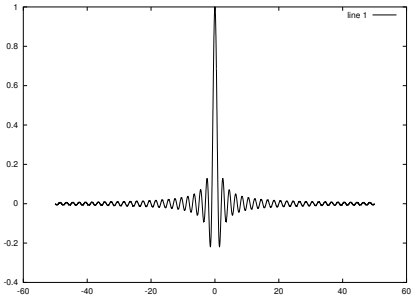
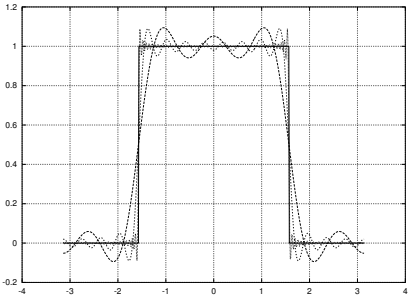




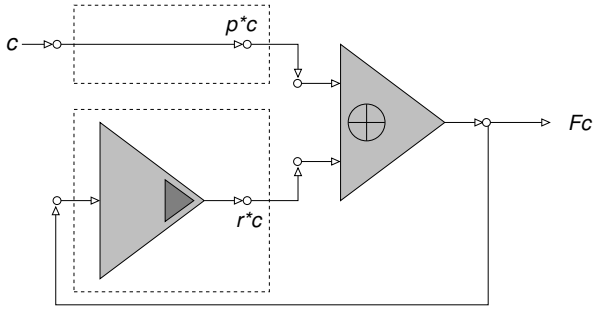
Digitale Signalverarbeitung

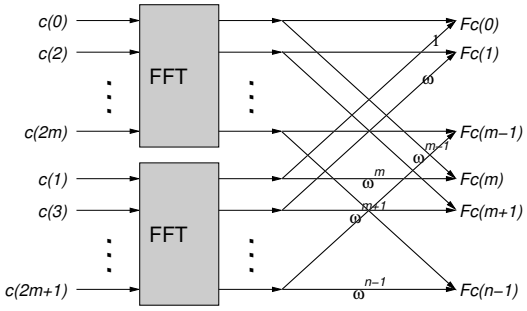
Vorlesung, zuerst gehalten im Sommersemester 2003, erweitert im Sommersemester 2006

Tomas Sauer

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

Version 2.0
Version 11.3.2010





Statt einer Leerseite . . . | 0

Die Wissenschaften sind nicht wie Minerva, welche vollständig bewaffnet dem Haupte Jupiters entsprang. Sie sind die Töchter der Zeit und bilden sich langsam, zuerst durch Sammlung der Methoden, welche die Erfahrung angibt, und später durch Entdeckung der Principien, die aus der Combination der Methoden sich folgern lassen.

[. . .]

Der Erste, der durch Zeichen jenes einfache Verhältniss $2 \times 2 = 4$ ausdrückte, erfand die Mathematik, jene mächtige Wissenschaft, welche wirklich den Menschen auf den Thron der Welt erhob.

J. A. Brillat–Savarin, *Physiologie des Geschmacks*

0 Inhaltsverzeichnis

1 Signale, Filter, Fourier **3**
1.1 Signale . . . 3
1.2 Fourieranalysis . . . 6
1.3 Der Abtastsatz . . . 15
1.4 Unschärfe . . . 18
1.5 Filter . . . 21

2 Ein größeres Repertoire an Filtern **30**
2.1 Die z–Transformation . . . 30
2.2 Rationale Filter und ihre Realisierung . . . 33
2.3 Stabilität . . . 38

3 Fourier – schnell und diskret **48**
3.1 Die diskrete Fouriertransformation . . . 48
3.2 Diskret versus diskretisiert . . . 53
3.3 Die schnelle Fouriertransformation . . . 56
3.4 Anwendungen der FFT . . . 59
3.5 Realisierung der FFT . . . 61
3.6 Undichte Fenster . . . 63
3.7 Fensterfunktionen . . . 66

4 Subband–Coding und Wavelets **71**
4.1 Allpass–Filter . . . 71
4.2 Upsampling, Downsampling und Filterbänke . . . 74
4.3 Zweikanal–Filterbänke . . . 81
4.4 Subband–Kaskaden, Subdivision und Wavelets . . . 88

5 Ecken, Kanten, Wavelets **98**
5.1 Polynome und Strang–Fix . . . 98
5.2 Waveletkoeffizienten glatter Funktionen . . . 108
5.3 Anwendungen . . . 113

6 Bilder **120**
6.1 Ein paar Grundlagen . . . 120
6.2 Einfache Filter für Bilder . . . 122
6.3 Tensorproduktwavelets für Bilder . . . 125
6.4 Viele Basen, Wörterbücher und der Nutzen der Gier . . . 128

7 Rauschen, Zufall und inverse Probleme **133**
7.1 Zufallsprozesse, Hauptkomponenten und Approximation 133
7.2 Regularisierung oder wie man unterbestimmte Probleme löst 136
7.3 Es rauscht mal wieder 141
7.4 Ein bisschen Variationsrechnung 142

Auf Silberwellen kommt gegangen
unsagbar süße Harmonie,
in eine Weise eingefangen,
unendlichfache Melodie

Chr. Morgenstern, *Phanta's Schloß*

1 Signale, Filter, Fourier

Die Objekte der digitalen Signalverarbeitung sind, wie der Name schon sagt, *Signale* und Methoden, diese zu modifizieren und zu analysieren. Das "Standardsignal" im Rahmen dieser Vorlesung wird ein *zeitabhängiges* Signal sein, das also nur von *einem*[1] Parameter abhängt.

1.1 Signale

Für uns soll vorerst ein Signal eine Abbildung $f : D \to \mathbb{R}$ sein, wobei $D \subseteq \mathbb{R}$ ist. Dabei betrachtet man vor allem die folgende Fälle des Definitionsbereichs D:

$D = \mathbb{R}$**:** Ein (prinzipiell) unbeschränktes *kontinuierliches* Signal.

$D = [a, b]$**:** Ein *zeitbeschränktes* kontinuierliches Signal. Mittels einer (affinen) Umskalierung des Intervalls können wir eigentlich immer gewährleisten, daß $a = 0$ und $b = 2\pi$ gilt. Ist außerdem $f(0) = f(2\pi)$, dann können wir das Signal *periodisch* zu einem unbeschränkten kontinuierlichen Signal fortsetzen, indem wir einfach

$$f\left(x + 2k\pi\right) := f(x), \qquad x \in [0, 2\pi], \quad k \in \mathbb{Z},$$

setzen. Analog können wir eine 2π–periodische Funktion auch als Funktion auf dem *Torus* $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z} \simeq [0, 2\pi]$ betrachten.

$D = \mathbb{Z}$**:** Ein *zeitdiskretes* Signal, also eine (doppeltunendliche) Folge. Wir werden also solche Gebilde bequemerweise als diskrete Funktionen schreiben.

Natürlich betrachtet man nicht beliebige, völlig unstrukturierte Signale, sondern solche, die gewissen mathematischen Voraussetzungen genügen[2], insbesondere in irgendeiner Form beschränkt sind. Daher erst einmal ein paar Definitionen.

[1] Im Gegensatz zu Bildern, die normalerweise eine Funktion von *zwei* Parametern sind, deren Wert ein Grauwert oder ein RGB–Wert ist.

[2] Und in Lehrbüchern für Signalverarbeitung oft einfach gemacht werden, ohne gesondert darauf hinzuweisen.

Definition 1.1 (Signalräume) *Wir bezeichnen mit $L(\mathbb{R})$ die Gesamtheit aller reellwertigen Funktionen*[3] *und mit $\ell(\mathbb{Z})$ alle reellen Folgen und definieren die folgenden Räume:*

1. Quadratsummierbare *Funktionen*

$$L_2(\mathbb{R}) := \left\{ f \in L(\mathbb{R}) \ : \ \infty > \|f\|_2 := \left(\int_{\mathbb{R}} |f(t)|^2 \, dt \right)^{1/2} \right\}$$

und Folgen

$$\ell_2(\mathbb{Z}) := \left\{ c \in \ell(\mathbb{Z}) \ : \ \infty > \|c\|_2 := \left(\sum_{k\in\mathbb{Z}} |c(k)|^2 \right)^{1/2} \right\};$$

man spricht in diesem Fall oftmals auch von endlicher Energie*, da die Gesamtenergie eines (diskreten oder kontinuierlichen) Signals gerade $\|f\|_2$ ist*[4].

2. (absolut)[5]summierbare *Funktionen und Folgen, definiert durch die Normen*

$$\|f\|_1 := \int_{\mathbb{R}} |f(t)| \, dt \qquad \textit{bzw.} \qquad \|c\|_1 := \sum_{k\in\mathbb{Z}} |c(k)| \, .$$

3. Beschränkte *Funktionen und Folgen unter Verwendung der Normen*

$$\|f\|_\infty := \sup_{t\in\mathbb{R}} |f(t)| \qquad \textit{bzw.} \qquad \|c\|_\infty := \sup_{k\in\mathbb{R}} |c(k)| \, .$$

Achtung: *Im kontinuierlichen Fall ist das Supremum ein* wesentliches Supremum*, das heißt, Mengen vom Maß 0 dürfen von der Supremumsbildung ausgeschlossen werden*[6].

4. *Funktionen und Folgen mit* endlichem Träger*, für die es ein $N \in \mathbb{N}$ gibt, so daß*

$$\left.\begin{array}{l} \{t \in \mathbb{R} \ : \ f(t) \neq 0\} \\ \{k \in \mathbb{Z} \ : \ c(k) \neq 0\} \end{array}\right\} \subseteq [-N, N] \, .$$

Auch hier sind im kontinuierlichen Fall wieder Nullmengen auszunehmen. Solche Funktionen schreiben wir als $L_{00}(\mathbb{R})$ bzw. $\ell_{00}(\mathbb{Z})$.

---

[3] Man könnte auch lokale Integrierbarkeit verlangen.

[4] Ich werde notationell **nicht** zwischen diskreten und kontinuierlichen Signalen unterscheiden, der Sinn wird sich ohnehin zumeist aus dem Kontext ergeben.

[5] Das "absolut" kann man sich eigentlich auch schenken, wenn man bedenkt, daß konvergente Reihen eigentlich immer absolut konvergent sein sollten, weil sonst der Grenzwert nicht wirklich vernünftig ist – Stichwort "Umordnungssatz".

[6] Preisfrage: Was ist dann $\sup \chi_{\mathbb{Q}}$?

