbrüchen ziemlich schnell und natürlich auftaucht, ist welche rationalen Objekte eigentlich als Konvergenten auftreten können. Genauer:

> *Für welche Folgen $c_n \in R^\star$, $n \in \mathbb{N}_0$, gibt es Kettenbrüche, die diese Folge als Konvergente haben?*

Laut [23] wurde diese Frage bereits 1775 von D. Bernoulli[70] in [2] beantwortet, und zwar für Kettenbrüche der Form

$$r_0 + \frac{s_1|}{|r_1} + \frac{s_2|}{|r_2} + \cdots + \frac{s_n|}{|r_n}, \qquad r_j, s_j \in R^\star \setminus \{0\}. \tag{3.5}$$

Satz 3.7 (D. Bernoulli) *Eine Folge $c_n \in R^\star$ hat genau dann eine Kettenbruchentwicklung als*

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

wenn $c_{n+1} \neq c_n$ ist, $n \in \mathbb{N}_0$. In diesem Fall ist

$$r_n = \frac{1}{q_{n-1}}\,\frac{c_n - c_{n-2}}{c_{n-2} - c_{n-1}}, \qquad s_n = \frac{1}{q_{n-2}}\,\frac{c_{n-1} - c_n}{c_{n-2} - c_{n-1}} \tag{3.6}$$

[68] Vielen Dank an dieser Stelle an H. M. Möller, der mich auf diese Tatsache hingewiesen hat.

[69] Siehe beispielsweise [8] oder [26].

[70] Daniel Bernoulli, 1700-1782, Sohn von Johann, Bruder von Nicolaus II und Neffe von Jacob Bernoulli, also mittendrin im berühmten Bernoulli–Clan. Promovierte, obwohl sein Vater ihn eigentlich als Kaufmann sehen wollte, in *Medizin*, und zwar über die Mechanik des Atmens. Neben Mathematik und Physik auch immer wieder Anwendungen der Mathematik in der Medizin.

Beweis: Der Beweis basiert auf einer Rekursionsformel für die Konvergenten

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

von Kettenbrüchen der Form (3.5), und zwar

$$\begin{array}{rcl} p_k &=& r_k\, p_{k-1} + s_k\, p_{k-2} \\ q_k &=& r_k\, q_{k-1} + s_k\, q_{k-2} \end{array}, \qquad \begin{array}{ll} p_{-1} = 1, & p_0 = r_0 \\ q_{-1} = 0, & q_0 = 1. \end{array} \tag{3.7}$$

Diese Rekursionsformel ergibt sich ganz analog zu (2.4) in Satz 2.4 per Induktion über k; der Fall $k = 0$ ist dabei die Definition von p_0 und q_0 und den Fall $k = 1$ rechnet man einfach nach:

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

Für den Induktionsschritt $k \to k+1$ setzen wir wieder

$$\frac{p'_k}{q'_k} = r_1 + \frac{s_2|}{|r_2} + \cdots + \frac{s_{k+1}|}{|r_{k+1}},$$

was uns dann sofort

$$\frac{p_{k+1}}{q_{k+1}} = r_0 + \frac{s_1}{r_1 + \dfrac{s_2|}{|r_2} + \cdots + \dfrac{s_{k+1}|}{|r_{k+1}}} = r_0 + \frac{s_1\, q'_k}{p'_k} = \frac{r_0\, p'_k + s_1\, q'_k}{p'_k}$$

liefert, und mit der "verschobenen" Induktionshypothese erhalten wir wieder, daß

$$\begin{array}{rcl} p_{k+1} &=& r_0 \left(r_{k+1} p'_{k-1} + s_{k+1}\, p'_{k-2}\right) + s_1 \left(r_{k+1} q'_{k-1} + s_{k+1}\, q'_{k-2}\right) \\ &=& r_{k+1} \left(r_0\, p'_{k-1} + s_1\, q'_{k-1}\right) + s_{k+1} \left(r_0\, p'_{k-2} + s_1\, q'_{k-2}\right) = r_{k+1}\, p_k + s_{k+1}\, p_{k-1} \\ q_{k+1} &=& p'_k = r_{k+1}\, p'_{k-1} + s_{k+1}\, p'_{k-2} = r_{k+1}\, q_k + s_{k+1}\, q_{k-1}, \end{array}$$

womit (3.7) bewiesen ist. Multiplizieren wir die erste Zeile wieder mit $-q_{k-1}$, die zweite mit p_{k-1} und addieren wir das Ganze, dann erhalten wir, daß

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

also

$$p_{k-1}\, q_k - p_k\, q_{k-1} = (-1)^k \prod_{j=1}^{k} s_j. \tag{3.8}$$

Daraus folgt dann auch schon eine Richtung unseres Satzes: Ist c_n, $n \in \mathbb{N}$, eine Folge von Konvergenten, dann ist

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

da alle $s_j \neq 0$ angenommen wurden[71].

Für die Umkehrung setzen wir mit Hilfe der Rekursionsformel (3.7)

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

an, was eindutig nach r_n, s_n auflösbar ist, da

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

ist – und zwar nach Voraussetzung für die c_k bzw. Induktion für die q_k, $k = n-1, n-2$.

Nach der Cramerschen Regeln sind nun

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

Ersetzen wir nun r_n, s_n durch $r'_n = a\, r_n$, $s'_n = a\, s_n$ für ein beliebiges $a \in R \setminus \{0\}$, dann ist natürlich immer noch

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

was uns mit durch $a = 1/q_n$ auch schon(3.6) liefert. □

Die letzte Bemerkung in diesem Beweis bringt uns dann auch wieder zurück zu unseren "normierten" Kettenbrüchen $[r_0; r_1, \ldots, r_n]$, bei denen ja $s_1 = \cdots = s_n = 1$ ist. Denn wenn wir in dem "Divisionsargument" $a = 1/s_n$ setzen, dann erhalten wir ja

$$r'_n = \frac{q_{n-2}}{q_{n-1}} \frac{c_n - c_{n-2}}{c_{n-1} - c_n}, \qquad s'_n = 1$$

und somit die Entwicklung in "unserer" Form von Kettenbrüchen.

[71] Man sieht sofort aus der Definition (3.5), daß ein Kettenbruch mit $s_k = 0$, $k \leq n$, eine Kettenbruch der Länge $k - 1 < n$ ist und dann stimmen natürlich alle weiteren Konvergenten überein.

Korollar 3.8 (Bernoulli normiert) *Erfüllt die Folge $c_n in R^\star$, $n \in \mathbb{N}_0$, die Bedingung $c_n \neq c_{n-1}$, dann ist*

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

wobei

$$r_n = \frac{q_{n-2}}{q_{n-1}} \frac{c_n - c_{n-2}}{c_{n-1} - c_n}, \qquad n \geq 2, \qquad r_{-1} = 0,\ r_0 = c_0,\ r_1 = \frac{1}{c_1 - c_0}. \tag{3.9}$$

Beweis: Wir können (3.9) auch direkt, nämlich aus (3.3) bekommen, indem wir nach geigneten Termen auflösen:

$$c_{n-1} - c_n = \frac{(-1)^n}{q_{n-1}\, q_n} \qquad \Rightarrow \qquad q_n = \frac{(-1)^n}{q_{n-1}\ (c_{n-1} - c_n)}; \tag{3.10}$$

nochmals Auflösen und Einsetzen von (3.10) führt dann auch schon zu

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

und das ist ja (3.9). □

Bemerkung 3.9 (Kettenbruchentwicklungen)

1. *Die obige Beobachtung zeigt, daß n $R^\star$ die allgemeine Kettenbruchentwicklung (3.5)* nicht *mehr eindeutig ist. Das führt zum Begriff der* äquivalenten *Kettenbrüche: Zwei Kettenbrüche heißen* äquivalent*, wenn alle ihre Konvergenten übereinstimmen.*

2. *Die Kettenbruchentwicklung aus Korollar 3.8, also die mit $s_n = 1$, $n \in \mathbb{N}$, spielt in ihrer äquivalenten Familie*[72] *von Kettenbruchentwicklungen eine besondere Rolle! Sie sind nämlich diejenigen, bei denen die Komponenten der Konvergente irreduzibel sind, die Konvergente also in* gekürzter *Form vorliegt, wie man wieder aus (3.3) sieht – das Argument ist genau dasselbe wie bei Satz 2.17.*

3. *Bei allgemeinen Kettenbruchentwicklungen sind gemeinsame Teiler von Zähler und Nenner hingegen nicht auszuschließen, siehe (3.8).*

Mit Hilfe des Satzes von Bernoulli können wir nun beispielsweise Kettenbruchentwicklungen von (Potenz-)Reihen berechnen. Sehen wir uns doch einmal anhand eines Beispiels an, wie das läuft.

Beispiel 3.10 *Die Funktion $f(x) = e^x$ hat die Potenzreihenentwicklung*

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

[72] Man könnte auch Äquivalenzklassen äquivalenter Kettenbrüche bilden, aber so nötig haben wir's nun auch wieder nicht.

und wir können die Kettenbruchentwicklung der Partialsummen

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

bestimmen – das klappt nach Korollar 3.8, da $c_n - c_{n-1} = \frac{x^n}{n!} \neq 0$ *ist*[73]*. Die ersten beiden Werte,* $r_0 = 1$*,* $r_1 = 1/x$ *und somit auch*[74] $q_0 = 1$*,* $q_1 = 1/x$ *liefern uns zusammen mit*

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

die Werte

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

und so weiter und so fort. Ein bißchen schöner wird das Ganze, wenn man $e^{1/x}$ *entwickelt, also* x *durch* x^{-1} *ersetzt.*

3.4 Orthogonale Polynome, Kettenbrüche und Gauß

In diesem Kapitel werden wir uns einmal den engen Bezug zwischen Kettenbrüchen und orthogonalen Polynomen ansehen, ein Bezug, der im wesentlichen auf der Drei–Term–Rekursionsformel (3.2) beruht, und der bereits von Gauß selbst [9] zur Herleitung der in der Numerik so beliebten *Gauß–Quadratur*, siehe z.B. [10, 15, 25], verwendet wurde. Das zweite Hilfsmittel bei Gauß' Beweis war die Entwicklung einer bestimmten Reihe nach Kettenbrüchen, und zwar mittels des Satzes von Bernoulli, also genau so, wie wir es im vorherigen Kapitel gesehen haben.

Wir werden jetzt deutlich spezifischer als in den vorherigen Kapiteln und betrachten den Ring $R = \Pi = \mathbb{R}[x]$ der (univariaten) Polynome mit reellen Koeffizienten, sowie die Vektorräume

$$\Pi_n = \text{span}\,\{1, x, \ldots, x^n\} = \{f \in \Pi \,:\, \deg f \leq n\}$$

[73] Hier bedeutet "$\neq 0$", daß ein Polynom nicht das *Nullpolynom* ist.

[74] Nicht vergessen: Wenn die r_j rational sind, dann sind es auch die Zähler und Nenner der Konvergenten!

der Polynome vom Grad $\leq n$, $n \in \mathbb{N}$. Was wir noch brauchen ist ein Skalarprodukt

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

also eine symmetrische[75] positive[76] Bilinearform[77], um den Begriff der Orthogonalität einzuführen. Dieses Skalarprodukt soll von einem *strikt quadratpositiven linearen Funktional* herrühren, also $\langle f, g \rangle = L\,(fg)$ sein, wobei

$$L \ : \ \Pi \to \mathbb{R}, \qquad L\left(f^2\right) > 0, \quad f \in \Pi. \tag{3.11}$$

Definition 3.11 *Die* Momentenfolge μ_n, $n \in \mathbb{N}$, *zu einem inneren Produkt* $\langle \cdot, \cdot \rangle$ *ist definiert als die "Integrale" der Monome:*

$$\mu_n = \langle 1, (\cdot)^n \rangle, \qquad n \in \mathbb{N}. \tag{3.12}$$

Die Momentenmatrix *ist die symmetrische positiv definite doppeltunendliche Matrix*

$$M = \left[\left\langle (\cdot)^j, (\cdot)^k \right\rangle \ : \ j, k \in \mathbb{N} \right] = \left[\mu_{j+k} \ : \ j, k \in \mathbb{N} \right]. \tag{3.13}$$

Eine Matrix A der Form $a_{j,k} = a_{j+k}$ bezeichnet man auch als Hankelmatrix *zur Folge $a = (a_n \ : \ n \in \mathbb{N})$.*

Auf Π sind innere Produkte, die durch quadratpositive Funktionale gegeben sind, und Momentenmatrizen äquivalent. Natürlich definiert jedes innere Produkt eine Momentenmatrix, und umgekehrt setzen wir für

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

ganz einfach

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

und die Hankelstruktur sorgt dafür, daß $\langle f, g \rangle = L(fg)$ ist.

Ein Folge $f_n \in \Pi_n \setminus \{0\}$, $n \in \mathbb{N}$, von Polynomen heißt *orthogonale Polynomfolge*, wenn

$$\langle f_n, \Pi_{n-1} \rangle = 0, \qquad \text{d.h.} \qquad \langle f_n, f \rangle = 0, \quad f \in \Pi_{n-1}. \tag{3.14}$$

Das Polynom f_n bezeichnet man als *orthogonales Polynom* der Ordnung n – diese Polynome sind bis auf Normierung eindeutig. Was orthogonale Polynome nun mit Kettenbrüchen gemeinsam haben ist die Tatsache, daß die ein Drei–Term–Rekursionsformel erfüllen.

[75] Das heißt $\langle f, g \rangle = \langle g, f \rangle$.
[76] Also $\langle f, f \rangle > 0$ für $f \neq 0$.
[77] Das Skalarprodukt ist linear in jeder der beiden Komponenten.

Satz 3.12 *Eine Polynomfolge f_n, $n \in \mathbb{N}$, ist genau dann eine orthogonale Polynomfolge, wenn es reelle Koeffizienten $\alpha_n > 0$, $\beta_n \in \mathbb{R}$ und $\gamma_n > 0$, $n \in \mathbb{N}$, gibt, so daß*

$$f_n = (\alpha_n x + \beta_n)\, f_{n-1} - \gamma_n\, f_{n-2}, \qquad n \in \mathbb{N}, \qquad f_0 = 1, \qquad f_{-1} = 0, \tag{3.15}$$

gilt.

Bemerkung 3.13 *Die Forderung $\alpha_n, \gamma_n > 0$ könnte man auch zu $\alpha_n\, \gamma_n > 0$ "aufweichen", denn wenn beide negativ sind, dann passiert auch nicht allzuviel.*

Beweis: Sei f_n, $n \in \mathbb{N}$, eine orthogonale Polynomfolge. Wir werden per Induktion über n zeigen, daß das Polynom

$$g_{n+1}(x) = x\, f_n(x) - \underbrace{\frac{\langle x f_n, f_n\rangle}{\langle f_n, f_n\rangle}}_{=:\beta_n'} f_n - \underbrace{\frac{\sqrt{\langle g_n, g_n\rangle\, \langle f_n, f_n\rangle}}{\langle f_{n-1}, f_{n-1}\rangle}}_{=:\gamma_n' > 0} f_{n-1}(x), \qquad x \in \mathbb{R}, \tag{3.16}$$

von Null verschieden und orthogonal zu Π_n ist und damit ein nichttriviales Vielfaches von f_n sein muß. Für $n = 0$ erhalten wir so, daß

$$g_1(x) = x\, f_0(x) - \langle x, 1\rangle f_0 \qquad \Rightarrow \qquad \langle g_1, f_0\rangle = \langle g_1, 1\rangle = \langle x, 1\rangle - \langle x, 1\rangle = 0,$$

und für allgemeines n stellen wir zuerst einmal fest, daß für $f \in \Pi_{n-2}$

$$\langle g_{n+1}, f\rangle = \langle f_n, xf\rangle - \beta_n'\, \langle f_n, f\rangle - \gamma_n' \langle f_{n-1}, f\rangle = 0$$

ist. Unter Verwendung unserer Induktionshypothese erhalten wir außerdem, daß $g_n = \lambda_n f_n$ mit[78]

$$\langle g_n, g_n\rangle = \lambda_n^2\, \langle f_n, f_n\rangle \qquad \Rightarrow \qquad \lambda_n = \sqrt{\frac{\langle g_n, g_n\rangle}{\langle f_n, f_n\rangle}}$$

und landen so bei

$$\begin{aligned}
\langle g_{n+1}, f_{n-1}\rangle &= \langle x\ f_n, f_{n-1}\rangle - \beta_n'\ \langle f_n, f_{n-1}\rangle - \gamma_n' \langle f_{n-1}, f_{n-1}\rangle \\
&= \langle f_n, x\ f_{n-1}\rangle - \gamma_n'\, \langle f_{n-1}, f_{n-1}\rangle = \big\langle f_n, g_n + \beta_{n-1}' f_{n-1} + \gamma_{n-1}' f_{n-2}\big\rangle - \gamma_n'\, \langle f_{n-1}, f_{n-1}\rangle \\
&= \sqrt{\frac{\langle g_n, g_n\rangle}{\langle f_n, f_n\rangle}} \langle f_n, f_n\rangle - \gamma_n' \langle f_{n-1}, f_{n-1}\rangle = 0,
\end{aligned}$$

und

$$\langle g_{n+1}, f_n\rangle = \langle x f_n, f_n\rangle - \beta_n'\, \langle f_n, f_n\rangle - \gamma_n'\, \langle f_n, f_{n-1}\rangle = 0.$$

Damit ist (3.16) bewiesen und die Koeffizienten sind von der Form

$$\alpha_n \in \mathbb{R}_+, \qquad \beta_n = -\alpha_n\, \beta_n', \qquad \gamma_n = \alpha_n\, \gamma_n',$$

[78] Wir wählen hier gleich die positive Wurzel, $-\lambda_n$ täte es natürlich ganz genauso!

wobei $\alpha_n > 0$ ein freier Normierungsparameter ist.

Sei nun umgekehrt f_n eine Folge von Polynomen, die die Rekursionsformel (3.15) erfüllt und setzen wir der Einfachheit halber $\alpha_n = 1$, so daß wir ein Folge *monischer* Polynome $f_n(x) = x^n + \widetilde{f}_n(x)$ erhalten. Nehmen wir außerdem induktiv an, wir hätten das Skalarprodukt schon auf $\Pi_{n-1} \times \Pi_{n-1}$ bestimmt, das heißt, wir kennen die Momente $\mu_0, \ldots, \mu_{2n-2}$. Nun betrachten wir

$$f_n(x) = x\, f_{n-1}(x) + \beta_n\, f_{n-1}(x) - \gamma_n\, f_{n-2}(x)$$

und bemerken, daß für $f \in \Pi_{n-3}$ das Skalarprodukt mit f_n bereits definiert ist, und zwar

$$\langle f_n, f\rangle := \langle f_{n-1}, x\, f\rangle + \beta_n \langle f_{n-1}, f\rangle - \gamma_n \langle f_{n-2}, f\rangle.$$

Auf der anderen Seite liefern die zusätzlichen Orthogonalitätsbedingungen

$$\begin{aligned} 0 &= \left\langle f_n, x^{n-2}\right\rangle = \left\langle f_{n-1}, x^{n-1}\right\rangle - \gamma_n \left\langle f_{n-2}, x^{n-2}\right\rangle \\ &= \mu_{2n-2} + \left\langle \widetilde{f}_{n-1}, x^{n-1}\right\rangle - \gamma_n \left\langle f_{n-2}, x^{n-2}\right\rangle \\ &=: \mu_{2n-2} + \sum_{j=0}^{2n-3} a_{n,j}\, \mu_j \end{aligned} \tag{3.17}$$

und

$$\begin{aligned} 0 &= \left\langle f_n, x^{n-1}\right\rangle = \langle f_{n-1}, x^n\rangle + \beta_n \left\langle f_{n-1}, x^{n-1}\right\rangle + \gamma_n \left\langle f_{n-2}, x^{n-1}\right\rangle \\ &= \mu_{2n-1} + \beta_n \mu_{2n-2} + \sum_{j=0}^{2n-3} b_{n,j}\, \mu_j, \end{aligned}$$

wodurch die Momente und damit auch das innere Produkt eindeutig bis auf Normalisierung, das heißt, bis auf die Wahl von $\mu_0 > 0$, festgelegt sind:

$$\begin{aligned} \mu_1 &= -\beta_1 \mu_0 \\ \mu_2 &= -a_{2,0}\, \mu_0 - a_{2,1}\, \mu_1 \\ \mu_3 &= -\beta_2 \mu_2 - b_{2,0}\, \mu_0 - b_{2,1}\, \mu_1 \\ &\vdots \\ \mu_{2n-2} &= -\sum_{j=0}^{2n-3} a_{n,j}\, \mu_j \\ \mu_{2n-1} &= -\beta_n \mu_{2n-2} - \sum_{j=0}^{2n-3} b_{n,j}\, \mu_j. \end{aligned}$$

Schließlich ist

$$\langle f_n, f_n\rangle = \langle f_n, x f_{n-1}\rangle = \langle f_n, x^n\rangle = \mu_n + \left\langle \widetilde{f}_n, x^n\right\rangle \tag{3.18}$$

Wir zeigen per Induktion über n, daß dieser Ausdruck positiv sein muß und daß damit unser inneres Produkt auch wohldefiniert, also positiv definit ist. Der Fall $n = 0$ ist dabei gerade

$\mu_0 > 0$, und der Rest der Induktion ist fast geschenkt: Ersetzt man n in (3.17) durch $n+1$, dann liefert die Induktion sofort, daß

$$\langle f_n, f_n\rangle = \langle f_n, x^n\rangle = \mu_n + \left\langle \widetilde{f}_n, x^n \right\rangle = \gamma_n \left\langle f_{n-1}, x^{n-1}\right\rangle = \gamma_n \langle f_{n-1}, f_{n-1}\rangle > 0, \qquad (3.19)$$

also ist die Bilinearform mit den oben definierten Monomen strikt positiv und damit ein Skalarprodukt. □

Bemerkung 3.14 *Mit etwas scharfem Hinschauen sieht man in (3.19) sogar eine explizite Formel für $\langle f_n, f_n\rangle$, nämlich*

$$\langle f_n, f_n\rangle = \gamma_n \langle f_{n-1}, f_{n-1}\rangle = \gamma_n\gamma_{n-1} \langle f_{n-2}, f_{n-2}\rangle = \cdots = \left(\prod_{j=1}^{n} \gamma_j\right) \langle f_0, f_0\rangle = \mu_0 \prod_{j=1}^{n} \gamma_j.$$

Und so können wir orthogonale Polynome *immer* als Konvergenten von Kettenbrüchen erhalten – der Beweis besteht lediglich aus dem Vergleich der Drei–Term–Rekursion für orthogonale Polynome mit der Rekursionsformel für die Konvergenten.