Bemerkung 1.2 *Die Räume $L_1(\mathbb{R})$, $L_2(\mathbb{R})$, $L_\infty(\mathbb{R})$ sowie $\ell_1(\mathbb{Z})$, $\ell_2(\mathbb{Z})$ und $\ell_\infty(\mathbb{Z})$ sind* Banachräume*, also vollständig*[7] *und $L_{00}(\mathbb{R})$ bzw. $\ell_{00}(\mathbb{Z})$ ist* dicht *in $L_1(\mathbb{R})$ und $L_2(\mathbb{R})$ bzw. $\ell_1(\mathbb{Z})$ und $\ell_2(\mathbb{Z})$, aber nicht in $L_\infty(\mathbb{R})$ bzw. $\ell_\infty(\mathbb{Z})$.*

Definition 1.3 (δ–Puls, Dirac–Puls) *Ein ganz besonderes Signal ist der "Puls" $\delta \in \ell_{00}(\mathbb{Z})$, definiert durch*[8]

$$\delta(k) = \delta_{0k} = \begin{cases} 1, & k = 0, \\ 0, & k \neq 0. \end{cases}$$

Man spricht hier oftmals auch von einer "Funktion" namens Dirac[9]*–δ, obwohl es sich dabei eigentlich um eine* Distribution *handelt.*

Nachdem man auf Rechnern eigentlich ja nur diskret arbeiten kann, empfiehlt es sich natürlich, ein eventuell vorhandenes kontinuierliches Signal[10] in ein zeitdiskretes Signal umzuwandeln; daß man eigentlich auch noch die kontinuierlichen Werte $f(t)$ in diskrete Werte umwandeln müsste – man spricht hier von *Quantisierung* – schenken wir uns an dieser Stelle und argumentieren wie in [32], daß Quantisierungseffekte bei modernen doppeltgenauen Gleitkommaarithmetiken, siehe [29], kaum ins Gewicht fallen. Um von einem kontinuierlichen zu einem diskreten Signal überzugehen, definiert man den *Abtastoperator* $S_h : L(\mathbb{R}) \to \ell(\mathbb{Z})$ mit *Schrittweite* h als

$$(S_h f)(k) := f(hk), \qquad k \in \mathbb{Z}. \tag{1.1}$$

Bemerkung 1.4 *Eigentlich ergibt S_h für nichtstetige Funktionen gar keinen Sinn, da f nur auf der* Nullmenge *$h\mathbb{Z}$ betrachtet wird. Das kann man beheben, indem man f als stückweise stetig annimmt, oder aber die Werte von f an der Stelle hk, $k \in \mathbb{Z}$, als*

$$f(hk) = \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_{-\varepsilon}^{\varepsilon} f(hk + t)\, dt$$

festlegt *und so zum* Lebesgue–Punkt *macht, siehe [1].*

Dies führt uns zur ersten interessanten *mathematischen* Frage, nämlich wann dieser Prozess umkehrbar ist, also unter welchen Voraussetzungen die Funktion f aus $S_h f$ rekonstruierbar ist. Diese Aussage, die eine Beziehung zwischen f und h herstellen wird, ist der berühmte *Abtastsatz von Shannon*, den wir nun beweisen wollen. Doch zuerst brauchen wir ein bißchen Grundlagen.

[7] Zur Erinnerung: Das bedeutet, daß der Grenzwert jeder Cauchy–Folge auch zum Raum gehört.

[8] Die Notation δ_{jk} ist das "Kronecker–Delta" (also eigentlich "Kronecker–δ"), das, man glaubt es kaum, von Leopold Kronecker eingeführt wurde.

[9] Paul Adrien Maurice Dirac, 1902–1984, mathematischer Physiker mit wichtigen Beiträgen zur Quantenmechanik.

[10] Beispielsweise die Schalldruckwerte an einem Mikrophon.

1.2 Fourieranalysis

Ein wichtiges Hilfsmittel bei der Betrachtung von Wavelets, aber auch in der Signalverarbeitung generell ist, vor allem in L_2 die *Fouriertransformierte* einer Funktion. Wir werden hier im wesentlichen den *Kalkül* der Fourier[11]–Analysis bereitstellen und uns weniger um ihre theoretischen Konzepte kümmern; für diese sei beispielsweise auf [33] verwiesen. Bei der Definition müssen und werden wir $f \in L_1(\mathbb{R})$ voraussetzen, was bei Funktionen mit *kompaktem* Träger sowieso einfacher ist.

Übung 1.1 Zeigen Sie: Ist $f \in L_2(\mathbb{R}) \cap L_{00}(\mathbb{R})$, dann ist $f \in L_1(\mathbb{R})$. ♢

Definition 1.5 *Für $f \in L_1(\mathbb{R})$ definieren wir die* Fouriertransformierte $\widehat{f} : \mathbb{R} \to \mathbb{C}$ *als*

$$\widehat{f}(\xi) := f^\wedge(\xi) := \int_{\mathbb{R}} f(t)\, e^{-i\xi t}\, dt, \qquad \xi \in \mathbb{R}, \tag{1.2}$$

und die Fouriertransformierte einer Folge $c \in \ell_1(\mathbb{Z})$ als diskretes Gegenstück

$$\widehat{c}(\xi) := c^\wedge(\xi) := \sum_{k\in\mathbb{Z}} c(k)\, e^{-ik\xi}, \qquad \xi \in \mathbb{R}. \tag{1.3}$$

Bemerkung 1.6

1. *In ihrer physikalischen oder technischen Interpretation liefert die Fouriertransformierte eines "Signals" (das man als Aplitudenfunktion der Zeit ansieht), den Anteil der entsprechenden Frequenz an diesem Signal.*

2. *Die Bedingung $f \in L_1(\mathbb{R})$ garantiert, daß die $\widehat{f}(\xi)$ für alle $\xi \in \mathbb{R}$ existiert:*

$$\left|\widehat{f}(\xi)\right| \le \int_{\mathbb{R}} |f(t)| \underbrace{\left|e^{-i\xi t}\right|}_{=1} dt = \|f\|_1 . \tag{1.4}$$

   Allerdings ist das "nur" hinreichend, aber eben nicht notwendig für die Existenz der Fouriertransformierten.

3. *Manchmal wird die Fouriertransformierte auch noch mit dem Vorfaktor $(2\pi)^{1/2}$ versehen, wir werden bald sehen, warum. Man sollte also bei der Verwendung von Literatur immer gut aufpassen, welche Normierung dort gewählt ist, sonst kann so ein konstanter Faktor für üble Fehler sorgen.*

4. *Man kann die Fouriertransformierte auch für allgemeinere "Funktionen"klassen als L_1 definieren, beispielsweise für* temperierte Distributionen*, siehe z.B. [33, 65].*

[11] Jean Baptiste Fourier, 1768–1830, französischer Mathematiker und Politiker. Er war nicht nur Mitglied der "Acadèmie des Sciences", sondern (vorher) auch Teilnehmer an der Ägypten–Expedtion von Napoleon (Bonaparte) als wissenschaftlicher Berater und Gouverneur des Departement Isére mit Hauptstadt Grenoble. In den beiden letzteren Eigenschaften trug er nicht unwesentlich (als Förderer von Champollion) zur Entzifferung der Hieroglyphen bei, siehe [12].

5. *Außerdem gibt es die Fouriertransformation nicht nur auf* $\mathbb{R}$ *oder* $\mathbb{R}^n$ *sondern auf lokal kompakten abelschen Gruppen unter Verwendung des Haar–Maßes; dann sieht man, daß die Fouriertransformierte auf der* dualen Gruppe *definiert ist. Das soll uns aber hier nicht stören, in unserem einfachen aber bedeutenden Spezialfall spielt* $\mathbb{R}$ *beide Rollen.*

6. *Die zwei bedeutendsten Gruppenoperationen*[12] *auf* $\mathbb{R}$ *sind die* Translation *und die* Skalierung *die durch die beiden Operatoren* τ_y *und* σ_h*, definiert als*

$$\tau_y f = f(\cdot + y) \qquad \textit{und} \qquad \sigma_h f = f(h\cdot)$$

realisiert werden sollen.

Definition 1.7 (Faltung) *Zu* $f, g \in L(\mathbb{R})$ *und* $c, d \in \ell(\mathbb{Z})$ *definieren wir*[13] *die* Faltungen

$$f * g := \int_{\mathbb{R}} f(\cdot - t)\, g(t)\, dt, \qquad * : L(\mathbb{R}) \times L(\mathbb{R}) \to L(\mathbb{R}), \tag{1.5}$$

sowie

$$c * d := \sum_{k\in\mathbb{Z}} c(\cdot - k)\, d(k), \qquad * : \ell(\mathbb{R}) \times \ell(\mathbb{R}) \to \ell(\mathbb{R}), \tag{1.6}$$

und

$$c * f := f * c := \sum_{k\in\mathbb{Z}} f(\cdot - k)\, d(k), \qquad * : L(\mathbb{R}) \times \ell(\mathbb{R}) \to L(\mathbb{R}). \tag{1.7}$$

Die Faltung (1.5) bezeichnet man als kontinuierlich*, die in (1.6) als* diskret *und die in (1.7) als* semidiskret*.*

Als nächstes stellen wir ein paar einfache Eigenschaften der Fouriertransformierten zusammen – daß die Fouriertransformierte linear in f bzw. c ist, das braucht nicht mehr besonders betont werden.

Satz 1.8 (Eigenschaften der Fouriertransformierten) *Es gelten die folgenden Aussagen:*

1. *Für* $f \in L_1(\mathbb{R})$ *und* $y \in \mathbb{R}$ *ist*

$$(\tau_y f)^\wedge(\xi) = e^{iy\xi}\, \widehat{f}(\xi), \qquad \xi \in \mathbb{R}. \tag{1.8}$$

2. *Für* $f \in L_1(\mathbb{R})$ *und* $h > 0$ *ist*

$$(\sigma_h f)^\wedge(\xi) = \frac{\widehat{f}\,(h^{-1}\xi)}{h}, \qquad \xi \in \mathbb{R}. \tag{1.9}$$

3. *Für* $f, g \in L_1(\mathbb{R})$ *bzw.* $c, d \in \ell_1(\mathbb{Z})$ *ist* $f * g \in L_1(\mathbb{R})$ *bzw.* $c * d \in \ell_1(\mathbb{Z})$ *und es gilt für* $\xi \in \mathbb{R}$

$$(f * g)^\wedge(\xi) = \widehat{f}(\xi)\, \widehat{g}(\xi), \qquad \textit{bzw.} \qquad (c * d)^\wedge(\xi) = \widehat{c}(\xi)\, \widehat{d}(\xi). \tag{1.10}$$

[12] Das heißt, sie sind insbesondere invertierbar.

[13] Vorbehaltlich Existenz der unendlichen Summen.

4. *Für $f \in L_1(\mathbb{R})$ und $c \in \ell_1(\mathbb{Z})$ ist $f * c \in L_1(\mathbb{R})$ und*

$$(f * c)^\wedge(\xi) = \widehat{f}(\xi)\,\widehat{c}(\xi), \xi \in \mathbb{R}. \tag{1.11}$$

5. *Sind $f, f' \in L_1(\mathbb{R})$, dann gilt*

$$\left(\frac{d}{dx}f\right)^\wedge(\xi) = i\xi\,\widehat{f}(\xi), \qquad \xi \in \mathbb{R}. \tag{1.12}$$

6. *Sind $f, xf \in L_1(\mathbb{R})$, dann ist $\widehat{f}$ differenzierbar und es gilt*

$$\frac{d}{d\xi}\widehat{f}(\xi) = (-ix\,f)^\wedge(\xi), \qquad \xi \in \mathbb{R}. \tag{1.13}$$

7. *Sind $f, \widehat{f} \in L_1(\mathbb{R})$, dann ist*

$$f(x) = \left(\widehat{f}\right)^\vee(x) := \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(\vartheta)\,e^{ix\vartheta}\,d\vartheta \tag{1.14}$$

Die Operation $f \mapsto f^\vee := \frac{1}{2\pi}f^\wedge(-\cdot)$ bezeichnet man als inverse Fouriertransformation[14].

Beweis: Für 1) berechnen wir

$$\begin{aligned}(\tau_y f)^\wedge(\xi) &= \int_{\mathbb{R}} f(t+y)\,e^{-i\xi t}\,dt = \int_{\mathbb{R}} f(t)\,e^{-i\xi(t-y)}\,dt = e^{iy\xi}\int_{\mathbb{R}} f(t)\,e^{-i\xi t}\,dt \\ &= e^{iy\xi}\,\widehat{f}(\xi),\end{aligned}$$

während 2) ganz ähnlich mit

$$(\sigma_h f)^\wedge(\xi) = \int_{\mathbb{R}} f(ht)\,e^{-i\xi t}\,dt = \frac{1}{h}\int_{\mathbb{R}} f(t)\,e^{-i(\xi/h)t}\,dt = \frac{\widehat{f}\left(\frac{\xi}{h}\right)}{h}$$

bewiesen wird. Die erste Aussage von 3) folgt aus

$$\|f * g\|_1 = \int_{\mathbb{R}}\left|\int_{\mathbb{R}} f(t)g(s-t)\,dt\right|\,ds \le \int_{\mathbb{R}}\int_{\mathbb{R}} |f(t)g(s)|\,dt\,ds = \|f\|_1\,\|g\|_1$$

bzw.

$$\|c * d\|_1 = \sum_{j\in\mathbb{Z}}\left|\sum_{k\in\mathbb{Z}} c(k)\,d(j-k)\right| \le \sum_{j,k\in\mathbb{Z}} |c(k)\,d(j)| = \|c\|_1\,\|c\|_1,$$

der zweite, etwas interessantere Teil hingegen aus

$$\begin{aligned}(f * g)^\wedge(\xi) &= \int_{\mathbb{R}}\left(\int_{\mathbb{R}} f(s)g(t-s)\,ds\right)e^{-i\xi t}dt = \int_{\mathbb{R}}\int_{\mathbb{R}} f(s)\,e^{i\xi s}\,g(t-s)\,e^{i\xi(t-s)}\,dsdt \\ &= \widehat{f}(\xi)\,\widehat{g}(\xi),\end{aligned}$$

[14] Die Gründe dafür sind ja wohl offensichtlich.

bzw.

$$\begin{aligned}(c*d)^{\wedge}(\xi) &= \sum_{j\in\mathbb{Z}}\left(\sum_{k\in\mathbb{Z}} c(k)\,d(j-k)\right)e^{-ij\xi} = \sum_{j,k\in\mathbb{Z}} c(k)\,e^{-ik\xi}\,d(j-k)\,e^{-i(j-k)\xi}\\ &= \widehat{f}(\xi)\,\widehat{g}(\xi).\end{aligned}$$

4) folgt aus Übung 1.2 und

$$\begin{aligned}(f*c)^{\wedge}(\xi) &= \int_{\mathbb{R}}\sum_{k\in\mathbb{Z}} f(t-k)\,c(k)e^{-i\xi t}\,dt = \int_{\mathbb{R}}\sum_{k\in\mathbb{Z}} f(t-k)\;e^{-i\xi(t-k)}\,c(k)\,e^{-ik\xi}\\ &= \widehat{f}(\xi)\,\widehat{c}(\xi).\end{aligned}$$

oder auch direkt unter Verwendung von (1.8). Für 5 verwenden wir partielle Integration[15], um

$$(f')^{\wedge}(\xi) = \int_{\mathbb{R}}\frac{df}{dt}(t)e^{-i\xi t}\,dt = -\int_{\mathbb{R}} f(t)\frac{d}{dt}e^{-i\xi t}\,dt = i\xi\int_{\mathbb{R}} f(t)e^{-i\xi t}\,dt = i\xi\,\widehat{f}(\xi).$$

6) erhalten wir, indem wir für $h>0$ den Differenzenquotient

$$\frac{\widehat{f}(\xi+h)-\widehat{f}(\xi)}{h} = \int_{\mathbb{R}} f(t)\,\frac{e^{-i(\xi+h)t}-e^{-i\xi t}}{h}\,dt = \int_{\mathbb{R}} f(t)\,e^{-i\xi t}\,\frac{e^{-iht}-1}{h}\,dt$$

betrachten; das Integral existiert, weil $xf\in L_1(\mathbb{R})$ und da

$$\lim_{h\to 0}\frac{e^{-iht}-1}{h} = \lim_{h\to 0}(-it)\,e^{-iht} = -it$$

ist, folgt (1.13). Der Beweise von 7) ist ein klein wenig aufwendiger und verwendet die *Fejér–Kerne*

$$F_\lambda := \lambda F(\lambda\cdot), \quad \lambda>0, \qquad F(x) := \frac{1}{2\pi}\int_{-1}^{1}(1-|t|)\;e^{ixt}\,dt, \quad x\in\mathbb{R},$$

auf $\mathbb{R}$, die die Eigenschaft haben, daß für jedes $f\in L_1(\mathbb{R})$

$$\lim_{\lambda\to\infty}\|f-f*F_\lambda\| = 0, \tag{1.15}$$

siehe [33, S. 124–126], also auch $f*F_\lambda\to f$ punktweise fast überall[16]. Dann ist aber für $x\in\mathbb{R}$

$$f*F_\lambda(x) = \frac{1}{2\pi}\int_{\mathbb{R}} f(t)\,\left(\lambda\int_{-1}^{1}(1-|\vartheta|)\,e^{i(x-t)\lambda\vartheta}\,d\vartheta\right)\,dt$$

[15] Daß dies gerechtfertigt ist, liegt an der Tatsache, daß für $f\in L_1(\mathbb{R})$ immer $\lim_{x\to\pm\infty}|f(x)|=0$ sein muß und daß die stetigen Funktionen mit kompaktem Träger bezüglich der Norm $\|\cdot\|_1$ *dicht* in $L_1(\mathbb{R})$ sind. Deswegen muß man sich um "Randwerte" hier nicht kümmern.

[16] Zumindest für eine Teilfolge, siehe [16, S. 96].

$$
\begin{aligned}
&= \frac{1}{2\pi}\int_{\mathbb{R}} f(t)\int_{-\lambda}^{\lambda}\left(1-\frac{|\vartheta|}{\lambda}\right)e^{i(x-t)\vartheta}\,d\vartheta\,dt\\
&= \frac{1}{2\pi}\int_{-\lambda}^{\lambda}\left(1-\frac{|\vartheta|}{\lambda}\right)\underbrace{\int_{\mathbb{R}} f(t)\,e^{-it\vartheta}\,dt}_{=\widehat{f}(\vartheta)}\,e^{ix\vartheta}\,d\vartheta\\
&= \frac{1}{2\pi}\int_{-\lambda}^{\lambda}\left(1-\frac{|\vartheta|}{\lambda}\right)\widehat{f}(\vartheta)\,e^{ix\vartheta}\,d\vartheta\\
&= \underbrace{\frac{1}{2\pi}\int\limits_{0\le|\vartheta|\le\sqrt{\lambda}}\overbrace{\left(1-\frac{|\vartheta|}{\lambda}\right)}^{1-1/\sqrt{\lambda}\le\cdot\le 1}\widehat{f}(\vartheta)\,e^{ix\vartheta}\,d\vartheta}_{\to\frac{1}{2\pi}\int_{\mathbb{R}}\widehat{f}(\vartheta)\,e^{ix\vartheta}\,d\vartheta}+\underbrace{\frac{1}{2\pi}\int\limits_{\sqrt{\lambda}\le|\vartheta|\le\lambda}\left(1-\frac{|\vartheta|}{\lambda}\right)\widehat{f}(\vartheta)\,e^{ix\vartheta}\,d\vartheta}_{\to 0},
\end{aligned}
$$

weil $\widehat{f}\in L_1(\mathbb{R})$. □

Übung 1.2 Zeigen Sie, daß für $f\in L_1(\mathbb{R})$ und $c\in\ell_1(\mathbb{Z})$ die Ungleichung

$$\|f*c\|_1\le\|f\|_1\,\|c\|_1$$

gilt. ♢

Übung 1.3 Beweisen Sie *ohne* Verwendung von (1.12) die folgende Aussage: Sind $f,f'\in L_1(\mathbb{R})$, dann ist $(f')^\wedge(0)=0$.
Hinweis: Partielle Integration. ♢

Übung 1.4 Zeigen Sie, daß sich für $f\in L_1(\mathbb{R})$ und $h\neq 0$ die Identität

$$(\sigma_h f)^\wedge=\frac{\widehat{f}\,(h^{-1}\cdot)}{|h|}\tag{1.9$'$}$$

ergibt. ♢

Beispiel 1.9 *Berechnen wir doch mal zu Übungszwecken so eine Fouriertransformierte, und zwar die der kardinalen B–Splines N_j, $j\in\mathbb{N}_0$, definiert durch $N_0=\chi$ und $N_j=\chi*N_{j-1}$, $j\in\mathbb{N}$. Insbesondere ist also*

$$\widehat{N_0}(\xi)=\widehat{\chi}(\xi)=\int_{\mathbb{R}}\chi(t)\,e^{-i\xi t}\,dt=\int_0^1 e^{-i\xi t}\,dt=\left.\frac{e^{-i\xi t}}{-i\xi}\right|_{t=0}^{1}=\frac{1-e^{-i\xi}}{i\xi}$$

und somit, nach (1.10),

$$\widehat{N_j}(\xi)=(\widehat{\chi}(\xi))^{j+1}=\left(\frac{1-e^{-i\xi}}{i\xi}\right)^{j+1}.$$

Übung 1.5 Die *zentrierten B–Splines* M_j, $j \in \mathbb{N}_0$, sind definiert als

$$M_j = \underbrace{\chi_{[-1/2,1/2]} * \cdots * \chi_{[-1/2,1/2]}}_{j+1}.$$

Zeigen Sie:

1. Diese Funktionen sind gerade: $M_j(-x) = M_j(x)$, $x \in \mathbb{R}$.
2. Für $j \in \mathbb{N}_0$ ist
$$\widehat{M_j}(\xi) = \left(\frac{\sin \xi/2}{\xi/2}\right)^{j+1}, \qquad \xi \in \mathbb{R}.$$

♢

Ist $f \in L_1(\mathbb{R})$, dann ist für $\xi, \eta \in \mathbb{R}$

$$\left|\widehat{f}\,(\xi+\eta) - \widehat{f}(\xi)\right| \le \int_{\mathbb{R}} |f(t)| \underbrace{\left|e^{-i\xi t}\right|}_{=1} \left|e^{-i\eta t} - 1\right| \, dt,$$

was auf der rechten Seite unabhängig von ξ ist und mit $\eta \to 0$ gegen Null konvergiert, denn für jedes $\varepsilon > 0$ gibt es ein $N > 0$, so daß

$$\int_{|t|>N} |f(t)| \, dt < \varepsilon$$

ist, während wir, durch Wahl eines hinreichend kleinen Wertes von η, die Funktion $|e^{-i\eta t} - 1|$ auf $[-N, N]$ so klein machen können, wie wir wollen. Der langen Rede kurzer Sinn:

> *Ist $f \in L_1(\mathbb{R})$, so ist $\widehat{f} \in C_u(\mathbb{R})$, dem Vektorraum der gleichmäßig stetigen und gleichmäßig beschränkten*[17] *Funktionen auf $\mathbb{R}$.*

Außerdem kann man sogar sagen, wie sich die Fouriertransformierte für $|\xi| \to \infty$ benimmt, das ist das klassische Riemann[18]–Lebesgue[19]–Lemma.