Korollar 3.15 *Die orthogonalen Polynome mit den Parametern $\alpha_n, \beta_n, \gamma_n$ in der Rekursionsformel erhält man als Nenner der Konvergenten zu den Kettenbrüchen*

$$\frac{-\gamma_1|}{|(\alpha_1 x + \beta_1)} - \frac{\gamma_2|}{|(\alpha_2 x + \beta_2)} - \frac{\gamma_3|}{|(\alpha_3 x + \beta_3)} + \cdots$$

beziehungweise

$$\left[0; -\frac{\alpha_1\, x + \beta_1}{\gamma_1}, -\frac{\alpha_2\, x + \beta_2}{\gamma_2}, \dots\right].$$

Umgekehrt sind die Nenner der Konvergenten aller Kettenbrüche der Form

$$\left[0; -\alpha_1 x + \beta_1, -\alpha_2 x + \beta_2, \dots\right], \qquad \alpha_j > 0,\ \beta \in \mathbb{R},$$

ein orthogonales Polynomsystem für ein passendes Skalarprodukt $\langle\cdot,\cdot\rangle$.

Wir haben jetzt also festgestellt, daß jede orthogonale Polynomfolge zu einem strikt quadratpositiven Funktional als Nenner der Konvergenten einer Folge von Kettenbrüchen geschrieben werden kann – aber wozu gehören denn dann diese Kettenbrüche? Dazu betrachten wir *Laurentreihen*.

Definition 3.16 (Laurenteihen und Konvergenz)

1. *Die* Laurentreihe $\lambda(x)$ *zu einer Folge $(\lambda_j \,:\, j \in \mathbb{N}_0)$ ist definiert als*

$$\lambda(x) = \sum_{j=0}^{\infty} \lambda_j\, x^{-j}.$$

2. *Eine Folge $\lambda_n(x)$ von Laurentreihen* konvergiert *gegen $\lambda^*(x)$, wenn es für alle $k \in \mathbb{N}_0$ ein $n_0 \in \mathbb{N}$ gibt, so daß für alle $n \geq n_0$*

$$\lambda_n(x) - \lambda^* x = x^{-k-1}\,\widetilde{\lambda}_n(x), \qquad \textit{d.h.} \qquad \lambda_{n,j} = \lambda_j^*, \quad j = 0, \ldots, k-1.$$

Mit anderen Worten: Konvergenz bedeutet, daß ab einem bestimmten Punkt die ersten k Terme der Laurentreihen mit den ersten k Termen der Grenzreihe übereinstimmen, und das für alle $k \in \mathbb{N}_0$. Dabei ist zu beachten, daß wir hier mit *formalen* Potenzreihen arbeiten, und also nicht dafür interessieren, ob die Potenzreihe im normalen analytischen Sinn an einem Punkt konvergiert und damit dort eine analytische Funktion darstellt.

Die erste, sehr einfache, aber erstaunlich fundamentale Beobachtung ist, daß wir jedes reziproke Polynom als Laurentreihe entwickeln können, wobei ein wesentlicher Teil der Koeffizienten Null ist.

Lemma 3.17 *Ist $p \in \Pi_n$ mit $p_n \neq 0$, dann ist*

$$\frac{1}{p(x)} = \sum_{j=n}^{\infty} \lambda_j\, x^{-j} =: \lambda(x).$$

Beweis: Wir schreiben $p(x) = p_0 + p_1\, x + \cdots + p_n\, x^n$ und setzen $1/p(x) = \lambda(x)$ an, also

$$\begin{aligned} 1 &= p(x)\,\lambda(x) = \left(\sum_{j=0}^{n} p_j\, x^j\right)\left(\sum_{k=0}^{\infty} \lambda_j\, x^{-k}\right) = \sum_{j=0}^{n}\sum_{k=0}^{\infty} p_j\,\lambda_k\, x^{j-k} \\ &= \sum_{j=-\infty}^{n} x^j \sum_{k-\ell=j} p_k\,\lambda_\ell = \sum_{j=-\infty}^{n} x^j \sum_{\ell=-j}^{n-j} p_{j+\ell}\lambda_\ell, \end{aligned}$$

wobei $\lambda_{-n} = \cdots = \lambda_{-1} = 0$. Durch Koeffizientenvergleich erhalten wir, daß

$$\sum_{k=-j}^{n-j} p_{j+k}\lambda_k = \delta_{j,0} = \begin{cases} 0, & j \neq 0, \\ 1, & j = 0, \end{cases}$$

also insbesondere

$$\begin{aligned} 0 &= p_n\,\lambda_0 \\ 0 &= p_{n-1}\,\lambda_0 + p_n\,\lambda_1 \\ &\vdots \\ 0 &= p_1\,\lambda_0 + \cdots + p_n\,\lambda_{n-1}, \end{aligned}$$

was in Matrixform geschrieben und unter Verwendung von $p_n \neq 0$

$$0 = \begin{bmatrix} p_n & & \\ \vdots & \ddots & \\ p_1 & \cdots & p_n \end{bmatrix} \begin{bmatrix} \lambda_0 \\ \vdots \\ \lambda_{n-1} \end{bmatrix} \qquad \Rightarrow \qquad \lambda_0 = \cdots = \lambda_{n-1} = 0$$

liefert. Die übrigen Koeffizienten erhalten wir dann durch sukzessives Lösen des Gleichungssystems

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \end{bmatrix} = \begin{bmatrix} p_n & & & \\ \vdots & \ddots & & \\ p_0 & \cdots & p_n & \\ & \ddots & & \ddots \end{bmatrix} \begin{bmatrix} \lambda_n \\ \lambda_{n+1} \\ \vdots \end{bmatrix},$$

was $\lambda_n, \lambda_{n+1}, \ldots$ *eindeutig* bestimmt. □

Definition 3.18 *Ein undendlicher Kettenbruch* $[0; r_1, r_2, \ldots]$, $r_j \in \Pi \setminus \Pi_0$ *heißt* konvergent, *wenn es eine Laurentreihe* $\lambda(x)$ *gibt, so daß*

$$\lim_{n\to\infty} \frac{p_n(x)}{q_n(x)} = \lambda(x)$$

ist.

Bemerkung 3.19 (Konvergenz von Kettenbrüchen)

1. *Definition 3.18 macht Sinn*[79]*! Da* $p_0 = 0$ *und* $p_1 = 1$ *ist, ist auch der* $\deg q_n > \deg p_n$ *und damit ist nach Lemma 3.17*

$$\frac{p_n(x)}{q_n(x)} = p_n(x) \sum_{j=\deg q_n}^{\infty} \lambda_j\, x^{-j} = \sum_{j=\deg q_n - \deg p_n}^{\infty} \widetilde{\lambda}_j\, x^{-j}$$

   auch jede Konvergente als Laurentreihe darstellbar.

2. *Man könnte die rationalen Funktionen auch nach positiven Potenzen von* x *entwickeln – das gibt dann die Taylorreihe der Funktion. Allerdings braucht man dann etwas andere Kettenbrüche, siehe [23], bei denen auch etwas im Zähler steht.*

Und tatsächlich gibt es jede Menge von konvergenten Kettenbrüchen, nämlich insbesondere die, die wir bereits von den orthogonalen Polynomen her kennen.

Satz 3.20 *Jeder Kettenbruch der Form* $[0; r_1, \ldots]$, $r_j \in \Pi$, $\deg r_j \geq 1$, $j \in \mathbb{N}$, *konvergiert gegen eine Laurentreihe* $\lambda(x)$ *und dabei ist*

$$\lambda(x) - \frac{p_n(x)}{q_n(x)} = O\left(x^{-d_{n+1}-d_n}\right) \tag{3.20}$$

das heißt,

$$\frac{p_n(x)}{q_n(x)} = \lambda_0 + \cdots + \lambda_{d_{n+1}+d_n-1}\, x^{-d_{n+1}-d_n+1} + \cdots \tag{3.21}$$

wobei $d_n := \deg q_n$, $n \in \mathbb{N}_0$.

[79] Was hätte man auch sonst erwarten sollen.

Beweis: In der formalen Laurentreihe

$$\lambda(x) - \frac{p_n(x)}{q_n(x)} = \sum_{j=n}^{\infty} \left(\frac{p_{j+1}(x)}{q_{j+1}(x)} - \frac{p_j(x)}{q_j(x)} \right) = \sum_{j=n}^{\infty} \frac{(-1)^j}{q_{j+1}(x)\, q_j(x)} = \sum_{j=d_{n+1}+d_n}^{\infty} \gamma_j\, x^{-j} =: \gamma(x)$$

sind alle Koeffizienten γ_j wohldefiniert, da γ_j ja nur von endlich vielen q_k abhängt. Damit ist aber Konvergenz klar, da

$$\frac{p_{n+k}(x)}{q_{n+k}(x)} - \frac{p_n(x)}{q_n(x)} = O\left(x^{-d_n - d_{n+1}}\right), \qquad k \in \mathbb{N},$$

und wir somit eine "Cauchyfolge" von Laurentreihen haben. Das überträgt sich auch auf die "Grenzreihe" $\lambda(x)$, was uns auch schon (3.20) liefert. □

Zurück zu unseren orthogonalen Polynomen heißt das dann insbesondere, daß Kettenbrüche mit *linearen* Koeffizienten immer konvergieren und zwar von einer sehr einfachen Ordnung.

Korollar 3.21 *Jeder Kettenbruch der Form* $[0; r_1, \ldots]$, $r_j \in \Pi_1 \setminus \Pi_0$, $j \in \mathbb{N}$, *konvergiert gegen eine Laurentreihe* $\lambda(x)$ *wobei*

$$\lambda(x) - \frac{p_n(x)}{q_n(x)} = O\left(x^{-2n-1}\right). \tag{3.22}$$

Diese schnell konvergierenden Kettenbrüche passen also besonders gut zur Laurentreihe λ, daher wollen wir sie uns mal etwas genauer ansehen. Man könnte die Theorie sogar wesentlich allgemeiner, also für relativ beliebige Kettenbrüche mit $r_j \in \Pi \setminus \Pi_0$, entwickeln[80], aber wir wollen uns jetzt auf Kettenbrüche mit *linearen* Faktoren beschränken, also $r_j(x) = \alpha_j x + \beta_j$, $\alpha_j \neq 0$, bei denen $\deg q_n = \deg p_n + 1 = n$ ist. Und dann geben wir den schnell konvergenten unter ihnen einen besonderen Namen.

Definition 3.22 *Der unendliche Kettenbruch* $[0; r_1, \ldots]$, $r_j \in \Pi_1 \setminus \Pi_0$ *heißt* assoziiert *zur Laurentreihe* $\lambda(x)$, *wenn*

$$\lambda(x) - \frac{p_n(x)}{q_n(x)} = O\left(x^{-2n-1}\right), \qquad n \in \mathbb{N},$$

also

$$\frac{p_n(x)}{q_n(x)} = \sum_{j=0}^{2n} \lambda_j\, x^{-j} + \sum_{j=2n+1}^{\infty} \gamma_{n,j}\, x^{-j}, \qquad n \in \mathbb{N}. \tag{3.23}$$

Es wäre ziemlich vermessen, wenn wir uns einbilden würden, daß *alle* Laurentreihen einen assoziierten Kettenbruch hätten, aber es wird sich herausstellen, daß eine Beschreibung der Laurentreihen, für die das gilt, sogar noch interessanter ist, denn da tauchen wieder die Hankelmatrizen auf, die wir ja bereits aus Definition 3.11, Gleichung (3.13) kennen[81]

[80] Was man nicht wirklich beweist, das kann man natürlich besonders einfach behaupten.

[81] Oder kennen sollten.

Satz 3.23 *Eine Laurentreihe $\lambda(x)$ besitzt genau dann einen assoziierten Kettenbruch $[0; r_1, \ldots]$, $r_j \in \Pi_1 \setminus \Pi_0$, wenn $\lambda_0 = 0$ und*

$$\det \Lambda_n \neq 0, \qquad \Lambda_n = \begin{bmatrix} \lambda_1 & \lambda_2 & \ldots & \lambda_n \\ \lambda_2 & \lambda_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_{2n-2} \\ \lambda_n & \ldots & \lambda_{2n-2} & \lambda_{2n-1} \end{bmatrix}, \qquad n \in \mathbb{N}. \tag{3.24}$$

Beweis: Der Kettenbruch ist genau dann assoziiert, wenn für alle $n \in \mathbb{N}$

$$\begin{array}{rcl} \dfrac{p_n(x)}{q_n(x)} & = & \lambda_0 + \cdots + \lambda_{2n} x^{-2n} \quad + \gamma_{n,2n+1} x^{-2n-1} \quad + \cdots \\ \dfrac{p_{n+1}(x)}{q_{n+1}(x)} & = & \lambda_0 + \cdots + \lambda_{2n} x^{-2n} \quad + \lambda_{2n+1} x^{-2n-1} \quad + \cdots \end{array} \tag{3.25}$$

Subtrahiert man die zweite Gleichung in (3.25) von der ersten und bedenkt man, daß dies gleich $(-1)^n / \left(q_{n+1}\, q_n\right)$ ist, dann ergibt sich, daß

$$\frac{(-1)^n}{q_{n+1}(x)\, q_n(x)} = \left(\gamma_{n,2n+1} - \lambda_{2n+1}\right) x^{-2n-1} + \cdots$$

Nun ist, wie man leicht durch induktive Anwendung der Rekursionsformel (3.2) sieht,

$$q_n(x) = \left(\prod_{j=1}^{n} \alpha_j\right) x^n + \cdots \qquad \Rightarrow \qquad q_{n+1}(x)\, q_n(x) = \alpha_{n+1} \left(\prod_{j=1}^{n} \alpha_j\right)^2 x^{2n+1} + \cdots \tag{3.26}$$

und ein Koeffizientenvergleich liefert

$$\alpha_{n+1} = \frac{(-1)^n \left(\gamma_{n,2n+1} - \lambda_{2n+1}\right)}{\left(\alpha_1 \cdots \alpha_n\right)^2}. \tag{3.27}$$

Als nächstes multiplizieren wir die erste Zeile von (3.25) mit $q_n(x)$, was[82] zu

$$\begin{aligned} p_n(x) &= \left(\sum_{j=0}^{2n} \lambda_j\, x^{-j} + \sum_{j=2n+1}^{\infty} \gamma_{n,j}\, x^{-j}\right) \left(\sum_{k=0}^{n} q_{n,k}\, x^k\right) \\ &= \sum_{j=0}^{2n} \sum_{k=0}^{n} \lambda_j\, q_{n,k}\, x^{k-j} + \sum_{j=2n+1}^{\infty} \sum_{k=0}^{n} \gamma_{n,j}\, q_{n,k}\, x^{k-j} = \sum_{k=0}^{n} \sum_{j=k-2n}^{k} \lambda_{k-j}\, q_{n,k} x^j + O\left(x^{-n-1}\right) \\ &= \sum_{-n \le k-2n \le j \le k \le n} \lambda_{k-j}\, q_{n,k} x^j + O\left(x^{-n-1}\right) = \sum_{-n \le j \le n} \sum_{j \le k \le j+2n} \lambda_{k-j}\, q_{n,k} x^j + O\left(x^{-n-1}\right) \\ &= \sum_{j=-n}^{n} x^j \sum_{k=j}^{j+2n} \lambda_{k-j}\, q_{n,k} + O\left(x^{-n-1}\right) = \sum_{j=-n}^{n} x^{-j} \sum_{k=0}^{n} \lambda_{j+k} q_k + O\left(x^{-n-1}\right), \end{aligned}$$

[82] Unter Ausnutzung der Konvention $0 = \lambda_j = p_k$, $j, k < 0$ bzw. $k > n$.

also

$$p_n(x) = \sum_{j=0}^{n} \eta_{-j} x^j + \sum_{j=1}^{n} \eta_j x^{-j} + \eta_{n+1} x^{-n-1} + O\left(x^{-n-2}\right) \tag{3.28}$$

führt, wobei

$$\begin{bmatrix} \eta_{-n} \\ \vdots \\ \eta_0 \end{bmatrix} = \begin{bmatrix} \lambda_0 & & \\ \vdots & \ddots & \\ \lambda_n & \dots & \lambda_0 \end{bmatrix} \begin{bmatrix} q_{n,n} \\ \vdots \\ q_{n,0} \end{bmatrix} \tag{3.29}$$

und[83]

$$\begin{bmatrix} \eta_1 \\ \vdots \\ \eta_n \\ \eta_{n+1} \end{bmatrix} = \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_{n+1} \\ \lambda_2 & \lambda_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_{2n-1} \\ \lambda_n & \dots & \lambda_{2n-1} & \lambda_{2n} \\ \lambda_{n+1} & \dots & \lambda_{2n} & \gamma_{n,2n+1} \end{bmatrix} \begin{bmatrix} q_{n,0} \\ \vdots \\ q_{n,n} \end{bmatrix}. \tag{3.30}$$

Nachdem die linke Seite von (3.28) ein Polynom ist, folgt aus einem Koeffizientenvergleich, daß $\eta_1 = \dots = \eta_{n+1} = 0$ sein muß. Also hat die Matrix in (3.30) Determinante 0. Bestimmen wir nun aus (3.27)

$$\gamma_{n,2n+1} = (-1)^n \alpha_{n+1} \prod_{j=1}^{n} \alpha_j + \lambda_{2n+1}$$

und setzen dies in (3.30) ein, dann erhalten wir, daß

$$\begin{aligned} 0 &= \det \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_{n+1} \\ \lambda_2 & \lambda_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_{2n-1} \\ \lambda_n & \dots & \lambda_{2n-1} & \lambda_{2n} \\ \lambda_{n+1} & \dots & \lambda_{2n} & \gamma_{n,2n+1} \end{bmatrix} \\ &= \det \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_{n+1} \\ \lambda_2 & \lambda_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_{2n-1} \\ \lambda_{n+1} & \dots & \lambda_{2n} & \lambda_{2n+1} \end{bmatrix} + \alpha_{n+1} \prod_{j=1}^{n} \alpha_j^2 \det \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & 0 \\ \lambda_2 & \lambda_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \lambda_n & \dots & \lambda_{2n-1} & 0 \\ \lambda_{n+1} & \dots & \lambda_{2n} & (-1)^n \end{bmatrix} \\ &= \det \Lambda_{n+1} + \alpha_{n+1} \left(\prod_{j=1}^{n} \alpha_j^2 \right) \det \Lambda_n \end{aligned}$$

also

$$\alpha_{n+1} = -\left(\prod_{j=1}^{n} \alpha_j^2 \right)^{-1} \frac{\det \Lambda_{n+1}}{\det \Lambda_n}. \tag{3.31}$$

[83] Die Bildungsregel für η_{n+1} ist offensichtlich, wenn man einmal verstanden hat, wie $\eta_1, \dots, \eta_n$ gebildet werden.

Fassen wir zusammen: Ein Kettenbruch mit linearen Faktorn ist genau dann zu einer Laurentreihe assoziiert, wenn diese Faktoren[84] die Gleichung (3.31) erfüllen – und das ist per Induktion dazu äquivalent, daß $\det \Lambda_n \neq 0$, $n \in \mathbb{N}$. Die Bedingung an λ_0 is einfacher: man berücksichtigt lediglich in (3.29), daß p_n vom Grad $n-1$ ist, also $0 = \eta_{-n} = \lambda_0 \, q_{n,n}$, wohingegen q_n ein Polynom vom Grad n ist, was $q_{n,n} \neq 0$ heißt. □

Nun ist Satz 3.29 an sich schon eine schöne Sache mit einem ganz netten Beweis, aber die volle Pracht dieser Beobachtung findet man erst im nächsten Resultat, das uns orthogonale Polynome und Kettenbrüche so richtig zusammenbringt – und gleichzeit die implizite Gauß'sche Definition der orthogonalen Polynome liefert.

Satz 3.24 *Sei μ die Momentenfolge zu einem quadratpositiven linearen Funktional. Dann sind die orthgonalen Polynome zu diesem Funktional die Nenner q_n, $n \in \mathbb{N}$ der Konvergenten des Kettenbruchs, der zu*

$$\mu(x) = \sum_{j=1}^{\infty} \mu_{j-1} x^{-j}$$

assoziierten Laurentreihe.

Beweis: Die Matrizen $\Lambda_n = M_{n-1}$, $n \in \mathbb{N}$, sind strikt positiv definit und haben damit sogar strikt positive Determinante. Also existiert ein assoziierter Kettenbruch. Wegen (3.30) und dem Koeffizientenvergleich in (3.28) ist außerdem

$$\begin{aligned} 0 &= \begin{bmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_{n+1} \\ \lambda_2 & \lambda_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda_{2n-1} \\ \lambda_n & \dots & \lambda_{2n-1} & \lambda_{2n} \end{bmatrix} \begin{bmatrix} q_{n,0} \\ \vdots \\ q_{n,n} \end{bmatrix} = \begin{bmatrix} \mu_0 & \mu_1 & \dots & \mu_n \\ \mu_1 & \mu_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mu_{2n-2} \\ \mu_{n-1} & \dots & \mu_{2n-2} & \mu_{2n-1} \end{bmatrix} \begin{bmatrix} q_{n,0} \\ \vdots \\ q_{n,n} \end{bmatrix} \\ &= \begin{bmatrix} \langle 1, q \rangle \\ \vdots \\ \langle (\cdot)^{n-1} \, q \rangle \end{bmatrix}, \end{aligned}$$

was uns Orthogonalität liefert. Und nach (3.31) haben sogar die Koeffizienten α_j in der Rekursionsformel das richtige Vorzeichen für ein orthogonales Polynom. □

Kleine Bemerkung am Rande: Wir haben jetzt also sogar eine Formel für die Rekursionsformel, indem wir α_n nach (3.31) bestimmen und β_n über den Koeffizientenvektor[85] $q_n = [q_{n,j} \, : \, j = 0, \dots, n]$ von q_n

$$\begin{aligned} 0 &= \langle q_{n+1}, q_n \rangle = \alpha_{n+1} \langle (\cdot) q_n, q_n \rangle + \beta_{n+1} \langle q_n, q_n \rangle + \langle q_{n-1}, q_n \rangle \\ &= \alpha_{n+1} \, q_n^T \begin{bmatrix} \mu_1 & \dots & \mu_{n+1} \\ \vdots & \ddots & \vdots \\ \mu_{n+1} & \dots & \mu_{2n+1} \end{bmatrix} q_n + \beta_{n+1} \, q_n^T M_n q_n \end{aligned}$$

[84] Was wir ganz genau hier investiert und wieder bekommen haben, ist die Assoziiertheitseigenschaft, also die gute Approximationsrate, für die n–te und $(n+1)$–te Konvergente.

[85] Es ist eigentlich gar nicht so abwegig, dasselbe Symbol zu verwenden, schließlich werden am Rechner Polynome ja normalerweise auch als Koeffizientenvektoren gespeichert und verarbeitet.

als

$$\beta_{n+1} = -\alpha_{n+1}\,\frac{q_n^T \widetilde{M}_n q_n}{q_n^T M_n q_n}, \qquad \widetilde{M}_n = \begin{bmatrix} \mu_1 & \cdots & \mu_{n+1} \\ \vdots & \ddots & \vdots \\ \mu_{n+1} & \cdots & \mu_{2n+1} \end{bmatrix}. \tag{3.32}$$

Was hat das alles nun mit Gauß zu tun? Ganz einfach, Kettenbrüche sind der Schlüssel bei der Bestimmung der sogenannten Gauß–Quadratur [9], also der Bestimmung von Gewichten ω_j und Punkten (oder Knoten) x_j, $j = 0, \ldots, n$, so daß

$$0 = L(f) - \Omega(f) = L(f) - \sum_{j=0}^{n} \omega_j\, f\,(x_j)\,, \qquad f \in \Pi_{2n+1}, \tag{3.33}$$

wobei L wieder unser quadratpositives lineares Funktional ist. Die *Quadraturformel* Ω in (3.33) hat den maximalen *Exaktheitsgrad* $2n+1$, denn das Polynom $f(x) = (x - x_0)\cdots(x - x_n) \in \Pi_{2n+2}$ erfüllt ja

$$L\left(f^2\right) > 0 = \sum_{j=0}^{n} \omega_j\, f\,(x_j)\,,$$

also kann (3.33) für dieses f nicht mehr gelten. Zu gegebenen Punkten $x_0, \ldots, x_n$ bzw. einem gegebenen Polynom[86] $w(x) = (x - x_0)\cdots(x - x_n)$ bestimmt man die Gewichte ω_j, $j = 0, \ldots$, als

$$\omega_j = L\,(\ell_j)\,, \qquad \ell_j = \prod_{k \neq j} \frac{\cdot - x_k}{x_j - x_k} = \frac{w}{w'\,(x_j)\;(\cdot - x_j)}, \qquad j = 0, \ldots, n. \tag{3.34}$$

Übung 3.1 Beweisen Sie die Formel (3.34) für die Polynome ℓ_j. ♢

Schreiben wir nun $w(x) = w_0 + w_1\,x + \cdots + w_n x^n + x^{n+1}$, dann erhalten wir, daß[87]

$$\begin{aligned}
w'\,(x_j)\;\ell_j(x) &= \frac{w(x)}{x - x_j} = \frac{w(x) - w\,(x_j)}{x - x_j} \\
&= \frac{w_1\,(x - x_j) + \cdots + w_n\left(x^n - x_j^n\right) + \left(x^{n+1} - x_j^{n+1}\right)}{x - x_j} \\
&= \sum_{k=1}^{n+1} w_k \frac{x^k - x_j^k}{x - x_j} = \sum_{k=1}^{n+1} w_k \sum_{m=0}^{k-1} x^m\, x_j^{k-1-m} \\
&= \begin{array}{ccccccccc}
x^n & + & x_j\,x^{n-1} & + & x_j^2\,x^{n-2} & + & \cdots & + & x_j^n \\
 & + & w_n\,x^{n-1} & + & w_n\,x_j\,x^{n-2} & + & \cdots & + & w_n\,x_j^{n-1} \\
 & & & + & w_{n-1}\,x^{n-2} & + & \cdots & + & w_{n-1}\,x_j^{n-2} \\
 & & & & & & \ddots & & \vdots \\
 & & & & & & & + & w_1
\end{array}
\end{aligned}$$

[86] Diese Beziehung zwischen Punkten und Polynomen ist natürlich *nicht* 1-1, denn es geht ja immer um Polynome mit einfachen reellen Nullstellen!

[87] Und was jetzt kommt, ist *original* Gauß'sche Rechenkunst, allerdings moderner geschrieben.

$$= \sum_{k=0}^{n} x^k \frac{w(x_j)}{x_j^{k+1}} + O\left(x_j^{-1}\right),$$

also

$$\begin{aligned} w_j'(x_j)\, L(\ell_j) &= \mu_n + \mu_{n-1}(x_j + w_n) + \cdots + \mu_0\left(x_j^n + w_n x_j^{n-1} + \cdots + w_1\right) \\ &= \sum_{k=0}^{n} \mu_k \left(x_j^{n-k} + \sum_{m=k+1}^{n} w_k\, x_j^{n-k} \right) =: \widetilde{w}(x_j), \qquad \widetilde{w} \in \Pi_n, \end{aligned}$$

und somit

$$\omega_j = \frac{\widetilde{w}(x_j)}{w'(x_j)}. \tag{3.35}$$

Diese Formel ermöglicht die Berechnung der Quadraturgewichte *direkt* aus den Momenten! Nun seien $\lambda_k = \Omega\left((\cdot)^k\right)$ die Momente der Quadraturformel, $\theta_k = \mu_k - \lambda_k$ und $\lambda(x), \theta(x)$ die zugehörigen Laurentreihen. Unter Verwendung der (formalen) Identität[88]

$$\frac{1}{x-\xi} = \sum_{j=1}^{\infty} \frac{\xi^{j-1}}{x^j} \tag{3.36}$$

fällt somit auf, daß

$$\lambda(x) = \sum_{k=1}^{\infty} \frac{\Omega\left((\cdot)^{k-1}\right)}{x^k} = \sum_{k=1}^{\infty} x^{-k} \sum_{j=0}^{n} \omega_j\, x_j^{k-1} = \sum_{j=0}^{n} \omega_j \sum_{k=1}^{\infty} x_j^{k-1} x^k = \sum_{j=0}^{n} \frac{\omega_j}{x - x_j},$$

weswegen

$$\theta(x) = \mu(x) - \lambda(x) = \mu(x) - \sum_{j=0}^{n} \frac{\omega_j}{x - x_j} \tag{3.37}$$

sein muss. Nach Konstruktion ist die Quadraturformel *interpolatorisch* und damit $\theta_0 = \cdots = \theta_n = 0$ und damit ist

$$O\left(x^{-1}\right) = w(x)\,\theta(x) = w(x)\,\mu(x) - \sum_{j=0}^{n} \omega_j \frac{w(x)}{x - x_j} = w(x)\,\mu(x) - \underbrace{\sum_{j=0}^{n} \omega_j\, w'(x_j)\, \ell_j(x)}_{\in \Pi_n}.$$

Und jetzt sind wir im Geschäft – hier zaubert Gauß nun in [9] das Kettenbruchkaninchen aus dem Ärmel: ist nun $w(x) = q_{n+1}(x)$ der Nenner der $n+1$–ten Konvergente[89] von $\mu(x)$, dann ist

$$\mu(x) = \frac{p_{n+1}(x)}{q_{n+1}(x)} + O\left(x^{-2n-3}\right) \qquad \Rightarrow \qquad q_{n+1}(x)\,\mu(x) = p_{n+1}(x) + O\left(x^{-n-2}\right),$$

[88] **Beweis:** Durchmultiplizieren und Koeffizientenvergleich, was auch sonst?

[89] Und die existiert ja nach Satz 3.24

und somit

$$w(x)\,\theta(x) = \underbrace{p_{n+1}(x) - \sum_{j=0}^{n} \omega_j\, w'\left(x_j\right)\, \ell_j(x)}_{=:p(x)} + O\left(x^{-n-2}\right) = O\left(x^{-1}\right),$$

also $p = 0$ und daher

$$w(x)\,\theta(x) = O\left(x^{-n-2}\right) \qquad \Rightarrow \qquad \theta(x) = \frac{O\left(x^{-n-2}\right)}{w(x)} = O\left(x^{-2n-3}\right),$$

weswegen

$$0 = \theta_0 = \cdots = \theta_{2n+1} \tag{3.38}$$

ist – die Quadraturformel mit den Nullstellen von q_{n+1} hat den gewünschten Exaktheitsgrad. Ach ja – natürlich muß q_{n+1} auch wirklich einfache und reelle Nullstellen haben, denn sonst macht das Ganze ja keinen Spass! Glücklicherweise wird das aber von der nachfolgenden Proposition garantiert.

Proposition 3.25 *Erfüllt eine Polynomfolge eine Rekursionsformel wir in (3.15), dann hat jedes Polynom f_n einfache relle Nullstellen.*

Wir können uns leider das Leben nicht ganz einfach machen, indem wir uns zuerst auf Satz 3.12 berufen und dann darauf, daß orthogonale Polynome immer einfache Nullstellen haben, denn in diesem Beweis verwendet man normalerweise eine Integraldarstellung des inneren Produkts. Man kann das zwar umgehen, braucht aber dann immer noch die nichttriviale Aussage, daß sich jedes *positive Polynom*[90] als positive Summe von Quadraten darstellen läßt. Nachdem wir allerdings später sowieso die Sturmschen Ketten brauchen werden, können wir den kleinen Abstecher auch gleich an dieser Stelle machen.

3.5 Sturmsche Ketten

Sturmsche Ketten liefern eine Methode, die Anzahl der Nullstellen eines Polynoms in einem bestimmten Intervall zu ermitteln, und zwar indem man nur Vorzeichenwechsel einer endlichen Zahlenfolge zählt. Das macht sie zu einem beliebten Hilfsmittel in der Numerik univariater Polynome, siehe z.B. [25]. Wir halten uns hier, was die Begriffsbildung angeht aber an [7].

Definition 3.26 *Eine endliche Folge $f_0, \ldots, f_n$ von Polynomen heißt* Sturmsche Kette *für ein Intervall*[91] *I, wenn*

1. *an jeder Nullstelle von f_k die Polynome f_{k+1} und f_{k-1} echt unterschiedliches Vorzeichen haben:*

$$f_k(x) = 0 \qquad \Rightarrow \qquad f_{k-1}(x)\, f_{k+1}(x) < 0, \qquad k = 1, \ldots, n-1. \tag{3.39}$$

[90] Also jedes *reelle* Polynom, das ≥ 0 auf ganz $\mathbb{R}$ und an mindestens einer Stelle > 0 ist.

[91] Offen oder abgeschlossen, beschränkt oder unbeschränkt.

2. *das Polynom f_0 keine Nullstelle in I hat.*

Was das mit Nullstellen zu tun hat, wird ziemlich schnell klar, wenn man für $x \in \mathbb{R}$ die Anzahl $V(x)$ der *echten* Vorzeichenwechsel[92] in dem Vektor $(f_0(x), \ldots, f_n(x))$ betrachtet und x variieren läßt. Solange $f_j(x) \neq 0$, $j = 0, \ldots, n$, hat $V(x \pm \varepsilon)$ genau denselben Wert für hinreichend kleines ε. Hat nun aber f_k, $1 < k < n$, eine Nullstelle an x, dann hat wegen (3.39) entweder f_{k+1} oder f_{k-1} dasselbe Vorzeichen wie f_k auf $[x-\varepsilon, x]$ und dasselbe gilt auch auf $[x, x+\varepsilon]$. Mit anderen Worten: nur wenn f_n relativ zu f_{n-1} das Vorzeichen wechselt, verändert sich auch $V(x)$. Haben f_{n-1} und f_n nun einen *gemeinsamen* Vorzeichenwechsel an x, dann bleibt V auf $[x-\varepsilon, x+\varepsilon]$ unverändert, andernfalls steigt oder fällt die Anzahl der Vorzeichenwechsel, je nachdem, ob f_{n-1} und f_n an $x-\varepsilon$ gleiches oder entgegengesetztes Vorzeichen hatten:

	$x-\varepsilon$	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	$k-1$

Das liefert uns die folgende Beobachtung.

Satz 3.27 *Seien*[93]

$$\sigma_+(f, I) := \{x \in I \ : \ f(x-\varepsilon) > f(x) = 0 > f(x+\varepsilon)\},$$

und

$$\sigma_-(f, I) := \{x \in I \ : \ f(x-\varepsilon) < f(x) = 0 < f(x+\varepsilon)\},$$

dann ist, mit $I = [a, b)$,

$$\sigma_+\left(\frac{f_n}{f_{n-1}}, I\right) - \sigma_-\left(\frac{f_n}{f_{n-1}}, I\right) = V(b) - V(a). \tag{3.40}$$

Und in der Tat haben wir nun auch schon alles beisammen, um die Einfachheit der Nullstellen zu beweisen – und zwar nur unter Verwendung der Rekursionsformel.

Proposition 3.28 *Für jede Polynomfolge f_n, $n \in \mathbb{N}_0$, definiert durch die Rekursionsformel*[94]

$$f_0 = 1, \qquad f_{n+1}(x) = (x + \beta_n)\, f_n(x) + \gamma_n\, f_{n-1}(x), \qquad \gamma_n < 0, \qquad n \in \mathbb{N}_0,$$

gilt:

1. *Jede endliche Folge $f_0, \ldots, f_n$ ist eine Sturmsche Kette für jedes Intervall $I \subseteq \mathbb{R}$.*

[92] Also die Anzahl der Vorzeichenwechsel, nachdem die Nullen getrichen wurden.

[93] Wir schreiben das etwas schlampig: $x - \varepsilon$ beinhaltet immer "für alle $\varepsilon > 0$, die hinreichend klein sind".

[94] Und diese Rekursionen sind nach Satz 3.12 genau die Rekursionen für *monische* orthogonale Polynome bzüglich eines strikt quadratpositiven linearen Funktionals.

2. *Das Polynom f_n hat genau n einfache relle Nullstellen, d.h.,*

$$\#Z_{\mathbb{R}}(f_n) = n, \qquad Z_I(f) = \{x \in I \,:\, f(x) = 0\}.$$

Beweis: Daß f_0 keine Nullstellen hat ist einleuchtend. Ist nun $f_n(x) = 0$, dann liefert die Rekursionsformel, daß

$$f_{n+1}(x) = \gamma_n f_{n-1}(x)$$

und damit ist entweder $f_{n+1}(x) f_{n-1}(x) < 0$ oder $f_{n+1}(x) = f_n(x) = f_{n-1}(x) = 0$. Im letzteren Fall wäre dann aber auch

$$f_{n-2} = \frac{f_n - (x + \beta_{n-1}) f_{n-1}}{\gamma_{n-1}} = 0$$

und mit demselben Argument ebenfalls $0 = f_{n-3}(x) = \cdots = f_0(x)$, was ein Widerspruch zu $f_0 = 1$ wäre. Also ist $f_0, \ldots, f_n$ eine Sturmsche Kette.

Da wir es also mit einer Sturmschen Kette zu tun haben, können wir Satz 3.27 verwenden. Da σ_+ und σ_- nur einen Teil der Nullstellen von f_n erfassen, ist offensichtlich für $I = [a, b)$

$$\left|\sigma_+\left(\frac{f_n}{f_{n-1}}, I\right) - \sigma_-\left(\frac{f_n}{f_{n-1}}, I\right)\right| \le \sigma_+\left(\frac{f_n}{f_{n-1}}, I\right) + \sigma_-\left(\frac{f_n}{f_{n-1}}, I\right) \le \#Z_{\mathbb{R}}(f_n) \le n. \tag{3.41}$$

Nun sind die Polynome aber alle monisch, d.h. $f_k(x) = x^k + \cdots$ und somit ist

$$\lim_{x \to -\infty} f_k(x) = (-1)^k \infty, \qquad \lim_{x \to +\infty} f_k(x) = \infty,$$

also

$$\lim_{a \to -\infty} V(a) = n, \qquad \lim_{b \to +\infty} V(b) = 0,$$

und somit für hinreichend kleines a und hinreichend großes

$$n = |V(b) - V(a)| = \left|\sigma_+\left(\frac{f_n}{f_{n-1}}, I\right) - \sigma_-\left(\frac{f_n}{f_{n-1}}, I\right)\right|.$$

Setzen wir diese Identität in (3.41), so erhalten wir, daß

$$n \le \#Z_{\mathbb{R}}(f_n) \le n \qquad \Rightarrow \qquad \#Z_{\mathbb{R}}(f_n) = n,$$

wie behauptet. □

Eigentlich sagt uns der Beweis sogar noch mehr! Denn

$$-n = V(b) - V(a) = \sigma_+\left(\frac{f_n}{f_{n-1}}, \mathbb{R}\right) - \sigma_-\left(\frac{f_n}{f_{n-1}}, \mathbb{R}\right)$$

ist nur dadurch zu erreichen, daß

$$\sigma_+\left(\frac{f_n}{f_{n-1}}, \mathbb{R}\right) = 0 \qquad \text{und} \qquad \sigma_-\left(\frac{f_n}{f_{n-1}}, \mathbb{R}\right) = n$$

ist. Also sind alle Vorzeichenwechsel von f_n/f_{n-1} Vorzeichenwechsel von $-$ nach $+$. Das kann aber nur dadurch erreicht werden, daß zwischen zwei Vorzeichenwechseln von f_n auch f_{n-1} sein Vorzeichen wechselt. Anders gesagt:

Die Nullstellen der (orthogonalen) Polynome aus Proposition 3.28 sind geschachtelt: Zwischen je zwei aufeinanderfolgenden Nullstellen von f_n liegt immer eine Nullstelle von f_{n-1}.