Proposition 1.10 (Riemann–Lebesgue–Lemma)
Ist $f \in L_1(\mathbb{R})$, so ist

$$\lim_{\xi \to \pm\infty} \widehat{f}(\xi) = 0. \tag{1.16}$$

[17] Siehe (1.4).

[18] Georg Friedrich Bernhard Riemann, 1826–1866, Schüler von Gauss mit Beiträgen zu Analysis, Algebra, Geometrie.

[19] Henri Lebesgue, 1875–1941, sein bedeutendstes Werk war seine Dissertation "Intégrale, longueur, aire" (1902).

Beweis: Ist auch $f' \in L_1(\mathbb{R})$ so folgt (1.16) sofort mittels (1.12) und (1.4):

$$\|f'\|_1 \geq \left|(f')^\wedge(\xi)\right| = |\xi| \left|\widehat{f}(\xi)\right|, \qquad \xi \in \mathbb{R},$$

also $\left|\widehat{f}(\xi)\right| \leq \|f'\|_1 / |\xi| \to 0$ für $|\xi| \to \infty$. Für beliebiges $f \in L_1(\mathbb{R})$ und differenzierbares $g \in L_1(\mathbb{R})$ mit[20] mit $\|f - g\|_1 \leq \varepsilon$ ist

$$\|f - g\|_1 \geq \left|\widehat{f}(\xi) - \widehat{g}(\xi)\right| \geq \left|\widehat{f}(\xi)\right| - |\widehat{g}(\xi)|,$$

also

$$\lim_{|\xi|\to\infty} \left|\widehat{f}(\xi)\right| \leq \lim_{|\xi|\to\infty} |\widehat{g}(\xi)| + \|f - g\|_1 \leq \varepsilon$$

und da man ε beliebig klein wählen kann, folgt die Behauptung. □

Wie sieht es nun auf anderen L_p–Räumen, $p \neq 1$, insbesondere mit $L_2(\mathbb{R})$ aus[21]? Hier nutzt man aus, daß $L_1(\mathbb{R}) \cap L_p(\mathbb{R})$ eine *dichte Teilmenge* von $L_p(\mathbb{R})$ ist. Für L_2 gibt es nun noch eine besonders schöne Eigenschaft, nämlich eine Isometrie, für die Parseval[22] bzw. Plancherel[23] die Namensgeber sind.

Satz 1.11 (Parseval/Plancherel) *Für $f, g \in L_1(\mathbb{R}) \cap L_2(\mathbb{R})$ ist*

$$\int_{\mathbb{R}} f(t)\, g(t)\, dt = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{f}(\vartheta)\, \overline{\widehat{g}(\vartheta)}\, d\vartheta, \tag{1.17}$$

also insbesondere, mit $f = g$,

$$\|f\|_2 = \frac{1}{\sqrt{2\pi}} \left\|\widehat{f}\right\|_2. \tag{1.18}$$

Diese Aussage hilft uns nun, die Definition der Fouriertransformierten auf $L_2(\mathbb{R})$ zu übertragen: Zu $f \in L_2(\mathbb{R})$ betrachtet man eine Folge

$$f_n := \chi_{[-n,n]} \cdot f \quad \in L_1(\mathbb{R}) \cap L_2(\mathbb{R}), \qquad n \in \mathbb{N},$$

die für $n \to \infty$ in der Norm $\|\cdot\|_2$ gegen f konvergiert. Da

$$\left\|\widehat{f}_{n+k} - \widehat{f}_n\right\|_2 = \left\|(f_{n+k} - f)^\wedge\right\|_2 = \sqrt{2\pi}\, \|f_{n+k} - f_n\|_2, \qquad k, n \in \mathbb{N},$$

[20] Man beachte, daß sogar die *unendlich oft stetig differenzierbaren Funktionen mit kompaktem Träger* eine dichte Teilmenge von $L_1(\mathbb{R})$ bilden.

[21] $L_2(\mathbb{R})$ spielt in der Signalverarbeitung schon deswegen so eine wesentliche Rolle, weil das gerade die Signale (und die sind normalerweise nicht unbedingt stetig) mit *endlicher Energie* sind – eine ziemlich realistische Annahmen, oder nicht?

[22] Marc–Antoine Parseval des Chênes, 1755–1836, Zeitgenosse von Fourier, der ziemlich heftig in die Wirren der französischen Revolution verwickelt wurde, publizierte überhaupt nur 5 (in Worten: "fünf") Arbeiten, die er aber allesamt der *Académie des Sciences* vorlegte.

[23] Leider keine biografischen Daten.

ist $\widehat{f_n}$ eine Cauchyfolge und konvergiert gegen eine Funktion in $L_2(\mathbb{R})$, die wir $\widehat{f}$ nennen wollen.
Beweis von Satz 1.11: Wir definieren

$$h(x) = \int_{\mathbb{R}} f(t)\, g\,(t-x)\; dt = (f * g(-\cdot))\,(x), \qquad x \in \mathbb{R},$$

und erhalten, daß $h(0) = \int fg$. Außerdem ist

$$\widehat{h}(\xi) = \widehat{f}(\xi)\; \underbrace{(g(-\cdot))^{\wedge}\,(\xi)}_{=\overline{\widehat{g}(\xi)}} = \widehat{f}(\xi)\,\overline{\widehat{g}(\xi)}, \qquad \xi \in \mathbb{R}.$$

Sind nun f und g so "brav", daß $\widehat{f}, \widehat{g} \in L_2(\mathbb{R})$ ist[24], dann ist mit (1.14)

$$\frac{1}{2\pi} \int_{\mathbb{R}} \widehat{f}(\vartheta)\,\overline{\widehat{g}(\vartheta)}\, d\vartheta = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{h}(\vartheta)\, e^{i0\theta}\, d\vartheta = h(0) = \int_{\mathbb{R}} f(t)\, g(t)\, dt,$$

was (1.17) liefert. Und die *Plancherel–Identität* (1.18) ist dann eine unmittelbare Konsequenz aus der *Parseval–Formel* (1.17). □

Jetzt machen wir als nächstes einen kleinen Abstecher in die Welt der Fourieranalysis auf dem Torus $\mathbb{T}$, bei dem wir den Begriff der *Fourierreihe* kennelernen und mit den bisherigen Fourierismen in Beziehung bringen werden. Tatsächlich wird nämlich der Beweis des Shannonschen Abtastsatzes, Satz 1.16 ebenfalls auf der Interaktion zwischen Fourierreihen und der Fouriertransformierten basieren. Doch dazu sollten wir erst einmal die Fourierreihe einer Funktion definieren.

Definition 1.12 *Zu $f \in L_1(\mathbb{T})$[25] sind die* Fourierkoeffizienten *definiert als*

$$f_k := \frac{1}{2\pi} \int_{\mathbb{T}} f(t)\, e^{-ikt}\, dt, \qquad k \in \mathbb{Z},$$

und die Fourierreihe[26] *zu f als*

$$\mathscr{F} f := \sum_{k\in\mathbb{Z}} f_k\, e^{ik\cdot}. \tag{1.19}$$

Eigentlich sollte uns die trigonometische Reihe in (1.19) bekannt vorkommen: Definieren wir nämlich zu $f \in L_1(\mathbb{T})$ mit Fourierkoeffizienten f_k, $k \in \mathbb{Z}$, die Folge

$$c_f(k) = f_k, \qquad k \in \mathbb{Z},$$

dann ist $c_f \in \ell_1(\mathbb{Z})$ und $\mathscr{F} f = \widehat{c_f}$. Da außerdem für $j, k \in \mathbb{Z}$

$$\int_{\mathbb{T}} e^{-ijt} e^{ikt}\, dt \;=\; \int_{-\pi}^{\pi} e^{i(k-j)t}\, dt = \begin{cases} 2\pi, & j = k, \\ \frac{1}{i(k-j)} e^{i(k-j)t}\Big|_{t=-\pi}^{\pi} = 0, & j \neq k, \end{cases}$$

[24] Das ist beispielsweise der Fall, wenn f und g differenzierbar sind; dies folgt aus (1.12) und dem Riemann–Lebesgue–Lemma, Proposition 1.10.

[25] Man bemerke, daß $L_p(\mathbb{T}) \subset L_1(\mathbb{T})$ für $1 < p \leq \infty$ ist.

[26] Eine sogenannte *trigonometrische Reihe*.

erhalten wir für $k \in \mathbb{Z}$, daß

$$c(k) = \frac{1}{2\pi} \int_{\mathbb{T}} \sum_{j\in\mathbb{Z}} c(j) e^{ijt}\, e^{-ikt}\, dt = \frac{1}{2\pi} \int_{\mathbb{T}} \widehat{c}(\theta)\, e^{ik\theta}\, d\theta =: \left(\widehat{c}\right)^{\vee}(k). \tag{1.20}$$

Mit anderen Worten: Wir haben eine *Inverse* zur Fouriertransformierten einer Folge gefunden. Daß diese (formale) Rechnung auch wirklich in Ordnung ist kann man sich übrigens leicht überlegen.

Übung 1.6 Zeigen Sie, daß für $c \in \ell_1(\mathbb{Z})$ auch $\widehat{c} \in L_1(\mathbb{T})$ gilt. ♢

Der folgende Satz stellt eine weitere zentrale Beziehung zwischen Fourierreihen un der Fouriertransformierten her und liefert eine Identität, die sich schon oft genug als hilfreich erwiesen hat, nämlich die Poissonsche[27] Summenformel.