3.6 Padé–Approximation

Man kann auch zwischen Padé–Approximation und Kettenbrüchen einen Bezug herstellen; eigentlich haben wir sogar bereits bei unserer Gauß–Quadratur eine Form von Padé–Approximation betrieben. Wir wollen hier aber nur kurz skizzieren, worum es dabei eigentlich geht, für Details sei auf [19] und [23] verwiesen. Wir betrachten formale Potenzreihen der Form

$$f(x) = f_0 + f_1\,x + f_2\,x^2 + \cdots = \sum_{j=0}^{\infty} f_j\,x^j, \qquad f_0 \neq 0$$

und versuchen, diese von möglichst hoher Ordnung durch eine rationale Funktion

$$r_{m,n}(x) = \frac{p_m(x)}{q_n(x)}, \qquad p_m \in \Pi_m,\ q_n \in \Pi_n,$$

anzunähern. Da wir immer eine Konstante in $r_{m,n}$ kürzen können ohne die rationale Funtkion zu verändern, haben wir also insgesamt $m + n + 1$ freie Parameter, die so bestimmt werden sollen, daß die ersten $m + n + 1$ Koeffizienten von f "erwischt" werden, daß also

$$q_n\,f(x) - p_m(x) = O\left(x^{n+m+1}\right)$$

gilt. Der Ansatz sollte uns ziemlich bekannt vorkommen, denn das war ja gerade der Job, den unsere Kettenbrüche bei den Laurentreihen gemacht haben, die von den Momenten gebildet wurden. Die Tabelle der rationalen Funktionen $r_{m,n}$ bezeichnet man als *Padétafel* und tatsächlich findet man in dieser Tafel auch Konvergenten von Kettenbrüchen, nämlich

$$r_{0,1}, r_{1,1}, r_{1,2}, r_{2,2}, r_{2,3}, \ldots$$

siehe [23, S. 256ff]. Weiter ins Detail zu gehen würde aber zu weit führen und keine Zeit mehr für die ebenfalls interessante Signalverarbeitung lassen.

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*

4 Signalverarbeitung, Hurwitz und Stieltjes

Sogar bei der Signalverarbeitung kann man Kettenbrüche nicht vermeiden. Hier werden sie in Form eines recht klassischen Satzes von Stieltjes aus [7] in Zusammenhang mit *Hurwitz–Polynomen* auftauchen, die wiederum eng mit der Stabilität von IIR–Filtern verwandt sind. Was das alles bedeutet? Ein bisschen Geduld noch . . .

4.1 Signale und Filter

Ein *zeitdiskretes* Signal s ist nichts anderes als eine doppeltunendliche Folge oder ein doppeltunendlicher Vektor

$$\sigma = (\sigma_j \ : \ j \in \mathbb{Z}) \in \ell(\mathbb{Z}).$$

Natürlich haben realistische Signale einen Anfang und ein Ende, also *endlichen Träger*[95], zumindest aber *endliche Energie*

$$\|\sigma\|_2 = \left(\sum_{j \in \mathbb{Z}} |\sigma_j|^2 \right)^{1/2},$$

aber es ist wesentlich angenehmer und praktischer, Signale in dieser unbeschränkten Form darzustellen. Ein *Filter* ist zuerst einmal nur ein Operator $F : \ell(\mathbb{Z}) \to \ell(\mathbb{Z})$, der Signale auf Signale abbildet. Trotzdem schränkt man sich in der Signalverarbeitung sehr schnell ein, nämlich auf sogenannte LTI–Filter (**L**inear **T**ime **I**nvariant), die, wie der Name schon sagt, zwei Eigenschaften aufweisen:

Linearität: Der Filter F ist ein *linearer* Operator von $\ell(\mathbb{Z})$ nach $\ell(\mathbb{Z})$.

Zeitinvarianz: Was passiert ist unabhängig davon, wann es passiert, das heißt, verschiebt man ein Signal in der (diskreten) Zeit um einen bestimmten Faktor so ist das gefilterte Signal bis auf *dieselbe* Zeitverschiebung wieder identisch:

$$\sigma'_j = \sigma_{j+k}, \quad j \in \mathbb{Z} \qquad \Rightarrow \qquad (F\sigma')_j = (F\sigma)_{j+k}, \quad j \in \mathbb{Z}.$$

[95] Man könnte auch sagen “kompakten Träger”, aber bei diskreten Signalen ist das nicht so wild.

Zeitinvarianz läßt sich schöner mit Hilfe des *Translationsoperators* τ schreiben, der durch $(\tau\sigma)_j = \sigma_{j+1}$, $j \in \mathbb{Z}$, definiert ist. Ein Filter F ist dann nmlich genau dann zeitinvariant, wenn er mit der Translation kommutiert, also wenn

$$\tau F = F\tau \tag{4.1}$$

gilt. Jeder lineare Filter läßt sich nun als doppeltunendliche Matrix schreiben, $F = [F_{jk} \,:\, j,k \in \mathbb{Z}]$, mit der normalen Multiplikation

$$(Ff)_j = \sum_{k\in\mathbb{Z}} F_{jk}\, f_k, \qquad j \in \mathbb{Z}.$$

Ist nun F ein LTI–Filter, dann ist

$$[F_{j+1,k} \,:\, j,k \in \mathbb{Z}] = \tau F = F\tau = [F_{j,k-1} \,:\, j,k \in \mathbb{Z}]$$

Da die beiden Matrizen denselben Operator liefern sollen, müssen sie in allen Komponenten übereinstimmen, weswegen $F_{j+1,k} = F_{j,k+1}$ bzw. nach Iteration

$$F_{j+\ell,k} = F_{j,k-\ell}, \qquad \ell \in \mathbb{Z},$$

sein muss. Dies ist offensichtlich erfüllt, wenn $F_{jk} = f_{j-k}$ für $f \in \ell(\mathbb{Z})$ ist, aber es gilt auch die Umkehrung: Ist $j-k = \ell-m$, dann ist $j-\ell = k-m$ und

$$F_{jk} = F_{\ell+(j-\ell),k} = F_{\ell,k-(k-m)} = F_{\ell,m}$$

und damit hängt F_{jk} nur von $j-k$ ab. Das können wir folgendermaßen zusammenfassen.

Proposition 4.1 *Ein Operator $F : \ell(\mathbb{Z}) \to \ell(\mathbb{Z})$ ist genau dann ein LTI–Filter wenn es einen Vektor $f \in \ell(\mathbb{Z})$ gibt, so daß $F_{jk} = f_{j-k}$, $j,k \in \mathbb{Z}$, ist. In diesem Fall ist*

$$(F\sigma)_j = \sum_{k\in\mathbb{Z}} f_{j-k}\, \sigma_k, \qquad j \in \mathbb{Z}. \tag{4.2}$$

Die Summe in (4.2) bezeichnet man als *Faltung* $f*\sigma$ von f mit σ. Als nächstes noch ein bißchen Terminologie.

Definition 4.2 (Puls, Filtertypen und z–Transformation)

1. *Das* Pulssignal δ *hat die Form*[96] $\delta_j = \delta_{j0}$.
2. *Die* Impulsantwort *eines Filters F ist $F\delta$.*
3. *Der* Träger *eines Signals oder Vektors $\sigma \in \ell(\mathbb{Z})$ ist*

$$\operatorname{supp}\sigma = \{j \in \mathbb{Z} \,:\, \sigma_j \neq 0\}$$

[96] Einmal steht δ für das Signal, einmal für das Kronecker–δ.

4. *Ein Filter heißt* FIR–Filter *(**F**inite **I**mpulse **R**esponse), wenn er ein LTI–Filter ist und seine Impulsantwort endlichen Träger*[97] *hat:*

$$F\delta \in \ell_{00}(\mathbb{Z}); \qquad \sigma \in \ell_{00}(\mathbb{Z}) = \{\sigma \in \ell(\mathbb{Z}) \ : \ \#\operatorname{supp}\sigma < \infty\}\,.$$

Ansonsten spricht man von einem IIR–Filter[98].

5. *Die z–*Transformation *eines Vektors oder Signals $f \in \ell(\mathbb{Z})$ ist*

$$f(z) = \sum_{k\in\mathbb{Z}} f_k\, z^{-k}, \qquad z \in \mathbb{C}^\times = \mathbb{C} \setminus \{0\}.$$

Der Grund für die Einführung der z–Transformation ist schnell erzählt: Für beliebige Folgen $f, g \in \ell(\mathbb{Z})$ ist

$$(f * g)(z) = \sum_{j\in\mathbb{Z}} \left(\sum_{k\in\mathbb{Z}} f_{j-k}\, g_k \right) z^{-j} = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} f_j\, g_k\, z^{-j-k} = \left(\sum_{j\in\mathbb{Z}} f_j z^{-j} \right) \left(\sum_{k\in\mathbb{Z}} g_k z^{-k} \right)$$

und somit

$$(f * g)(z) = f(z)\, g(z), \tag{4.3}$$

Faltungen werden also zu Produkten der z–Transformationen. Insbesondere ist also auch für jeden LTI–Filter F

$$(F\sigma)(z) = f(z)\, \sigma(z). \tag{4.4}$$

Daß wir hier multiplizieren können ist aber nur ein Teil der Geschichte! Fast noch bedeutsamer ist die Tatsache, daß man auf diese Art und Weise auch eine extrem *schnelle* Filterung implementieren kann, indem man in (4.4) $z = e^{-i\omega}$ setzt, diskretisiert und die *schnelle Fouriertransformation* (FFT) nutzt, siehe z.B. [21, 27, 29]. `Matlab` und `Octave` besitzen beispielsweise Routinen für diesen Zweck, in `Octave` hört diese Routine auf den Namen `fftfilt`. Grob gesprochen kann man so die Komplexität der Filterung mit einem Filter der Länge[99] N von $O(N^2)$ auf den deutlich besseren und wahrscheinlich optimalen[100] Wert $O(N \log N)$ reduzieren.

Ist nun F ein FIR–Filter, dann ist seine z–Transformation von der Form

$$f(z) = \sum_{j=n_0}^{n_1} f_j z^j, \qquad n_0 \le n_1 \in \mathbb{Z};$$

[97] Und damit endliche Dauer.

[98] Wofür das "I" wohl stehen wird?

[99] Die Filterlänge ist die Differenz zwischen dem größten und kleinsten Index zu "*Taps*", das sind die von Null verschienenen Filterkoeffizienten; diese Größe legt gleichzeitig fest, welche Dimension ein Puffer für die Eingabedaten haben muß.

[100] So weit ich weiß existiert kein *Beweis*, daß die Komplexität der FFT wirklich optimal für diesen Job ist – aber seit der "Erfindung" in [5], siehe auch [3, 4] über die "historische" Entwicklung, hat niemand etwas besseres gefunden. Und wenn man sieht, daß die FFT in unheimlich vielen Bereichen, von der Multiplikation bestimmter Matrizen bis hin zur Multiplikation ganzer Zahlen [28], eingesetzt wird, dann will das schon etwas heißen.

so eine endliche Summe, in der positive und negative Potenzen von z vorkommen können bezeichnet man als *Laurentpolynom*. Ist $n_0 \geq 0$, also supp $f \subseteq \mathbb{N}_0$, so nennt man den Filter *kausal*, denn dann ist für $j \in \mathbb{Z}$

$$(F\sigma)_j = \sum_{k\in\mathbb{Z}} f_{j-k}\,\sigma_k = \sum_{k\in\mathbb{Z}} f_k\,\sigma_{j-k} = \sum_{k\in\mathbb{N}_0} f_k\,\sigma_{j-k}$$

und das gefilterte Signal zum Zeitpunkt j hängt nur von den Werten von σ in der *Vergangenheit* ab – was man auch von einem realisierbaren Filter erwarten sollte, dem normalerweise ja die Fähigkeit fehlt, in die Zukunft zu sehen.

4.2 Rationale Filter und Stabilität

Eines sollte uns bei unserem Einstieg in die Signalverarbeitung inzwischen klargeworden sein: FIR–Filter sind eine feine Sache, da sie wirklich realisierbar sind, ganz egal, ob sie nun kausal sind oder nicht, zumindest, wenn man eine verzögerte Ausgabe zulässt. In der Tat kann man

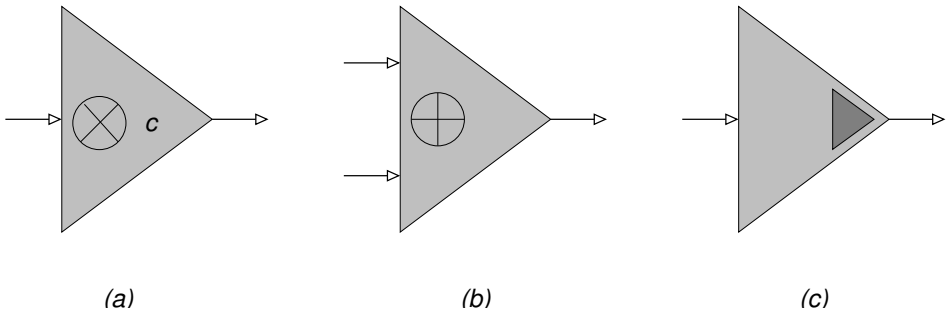


Abbildung 4.1: Symbolische Darstellung der drei Bausteine: Multiplizierer (a), Addierer (b) und Verzögerer (c).

jeden Filter mit drei Bausteinen, siehe Abb. 4.1, realisieren: Multiplizierern, Addierern[101] und Verzögerungsgliedern, die dafür zuständig sind, daß die "zeitverschobenen" Komponenten des Signals dem Filter zugeführt werden, wenn man sie braucht. Das Blockschaltbild zur Realisierung eines kausalen Filters aus diesen drei Bausteinen ist dann in Abb. 4.2 dargestellt.

Andererseits bieten FIR–Filter nicht genug Flexibilität, insbesondere, wenn man "steilflankige" Bandpassfilter realisieren möchte, die nur ein gewisses Frequenzband durchlassen und einen scharfen Übergang zwischen Durchlass- und Sperrbereich aufweisen; bei der "optimalen" Näherung durch FIR–Filter tritt ein sehr unerfreuliches Oszillierungsphänomen, das sogenannte *Gibbs–Phänomen* auf, siehe Abb. 4.3. Wie Abb. 4.3 handelt es sich zwar nicht wirklich um ein absolut unvermeidbares Problem, was man sich bei seiner Verhinderung allerdings einhandelt ist ein deutlicher Verlust an "Steilflankigkeit".

[101] Was diese beiden machen, sollte aus dem Namen hervorgehen.

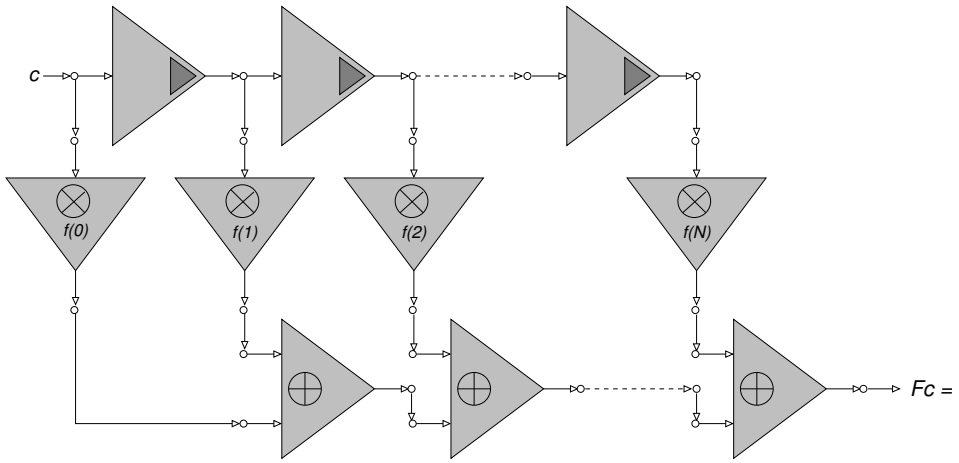


Abbildung 4.2: Ein FIR–Filter als Kaskade der Bausteine aus Abb. 4.1. Die Verzögerer sorgen für die Translationen, die Multiplizierer für die Gewichtung und die Addierer summieren den ganzen Kram auf.

Deswegen versucht man, die Klasse an zulässigen Filtern zu erweitern, indem man *rationale* Filter der Form

$$(F\sigma)(z) = f(z)\,\sigma(z) = \frac{p(z)}{q(z)}\,\sigma(z), \qquad p(z) = \sum_{j\in\mathbb{N}_0} p_j\, z^{-j}, \quad q(z) = \sum_{j\in\mathbb{N}_0} q_j\, z^{-j},$$

verwendet, deren z–Transformierte der Quotient zweier Laurentpolynome ist. Indem wir Zähler und Nenner wenn nötig mit einer Potenz von z und einer passenden Konstanten multiplizieren, können wir immer annehmen, daß[102] $q(z) = 1 + q_1 z^{-1} + \cdots + q_n z^{-n}$, $q_n \neq 0$, für ein passendes n, also $q(z) = z^{-n}\widehat{q}(z)$, wobei $\widehat{q}(z) = q_n + q_{n-1}\,z + \cdots + z^n$ ein Polynom ist. Nach Lemma 3.17 ist also

$$\frac{1}{q(z)} = z^n\,\frac{1}{\widehat{q}(z)} = z^n\,\sum_{j=n}^{\infty} \lambda_j\, z^{-j} = \sum_{j=0}^{\infty} \lambda_j\, z^{-j}, \qquad \lambda \in \ell(\mathbb{Z}),$$

so daß

$$f(z) = \sum_{j=0}^{\infty} f_j\, z^{-j} \qquad \Rightarrow \qquad f \in \ell(\mathbb{Z}), \quad \operatorname{supp} f \subseteq \mathbb{N}_0,$$

ist. Wir sollten also nicht erwarten oder auch nur hoffen, daß f noch ein FIR–Filter ist. Trotzdem kann man F noch einfach realisieren: dazu formen wir die Definition von $F\sigma$ in

$$p(z)\sigma(z) = (F\sigma)(z)\,q(z) = (F\sigma)(z) + z^{-1}\,\widetilde{q}(z)\;(F\sigma)(z), \qquad \widetilde{q}(z) = q_1 + \cdot + q_n z^{-n},$$

also

$$(F\sigma)(z) = p(z)\sigma(z) - \left[z^{-1}\,(F\sigma)(z)\right]\;\widetilde{q}(z) = p(z)\sigma(z) - q(z)\;\left(\tau^{-1}\,F\sigma\right)(z) \tag{4.5}$$

[102] Die Konstante brauchen wir, um den konstanten Term von q auf 1 zu normieren.

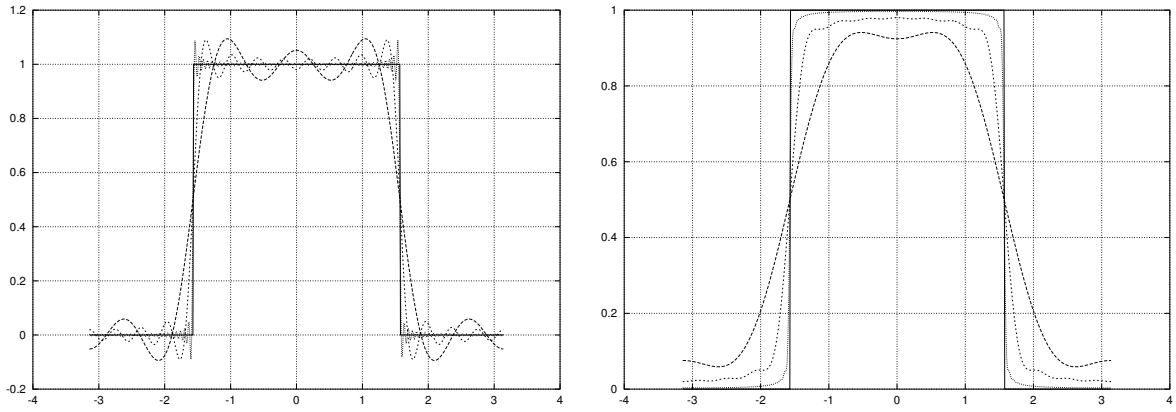


Abbildung 4.3: *Links:* Approximation eines Bandpassfilters durch Partialsummen der Fourierreihe für $n = 5, 15, 100$ (Werte eher zufällig) zeigt das Gibbs–Phänonem. Man beachte, daß die "Überschiesser" nur schmaler, nicht aber kleiner werden.

Rechts: Approximation durch ein anderes Approximationsverfahren, nämlich die *Fejérschen Mittel*. Diese haben zwar eine größere Abweichung vom Bandpass als die Partialsummen (Partialsummen sind eine Bestapproximation), verzichten dafür aber auf wilde Oszillationen.

da

$$z^{-1}\,(F\sigma)\,(z) = \sum_{j\in\mathbb{Z}} (F\sigma)_j\; z^{-j-1} = \sum_{j\in\mathbb{Z}} (F\sigma)_{j-1}\; z^{-j} = \left(\tau^{-1}\,F\sigma\right)(z).$$

Nun ist aber $\widetilde{q}$ ein *kausaler* FIR–Filter und interessiert sich zum Zeitpunkt j nur dafür, was $\tau^{-1}\,F\sigma$ bis zum Zeitpunkt j für Werte hatte, also für die Werte, die $F\sigma$ bis zum Zeitpunkt $j-1$ hatte – und die sind aber bekannt. Anders gesagt: Wir berechnen $F\sigma$ durch Filterung von σ mit p und Feedback unter Verwendung von $\widetilde{q}$; dies ist in Abb 4.4 dargestellt, für Details der Herleitung siehe z.B. [12, 13, 27]. Alles was uns an dieser Stelle interessieren soll ist die Tatsache, daß rationale Filter eine auch praktisch relevante Sache sind, da man sie mit den drei Bausteinen tatsächlich realisieren kann.