Satz 1.13 (Poisson–Summenformel) *Für $f \in L_1(\mathbb{R})$ ist*

$$\sum_{k\in\mathbb{Z}} f\left(2k\pi\right) = \frac{1}{2\pi} \sum_{k\in\mathbb{Z}} \widehat{f}\,(k) \qquad \textit{und} \qquad \sum_{k\in\mathbb{Z}} f(k) = \sum_{k\in\mathbb{Z}} \widehat{f}\,(2k\pi)\,. \tag{1.21}$$

Beweis: Wir setzen

$$g(x) = \sum_{k\in\mathbb{Z}} f\left(x + 2k\pi\right), \qquad x \in \mathbb{R}, \tag{1.22}$$

und bemerken, daß $g\left(x + 2\pi\right) = g(x)$, also g eine 2π–periodische Funktion ist, und wegen

$$\begin{aligned} \|g\|_{\mathbb{T},1} &= \int_{\mathbb{T}} |g(t)|\, dt = \int_{\mathbb{T}} \left| \sum_{k\in\mathbb{Z}} f\left(t + 2k\pi\right) \right| dt \le \sum_{k\in\mathbb{Z}} \int_0^{2\pi} \left| f\left(t + 2k\pi\right) \right| dt \\ &= \int_{\mathbb{R}} |f\left(t\right)|\, dt = \|f\|_{\mathbb{R},1} \end{aligned}$$

ist $g \in L_1(\mathbb{T})$ und insbesondere wohldefiniert – die Summe in (1.22) divergiert nicht allzu unmotiviert. Die Fourierkoeffizienten g_k von g haben dann die Form

$$g_k = \frac{1}{2\pi} \int_{\mathbb{T}} g(t)\, e^{-ikt}\, dt = \frac{1}{2\pi} \int_{\mathbb{T}} \sum_{\ell\in\mathbb{Z}} f\left(t + 2\ell\pi\right)\, e^{-ikt}\, dt = \frac{1}{2\pi} \int_{\mathbb{R}} f(t)\, e^{-ikt}\, dt = \frac{1}{2\pi}\, \widehat{f}(k)$$

und, angenommen die Partialsummen der Fourierreihe von g würden konvergieren[28], erhalten so, daß

$$\frac{1}{2\pi} \sum_{k\in\mathbb{Z}} \widehat{f}(k) = \sum_{k\in\mathbb{Z}} g_k \underbrace{e^{ik0}}_{=1} = g(0) = \sum_{k\in\mathbb{Z}} f\left(0 + 2k\pi\right) = \sum_{k\in\mathbb{Z}} f\left(2k\pi\right),$$

was die erste Identität liefert. Mit deren Hilfe und (1.9) ergibt sich dann, daß

$$\sum_{k\in\mathbb{Z}} f(k) = \sum_{k\in\mathbb{Z}} \left(\sigma_{(2\pi)^{-1}} f\right)\left(2k\pi\right) = \frac{1}{2\pi} \sum_{k\in\mathbb{Z}} \left(\sigma_{(2\pi)^{-1}} f\right)^{\wedge}(k) = \sum_{k\in\mathbb{Z}} \widehat{f}\,(2k\pi)\,.$$

□

[27] Siméon Denis Poisson, 1781–1840, studierte bei Laplace und Legendre, Beiträge zur Fourier–Analysis und Wahrscheinlichkeitstheorie ("Poisson–Verteilung"), schrieb zwischen 300 und 400 Arbeiten, auch über Elektrizität, Magnetismus und Astronomie.

[28] Ansonsten müssten wir ein Summationsverfahren, beispielsweise den Féjer–Kern verwenden.

1.3 Der Abtastsatz

Jetzt können wir auch schon unsere erste "große" Frage beantworten, nämlich welche Funktionen aus ihrer Abtastfolge eindeutig rekonstruiert werden können. Dazu zwei Begriffe.

Definition 1.14

1. *Eine Funktion* $f \in L_1(\mathbb{R})$ *heißt* "bandbeschränkt *mit* Bandbreite T" *oder kurz* "T–bandbeschränkt"*, wenn*
$$\widehat{f}(\xi) = 0, \qquad \xi \notin [-T, T].$$
2. *Die* Sinus Cardinalis- *oder* sinc*–Funktion*[29] *ist definiert als*
$$\operatorname{sinc}(x) := \frac{\sin \pi x}{\pi x}, \qquad x \in \mathbb{R}.$$

Bemerkung 1.15 *Wegen*

$$\operatorname{sinc} 0 = \lim_{x \to 0} \frac{\sin \pi x}{\pi x} = \lim_{x \to 0} \frac{\cos x}{1} = \cos 0 = 1$$

ist

$$(S_1 \operatorname{sinc})(k) = \operatorname{sinc}(k) = \delta(k), \qquad k \in \mathbb{Z}. \tag{1.23}$$

Daher stammt dann auch der Name "kardinal"*: Die Funktion hat an* $\mathbb{Z}$ *die Werte* $0/1$.

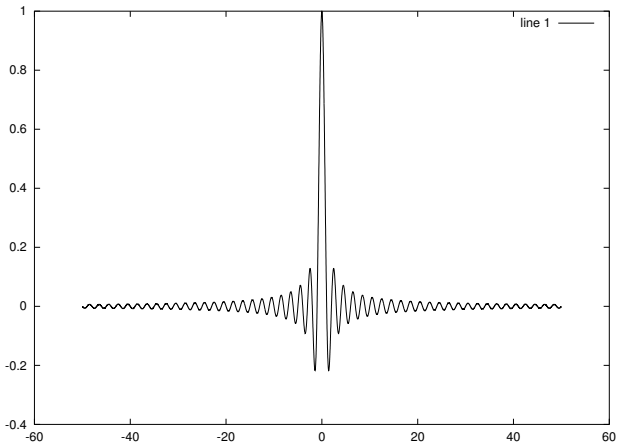


Abbildung 1.1: Die Funktion sinc .

Das nächste Resultat, der Shannonsche[30] Abtastsatz[31], sagt uns nun, daß man bandbeschränkte Funktionen rekonstruierbar abtasten kann.

[29] In Ingenieurskreisen auch als "si"–Funktion bezeichnet.

[30] Claude Elwood Shannon, 1916–2001, Elektroingenieur und Mathematiker, Erfinder des Wortes "bit" und Entwickler von Schachprogrammen (und zwar um 1950).

[31] Der eigentlich gar nicht von Shannon ist, siehe Bemerkung 1.17.

Satz 1.16 (Abtastsatz von Shannon) *Ist $f \in L_1(\mathbb{R})$ eine T–bandbeschränkte Funktion und ist $h < h^* = \frac{\pi}{T}$, dann ist*

$$f = (S_h f * \text{sinc}\,)\left(h^{-1}\cdot\right) = \sum_{k\in\mathbb{Z}} f\,(hk)\ \frac{\sin\pi\,(x/h-k)}{\pi\,(x/h-k)}. \tag{1.24}$$

Bemerkung 1.17 *Die Kopplung von Bandbreite T der Funktion und Auflösung h ist ein wirklich zentrales Konzept der digitalen Signalverarbeitung. Daher noch ein paar Anmerkungen:*

1. *In vielen Büchern heißt es, daß die* Abtastfrequenz $1/h$*, oftmals auch als* "Nyquist–Frequenz" *bezeichnte, die* Hälfte *der maximalen Frequenz T sein soll, siehe z.B. [32]. Solche Konstanten kommen von einer etwas anderen Normierung von Frequenzen (mit 2π) oder auch der Fouriertransformierten.*

2. *Die* sinc*–Funktion ist kein wirklich guter Weg zur Rekonstruktion eines Signals. Sie hat ja keinen endlichen Träger und fällt nur linear*[32] *ab, so daß man mit Abschneiden und den dabei auftretenden Fehlern* sehr *vorsichtig sein muß.*

3. *In vielen Fällen wählt man die Abtastrate h einfach als h^*/k für ein ganzzahliges k, was nichts anderes ist als die Abtastfrequenz k–mal so groß wie nötig anzusetzen. In diesem Fall spricht man von "k–*fachem Oversampling"[33].

4. *Die Beweise des Abtastsatzes aus Bücher der elektrotechnischen Literatur, z.B. [23] oder [54], und die dort verwendeten Argumentationen, sind oftmals (zumindest für Mathematiker) nur schwer nachzuvollziehen. Daß es auch anders geht sieht man in [38]. Der jetzt folgende Beweis ist eine Modifikation des Beweises aus [32].*

5. *Auch die Herkunft des Satzes ist nicht so ganz klar. Laut [38] wurde er zuerst (theoretisch) 1935 von Whittaker bewiesen [64] und 1949 von Shannon im Kontext der Signalverarbeitung wiederentdeckt [55].*

Beweis von Satz 1.16: Wegen der Bandbeschränktheit ist $\widehat{f} \in C_u(\mathbb{R}) \cap L_{00}(\mathbb{R})$, also $\widehat{f} \in L_1(\mathbb{R})$. Für $h > 0$ und $k \in \mathbb{Z}$ ist gemäß der *diskreten* Beziehung (1.20)

$$S_h f(k) = \left((S_h f)^\wedge\right)^\vee (k) = \frac{1}{2\pi}\int_{\mathbb{T}} (S_h f)^\wedge\,(\theta)\ e^{ik\theta}\,d\theta, \tag{1.25}$$

aber eben auch

$$\begin{aligned} S_h f(k) &= f\,(hk) = \widehat{f}^\vee(hk) = \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(\theta)\ e^{ihk\theta}\,d\theta = \frac{1}{2\pi}\sum_{j\in\mathbb{Z}} \int\limits_{h^{-1}([-\pi,\pi]+2\pi j)} \widehat{f}(\theta)\ e^{ihk\theta}\,d\theta \\ &= \frac{1}{2\pi}\sum_{j\in\mathbb{Z}} h^{-1} \int\limits_{-\pi+2j\pi}^{\pi+2\pi j} \widehat{f}\left(h^{-1}\theta\right)\ e^{ik\theta}\,d\theta = \frac{1}{2\pi h}\sum_{j\in\mathbb{Z}} \int_{-\pi}^{\pi} \widehat{f}\left(h^{-1}(\theta+2\pi j)\right)\ e^{ik\theta}\,d\theta \end{aligned}$$

[32] Also wie $1/x$.