Allerdings kann sich das Feedback–System $\widetilde{q}$ auch ziemlich danebenbenehmen! Dazu entwickeln wir nochmal $1/q$ als Laurentreihe,

$$\frac{1}{q(z)} = \sum_{j=0}^{\infty} \lambda_j\, z^{-j},$$

erhalten unter der Annahme, daß $\operatorname{supp} p \subseteq [0, m]$ die Identität

$$f(z) = \sum_{j=0}^{\infty}\sum_{k=0}^{m} p_k\,\lambda_j\, z^{-j-k} = \sum_{j=0}^{\infty}\left[\sum_{k=0}^{m} p_k\lambda_{j-k}\right] z^{-j} = (\lambda * p)\,(z)$$

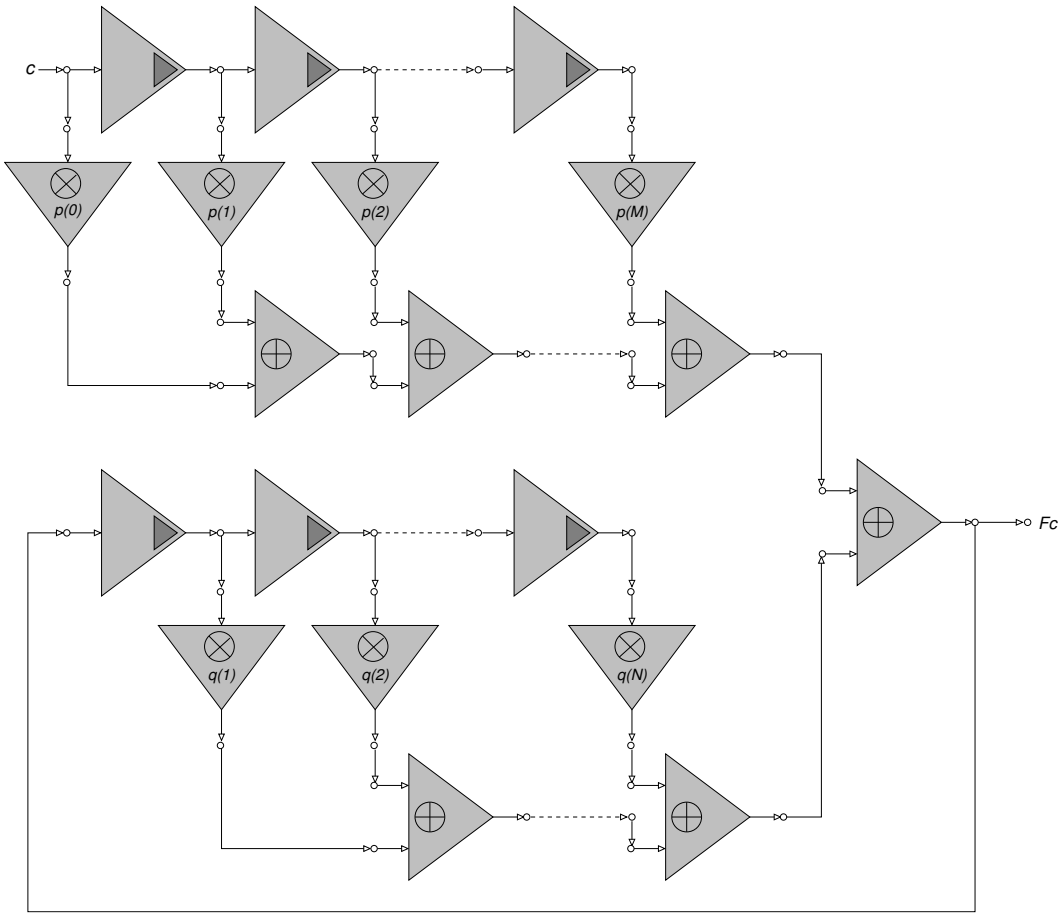


Abbildung 4.4: Ein rationaler Filter, realisiert mittels *delayed feedback*: Das mit p gefilterte Signal wird verzögert in den Filter q geschickt und die beiden Ergebnisse summiert.

und sehen uns an, wie sich λ_j und damit auch f_j für $j \to \infty$ verhält. In der Tat kann sich q *dämpfend* verhalten, wenn $\lambda_j \to 0$ für $j \to \infty$ oder aber *verstärkend*, wenn $|\lambda_j| \to \infty$ für $j \to \infty$. Da $f_j = (\lambda * p)_j$ ist, überträgt sich dieses Konvergenz- und Divergenzverhalten auch auf die Impulsantwort f. Und ein "guter" Filter hat besser eine abklingende Impulsantwort, denn sonst kann er eigentlich gar nicht realisiert werden: Ein derartiger Filter, dessen Impulsantwort divergiert, würde *unendliche* Energie benötigen. Außerdem wäre sein "Eigenleben", also das was im internen Feedback–Kreislauf passiert, irgendwann so stark, daß es alle Eingabedaten, alle weiteren Impulse, dominieren würde – der Filter würde nicht einmal auf die Außenwelt reagieren.

Definition 4.3 *Der LTI–Filter F heißt* stabil*, wenn*

$$\lim_{j \to -\infty} f_j = \lim_{j \to \infty} f_j = 0$$

ist.

Was aber bedeutet nun Stabilität für unser Nennerpolynom q Das wird klar, wenn wir uns einmal den einfachsten nichttrivialen Fall ansehen, nämlich $q(z) = 1 - \zeta\, z^{-1} = z^{-1}\,(z - \zeta)$, $\zeta \in \mathbb{C}^{\times}$. Erinnern wir uns an (3.36), dann ist[103]

$$\frac{1}{q(z)} = z\,\frac{1}{z-\zeta} = \sum_{j=0}^{\infty} \frac{\zeta^j}{z^j} \qquad \Rightarrow \qquad \lambda_j = \zeta^j,$$

und Stabilität ist äquivalent zu $|\zeta| < 1$, die Nullstelle ζ von $q(z)$ muß also *im Einheitskreis*[104]

$$\mathbb{D}^0 = \mathbb{D} \setminus \partial\mathbb{D}, \qquad \mathbb{D} = \{z \in \mathbb{C} \,:\, |z| \leq 1\}$$

liegen! Ist hingegen $|\zeta| > 1$, dann fliegt uns der Filter um die Ohren, nur für Nullstellen auf dem Einheitskreis, $|\zeta| = 1$, können wir keine richtig negativen, aber auch keine wie auch immer gearteten positiven Aussagen über die Impulsantwort machen. Wenn wir nun einen beliebigen Filter mit rationaler z–Transformierter haben, dann faktorisieren wir q in $q(z) = z^{-n}\,(z-\zeta_1)\cdots(z-\zeta_n)$ und verwenden die *Partialbruchzerlegung*

$$f(z) = \frac{p(z)}{q(z)} = \sum_{j=1}^{k} \frac{p_j(z)}{(z-\zeta_j)^{\alpha_j}}, \qquad \alpha_1 + \cdots + \alpha_k = n,$$

wobei α_j die *Vielfachheit* der Nullstelle ζ_j bezeichnet. Das Konvergenz-/Divergenzhalten wird dann vom betragsgrößten ζ_j entschieden: liegt es außerhalb des Einheitskreises, dann haben wir es mit Divergenz zu tun, liegt es innerhalb des Einheitskreises, dann können wir uns über Konvergenz freuen. Und das ist auch das wesentliche Resultat über die Stabilität rationaler Filter[105], ein vollständiger Beweis findet sich z.B. in [27].

Satz 4.4 *Ein rationaler Filter F mit z–Transformation $f(z) = p(z)/q(z)$ ist genau dann stabil wenn alle Nullstellen von q im Einheitskreis liegen.*

4.3 Fourier und Abtasten

Bevor wir uns an die Frage machen, wie man Polynome bekommt, die keine Nullstellen im Einheitskreis haben, zuerst noch eine kurze Bemerkung, warum der Einheitskreis

$$\partial\mathbb{D} = \{z \in \mathbb{C} \,:\, |z| = 1\} = \left\{e^{-i\theta} \,:\, \theta \in [-\pi, \pi]\right\}$$

so eine wichtige Rolle spielt. Daß man statt der z–Transformation $\sigma(z)$ eines Signals auch die zugehörige trigonometrische Reihe oder *Fourierreihe*

$$\widehat{\sigma}(\theta) = \sigma\left(e^{i\theta}\right) = \sum_{k\in\mathbb{Z}} \sigma_k\, e^{-ik\theta} = \sum_{k\in\mathbb{Z}} \sigma_k\, \cos k\theta + i \sum_{k\in\mathbb{Z}} \sigma_k\, \sin k\theta$$

[103] Man kann es natürlich auch sehr einfach nochmals "beweisen".
[104] Wir bezeichnen mit $\mathbb{D}$ den *abgeschlossenen* Einheitskreis, mit $\mathbb{D}^0$ den offenen.
[105] Auch gerne als *rekursive Filter* bezeichnet, der Grund dafür sollte klar sein.

betrachten kann, ist genauso naheliegend wie die Tatsache, daß

$$(f * \sigma)^{\wedge}(\theta) = (f * \sigma)\left(e^{i\theta}\right) = f\left(e^{i\theta}\right)\, \sigma\left(e^{i\theta}\right) = \widehat{f}(\theta)\, \widehat{\sigma}(\theta). \tag{4.6}$$

Die (prinzipiell komplexwertige) Funktion $\widehat{f}(\theta)$ bezeichnet man als *Transferfunktion* des Filters; sie wird in Anwendungen meistens in der logarithmischen *Dezibel*[106]–Skala "dB" angegeben, d.h., man verwendet anstelle des Wertes y den Wert $10 \log_{10} y$ und schreibt "dB"dahinter[107]. Da Cosinus und Sinus gerade bzw. ungerade Funktionen sind, hat die Transferfunktion die Form

$$\widehat{f}(\theta) = f_0 + \sum_{k=1}^{\infty} (f_k + f_{-k})\, \cos k\theta + i \sum_{k=1}^{\infty} (f_k - f_{-k})\, \sin k\theta$$

und ist somit genau dann reell, wenn $f_k = f_{-k}$, als der Filter symmetrisch ist. Was für uns an dieser Stelle wichtig ist: Durch den Übergang von z–Transformationen zu trigonometrischen Polynomen können wir uns statt auf $\mathbb{C}^{\times}$ auf den Einheitskreis $\partial\mathbb{D}$ beschränken.

Außerdem werden Frequenzgänge in dieser Darstellung sehr viel natürlicher wiedergegeben! Ein Bandpassfilter ist jetzt eben wirklich von der Form $\widehat{f} = \chi_{[\omega_0,\omega_1]}$. Aber Moment einmal! Wo bitte liegen jetzt die Frequenzgänge von beispielsweise 3000-4000 Hz? Alles was wir haben sind Werte $\widehat{f}[0,\pi]$ – zumindest wenn wir eine relle Transferfunktion wollen. Und das sind bestenfalls "relative Frequenzen". Die "wirklichen", "absoluten" Frequenzen sind nämlich im Signal σ codiert, und zwar in der *Abtastfrequenz*. Wir haben bisher immer nur gesagt, daß σ ein zeitdiskretes Singal, also eine Folge, sein soll, und das bedeutet, daß

$$\sigma_k = s\,(t_0 + k\tau)\,, \qquad k \in \mathbb{Z}, \qquad t_0 \in \mathbb{R},\, \tau > 0,$$

eine *Abtastung* des Originalsignals s darstellt, wobei τ das Abtastintervall und $1/\tau$ die Abtastfrequenz ist. Und man kann es sich leicht vorstellen: je kleiner τ ist, je höher also die Abtastfrequenz ist, desto höher wird die *Frequenzauflösung* sein. Das kann man formalisieren und das führt zum berühmten Abtastsatz von Shannon[108], für den wir aber noch einen Begriff benötigen.

Definition 4.5 *Eine Funktion $f \in L_1(\mathbb{R})$ heißt* bandbeschränkt *mit Bandbreite T, wenn ihre* Fouriertransformation

$$\widehat{f}(\xi) = \int_{\mathbb{R}} f(t)\, e^{-i\xi t}\, dt$$

außerhalb von $[-T,T]$ verschwindet:

$$\widehat{f}(\xi) = 0, \qquad \xi \notin [-T,T].$$

[106] Trotz des fehlenden "l" angeblich nach Alexander Graham Bell benannt.

[107] Die Dezibel–Skala ist also logarithmisch! Erhöht man also die Lautstärke in einer Disco um wenige Dezibel, kann sich der Schalldruck sehr wohl vervielfachen, aber diese Bemerkung stößt dort sowieso auf taube Ohren ...

[108] Bzw. Shannon–Whittaker bzw. Shannon–Whittaker–Kotelnikov. Es sieht so aus, als hätte Whittaker [32] bereits 1915 dieses Resultat erhalten, siehe auch [33], allerdings eher theoretisch im Zusammenhang mit Interpolation, wohingegen Shannon deutlich später [30, 31] den Zusammenhang mit der Signalverarbeitung erkannte.

Nachdem der Wert $\widehat{f}(\xi)$ den Energieanteil der Frequenz ξ im kontinuierlichen Signal f angibt[109], und da $-\xi$ ja dieselbe Frequenz ist wie ξ, heißt "bandbeschränkt" also nichts anderes, als daß nur Frequenzen $\leq T$ im Signal f auftauchen. Der Shannonsche Abstastsatz sagt uns nun, daß wir bandbeschränkte Funktionen aus hinreichend feinen Abtastungen wieder vollständig rekonstruieren können.

Satz 4.6 (Abtastsatz) *Ist f eine T–bandbeschränkte Funktion, und ist $\tau < \tau^* = \frac{\pi}{T}$, dann ist*

$$f(x) = \sum_{k\in\mathbb{Z}} \sigma_k \, \frac{\sin \pi \left(x/\tau - k\right)}{\pi \left(x/\tau - k\right)}, \qquad \sigma_k = f\left(k\tau\right), \quad k \in \mathbb{Z}. \tag{4.7}$$

Die kritische Frequenz $1/\tau^* = T/\pi$ bzw. die Hälfte davon[110] bezeichnet man als *Nyquist–Frequenz*, die Funktion

$$g(x) = \frac{\sin \pi x}{\pi x} =: \operatorname{sinc} x, \qquad x \in \mathbb{R},$$

als *Sinus Cardinalis* oder kurz *"sinc–Funktion"*. Das "cardinalis" kommt daher, daß[111]

$$\operatorname{sinc} k = \delta_{0k} = \begin{cases} 1, & k = 0, \\ 0, & \text{sonst}, \end{cases}$$

ist, siehe Abb. 4.5. Den Beweis von Satz 4.6 findet man beispielsweise in [21], von wo auch der

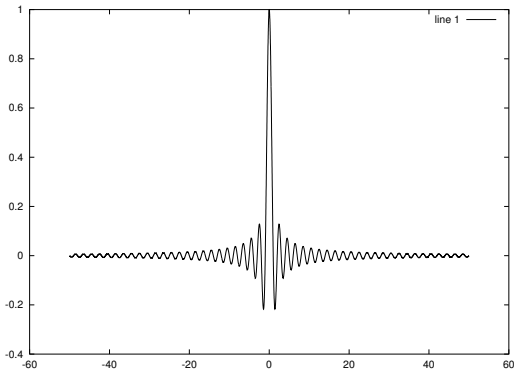


Abbildung 4.5: Die sinc–Funktion. Man sieht, daß sie für $|x| \to \infty$, aber halt eben nur wie $|x|^{-1}$ und das ist schon sehr langsam. Die erste Schwinungsperiode täuscht hier etwas!

in [27] angegebene stammt, was einem die Ingenieursliteratur, z.B. [16], vorsetzt, erfüllt nicht immer mathematische Ansprüche an Korrektheit.

[109] Da $\|f\|_2 = \left\|\widehat{f}\right\|_2$ ist, wie man leicht nachrechnet, geht auf diese Art und Weise auch keine Energie verloren.

[110] Das ist letztendlich eine Frage der Normierung von Frequenzen.

[111] An der Stelle 0 braucht man die Regel von L'Hospital.

Und damit können wir auch die Frage nach dem Frequenzbereich der der Digitalfilter beantworten: Die Werte $\theta \in [0, \pi]$ entsprechen den Frequenzen $[0, \tau^{-1}]$, also dem Frequenzbereich, der durch die Abtastung vorgegeben wird.

Und genau deswegen ist die Sache mit der Stabilität eben nicht so einfach: Der Filter soll einerseits auf $\partial\mathbb{D}$ einen vorgegebenen Frequenzgang so gut wie möglich approximieren und andererseits keine Nullstellen im Inneren haben – das ist eine Nebenbedingung an den Nenner des Filters, aber eben nur eine! Übrigens ist das eine leicht vertrackte Situation: Wir legen eine Funktion auf dem Einheitskreis fest, müssen aber für Ihre Stabilität das Verhalten der Funktion *im* Einheitskreis berücksichtigen, genauer, die Frage, ob unsere Funktion im Inneren des Einheitskreises Pole hat oder nicht. Dennoch besteht Hoffnung: Die Funktionentheorie schlägt sich ja schließlich fast dauernd mit derartigen Problemen herum!

4.4 Nullstellen von Polynomen

Kehren wir jetzt zurück zu unserer guten alten z–Transformation und betrachten wir also wieder Polynome. Die "guten" Filter sind also genau die, bei denen alle Nullstellen von $q(z)$ im Inneren des Einheitskreises bzw. alle Nullstellen von

$$q\left(z^{-1}\right) = \sum_{j=0}^{n} q_j\, z^j, \qquad z \in \mathbb{C}^\times,$$

außerhalb des Einheitskreises liegen und tatsächlich gibt es in der klassischen Literatur zur Funktionentheorie auch einiges an Resultaten, die genau diese Frage untersuchen: Wann hat eine komplexes Polynom $f \in \mathbb{C}[z]$ entweder *alle* oder *keine* Nullstellen im Einheitskreis. Ein Klassiker in dieser Richtung ist das Eneström–Kakeya–Theorem, das man beispielsweise in [6] findet und das uns eine *hinreichende* Bedingung liefert, wann ein Polynom keine Nullstelle im Einheitskreis hat.

Satz 4.7 (Eneström–Kakeya) *Ist $p_0 > p_1 > \cdots > p_n > 0$, dann hat das Polynom $p(z) = p_0 + \cdots + p_n\, z^n$ keine Nullstelle in $\mathbb{D}$.*

Beweis: [112] Für $z \in \mathbb{C}$ ist

$$(1-z)\, p(z) = p_0 + \sum_{j=1}^{n} (p_j - p_{j-1})\, z^j - p_n z^{n+1}$$

und somit für $|z| \leq 1$

$$\begin{aligned} |1-z|\ |p(z)| &= p_0 + \left| \sum_{j=1}^{n} (p_j - p_{j-1})\, z^j - p_n z^{n+1} \right| \\ &\geq\ p_0 - \sum_{j=1}^{n} |p_j - p_{j-1}|\ \left|z^j\right| - |p_n|\ \left|z^{n+1}\right| \geq p_0 + \sum_{j=1} (p_j - p_{j-1}) - p_n = 0 \end{aligned}$$

[112] Der Beweis hat zwar nicht wirklich was mit dem zu tun, was folgt, aber da er kurz und einfach ist, sehen wir ihn uns kurz an.

mit Gleichheit dann und nur dann, wenn $|z| = 1$ d.h. $z = e^{i\theta}$ ist und wenn alle Potenzen $z^j = e^{i\theta j}$ dasselbe Argument haben, also wenn $\theta = 0$ bzw. $z = 1$ ist. Da $p(1) = p_0 + \cdots + p_n > 0$ ist, kann p aber auch an $z = 1$ keine Nullstelle haben und somit ist $0 \notin p(\mathbb{D})$. □

Das ist ja schön und gut, aber halt eben auch nur ein hinreichendes Kriterium. Wie aber kann man[113] *charakterisieren*, ob ein Polynom keine Nullstellen im Einheitskreis hat? Zuerst einmal modifiziert man das Problem, und zwar beispielsweise[114] mittels der gebrochen rationalen Transformation

$$w = \frac{z+1}{z-1}, \qquad z = \frac{w+1}{w-1}.$$

Daß diese Transformationen wirklich Inverse voneinander sind, also die Transformation selbstinvers ist, sieht man sofort daran, daß sich beide in die Form $zw - z - w - 1 = 0$ bringen lassen. Was bedeutet das aber nun? Schreiben wir $w = u + iv$, dann ist

$$|z|^2 = \left|\frac{w+1}{w-1}\right|^2 = \frac{(u+1)^2+v^2}{(u-1)^2+v^2} \qquad \Rightarrow \qquad \begin{cases} |z| > 1, & u > 0, \\ |z| = 1, & u = 0, \\ |z| < 1, & u < 0. \end{cases}$$

Damit bildet also die Transformation $z \to w$ die komplexe Ebene auf sich selbst ab, und zwar so, daß $|z| < 1$ genau dann, wenn der Realteil $\Re w$ von w negativ ist. Ist also $p(z)$ ein Laurentpolynom, dann erhalten wir, daß

$$\begin{aligned} p(z) &= \sum_{j=0}^{n} p_j\, z^{-j} = \sum_{j=0}^{n} p_j \left(\frac{w+1}{w-1}\right)^{-j} = \left(\frac{1}{w+1}\right)^n \sum_{j=0}^{n} p_j\,(w-1)^j (w+1)^{n-j} \\ &= \left(\frac{1}{w+1}\right)^n \sum_{j=0}^{n} p_j^w\, w^j = (1+w)^{-n}\, p^w(w), \end{aligned}$$

wobei

$$(1+w)^{-1} = \left(1 + \frac{z+1}{z-1}\right)^{-1} = \left(\frac{2z}{z-1}\right)^{-1} = \frac{z-1}{2z}.$$

Ist nun z eine Nullstelle von p, mit[115] $0 < |z| < 1$, dann ist $w \neq 1$ und somit muß auch p^w an der zugehörigen Stelle w verschwinden und die liegt nun nach unseren Beobachtungen in der linken Halbebene! Das halten wir fest.