[33] Was ja bespielsweise von CD–Playern bekannt sein sollte.

$$= \frac{1}{2\pi h}\int_{-\pi}^{\pi}\left(\sum_{j\in\mathbb{Z}}\widehat{f}\left(h^{-1}(\theta+2\pi j)\right)\right)\,e^{ik\theta}d\theta, =: \frac{1}{2\pi}\int_{-\pi}^{\pi}F(\theta)\,e^{ik\theta},$$

das heißt

$$S_hf(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi}F(\theta)\,e^{ik\theta}, \qquad F = \frac{1}{h}\sum_{j\in\mathbb{Z}}\widehat{f}\left(h^{-1}(\cdot+2\pi j)\right) \tag{1.26}$$

wobei $F \in C(\mathbb{T}) \subset L_1(\mathbb{T})$ ist – die Funktion ist offensichtlich 2π–periodisch und wegen der Bandbeschränktheit von f ist für $\theta \in [0, 2\pi]$ die Summen nur endlich. Da die Exponentialfunktionen $e^{ik\cdot}$, $k \in \mathbb{Z}$, ein vollständiges Orthonormalsystem bilden, können wir aus (1.25) und (1.26) folgern, daß

$$h^{-1}\sum_{j\in\mathbb{Z}}\widehat{f}\left(h^{-1}(\theta+2\pi j)\right) = F(\theta) = (S_hf)^{\wedge}(\theta) = \sum_{j\in\mathbb{Z}}f(hj)\,e^{-ij\theta}, \qquad \theta\in\mathbb{T}. \tag{1.27}$$

Ist nun h so klein, daß

$$h^{-1}[-\pi,\pi] \supseteq [-T,T] \quad\Longleftrightarrow\quad [-\pi,\pi] \supseteq [-Th,Th] \quad\Longleftrightarrow\quad Th<\pi \quad\Longleftrightarrow\quad h<\frac{\pi}{T},$$

dann erhalten wir für $j > 0$ und $\theta \in [-\pi, \pi]$, daß

$$h^{-1}(\theta+2\pi j) > \frac{T}{\pi}(-\pi+2\pi j) \geq T(-1+2j) \geq T,$$

und analog für $j < 0$, daß $h^{-1}(\theta+2\pi j) < -T$. Das heißt aber, daß die Summe auf der linken Seite von (1.27) nur aus dem Term $j = 0$ besteht und wenn wir nun noch θ durch $h\xi$ ersetzen, dann erhalten wir, daß

$$\widehat{f}(\xi) = h\sum_{j\in\mathbb{Z}}f(hj)\,e^{-ijh\xi}$$

und somit ergibt sich, da $T < \pi/h$ und f T–bandbeschränkt ist,

$$\begin{aligned}
f(x) &= \frac{1}{2\pi}\int_{\mathbb{R}}\widehat{f}(\theta)\,e^{ix\theta}\,d\theta = \frac{1}{2\pi}\int_{-T}^{T}\widehat{f}(\theta)\,e^{ix\theta}\,d\theta = \frac{h}{2\pi}\int_{-T}^{T}\sum_{j\in\mathbb{Z}}f(hj)\,e^{i(x-jh)\theta}\,d\theta\\
&= \frac{h}{2\pi}\sum_{j\in\mathbb{Z}}f(hj)\int_{-\pi/h}^{\pi/h}e^{i(x-jh)\theta}\,d\theta = \frac{h}{2\pi}\sum_{j\in\mathbb{Z}}f(hj)\left[\frac{e^{i(x-jh)\theta}}{i(x-jh)}\Bigg|_{\theta=-\pi/h}^{\pi/h}\right]\\
&= \sum_{j\in\mathbb{Z}}f(hj)\underbrace{\frac{e^{i(x-jh)\pi/h}-e^{-i(x-jh)\pi/h}}{2i}}_{=\sin\pi(x/h-j)}\underbrace{\frac{h}{\pi}\frac{1}{(x-jh)}}_{=(\pi(x/h-j))^{-1}}\\
&= \sum_{j\in\mathbb{Z}}f(hj)\,\frac{\sin\pi(x/h-j)}{\pi(x/h-j)} = (S_hf * \mathrm{sinc})(\cdot/h),
\end{aligned}$$

was damit (1.24) liefert. □

Der Clou im Beweis von Satz 1.16 ist die Betrachtung der Identität

$$\sum_{k\in\mathbb{Z}} \widehat{f}\left(h^{-1}(\theta+2\pi k)\right) = h\sum_{k\in\mathbb{Z}} f(hk)\, e^{-ik\theta}, \qquad \xi\in\mathbb{T}, \tag{1.28}$$

die die *Periodisierung* der Fouriertransformierten einer Funktion mit der Fouriertransformierten der Abtastfolge verknüpft, und zwar *ohne* jedwede Forderung von Bandbeschränktheit[34] oder hinreichend feine Abtastung. Ist nun h so groß, daß die Funktion $\widehat{f}\,(h^{-1}\cdot)$ über das Intervall $[-\pi,\pi]$ "hinausragt", dann wird die Funktion durch die überlappenden Teile periodisch "verschmiert", siehe Abb. 1.2. Aus dieser Funktion läßt sich $\widehat{f}$ natürlich nicht mehr rekonstru-

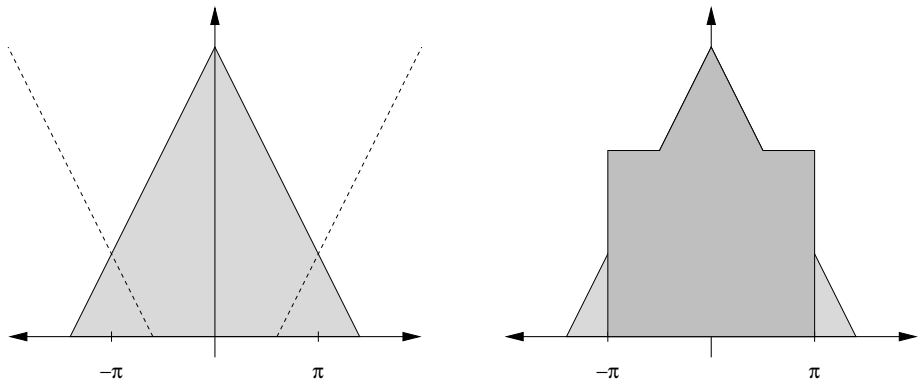


Abbildung 1.2: Periodisierung einer Funktion mit "zu großem" Träger, die schraffiert unterlegte Funktion wird dann periodisch fortgesetzt. Natürlich ist es nicht möglich, die Funktion links *eindeutig* aus der Periodisierung zu rekonstruieren.

ieren. Aber es ist sogar noch schlimmer: Durch die Überlagerung von Frequenzen die eigentlich nichts miteinander zu tun haben und nun modulo 2π betrachtet werden, kommt es bei der Rekonstruktion des Signals im Falle von Unterabtastung zu sehr unerwünschten Effekten, die man als *Aliasing* bezeichnet.

Anders wird die Sache, wenn man h so wählt, daß der Träger von $\widehat{f}$ *ganz* in $[-\pi,\pi]$ liegt, denn dann gibt es keine Überlappung und die 2π–periodisch Fourierreihe $(S_h f)^\wedge$ ist gleich der Periodisierung von $\widehat{f}$. Und was dann noch kam war reine Rechnerei ...

1.4 Unschärfe

Aufgrund des Abtastsatzes 1.16 sind offenbar bandbeschränkte Funktionen besonders wichtig, da sie durch durch Diskretisierung *verlustfrei* in Folgen umgewandelt werden können und so am Rechner verarbeitet werden können. Nur hat so ein Rechner natürlich auch die schlechte Angewohnheit, in endlicher Zeit nur *endliche* Folgen verarbeiten zu können. Es wäre also schön,

[34] Außer man möchte, daß die Summe auf der linken Seite von (1.28) existiert, dann sollte vielleicht doch zumindest $\widehat{f}\in L_1(\mathbb{R})$ sein.

wenn die diskretisierte Folge endlich wäre, was am einfachsten dadurch gewährleistet würde[35], wenn $f \in L_{00}(\mathbb{R})$ wäre. Genau das klappt aber *nicht*: Es gibt genau eine bandbeschränkte Funktion mit kompaktem Träger und das ist die Nullfunktion. Diese etwas unerfreuliche Tatsache ist eine Konsequenz der berühmten *Heisenbergschen*[36] *Unschärferelation*. Dazu definieren wir für $f \in L_2(\mathbb{R})$ die *Varianz*

$$V(f) := \left(\int_{\mathbb{R}} x^2 \, |f(x)|^2 \, dx \right)^{1/2},$$

also das *zweite Moment* von $|f|$. Die Größe $V(f)/\|f\|_2$ wird in [17] auch als *effektive Dauer* von f, die Größe $V(\widehat{f})/\|f\|_2$ als *effektive Bandbreite* bezeichnet. Ist f eine zeitbeschränkte Funktion, also $f(x) = 0$, $x \notin [-S, S]$, dann ist

$$V^2(f) = \int_{-S}^{S} x^2 \, |f(x)|^2 \, dx \le S^2 \underbrace{\int_{-S}^{S} |f(x)|^2 \, dx}_{=\|f\|_2}$$

und damit ist die effektive Dauer immer kleiner gleich der[37] wirklichen Dauer und dasselbe gilt natürlich auch für die Bandbreite. Dann gilt die folgende Aussage.

Satz 1.18 (Heisenbergsche Unschärferelation) *Für $f \in L_2(\mathbb{R})$ mit $x\, f, \xi\, \widehat{f}(\xi) \in L_2(\mathbb{R})$ ist*

$$V(f)\; V(\widehat{f}) \ge \frac{\|f\|_2^2}{4\pi}. \tag{1.29}$$

Ein Beweis findet sich beispielsweise in [38, S. 31–32]. Die Heisenbergsche Unschärferelation sagt uns, daß die effektive Dauer eines Signals und die effektive Bandbreite nicht *gleichzeitig* beliebig klein werden können. Und das heißt auch, daß die Bandbreite einer Funktion und ihrer Fouriertransformierten nicht gleichzeitig zu klein werden können: Ist nämlich f eine S–zeitbeschränkte und T–bandbeschränkte Funktion, dann liefern

$$V^2(f) = \int_{-S}^{S} x^2 \, |f(x)|^2 \, dx \le S^2 \|f\|_2^2 \qquad \text{und} \qquad V^2(\widehat{f}) \le T^2 \|\widehat{f}\|_2^2,$$

zusammen mit (1.29), daß $ST \ge (4\pi)^{-1}$ sein muß. Das allein schließt natürlich kompakte Träger nicht aus, nur sehr kleine kompakte Träger auf beiden Seiten. Aber trotzdem geht es nicht, es gilt sogar die wesentlich schärfere Aussage, daß eine von Null verschiedene bandbeschränkte Funktion auf auf keinem nichttrivialen Intervall verschwinden darf, ein Verhalten, das man ja auch von Polynomen kennt.

[35] Die vielen Konjunktive machen nicht gerade Hoffnung.

[36] Werner Karl Heisenberg, 1901–1976, "a rather stiff, tightly controlled, authoritarian figure", Vater der Quantenphysik. Sagte über sich selbst: *I learned optimism from Sommerfeld, mathematics at Gttingen, and physics from Bohr.*

[37] Wann tritt denn Gleichheit ein?