Satz 4.8 *Das Laurentpolynom $p(z)$ hat genau dann alle Nullstellen im Einheitskreis, wenn p^w alle Nullstellen in der linken Halbenene $\mathbb{H}_- := \{z \in \mathbb{C} \,:\, \Re z < 0\}$ hat.*

[113] Und zwar ohne das Ding zu faktorisieren, denn so ohne ist ja die Bestimmung der Nullstellen eines Polynoms auch wieder nicht!

[114] In [13] findet man die Transformation $w = (i+z)/(i-z)$, aber das sind nur Rotationen.

[115] Zur Erinnerung: *Laurent*polynome haben an $z = 0$ nichts verloren!

4.5 Hurwitz–Polynome und der Satz von Stieltjes

Eine weitere offensichtliche Beobachtung ist daß die Koeffizienten von p^w reell sind, wenn die von p reell sind. Und das führt uns zu der Klasse von Polynomen, die uns den Rest der Vorlesung interessieren soll. Ab sofort schreiben wir unsere Polynome wieder als Polynome in z, nur interessieren wir uns jetzt nicht mehr für den Einheitskreis, sondern für die linke Halbebene.

Definition 4.9 *Ein Polynom $f \in \mathbb{C}[z]$ heißt Hurwitz–Polynom, wenn es* relle *Koeffizienten hat und alle seine Nullstellen negativen Realteil haben.*

Bevor wir einige Information über Hurwitz–Polynome sammeln, wollen wir uns der drängenden Frage widmen, was die nun wieder mit Kettenbrüchen zu tun haben. Dazu zuerst einmal eine klassische Methode, Polynome zu zerlegen, indem man ein Polynom $f(z)$ als

$$f(z) = \sum_{j=0}^{n} f_j\, z^j = \sum_{j \le n/2} f_{2j}\, z^{2j} + \sum_{j < n/2} f_{2j+1}\, z^{2j+1} = h\left(z^2\right) + z g\left(z^2\right)$$

schreibt, wobei h die Koeffizienten von f mit geradem Index, g die Koeffizienten mit ungeradem Index enthält.

Definition 4.10 *Zwei reelle Polynome $p(x)$ und $q(x)$ mit $\deg p = \deg q = n$ oder $\deg p = n$ und $\deg q = n-1$* bilden ein positives Paar, *wenn ihre Nullstellen $x_1, \ldots, x_n$ und $x'_1, \ldots, x'_n$ bzw. $x'_1, \ldots, x'_{n-1}$ die Bedingung*

$$\begin{aligned} x'_1 < x_1 < x'_2 < \cdots < x'_n < x_n < 0, \qquad & q \in \Pi_n, \\ x_1 < x'_1 < x_2 < \cdots < x'_{n-1} < x_n < 0, \qquad & q \in \Pi_{n-1} \end{aligned} \tag{4.8}$$

erfüllen und die Leitkoeffizienten von p und q gleiches Vorzeichen haben[116].

Und positive Paare beschreiben nun gerade die Hurwitz–Polynome, werden aber andererseits auch durch Kettenbrüche charakterisiert.

Satz 4.11 (Stieltjes) *Für ein Polynom $f(z) = g\left(z^2\right) + z h\left(z^2\right)$ sind äquivalent:*

1. *f ist ein Hurwitz–Polynom.*
2. *Die Polynome g und h bilden ein positives Paar*[117].
3. *Es gibt eine Zahl $c_0 \ge 0$ und positive Zahlen c_j, d_j, $j = 1, \ldots, m$, so daß*

$$\frac{h(x)}{g(x)} = [c_0; d_1 x, c_1, d_2 x, c_2, \ldots, d_m x, c_m]\,, \tag{4.9}$$

wobei genau dann $c_0 = 0$ ist, wenn $\deg f$ ungerade ist.

[116] Das ist wieder einmal nur eine Normierungsfrage, den Nullstellen ist das Vorzeichen aber auch sowas von egal.

[117] Man beachte: der Grad von h kann hierbei kleiner sein als der von g.

Der Kettenbruch in (4.9) hat neben den positiven Koeffizienten auch eine amüsante Struktur zu bieten: es wechseln sich in den Teilnennern immer ein Polynom vom Grad 1 und ein Polynom vom Grad 0 ab. Um Satz 4.11 beweisen zu können, müssen wir natürlich ein bißchen mehr arbeiten, aber das Resultat sollte es uns wert sein. Bevor wir uns aber an die einzelnen Schritte des Beweises machen, wollen wir erst noch eine einfache Eigenschaft der Hurwitz Polynome festhalten, nämlich, daß alle Koeffizienten von f *strikt* dasselbe Vorzeichen haben müssen.

Lemma 4.12 *Sei f ein Hurwitz–Polynom vom Grad n und $f_n > 0$. Dann ist $f_j > 0$, $j = 0, \ldots, n$.*

Beweis: Wir faktorisieren f in

$$f(z) = f_n \prod_{j=1}^{n} (z - \zeta_j), \qquad \zeta_j \in \mathbb{H}_-.$$

Da in einem reellen Polynom alle Nullstellen auch konjugiert komplex auftreten müssen, enthält f entweder Faktoren der Form $(z + \alpha)$, $\alpha \in \mathbb{R}_+$, nämlich dann, wenn die Nullstelle reell ist oder aber von der Form

$$(z - \zeta)\left(z - \overline{\zeta}\right) = z^2 - \underbrace{\left(\zeta + \overline{\zeta}\right)}_{=\Re\zeta<0} z + \underbrace{\zeta\,\overline{\zeta}}_{=|\zeta|^2>0} = z^2 + \beta z + \gamma, \qquad \beta, \gamma \in \mathbb{R}_+,$$

so daß

$$f(z) = f_n \left[\prod_{j=0}^{k} (z + \alpha_j)\right] \left[\prod_{j=0}^{k'} \left(z^2 + \beta_j z + \gamma_j\right)\right]$$

nur positive Koeffizienten hat. □

4.6 Der Cauchy–Index und das Argumentenargument

Es wird Zeit, sich an die Sturmschen Ketten zu erinnern! Dabei haben wir für ein Intervall $I = [a, b]$ die Anzahl der *gewichteten* Vorzeichenwechsel $\Sigma_a^b f = \sigma(f, [a, b])$ einer Funktion f untersucht. In unserem Beweis von Proposition 3.28 haben wir dabei eine rationale Funktion f betrachtet, die als Quotient von zwei aufeinanderfolgenden orthogonalen Polynomen definiert war. So eine rationale Funktion hat aber nicht nur Nullstellen des Zählers, sondern auch Nullstellen des Nenners, also Pole, und auch diese Pole ermöglichen Vorzeichenwechsel, jetzt aber von $\pm\infty$ nach $\mp\infty$. Und die Anzahl dieser *singulären Vorzeichenwechsel*[118] bezeichnet man als den *Cauchy–Index* $I_a^b f$ von f auf $[a, b]$, wobei die Vorzeichenwechsel von $-\infty$ nach $+\infty$ positiv, die von $+\infty$ nach $-\infty$ hingegen negativ gezählt werden. Mit anderen Worten:

$$I_a^b f := -\Sigma_a^b f^{-1}. \tag{4.10}$$

Es erfordert nicht viel Phantasie sich vorzustellen, daß auch der Cauchy–Index sehr viel mit Sturmschen Ketten zu tun haben wird. Aber um den Beweis wir in [7] durchzuführen, brauchen wir zuerst ein klein wenig Funktionentheorie, siehe z.B. [6, Theorem 2, S. 175].

[118] Also Vorzeichenwechsel mittels einer Singularität – der Begriff ist nicht Standard, erscheint mir aber angemessen.

Satz 4.13 (Argumentenprinzip) *Ist f analytisch auf einem Gebiet $D \subset \mathbb{C}$ und γ eine positiv orientierte stückweise glatte geschlossene Kurve in D, die ein Gebiet $\Omega \subset D$ einschließt, dann ist*

$$\frac{1}{2\pi}\,\Delta_\gamma \arg f(z) = \#\left\{z \in \Omega \;:\; f(z) = 0\right\}.$$

wobei Δ_γ die Anzahl der Veränderungen im Argument entlang γ bezeichnet.

Nun sei f ein Hurwitz–Polynom und für $R > 0$ betrachten wir das Integral über die Kurve γ, die aus dem Intervall $[-Ri, Ri]$ und dem Halbkreis mit Radius R in $\mathbb{H}_+$ besteht, siehe Abb. 4.6. Also ist

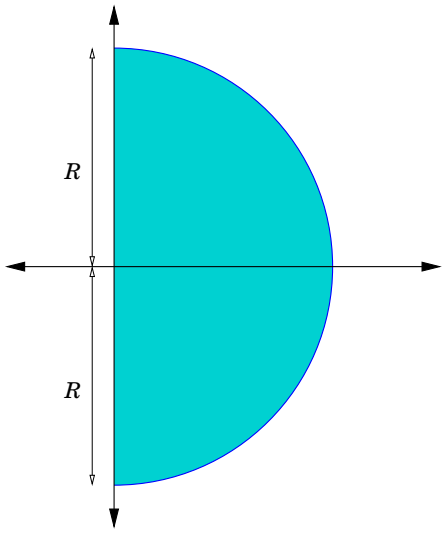


Abbildung 4.6: Das Integrationsgebiet, in dem sich keine Nullstellen nicht befinden, ganz egal, wie groß wir R wählen. Denn schließlich ist f ja ein Hurwitz–Polynom.

$$0 = \Delta_{-R}^{R} \arg f\,(ix) - \Delta_{-\pi}^{\pi} f\left(R\,e^{ix}\right).$$

Für hinreichend große Werte von R wird aber der Argumentenwechsel entlang des Halbkreises vom Leitterm $f_n\,z^n$ bestimmt, $n = \deg f$, und beträgt $n\pi$, also ist

$$\Delta_{-\infty}^{\infty} \arg f\,(ix) = \lim_{R\to\infty} \Delta_{-R}^{R} \arg f\,(ix) = n\pi. \tag{4.11}$$

Schreiben wir f in der etwas exzentrischen Form

$$f(z) = a_0\,z^n + b_0\,z^{n-1} + a_1\,z^{n2} + \cdots, \qquad a_0 \neq 0,$$

dann ist für $n = 2m$

$$\begin{aligned} f(ix) &= (-1)^m a_0\,x^n + i\,(-1)^{m-1}\,x^{n-1} + (-1)^{m-1} a_1\,x^{n-2} + \cdots \\ &= (-1)^m\left(a_0\,x^n - a_1\,x^{n-2} + a_2\,x^{n-4} + \cdots\right) + i\,(-1)^{m-1}\left(b_0\,x^{n-1} - b_1\,x^{n-3} + \cdots\right) \end{aligned}$$

und für $n = 2m + 1$

$$f(ix) = (-1)^m \left(b_0\, x^{n-1} - b_1\, x^{n-3} + \cdots\right) + i\,(-1)^m \left(a_0\, x^n - a_1\, x^{n-2} + \cdots\right),$$

in beiden Fällen ist also

$$f(ix) = p(x) + i\, q(x), \qquad x \in \mathbb{R}, \tag{4.12}$$

wobei

$$p(x) = \begin{cases} (-1)^m \left(a_0\, x^n - a_1\, x^{n-2} + \cdots + (-1)^m a_m\right), & n = 2m, \\ (-1)^m \left(b_0\, x^{n-1} - b_1\, x^{n-3} + \cdots + (-1)^m b_m\right), & n = 2m+1, \end{cases} \tag{4.13}$$

und

$$q(x) = \begin{cases} (-1)^{m-1} \left(b_0\, x^{n-1} - b_1\, x^{n-3} + \cdots + (-1)^{m-1}\, b_{m-1}\, x\right), & n = 2m, \\ (-1)^m \left(a_0\, x^n - a_1\, x^{n-2} + \cdots + (-1)^m\, a_m\, x\right), & n = 2m+1. \end{cases} \tag{4.14}$$

Das *Argument*[119] $\theta =: \arg z$ ist ja definiert durch

$$\Re z + i\,\Im z = z = |z|\, e^{i\theta} = |z|\,(\cos\theta + i\,\sin\theta) \qquad \Rightarrow \qquad \begin{cases} \cos\theta = \Re z/|z| \\ \sin\theta = \Im z/|z| \end{cases}$$

und somit ist

$$\tan\theta = \frac{\Im z}{\Re z}, \quad \cot\theta = \frac{\Re z}{\Im z} \qquad \Rightarrow \qquad \theta = \arctan\frac{\Im z}{\Re z} = \operatorname{arccot}\frac{\Re z}{\Im z}.$$

Angewandt auf (4.12) bedeutet das also, daß

$$\arg f(ix) = \arctan\frac{q(x)}{p(x)} = \operatorname{arccot}\frac{p(x)}{q(x)}$$

Nun entspricht aber jedes Inkrement des Arguments, also jeder "Umlauf" von $f(ix)$, einer Singularität des Tangens und deswegen ist

$$\frac{1}{\pi}\,\Delta_{-\infty}^{\infty} \arg f(ix) = \begin{cases} I_{-\infty}^{\infty} \dfrac{p(x)}{q(x)}, & n = 2m+1, \\ -I_{-\infty}^{\infty} \dfrac{q(x)}{p(x)}, & n = 2m, \end{cases}$$

und somit erhalten wir für Hurwitz–Polynoms unter Berücksichtigung von (4.11) die Charakterisierung

$$n = I_{-\infty}^{\infty} \frac{b_0\, x^{n-1} - b_1\, x^{n-3} + \cdots}{a_0\, x^n - a_1 x^{n-2} + \cdots} = -\Sigma_{-\infty}^{\infty} \frac{a_0\, x^n - a_1 x^{n-2} + \cdots}{b_0\, x^{n-1} - b_1\, x^{n-3} + \cdots}. \tag{4.15}$$

[119] Vielleicht hätte man den Begriff ja definieren sollen, bevor man sein Prinzip einführt?

Jetzt kehren wir wieder zurück zu unserer guten alten Zerlegung $f(z) = g(z^2) + z\,h(z^2)$ und betrachten zuerst einmal den Fall $n = 2m$. Dann ist

$$g(x) = f_n\,x^m + f_{n-2}\,x^{m-1} + \cdots + f_0, \qquad h(x) = f_{n-1}\,x^{m-1} + f_{n-3}\,x^{m-2} + \cdots + f_1, \tag{4.16}$$

also[120]

$$g\left(-z^2\right) = (-1)^m\left(a_0\,z^n - a_1\,z^{n-2} + \cdots\right), \qquad h\left(-z^2\right) = (-1)^m\left(b_0\,z^{n-2} - b_1\,z^{n-4} + \cdots\right),$$

womit wir also dank (4.15) bei

$$n = -I_{-\infty}^{\infty}\frac{z\,h\left(-z^2\right)}{g\left(-z^2\right)} \tag{4.17}$$

ankommen. Im Fall $n = 2m+1$ haben wir entsprechend

$$g(x) = f_{n-1}\,x^m + f_{n-3}\,x^{m-1} + \cdots + f_0, \qquad h(x) = f_n\,x^m + f_{n-2}\,x^{m-1} + \cdots + f_1 \tag{4.18}$$

und

$$n = -I_{-\infty}^{\infty}\frac{g\left(-z^2\right)}{z\,h\left(-z^2\right)}. \tag{4.19}$$

Als nächstes brauchen wir eine Eigenschaft des Cauchy–Index, die eigentlich nichts anderes als eine Umformulierung von Satz 3.27 ist.

Lemma 4.14 *Sei $a < c < b$ und ϕ eine beliebige Funktion. Dann ist*

$$I_a^b\phi = I_a^c\phi + I_c^b\phi + \eta_c\phi,$$

wobei

$$\eta_c\phi = \begin{cases} 1 \\ -1 \\ 0 \end{cases} \quad \textit{falls} \quad \lim_{x\to c_-}\phi(x) \begin{cases} = +\infty \\ = -\infty \\ \textit{sonst}. \end{cases}$$

Beweis: Ersetzt man ϕ durch ϕ^{-1} dann entspricht der Cauchy–Index dem Zählen der Vorzeichenwechsel wie in Satz 3.27 – singuläre Vorzeichenwechsel von ϕ sind "normale" Vorzeichenwechsel von ϕ^{-1} und umgekehrt. Ist nun gerade an c so ein Vorzeichenwechsel, dann wird der von den beiden "Teilindizes" nicht erkannt und muss durch das η_c explizit hinzugefügt werden. □

Damit können wir also (4.17) folgendermaßen entwickeln[121]

$$\begin{aligned} n &= -I_{-\infty}^{\infty}\frac{z\,h\left(-z^2\right)}{g\left(-z^2\right)} = -\left(I_{-\infty}^{0} + I_0^{\infty}\right)\frac{z\,h\left(-z^2\right)}{g\left(-z^2\right)} = -2\,I_{-\infty}^{0}\frac{z\,h\left(-z^2\right)}{g\left(-z^2\right)} \\ &= 2\,I_{-\infty}^{0}\frac{h\left(-z^2\right)}{g\left(-z^2\right)} = 2\,I_{-\infty}^{0}\frac{h(x)}{g(x)} = I_{-\infty}^{0}\frac{h(x)}{g(x)} - I_{-\infty}^{0}\frac{x\,h(x)}{g(x)} \\ &= I_{-\infty}^{0}\frac{h(x)}{g(x)} - I_{-\infty}^{0}\frac{x\,h(x)}{g(x)} + \underbrace{I_0^{\infty}\frac{h(x)}{g(x)} - I_0^{\infty}\frac{x\,h(x)}{g(x)}}_{=0} = I_{-\infty}^{\infty}\frac{h(x)}{g(x)} - I_{-\infty}^{\infty}\frac{x\,h(x)}{g(x)}. \end{aligned}$$

[120] Da $a_j = f_{n-2j}$ und $b_j = f_{n-1-2j}$ ist.

[121] Dabei ist zu beachten, daß der Faktor z im Zähler für den Cauchy–Index irrelevant ist, da das Nennerpolynom g ja $g(0) = f_0 \neq 0$ erfüllt, also kein η_0–Term auftreten kann.

Für $n = 2m + 1$ ergibt sich analog

$$n = I_{-\infty}^{\infty} \frac{g(x)}{x\,h(x)} - I_{-\infty}^{\infty} \frac{g(x)}{h(x)},$$

und somit

$$n = \begin{cases} I_{-\infty}^{\infty} \dfrac{h(x)}{g(x)} - I_{-\infty}^{\infty} \dfrac{x\,h(x)}{g(x)}, & n = 2m, \\ I_{-\infty}^{\infty} \dfrac{g(x)}{x\,h(x)} - I_{-\infty}^{\infty} \dfrac{g(x)}{h(x)}, & n = 2m+1. \end{cases} \tag{4.20}$$

Damit können wir auch schon einen Teil von Satz 4.11 angehen, und diese Aussage hat sogar einen eigenen Namen[122].