Proposition 1.19 *Ist $f \in L_1(\mathbb{R})$ eine bandbeschränkte Funktion und gibt es ein Intervall $[a,b]$, $a<b$, so daß $f(x)=0$, $x\in[a,b]$, dann ist $f=0$.*

Beweis: Wegen der Bandbeschränktheit ist $\widehat{f}\in L_1(\mathbb{R})$ und somit

$$f(x)=\widehat{f}^{\vee}(x)=\frac{1}{2\pi}\int_{\mathbb{R}}\widehat{f}(\theta)\;e^{ix\theta}\,d\theta=\frac{1}{2\pi}\int_{-T}^{T}\widehat{f}(\theta)\;e^{ix\theta}\,d\theta.$$

Damit ist für $x_0:=\frac{a+b}{2}$ und $n\geq 0$

$$0=f^{(n)}\,(x_0)=\frac{1}{2\pi}\int_{-T}^{T}\widehat{f}(\theta)\;\frac{d^n}{dx^n}\,e^{ix_0\theta}\,d\theta=\frac{1}{2\pi}\int_{-T}^{T}\widehat{f}(\theta)\;(i\theta)^n e^{ix_0\theta}\,d\theta \tag{1.30}$$

Andererseits ist, unter Verwendung der Reihenentwicklung $e^z=\sum_k\frac{z^k}{k!}$ und von (1.30) dann aber für $x\in\mathbb{R}$

$$\begin{aligned} f(x) &= \widehat{f}^{\vee}(x)=\frac{1}{2\pi}\int_{-T}^{T}\widehat{f}(\theta)\;e^{i(x-x_0)\theta}e^{ix_0\theta}\,d\theta=\frac{1}{2\pi}\int_{-T}^{T}\widehat{f}(\theta)\;\left(\sum_{k=0}^{\infty}\frac{(x-x_0)^k\;(i\theta)^k}{k!}\right)\;e^{ix_0\theta}\,d\theta\\ &= \frac{1}{2\pi}\sum_{k=0}^{\infty}\frac{(x-x_0)^k}{k!}\underbrace{\int_{-T}^{T}\widehat{f}(\theta)\;(i\theta)^n e^{ix_0\theta}\,d\theta}_{=0}=0, \end{aligned}$$

was genau die Behauptung ist. □

Beim Beweis von Proposition 1.19, der aus [38, Theorem 2.6] stammt, ist zuerst vielleicht nicht so ganz klar, wozu man eigentlich den kompakten Träger der Fouriertransformierten braucht. Das kommt aber einfach daher, daß die Integrale

$$\int_{\mathbb{R}}\widehat{f}(\theta)\;(i\theta)^n\;e^{ix\theta}\,d\theta,\qquad x\in\mathbb{R},\quad n\in\mathbb{N}_0,$$

nur dann existieren, wenn $\widehat{f}(\theta)\;(i\theta)^n$ *für alle* $n\in\mathbb{N}_0$ eine L_1–Funktion ist, wenn also $\widehat{f}$ hinreichend schnell abfällt. Tatsächlich ist *exponentieller* Abfall der Fouriertransformierten,

$$\widehat{f}\,e^{a|\cdot|}\in L_2(\mathbb{R})\qquad\text{für ein } a>0$$

mehr oder weniger äquivalent dazu, daß f im Streifen $\{x+iy\;:\;|y|<a\}\subset\mathbb{C}$ eine *analytische* Funktion ist[38]; das ist der berühmte *Satz von Paley*[39]*–Wiener*[40], siehe z.B. [33, S. 174].

[38] Und deren Taylorreihe bezüglich eines beliebigen Punkts konvergiert ja überall, also kann, wie auch bei einem Polynom diese Funktion auf keinem Teilintervall verschwinden, ein weiterer Beweis für Proposition 1.19.

[39] Raymond Edward Alan Christopher Paley, 1907–1933, Schüler von Littlewood, Zusammenarbeit mit Zygmund, Polyá und Wiener, wurde beim Skifahren in Kanada durch eine Lawine getötet.

[40] Norbert Wiener, 1894–1964, studierte zuerst Zoologie, später mathematische Philosophie, Dissertation über formale Logik. Bekannt sind auch seine Beiträge zu stochastischen Prozessen, insbesondere zur Brownschen Bewegung.

1.5 Filter

Jetzt aber endlich zu den Objekten der Begierde der digitalen Signalen, nämlich den Filtern, denn die sind es ja letztendlich, die gerade den Job der Verarbeitung der Signale erledigen sollen. Ein *Filter* ist nun nichts anderes als eine Abbildung von einem Signalraum in einen anderen; manchmal spricht man auch von einem *System*, das ein diskretes oder kontinuierliches Sigal in ein Signal gleicher Bauart (also wieder diskret oder kontinuierlich) umsetzt. Wir wollen uns hier, schon aufgrund der "Praxisnähe" mit *diskreten* Filtern befassen, also Filtern, die diskrete Signale in diskrete Signale überführen.

Definition 1.20 (Filter und Filtertypen) *Ein* Filter F *ist ein Operator, der* $c \in \ell(\mathbb{Z})$ *in* $Fc \in \ell(\mathbb{Z})$ *abbildet. Man spricht von*

1. energieerhaltenden *Filtern, wenn* $F : \ell_2(\mathbb{Z}) \to \ell_2(\mathbb{Z})$ *und*

$$1 = \|F\|_2 := \sup_{\|c\|_2=1} \|Fc\|_2$$

   ist.

2. linearen *Filtern, wenn*

$$F\left[\alpha c + \beta c'\right] = \alpha\, Fc + \beta\, Fc', \qquad \alpha, \beta \in \mathbb{R}, \quad c, c' \in \ell(\mathbb{Z}).$$

3. zeitinvarianten *Filtern, wenn der Operator* stationär *ist, d.h. seine Handlung nicht vom jeweiligen Zeitpunkt abhängt*[41]*:*

$$F\left[c\left(\cdot + k\right)\right] = \left[Fc\right]\left(\cdot + k\right), \qquad c \in \ell(\mathbb{Z}), \quad k \in \mathbb{Z}.$$

   Mit Hilfe des diskreten Translationsoperators τ*, definiert durch* $\tau c = \tau_1 c = c(\cdot + 1)$ *kann man das auch recht komfortabel als die "Kommutationseigenschaft"*

$$\tau F = F\tau \qquad \textit{bzw.} \qquad \tau_k F = F\tau_k, \quad \tau_k = \tau^k,$$

   schreiben.

4. *Ein Filter heißt* kausal*, wenn das Ergebnis zum Zeitpunkt* k *nur von den Eingaben* $c(j)$*,* $j \le k$*, abhängt, der Filter kann also nicht in die Zukunft sehen.*

Lineare und zeitinvariante Filter, nach [32] *"LTI–Filter"*, in [24] direkt als *"digitaler Filter"* eingeführt, sind die richtig schön strukturierten Filter. Sie haben nämlich die Eigenschaft, daß sie nur von der *Impulsantwort*

$$f := F\delta, \qquad \text{also} \qquad f(k) = \left[F\delta\right](k), \quad k \in \mathbb{Z},$$

[41] Einfachstes Beispiel: Wenn man den CD–Player eine Stunde später mit derselben CD anwirft, dann sollte auch dieselbe Musik rauskommen.

abhängen und das auch noch auf ziemlich strukturierte Art und Weise. Da man jedes Signal $c \in \ell(\mathbb{Z})$ formal als

$$c = \sum_{k\in\mathbb{Z}} c(k)\,\delta(\cdot - k) = \sum_{k\in\mathbb{Z}} c(k)\,\tau_{-k}\delta$$

schreiben kann, ist wegen der Linearität und der Zeitinvarianz

$$\begin{aligned} Fc &= F\left[\sum_{k\in\mathbb{Z}} c(k)\,\tau_{-k}\delta\right] = \sum_{k\in\mathbb{Z}} c(k)\,F\left[\tau_{-k}\delta\right] = \sum_{k\in\mathbb{Z}} c(k)\,\tau_{-k}F\delta = \sum_{k\in\mathbb{Z}} c(k)\,\tau_{-k}f \\ &= \sum_{k\in\mathbb{Z}} c(k)\,f\,(\cdot - k) = c * f, \end{aligned}$$

weswegen lineare, zeitinvariante Filter immer *Faltungen* mit der Impulsantwort sind. Damit wissen wir aber auch, wie so ein Filter oder System im *Frequenzbereich* funktioniert, nämlich ganz einfach als

$$(Fc)^{\wedge}(\xi) = (f * c)^{\wedge}(\xi) = \widehat{f}(\xi)\,\widehat{c}(\xi). \tag{1.31}$$

Im Frequenzbereich ist also ein LTI–Filter immer linear, also eine einfache Multiplikation zwischen der Fouriertransformierten des Filters, der sogenannten *Transferfunktion* des Filters und der Fouriertransformierten des Signals. Und oftmals werden auch Eigenschaften des Filters, wie beispielsweise Hoch-, Tief- oder Bandpass über deren Fouriertransformierte gefordert.

Was sind aber nun Filter, die man wirklich in der Praxis realisieren kann? Nun, zuerst sollten diese Filter natürlich einmal kausal sein, das heißt, daß für $k < 0$ der Impuls δ keinen Beitrag liefern darf, daß also

$$0 = [F\delta]\,(k) = f(k), \qquad k < 0,$$

ist. Ein Filter mit der Eigenschaft, daß $f(k) = 0$, $k > 0$, heißt im Übrigen *antikausal*[42]. Auch nichtkausale Filter können Sinn machen, und zwar dann, wenn das Signal zeitbeschränkt ist[43], gespeichert und in beide Richtungen abgearbeitet werden kann.

Von jetzt an, soll die Bezeichnung "digitaler Filter" immer für einen LTI–Filter stehen. Bevor wir uns mit der praktischen Realisierung befassen, noch zwei Begriffe.

Definition 1.21 (Weitere Filtertypen) *Ein digitaler Filter F heißt*

1. "FIR[44]–Filter" *oder* Filter mit endlicher Impulsantwort*, wenn die Impulsantwort endlichen Träger hat, wenn also $F\delta \in \ell_{00}(\mathbb{Z})$ liegt.*

2. "IIR[45]–Filter" *oder* Filter mit unendlicher Impulsantwort*, wenn die Impulsantwort* keinen *endlichen Träger hat.*

Jetzt können wir uns überlegen, wie man einen Digitalfilter, genauer einen *kausalen FIR–Filter* in der Praxis realisiert. Dazu braucht man "nur" drei Schaltglieder, nämlich

[42] So ein Filter wird dadurch realisiert, daß man die Zeit rückwärts laufen lässt.
[43] Beispielsweise bei Bildern, die sind zwar ausdehnungsbeschränkt, aber das läuft hier aufs gleiche raus.
[44] **F**inite **I**mpulse **R**esponse
[45] **I**nfinite **I**mpulse **R**esponse

- einen *Multiplizierer*, der eine Zahl mit einer *festen* Konstanten c multipliziert,
- einen *Addierer*, der zwei Zahlen miteinander addiert,
- einen *Verzögerer*, der einen Wert *eine* Zeiteinheit lang speichern kann.