Satz 4.15 (Hermite–Biehler) *Ein Polynom $f(z) = g(z^2) + z\,h(z^2)$ ist genau dann ein Hurwitz–Polynom wenn g und h ein positives Paar bilden.*

Beweis: Wir haben bereits gezeigt, daß f genau dann ein Hurwitz–Polynom ist, wenn (4.20) erfüllt ist. Nun müssen wir wohl oder übel zwei Fälle unterscheiden:
$n = 2m$: Das Polynom g im Nenner hat Grad m und damit höchstens m Nullstellen. Damit[123] muß wegen

$$2m = I_{-\infty}^{\infty} \frac{h(x)}{g(x)} - I_{-\infty}^{\infty} \frac{x\,h(x)}{g(x)} \qquad \Rightarrow \qquad I_{-\infty}^{\infty} \frac{h(x)}{g(x)} = -I_{-\infty}^{\infty} \frac{x\,h(x)}{g(x)} = m$$

der Quotient $h(x)/g(x)$ nur singuläre Vorzeichenwechsel von $-\infty$ nach $+\infty$, der Quotient $x\,h(x)/g(x)$ hingegen nur singuläre Vorzeichenwechsel von $+\infty$ nach $-\infty$ haben. Das ist aber genau dann möglich, wenn alle diese Sprünge an negativen x passieren und wenn zwischen je zwei solchen Sprüngen ein Vorzeichenwechsel, also eine Nullstelle von h liegt. Nun hat g aber gerade m solche Nullstellen $x_1, \ldots, x_m$ und h andererseits $m-1$ Nullstellen $x_1', \ldots, x_{m-1}'$ und nach dem, was wir gerade gezeigt haben, müssen sich diese als

$$x_1 < x_1' < x_2 < x_2{}' < \cdots < x_{m-1}' < x_m < 0$$

anordnen lassen. Nach (4.16) und Lemma 4.12 können wir außerdem davon ausgehen, daß g und h beide positiven Leitkoeffizienten haben[124] und somit sind sie ein positives Paar. Die Umkehrung erhält man, indem man die Beweisschritte einfach rückwärts durchgeht.
$n = 2m + 1$: Nun müssen die $n = 2m + 1$ singulären Vorzeichenwechsel dadurch erreicht werden, daß wir $m + 1$ Vorzeichenwechsel von $x\,h(x)$ und m Vorzeichenwechsel von $h(x)$ mit

[122] Um genau zu sein: laut [7] ist der folgende Satz ein *Spezialfall* des Hermite–Biehler–Theorems.

[123] Dieses Argument hatten wir schon einmal, nämlich beim Beweis von Proposition 3.28, als wir gezeigt haben, daß orthogonale Polynome die maximale Anzahl an reellen Nullstellen haben.

[124] Die Koeffizienten f_n und f_{n-1} müssen dasselbe Vorzeichen haben und wären sie negativ, dann multiplizieren wir halt f, g, h alle mit -1.

entgegengesetzten Paritäten haben. Das heißt aber nichts anderes, daß die $m+1$ Vorzeichenwechsel von $x\,h(x)$ an den Stellen $x'_1 < \cdots < x'_m < 0$ und eben an 0 erfolgen müssen[125] und zwischen diesen Vorzeichenwechseln müssen nun mit demselben Argument wie oben wieder Vorzeichenwechsel von g liegen, also

$$x'_1 < x_1 < x'_2 < \cdots < x'_m < x_m < 0,$$

wie behauptet. □

Aus der Identität (4.20) die ja dazu äquivalent ist, daß f ein Hurwitz–Polynom ist, bzw. g und h ein positives Paar bilden, kann man auch noch eine weitere Schlußfolgerung ziehen.

Proposition 4.16 *Zwei Polynome g und h, $\deg g = m$, bilden genau dann ein positives Paar, wenn*

$$m = I^{\infty}_{-\infty}\frac{h(x)}{g(x)} = -I^{\infty}_{-\infty}\frac{x\,h(x)}{g(x)} \tag{4.21}$$

und wenn im Fall $\deg g = \deg h$ zusätzlich noch

$$\epsilon_{\infty} = \lim_{x\to+\infty} \operatorname{sgn}\frac{h(x)}{g(x)} = 1 \tag{4.22}$$

ist.

Beweis: Daß (4.21) für $n = 2m$ direkt aus (4.20) folgt, haben wir ja schon gesehen, um aber auch für $n = 2m+1$ von (4.20) zu der Aussage von Proposition 4.16 zu gelangen, brauchen wir eine Identität für den Cauchy–Index einer rationalen Funktion f, deren Zählergrad größer als der Nennergrad ist, nämlich

$$I^{\infty}_{-\infty} f(x) + I^{\infty}_{-\infty} f^{-1}(x) = \frac{\epsilon_{\infty} - \epsilon_{-\infty}}{2}, \qquad \epsilon_{\pm\infty} = \lim_{x\to\pm\infty} \operatorname{sgn} f(x), \tag{4.23}$$

In der Tat ist der Ausdruck auf der linken Seite ja nichts anderes als die Anzahl aller singulären Vorzeichenwechsel von f zusammen mit den Vorzeichenwechseln von f, und diese summieren sich zu gerade zu 1 wenn $\epsilon_{\infty} = 1$ und $\epsilon_{-\infty} = -1$, zu -1, wenn die Vorzeichen andersrum verteilt sind und zu 0, wenn $\epsilon_{\infty} = \epsilon_{-\infty}$ ist.

Mit (4.23) können wir nämlich jetzt die zweite Zeile von (4.20) in

$$2m+1 = n = I^{\infty}_{-\infty}\frac{g\,(x)}{x\,h\,(x)} - I^{\infty}_{-\infty}\frac{g\,(x)}{h\,(x)} = I^{\infty}_{-\infty}\frac{h(x)}{g(x)} - \frac{1-1}{2} - I^{\infty}_{-\infty}\frac{x\,h(x)}{g(x)} + \frac{1+1}{2}$$

umschreiben, was uns also auch wieder (4.21) liefert. Daß die Leitkoeffizienten von g und h gleiches Vorzeichen haben[126], folgt für $n = 2m$, und damit $\deg h = \deg g - 1$, direkt aus (4.21), für $n = 2m+1$, also $\deg h = \deg g$, benötigt man eben die zusätzliche Annahme (4.22). □

Auf dem Weg zum Beweis des zweiten Teils von Satz 4.11 brauchen wir noch die folgende Hilfsaussage.

[125] Denn $x = 0$ ist ja die einzige Nullstelle, die beim Übergang von $h(x)$ zu $xh(x)$ dazukommt und letztere Funktion hat eine Nullstelle mehr.

[126] Was ja eine Bedingung für ein positives Paar ist!

Lemma 4.17 *Angenommen, die beiden Polynome g und h, $\deg g = m$ bilden ein positives Paar*[127] *und es gibt Konstanten c, d und Polynome $g_1, h_1 \in \Pi_{m-1}$, so daß*

$$\frac{h(x)}{g(x)} = c + \cfrac{1}{dx + \cfrac{g_1(x)}{h_1(x)}} = \left[c; dx, \frac{g_1(x)}{h_1(x)}\right]. \tag{4.24}$$

Dann sind c, d sowie g_1, h_1 eindeutig durch g, h bestimmt und es gilt:

1. *$c \geq 0$, $d > 0$,*
2. *$\deg g_1 = \deg h_1 = m - 1$,*
3. *g_1 und h_1 bilden ein positives Paar.*

Erfüllen umgekehrt die Zahlen c, d und die Polynome g_1, h_1 die obigen drei Bedingungen, und sind g, h durch (4.24) definiert, dann bilden g und h ein positives Paar.

Beweis: Wenn g, h ein positives Paar, dann hat g insbesondere m reelle Nullstellen wir erhalten unter Verwendung von (4.24) daß[128]

$$m = I_{-\infty}^{\infty} \frac{h(x)}{g(x)} = I_{-\infty}^{\infty} \left[c + \cfrac{1}{dx + \cfrac{g_1(x)}{h_1(x)}}\right] = I_{-\infty}^{\infty} \frac{h_1(x)}{dx\, h_1(x) + g_1(x)}. \tag{4.25}$$

Damit muß aber der Nenner ein Polynom vom Grad n sein, also $d \neq 0$ und $\deg h_1 = m - 1$, denn sonst kämen wir beim besten Willen über Grad $m - 1$ nicht hinaus. Wir können außerdem ohne Einschränkung annehmen, daß der Leitterm von h_1 positiv ist[129]. Nun sagt uns aber (4.25), daß beide rationale Funktionen, $h(x)/g(x)$ wie auch $h_1(x)/\left(dx\, h_1(x) + g_1(x)\right)$, ihr maximale Anzahl von singulären Vorzeichenwechseln von $-$ nach $+$ haben und somit für hinreichend kleines x strikt *negativ*, für hinreichend großes x hingegen strikt *positiv* sind. Damit ist

$$-1 = -\operatorname{sgn} d = \lim_{x \to -\infty} \frac{h_1(x)}{dx\, h_1(x) + g_1(x)}, \qquad 1 = \operatorname{sgn} d = \lim_{x \to -\infty} \frac{h_1(x)}{dx\, h_1(x) + g_1(x)},$$

woraus $d > 0$ folgt. Nach (4.25) hat h/g genau m singuläre Vorzeichenwechsel von $-\infty$ nach $+\infty$, zwischen denen wieder $m-1$ Vorzeichenwechsel von $+$ nach $-$ liegen müssen, und somit ist

$$-I_{-\infty}^{\infty} \left[d\, x + \frac{g_1(x)}{h_1(x)}\right] \geq m - 1; \tag{4.26}$$

[127] Das heißt insbesondere, daß $\deg h \in \{m - 1, m\}$.

[128] Hier erweist sich der Cauchy–Index als hilfreich und nützlich: im Gegensatz zu "normalen" Vorzeichenwechseln lassen sich singuläre Vorzeichenwechsel von Konstanten, die man zur Funktion addiert, nicht beeindrucken.

[129] Ansonsten mulitplizieren wir g_1 und h_1 beide mit -1.

da dieser Cauchy–Index höchstens $1-m$ sein kann[130] gilt in (4.26) Gleichheit und daher

$$m-1 = -I_{-\infty}^{\infty}\left[d\,x+\frac{g_1(x)}{h_1(x)}\right] = -I_{-\infty}^{\infty}\frac{g_1(x)}{h_1(x)}. \tag{4.27}$$

Aus der zweiten Identität in (4.21) sehen wir außerdem, daß

$$\begin{aligned} m &= -I_{-\infty}^{\infty}\frac{x\,h(x)}{g(x)} = -I_{-\infty}^{\infty}\left[c\,x+\frac{x}{d\,x+\dfrac{g_1(x)}{h_1(x)}}\right] = -I_{-\infty}^{\infty}\left[c\,x+\frac{1}{d+\dfrac{g_1(x)}{x\,h_1(x)}}\right] \\ &= -I_{-\infty}^{\infty}\left[\frac{1}{d+\dfrac{g_1(x)}{x\,h_1(x)}}\right] = I_{-\infty}^{\infty}\left[d+\frac{g_1(x)}{x\,h_1(x)}\right] = I_{-\infty}^{\infty}\frac{g_1(x)}{x\,h_1(x)} \end{aligned} \tag{4.28}$$

und somit ist auch $\deg g = m-1$, weil wieder einmal zwischen jeder Sprungstelle ein Vorzeichenwechsel liegen muss. Damit ist Punkt 2 auch schon erledigt.

Da die beiden Polynome g_1, h_1 denselben Grad haben, ist

$$\lim_{x\to\pm\infty}\frac{g_1(x)}{h_1(x)} = \mu \neq 0 \quad\Rightarrow\quad \lim_{x\to\pm\infty} d\,x+\frac{g_1(x)}{h_1(x)} = \pm\infty \quad\Rightarrow\quad \lim_{x\to\pm\infty}\frac{1}{d\,x+\dfrac{g_1(x)}{h_1(x)}} = 0$$

und somit nach (4.24)

$$c = \lim_{x\to\infty}\left[\frac{h(x)}{g(x)}-\frac{1}{d\,x+\dfrac{g_1(x)}{h_1(x)}}\right] = \lim_{x\to\infty}\frac{h(x)}{g(x)}\begin{cases} >0, & \deg g = \deg h, \\ =0, & \deg g > \deg h, \end{cases}$$

und damit ist auch Behauptung 1 bewiesen.

Fehlt noch, daß g_1 und h_1 wirklich ein positives Paar bilden. Dazu wenden wir (4.23) auf (4.28) an und erhalten, daß

$$I_{-\infty}^{\infty}\frac{x\,h_1(x)}{g_1(x)} = -m+\frac{\epsilon_\infty-\epsilon_{-\infty}}{2} = -m+\epsilon_\infty, \tag{4.29}$$

da

$$\lim_{x\to+\infty}\operatorname{sgn}\frac{h_1(x)}{g_1(x)} = \epsilon_\infty := \lim_{x\to+\infty}\operatorname{sgn}\frac{x\,h_1(x)}{g_1(x)} = -\lim_{x\to-\infty}\operatorname{sgn}\frac{x\,h_1(x)}{g_1(x)} = -\epsilon_{-\infty}.$$

Normieren wir also g_1 und h_1 so, daß $\epsilon_\infty = 1$ ist, dann liefert uns das zusammen mit (4.27) und (4.29) genau das, was wir brauchen, um Proposition 4.16 anwenden zu können – und siehe da, g_1 und h_1 bilden wirklich ein positives Paar.

[130] Schließlich ist ja $\deg h_1 = m-1$.

Für die Umkehrung liest man wieder alle Argumente in umgekehrter Reihenfolge – wir haben ja entweder Identitäten oder Charakterisierungen verwenden. □

Mit diesem Lemma ist der Beweis von Satz 4.11 kein großes Hexenwerk mehr, denn schließlich zeigt es uns ja, daß positive Paare unter eine "Doppelschritt" der Kettenbruchzerlegung in positive Paare überführt werden und umgekehrt. Und tatsächlich ist Satz 4.11 nur noch eine Kombination von Hermite–Biehler, Satz 4.15, und dem folgenden Resultat.

Satz 4.18 *Zwei Polynome g und h, $\deg g = m$, bilden genau dann ein positives Paar, wenn es*

$$c_0 \begin{cases} > 0, & \deg g = \deg h, \\ = 0, & \deg g = \deg h + 1, \end{cases} \qquad c_j, d_j \in \mathbb{R}_+, \quad j = 1, \dots, m,$$

gibt, so daß

$$\frac{h(x)}{g(x)} = [c_0; d_1\, x, c_1, \dots, d_m\, x, c_m] \tag{4.30}$$

ist.

Beweis: Dank Lemma 4.17 brauchen wir nur noch zu zeigen, daß es zu jedem positiven Paar g, h eine Zerlegung mit g_1, h_1 wie in (4.24) gibt. Ist nun $m = \deg g = \deg h$, dann können wir h mit Rest h_1 durch g teilen, also $h = c_0\, g + h_1$, wobei sogar $c_0 > 0$ ist[131] und $\deg h_1 = m - 1$. Damit ist

$$\frac{h(x)}{g(x)} = \frac{c\, g(x) + h_1(x)}{g(x)} = c_0 + \frac{h_1(x)}{g(x)} = c_0 + \frac{1}{\dfrac{g(x)}{h_1(x)}}.$$

Nun ist $\deg g = m = \deg h_1 + 1$, also $g(x) = d_1\, x\, h_1(x) + g_1(x)$, $\deg g_1 \le m - 1$, und damit

$$\frac{h(x)}{g(x)} = c_0 + \frac{1}{\dfrac{d_1 x\, h_1(x) + g_1(x)}{h(x)}} = c_0 + \frac{1}{d_1\, x + \dfrac{g_1(x)}{h_1(x)}}$$

und nach Lemma 4.17 ist $d_1 > 0$ und $\deg g_1 = \deg h_1 = m - 1$. Für $\deg h = \deg g - 1$ gilt genau dasselbe, nur eben mit $c = 0$ und daher $h_1 = h$. Was wir also gezeigt haben ist, daß in beiden Fällen

$$\frac{h(x)}{g(x)} = c_0 + \frac{1}{dx + \dfrac{1}{h_1(x)/g_1(x)}} = \left[c_0; d_1\, x, \frac{h_1(x)}{g_1(x)}\right], \qquad \deg g_1 = \deg h_1 = m - 1, \tag{4.31}$$

ist. Nun können wir h_1/g_1 aber als $\left[c_1; d_2\, x, \frac{h_2(x)}{g_2(x)}\right]$ mit $\deg g_2 = \deg h_2 = m - 2$ schreiben. Iterieren wir das in (4.31), dann erhalten wir, daß

$$\frac{h(x)}{g(x)} = \left[c_0; d_1\, x, c_1, \dots, d_j\, x, \frac{h_j(x)}{g_j(x)}\right], \qquad \deg g_j = \deg h_j = m - j, \qquad j = 1, \dots, m, \tag{4.32}$$

und $j = m$ zusammen mit $g_m, h_m \neq 0$ liefert uns $c_m \neq 0$ und somit (4.30). Die Umkehrung ist einfaches Ausmultiplizieren des Kettenbruchs. □

[131] Als positives Paar haben die Leitkoeffizienten von g und h dasselbe Vorzeichen!

4.7 Der Satz von Routh–Hurwitz

Der Satz von Routh–Hurwitz[132] liefert uns eine weitere Charakterisierung von von Hurwitz–Polynomen, diesmal über bestimmte Determinanten. Und da zu Determinanten immer Matrizen gehören[133] beginnen wir mit diesen.

Definition 4.19 *Sei $p \in \Pi$ ein Polynom vom Grad n. Die* Hurwitz–Matrix *zu p ist die $n \times n$–Matrix*

$$H_p := \begin{bmatrix} p_{n-1} & p_{n-3} & p_{n-5} & \dots & 0 \\ p_n & p_{n-2} & p_{n-4} & \dots & 0 \\ 0 & p_{n-1} & p_{n-3} & \dots & 0 \\ 0 & p_n & p_{n-2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & p_0 \end{bmatrix} \tag{4.33}$$

Beispiel 4.20 *Sehen wir uns doch mal ein paar Beispiele von Hurwitz–Matrizen an, nämlich für kleine Werte von n und $p(x) = p_0 + \dots + p_n\, x^n$:*

$n = 1$**:** *Für hier haben wir lediglich die 1×1–Matrix $H_p = [p_0]$.*

$n = 2$ *Die Hurwitz–Matrix ist in diesem Fall*

$$H_p = \begin{bmatrix} p_1 & 0 \\ p_2 & p_0 \end{bmatrix}$$

und enthält zum ersten Mal eine Null.

$n = 3$**:** *Jetzt erkennt man schön langsam ein bißchen mehr von der Struktur:*

$$H_p = \begin{bmatrix} p_2 & p_0 & 0 \\ p_3 & p_1 & 0 \\ 0 & p_2 & p_0 \end{bmatrix}$$

$n = 4$**:** *Liefert noch ein bißchen mehr Struktur*

$$H_p = \begin{bmatrix} p_3 & p_1 & 0 & 0 \\ p_4 & p_2 & p_0 & 0 \\ 0 & p_3 & p_1 & 0 \\ 0 & p_4 & p_2 & p_0 \end{bmatrix}$$

[132] Und hier ist nicht der Satz *"A PhD dissertation is a paper of the professor written under aggravating circumstances"* gemeint, der in [20] A. Hurwitz zugeschrieben wird.

[133] Oder war es umgekehrt?

Die Beispiele zeigen uns, daß wir wieder einmal zwischen geraden und ungeraden Werten von n unterscheiden müssen, und zwar

$$H_p = \begin{bmatrix} p_{n-1} & \dots & p_3 & p_1 & 0 & 0 & \dots & 0 & 0 \\ p_n & \dots & p_4 & p_2 & p_0 & 0 & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & p_{n-1} & p_{n-3} & p_{n-5} & \dots & p_1 & 0 \\ 0 & \dots & 0 & p_n & p_{n-2} & p_{n-4} & \dots & p_2 & p_0 \end{bmatrix}, \qquad n = 2m, \tag{4.34}$$

bzw.

$$H_p = \begin{bmatrix} p_{n-1} & \dots & p_2 & p_0 & 0 & \dots & 0 \\ p_n & \dots & p_3 & p_1 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & p_{n-1} & p_{n-3} & \dots & p_0 \end{bmatrix}, \qquad n = 2m+1. \tag{4.35}$$

Was wir jetzt noch brauchen ist der Begriff der Minore.

Definition 4.21 *Sei $A \in \mathbb{R}^{n\times n}$ und $I \subset \{1,\dots,n\}$. Die I–*Minore *von A ist definiert als*

$$m_I(A) = \det A(I,I) = \det\left[a_{jk} \ : \ j,k \in I\right],$$

und die j–te Hauptminore *als*

$$m_j(A) = m_{\{1,\dots,j\}}(A) = \det\left[a_{k\ell} \ : \ k,\ell = 1,\dots,j\right].$$

Satz 4.22 (Routh–Hurwitz) *Ein Polynom $f \in \Pi$ mit positivem Leitkoeffizient*[134] *ist genau dann ein Hurwitz–Polynom, wenn*

$$m_k\left(H_f\right) > 0, \qquad j = 1,\dots,\deg f. \tag{4.36}$$

Bevor wir uns an den Beweis machen, sehen wir uns mal die ersten Spezialfälle an: Für $n = 1$, also ein Polynom $f(x) = f_1 x + f_0$, $f_1 > 0$, sagt uns Satz 4.22, daß f genau dann ein Hurwitz–Polynom ist, wenn $0 < m_1\left(H_f\right) = f_0$, was sich leicht verifizieren läßt, da

$$f(x) = 0 \qquad \Leftrightarrow \qquad x = -\frac{f_0}{f_1}$$

ist. Ein bißchen interessanter wird es schon für $n = 2$ wo die Positivität der Hauptminoren von

$$H_f = \begin{bmatrix} f_1 & 0 \\ f_2 & f_0 \end{bmatrix} \qquad \text{zu} \qquad 0 < f_1, \quad 0 < f_0\, f_1 \qquad \Leftrightarrow \qquad 0 < f_0, f_1$$

führt. Und tatsächlich sind ja die Nullstellen von f die Werte

$$x = \frac{-f_1 \pm \sqrt{f_1^2 - 4 f_0 f_2}}{2 f_0} \qquad \Rightarrow \qquad \Re x < 0 \text{ für } 0 < f_0, f_1, f_2,$$

[134] Das ist bekanntlich der Koeffizient vor dem Monom höchster Ordnung, also $f_{\deg f}$.

da die Wurzel genau dann entweder imaginär oder $< f_1$ ist, wenn $f_0 f_2 > 0$, also $f_2 > 0$ ist. Also können wir auch hier das Routh–Hurwitz–Kriterium "zu Fuß" verfizieren. Interessanter wird es dann schon im Fall $n = 3$, wo alle Hauptminoren der Matrix

$$M_f = \begin{bmatrix} f_2 & f_0 & 0 \\ f_3 & f_1 & 0 \\ 0 & f_2 & f_0 \end{bmatrix}$$

positiv sein müssen, was zu $f_0, f_2 > 0$ und $f_1 f_2 > f_0 f_3$ äquivalent[135] ist.

4.8 Das Routh–Schema oder die Rückkehr der Sturmschen Kette

Der Ausgangspunkt für den Beweis von Satz 4.22 ist die Charakterisierung (4.15) der Hurwitz–Polynome, also

$$n = I_{-\infty}^{\infty} \frac{b_0\, x^{n-1} - b_1\, x^{n-3} + \cdots}{a_0\, x^n - a_1 x^{n-2} + \cdots} =: I_{-\infty}^{\infty} \frac{f_1(x)}{f_0(x)}. \tag{4.37}$$

Mit diesen beiden Polynomen, die keine gemeinsame Nullstelle haben[136], können wir nun eine Folge von Polynomen $f_2, \ldots, f_m$ durch Division mit Rest wie folgt konstruieren:

$$f_j(x) = q_j(x)\, f_{j+1}(x) - f_{j+2}, \qquad \deg f_{j+2} < \deg f_{j-1}. \tag{4.38}$$

Das ist der gute alte euklidische Algrithmus, der uns einen alten Bekannten liefert.

Lemma 4.23 *Sind f_0, f_1 zwei Polynome ohne gemeinsame Nullstelle und ist $f_m \in \Pi_0 \setminus \{0\}$ in der durch (4.38) gebildeten Folge, dann bilden $f_0, \ldots, f_m$ eine Sturmsche Kette*[137].

Beweis: Da die beiden Polynome keine gemeinsame Nullstelle haben, liefert der euklidische Algorithmus den größten gemeinsamen Teiler f_m als von Null verschiedene konstante Funktion. Was wir zeigen müssen, ist, daß an jeder Nullstelle von f_j die Polynome f_{j-1} und f_{j+1} umgekehrtes Vorzeichen haben; ersetzen wir in (4.38) j durch $j - 1$, dann liefert eine Umformung, daß an jeder Nullstelle x von f_j

$$0 = q_j(x)\, f_j(x) = f_{j-1}(x) + f_{j+1}(x)$$

sein muß – und damit ist wieder entweder $f_{j-1}(x) = f_{j+1}(x) = 0$ oder die beiden haben, wie gewünscht, unterschiedliches Vorzeichen. Wären andererseit aber $f_j(x) = f_{j+1}(x) = 0$, dann ist[138] nach (4.38) auch $f_{j+2}(x) = 0$ und, per Iteration, auch $f_m(x) = 0$, was natürlich nicht sein kann. □

[135] Aus der letzten Ungleichung folgt übrigens unmittelbar die Positivität von f_1.

[136] Den hätten sie eine, dann könnten wir den gemeinsamen linearen Faktor kürzen und das Nennerpolynom hätte in Wirklichkeit nur Grad $n - 1$ und damit auch maximal $n - 1$ Nullstellen, womit dann aber der Cauchy–Index $\leq n - 1$ wäre.

[137] Die jetzt im Gegensatz zu Definition 3.26 umgekehrt indiziert ist.

[138] Und dieses Argument sollte uns bekannt vorkommen – wir kennen es ja aus dem Beweis von Proposition 3.28.

Führen wir jetzt den euklidischen Algorithmus durch, dann erhalten wir die Polynome

$$\begin{aligned} f_2(x) &= \frac{a_0}{b_0} x\, f_1(x) - f_0(x) = c_0\, x^{n-2} - c_1\, x^{n-4} + \cdots \\ f_3(x) &= \frac{b_0}{c_0} x\, f_2(x) - f_1(x) = d_0\, x^{n-3} - d_1\, x^{n-5} + \cdots \\ f_j(x) &= a_0^j\, x^{n-j} - a_1^j\, x^{n-j-2} + \cdots = \frac{a_0^{j-2}}{a_0^{j-1}} x\, f_{j-1}(x) - f_{j-2}(x), \end{aligned} \tag{4.39}$$

wobei

$$a_k^0 = a_k, \qquad a_k^1 = b_k, \qquad a_k^j = \frac{a_0^{j-1}\, a_{k+1}^{j-2} - a_0^{j-2}\, a_{k+1}^{j-1}}{a_0^{j-1}}, \tag{4.40}$$

denn[139]

$$\begin{aligned} f_j(x) &= \frac{a_0^{j-2}}{a_0^{j-1}} x \left[\sum_{k=0}^{(n-j+1)/2} (-1)^k a_k^{j-1} x^{n-j+1-2k} \right] - \left[\sum_{k=0}^{(n-j)/2+1} (-1)^k a_k^{j-2} x^{n-j+2-2k} \right] \\ &= \sum_{k=1}^{(n-j)/2+1} (-1)^k \frac{a_0^{j-2}\, a_k^{j-1} - a_0^{j-1}\, a_k^{j-2}}{a_0^{j-1}}\, x^{n-j+2-2k} \\ &= \sum_{k=0}^{(n-j)/2} (-1)^k \underbrace{\frac{a_0^{j-1}\, a_{k+1}^{j-2} - a_0^{j-2}\, a_{k+1}^{j-1}}{a_0^{j-1}}}_{=a_k^j}\, x^{n-j-2k}. \end{aligned}$$

Prinzipiell kann es natürlich passieren, daß in einem Schritt

$$0 = a_0^j = \frac{a_0^{j-2}\, a_1^{j-1} - a_0^{j-1}\, a_1^{j-2}}{a_0^{j-1}}, \qquad a_0^{j-1} \neq 0$$

ist; in diesem Fall ersetzen wir a_1^{j-2} durch $a_1^{j-2} + \varepsilon$ mit einem hinreichend kleinen $\varepsilon > 0$. Selbst wenn wir das mehrfach machen würde, könnten wir letztendlich $\varepsilon \to 0$ gehen lassen. Dieser Prozess klappt, solange f keine Nullstelle auf der imaginären Achse hat, für Details siehe [7].

So können wir uns auf den *regulären* Fall beschränken, daß wir durch den Prozess (4.39) eine Sturmsche Kette der Länge n erhalten. Nun ist jedes Polynom mit geradem Index, $f_0, f_2, \ldots$, ein Polynom von derselben Parität[140] wie n und jedes mit ungeradem Index, $f_1, f_3, \ldots$, mit der entgegengesetzten Parität. Damit ist aber

$$\begin{aligned} V(-x) &= V\left(f_0(-x), f_1(-x), \ldots, f_{n-1}(-x), f_n(-x)\right) \\ &= \begin{cases} V\left(f_0(x), -f_1(x), \ldots, -f_{n-1}(x), f_n(x)\right), & n = 2m, \\ V\left(-f_0(x), f_1(x), \ldots, f_{n-1}(x), -f_n(x)\right), & n = 2m+1. \end{cases} \end{aligned}$$

[139] Hierbei sind Summationsgrenzen immer als ganzzahliger Anteil aufzufassen.

[140] Ein Polynom heißt *gerade*, wenn $f(-x) = f(x)$ und *ungerade*, wenn $f(-x) = -f(x)$.

und damit[141]

$$V(-\infty) + V(\infty) = n, \tag{4.41}$$

denn entweder gibt es von $f_j(\infty)$ nach $f_{j+1}(\infty)$ einen Vorzeichenwechsel oder von $\pm f_j(\infty)$ nach $\pm f_{j+1}(-\infty) = \mp f_{j+1}(\infty)$. Andererseits liefert uns (4.37) zusammen mit (4.10) und Satz 3.27, daß

$$n = I_{-\infty}^{\infty} \frac{f_1(x)}{f_0(x)} = -\Sigma_{-\infty}^{\infty} \frac{f_0(x)}{f_1(x)} = V(-\infty) - V(\infty),$$

also ist f genau dann ein Hurwitz–Polynom, wenn

$$0 = V(\infty) = V\left(a_0^j \,:\, j = 0, \ldots, n\right), \qquad n = V(-\infty). \tag{4.42}$$

Damit erhalten wir auch schon den folgenden Satz.

Satz 4.24 (Routh–Kriterium) *Das Polynomom $f(z)$ ist genau dann ein Hurwitz–Polynom, wenn alle Zahlen a_0^j, $j = 0, \ldots, n$, strikt dasselbe Vorzeichen haben.*

Bemerkung 4.25 *Nach (4.42) muß der Vektor, dessen Vorzeichenwechsel $V(\infty)$ bei einem Hurwitz–Polynom mindestens $n+1$ Einträge enthalten – wie sonst soll man auch auf n Vorzeichenwechsel kommen? Das heißt aber auch, daß der euklidische Algorithmus bei eine Hurwitz–Polynom keine "Sprünge" machen darf, daß also alle q_j gerade Grad 1 haben dürfen und nicht mehr. Oder, nochmals anders gesagt: Würden wir in (4.40) durch Null dividieren, dann hätten wir es auf keinen Fall mit einem Hurwitz–Polynom zu tun.*

Man kann nun die Koeffizienten[142] der Polynome $f_0, f_1, \ldots, f_n$ in einer Tabelle darstellen und erhält so das *Routh–Schema*

$$\begin{array}{ccc} a_0^0 & a_1^0 & \ldots \\ a_0^1 & a_1^1 & \ldots \\ \vdots & & \\ a_0^n & & \end{array}$$

das sich rekursiv über (4.40) bestimmen lässt. Das Routh–Kriterium aus Satz 4.24 sagt uns nun, daß wir Hurwitz–Polynome daran erkennen können, daß alle Einträge der ersten Spalte des Routh–Schema strikt dasselbe Vorzeichen haben, und das ist nun wirklich ein sehr einfaches Kriterium.

Beispiel 4.26 *Versuchen wir einmal, ein bißchen "Gefühl" für das Kriterium zu bekommen.*

1. *Für $n = 2$ und $f(z) = f_0 + f_1\, z + f_2\, z^2$ erhalten wir, daß $a_0^0 = f_2$, $a_1^0 = f_0$ und $a_0^1 = f_1$, also*

$$a_0^2 = \frac{a_0^1\, a_1^0}{a_1^0},$$

und wir haben es wieder genau dann mit einem Hurwitz–Polynom zu tun, wenn f_0, f_1, f_2 strikt dasselbe Vorzeichen haben.

[141] Hier steht ∞ für ein x, das so groß ist, daß alle $f_j(x)$ ihr "ultimatives" Vorzeichen angenommen haben, also keine Nullstelle mehr rechts von diesem Punkt haben.

[142] In diesen Koeffizienten steckt immer noch das alternierende Vorzeichen

2. *Ein etwas komplexeres Beispiel aus [7], bei dem man auch das "ε–Argument" in Aktion sehen kann, ist das Polynom $f(z) = z^4 + z^3 + 2z^2 + 2z + 1$, das uns das Schema*

$$\begin{array}{cccc} 1 & 2 & 1 & \\ 1 & 2 & & \\ \varepsilon & 1 & & \leftarrow \quad 0 \quad 1 \\ 2 - \dfrac{1}{\varepsilon} & & & \\ 1 & & & \end{array}$$

der Länge n liefert. Allerdings ist f kein *Hurwitz–Polynom, denn wenn wir $\varepsilon > 0$ wählen*[143]*, dann haben wir eine Vorzeichenverteilung $+, +, +, -, +$, für negatives ε hingegen $+, +, -, +, +$ und in beiden Fällen ist $V(\infty) = 2$. Das zeigt übrigens, daß f zwei Nullstellen in $\mathbb{H}_+$ hat!*

Der Weg vom Routh–Schema zu Satz 4.22 ist nun erfreulich kurz! Dafür bemerken wir zuerst, daß die Hurwitz–Matrix von der Form

$$H_f = \begin{bmatrix} b_0 & -b_1 & b_2 & \dots \\ a_0 & -a_1 & a_2 & \dots \\ 0 & b_0 & -b_1 & \dots \\ 0 & a_0 & -a_1 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

ist. Nun ziehen wir wie bei der Gauß–Elimination[144] das a_0/b_0–fache der ersten Zeile von der dritten ab, dasselbe mit der zweiten und vierten Spalte und so weiter, was uns eine Matrix der Form

$$H'_f = \begin{bmatrix} b_0 & -b_1 & b_2 & \dots \\ 0 & c_0 & -c_1 & \dots \\ 0 & b_0 & -b_1 & \dots \\ 0 & 0 & c_0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \qquad c_k = \frac{b_0\, a_{k+1} - a_0\, b_{k+1}}{b_0}$$

liefert. Diese Formel für die c_k sollte uns aber bekannt vorkommen, denn sie ist genau (4.40) und somit erhalten wir

$$H_f^{(1)} := H'_f = \begin{bmatrix} a_0^1 & a_1^1 & a_2^1 & \dots \\ 0 & a_0^2 & a_1^2 & \dots \\ 0 & a_0^1 & a_1^1 & \dots \\ 0 & 0 & a_0^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

und die Sache fängt langsam an, Spass zu machen. Denn jetzt subtrahieren wir natürlich das a_0^1/a_0^2–fache der zweiten Zeile von der dritten und verfahren entsprechend auch mit der vierten

[143] Zur Erinnerung: $|\varepsilon|$ ist immer noch klein!

[144] Muß man hier noch Literaturhinweise geben?

und fünften Zeile, und so weiter. Wieder taucht die Rekursion (4.40) auf und liefert uns die Matrix

$$H_f^{(2)} = \begin{bmatrix} a_0^1 & a_1^1 & a_2^1 & \dots \\ 0 & a_0^2 & a_1^2 & \dots \\ 0 & 0 & a_0^3 & \dots \\ 0 & 0 & a_0^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Vorausgesetzt, daß wir nirgendwo durch Null dividieren müssen, endet diese Iteration bei der oberen Dreiecksmatrix

$$H_f^{(n)} = \begin{bmatrix} a_0^1 & \dots & * \\ & \ddots & \vdots \\ & & a_0^n \end{bmatrix}$$

und da wir von der k–ten Zeile nur Vielfache der Zeilen $1, \dots, k-1$ abgezogen haben, stimmen die Hauptminoren von H_f und H_f^n überein:

$$m_k\left(H_f\right) = m_k\left(H_f^{(n)}\right) = \prod_{j=1}^{k} a_0^j, \qquad k = 1, \dots, n. \tag{4.43}$$

Beweis von Satz 4.22: Nach Satz 4.24 ist $f(z)$ mit $a_0^0 = f_n > 0$ genau dann ein Hurwitz–Polynom, wenn $a_0^j > 0$ ist, $j = 1, \dots, n$, was nach (4.43) wiederum dazu äquivalent ist, daß alle Hauptminoren von $H_f^{(n)}$ und damit auch alle Hauptminoren von H_f positiv sind. □

Uns ist in alten mæren
wunders viel geseit
von Helden lobebæren
von grôzer arebeit

Das Nibelungenlied

4 Literatur

[1] O. Becker, *Quellen und Studien zur Geschichte*, Math., Astron., Physik **B2** (1933), 311–333.

[2] D. Bernoulli, *Disquisitiones ulteriores de idole fractionum continuarum*, N. C. Pet. **20** (1775).

[3] J. W. Cooley, *The re–discovery of the Fast Fourier Transform*, Mikrochimica Acta **3** (1987), 33–45.

[4] _____, *How the FFT gained acceptance*, A History of Scientific Computing (S. G. Nash, ed.), ACM–Press and Addison–Wesley, 1990, pp. 133–140.

[5] J. W. Cooley and J. W. Tukey, *An algorithm for machine calculation of complex Fourier series*, Math. Comp. **19** (1965), 297–301.

[6] S. D. Fisher, *Complex variables*, Wadsworth & Brooks, 1990, Dover Reprint 1999.

[7] F. R. Gantmacher, *Matrix Theory. Vol. II*, Chelsea Publishing Company, 1959, Reprinted by AMS, 2000.

[8] J. von zur Gathen and J. Gerhard, *Modern computer algebra*, Cambridge University Press, 1999.

[9] C. F. Gauss, *Methodus nova integralium valores per approximationem inveniendi*, Commentationes societate regiae scientiarum Gottingensis recentiores **III** (1816).

[10] W. Gautschi, *Numerical analysis. an introduction*, Birkhäuser, 1997.

[11] W. Gröbner, *Algebraische Geometrie I*, B.I–Hochschultaschenbücher, no. 273, Bibliographisches Institut Mannheim, 1968.

[12] D. Ch. von Grüningen, *Digitale Signalverarbeitung*, VDE Verlag, AT Verlag, 1993.

[13] R. W. Hamming, *Digital filters*, Prentice–Hall, 1989, Republished by Dover Publications, 1998.

[14] D. R. Hofstadter, *Gödel, escher, bach: ein endloses geflochtenes band*, Klett–Cotta, 1985.

[15] E. Isaacson and H. B. Keller, *Analysis of Numerical Methods*, John Wiley & Sons, 1966.

[16] K. D. Kammeyer and K. Kroschel, *Digitale Signalverarbeitung*, Teubner Studienbücher Elektrotechnik, B. G. Teubner, Stuttgart, 1998.

[17] A. Ya. Khinchin, *Continued fractions*, 3rd ed., University of Chicago Press, 1964, Reprinted by Dover 1997.

[18] D. E. Knuth, *The art of computer programming. seminumerical algorithms*, 3rd ed., Addison–Wesley, 1998.

[19] G. G. Lorentz, M. v. Golitschek, and Y. Makovoz, *Constructive approximation. Advanced problems*, Grundlehren der mathematischen Wissenschaften, vol. 304, Springer, 1996.

[20] MacTutor, *The MacTutor History of Mathematics archive*, `http://www-groups.dcs.st-and.ac.uk/~history`, 2003, University of St. Andrews.

[21] S. Mallat, *A wavelet tour of signal processing*, 2. ed., Academic Press, 1999.

[22] O. Perron, *Die Lehre von den Kettenbrüchen I*, 3rd ed., B. G. Teubner, 1954.

[23] ———, *Die Lehre von den Kettenbrüchen II*, 3rd ed., B. G. Teubner, 1954.

[24] C. Sagan, *Unser kosmos*, Droemersche Verlagsanstalt Th. Knaur Nachf., 1989, Deutsche Taschenbuchausgabe.

[25] T. Sauer, *Numerische Mathematik II*, Vorlesungsskript, Friedrich–Alexander–Universität Erlangen–Nürnberg, Justus–Liebig–Universität Gießen, 2000, `http://www.math.uni-giessen.de/tomas.sauer`.

[26] ———, *Computeralgebra*, Vorlesungsskript, Justus–Liebig–Universität Gießen, 2001, `http://www.math.uni-giessen.de/tomas.sauer`.

[27] ———, *Digitale Signalverarbeitung*, Vorlesungsskript, Justus–Liebig–Universität Gießen, 2003, `http://www.math.uni-giessen.de/tomas.sauer`.

[28] A. Schönhage and V. Strassen, *Schnelle Multiplikation großer Zahlen*, Computing **7** (1971), 281–292.

[29] H. W. Schüßler, *Digitale Signalverarbeitung*, 3. ed., Springer, 1992.

[30] C. E. Shannon, *A mathematical theory of communication*, Bell System Tech. J. **27** (1948), 379–423.

[31] ———, *Communications in the presence of noise*, Proc. of the IRE **37** (1949), 10–21.

[32] E. T. Whittaker, *On the functions which are represented by the expansions of the interpolation–theory*, Edinb. R. S. Proc. **35** (1915), 181–194.

[33] J. Whittaker, *Interpolatory function theory*, Cambridge Tracts in Math. and Math. Physics, vol. 33, 1935.

Index 4