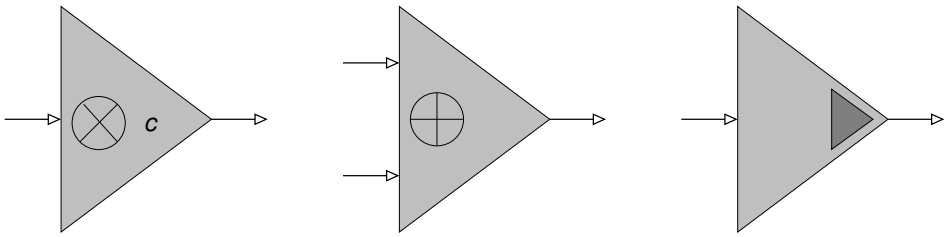


Abbildung 1.3: Symbolische Darstellung der drei Bausteine: Multiplizierer (a), Addierer (b) und Verzögerer (c).

Die Symbole für diese drei Bausteine sind in Abb. 1.3 dargestellt. Da ein Verzögerer, angewandt auf ein Signal $c \in \ell(\mathbb{Z})$ das Signal $\tau_{-1}c$ liefert und somit eine Kette von k Verzögerern das Signal $\tau_{-k}c$, kann man einen kausalen FIR–Filter F, der ja die Form

$$Fc = f * c = \sum_{k\in\mathbb{Z}} f(k)\, c\,(\cdot - k) = \sum_{k=0}^{N} f(k)\, \tau_{-k}c, \qquad \text{supp} f \subseteq [0, N],$$

hat, mit Hilfe von N unserer Bausteine darstellen: Die Werte $\tau_{-k}c$, $k = 0, \ldots, N$, werden von einer Kette von $N + 1$ Verzögerern abgegriffen, jeweils durch Multiplizierer mit den Werten $f(k)$ gewichtet und dann mit N Summierern aufsummiert. Diese Vorgehensweise ist in Abb. 1.4 dargestellt. Mit Hilfe von M Speichereinheiten könnte man übrigens auch einen Filter realisieren, der $\text{supp}\, F \subseteq [-M, N]$ erfüllt, aber das Ergebnis um M Zeiteinheiten *verzögert* ausgibt – womit sich dann *alle* FIR–Filter praktisch realisieren lassen.

Jetzt wollen wir uns mal ein einfaches Beispiel aus [24, Sec. 3.8] für Filterdesign ansehen und zwar einen symmetrischen, *nichtkausalen* FIR–Filter mit

$$f(0) = a, \qquad f(\pm 1) = b, \qquad f(\pm 2) = c.$$

Um die Symmetrie $f(k) = f(-k)$ und somit eine reelle Transferfunktion zu erhalten, ist es vorteilhaft, besser auf die Kausalität zu verzichten.

Übung 1.7 Zeigen Sie: Ist $f \in \ell_{00}(\mathbb{Z})$ der reelle Koeffizientenvektor eines Filters F, dann ist die Transferfunktion $\widehat{f}$ genau dann reell, wenn f symmetrisch ist und genau dann rein imaginär, wenn f antisymmetrisch ist, d.h. $f(k) = -f(-k)$. ♢

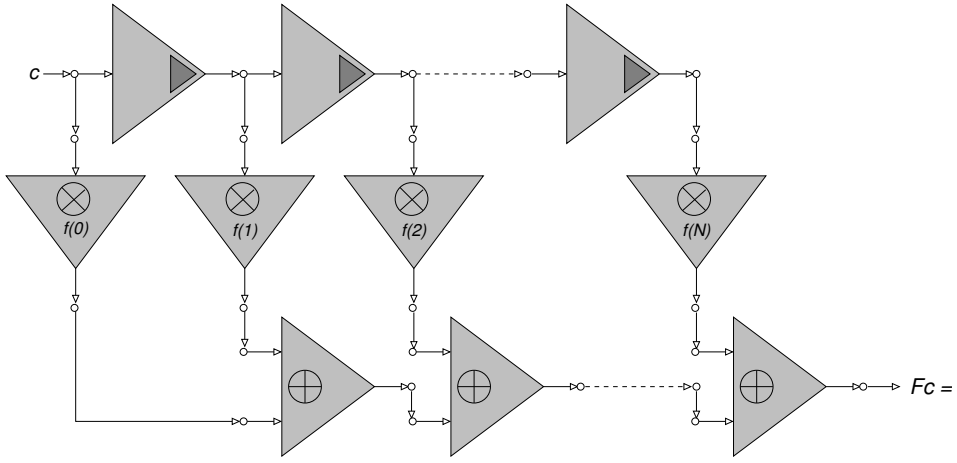


Abbildung 1.4: Ein FIR–Filter als Kaskade der Bausteine aus Abb. 1.3. Die Verzögerer sorgen für die Translationen, die Multiplizierer für die Gewichtung und die Addierer summieren den ganzen Kram auf.

Das *Design* des Filters wird nun meistens im *Frequenzbereich* festgelegt, das heißt, man stellt Forderungen an die Transferfunktion[46] $\widehat{f}(\xi)$. Die Forderung hier soll sein, daß der Filter im niederfrequenten Bereich voll durchlässig ist und im hochfrequenten Bereich sperrt, also

$$\widehat{f}(0) = 1, \qquad \widehat{f}(\pi) = 0. \tag{1.32}$$

Mit

$$\widehat{f}(\xi) = a + b\,\left(e^{-i\xi} + e^{i\xi}\right) + c\,\left(e^{-2i\xi} + e^{2i\xi}\right) = a + 2b\cos\xi + 2c\cos 2\xi$$

heißt (1.32), daß

$$a + 2b + 2c = 1, \qquad \text{und} \qquad a - 2b + 2c = 0,$$

also $b = \frac{1}{4}$ und $a = \frac{1}{2} - 2c$, also

$$\begin{aligned}
\widehat{f}(\xi) &= \frac{1}{2} - 2c + \frac{1}{2}\cos\xi + 2c\,\cos 2\xi = \frac{1}{2} - 2c + \frac{1}{2}\cos\xi + 2c\left(2\cos^2\xi - 1\right) \\
&= \frac{1}{2} - 4c + \frac{1}{2}\,\cos\xi + 4c\,\cos^2\xi = 4c\,\left[\cos^2\xi + \frac{1}{8c}\cos\xi + \frac{1}{8c} - 1\right] \\
&= 4\,(\cos\xi + 1)\left(c\,\cos\xi + \frac{1}{8} - c\right).
\end{aligned}$$

Der erste Faktor, $\cos\xi + 1$, sorgt dabei für die Nullstelle an $\xi = \pi$, der zweite degeneriert zu einer Konstanten, wenn $c = 0$ ist, also wird mit Sicherheit der Fall $c = 0$ besonders sein[47]

[46] Diese ist eine 2π–periodische Funktion auf $\mathbb{R}$, oder eben auch eine Funktion auf $\mathbb{T}$.

[47] Kein Wunder, denn dann hat ja das trigonometrische Polynom $\widehat{f}$ Grad 1 und nicht Grad 2.

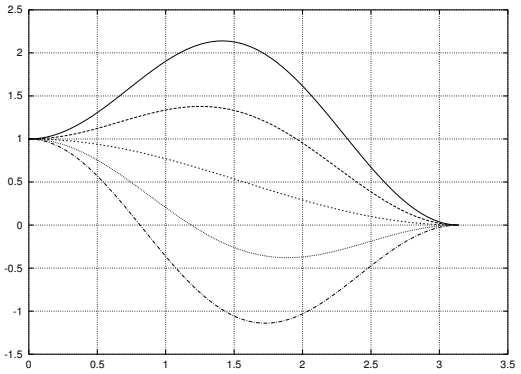


Abbildung 1.5: Beispiele für Transferfunktionen mit $c = -.4, -.2, 0, .2, .4$ (von oben nach unten). Besonders interessant ist der Fall $c = 0$, denn da bewegt sich der Filter nur zwischen 1 und 0 und stellt eine sogenannte *sigmoidale* Funktion dar.

Einige Beispiele für die Transferfunktion dieser Filter mit variierendem Parameter c finden sich in Abb. 1.5.

Man kann aber diesen *Designparameter* c nun auch so wählen, daß der resultierende Filter weitere Eigenschaften besitzt, beispielsweise:

- *Erhaltung eine weiteren Frequenz* ω, d.h.

  $$1 = \widehat{f}(\omega) = 4\left(\cos\omega + 1\right)\left(c\,\cos\omega + \frac{1}{8} - c\right)$$

  also

  $$c = \left(\frac{1}{4\left(\cos\omega + 1\right)} - \frac{1}{8}\right) / \left(\cos\omega - 1\right) = -\frac{1}{8\left(\cos\omega + 1\right)}$$

  was für $\omega \neq \pi$ immer lösbar ist.

- *So flach wie möglich an* $\xi = 0$, d.h., es verschwinden so viele Ableitungen von $\widehat{f}$ wie möglich an $\xi = 0$, so daß sich der Filter "soweit wie möglich" einer "charakteristischen Funktion" annähert. Da

  $$\frac{d}{d\xi}\widehat{f}(\xi) = -2b\,\sin\xi - 4c\sin 2\xi$$

  an $\xi = 0$ *immer* verschwindet, können wir also fordern, daß

  $$0 = \frac{d^2}{d\xi^2}\widehat{f}(0) = -2b\,\cos 0 - 8c\cos 0 \qquad \Longrightarrow \qquad c = -\frac{1}{4}\,b = -\frac{1}{16}.$$

- *"Balanciertheit"* oder *Antisymmetrie um* $\frac{\pi}{2}$, d.h.

  $$\widehat{f}\left(\frac{\pi}{2} - \xi\right) - \widehat{f}\left(\frac{\pi}{2}\right) = \widehat{f}\left(\frac{\pi}{2}\right) - \widehat{f}\left(\frac{\pi}{2} + \xi\right),$$

also

$$\frac{1}{2}\left[\widehat{f}\left(\frac{\pi}{2}-\xi\right)+\widehat{f}\left(\frac{\pi}{2}+\xi\right)\right]=\widehat{f}\left(\frac{\pi}{2}\right)$$

und somit insbesondere ($\xi=\frac{\pi}{2}$)

$$\frac{1}{2}=\widehat{f}\left(\frac{\pi}{2}\right)=a+2b\underbrace{\cos\frac{\pi}{2}}_{=0}+2c\underbrace{\cos\pi}_{=-1}\qquad\Longrightarrow\qquad a=\frac{1}{2}+2c$$

was zusammen mit $a=\frac{1}{2}-2c$ die Bedingungen $a=\frac{1}{2}$ und $c=0$ liefert.

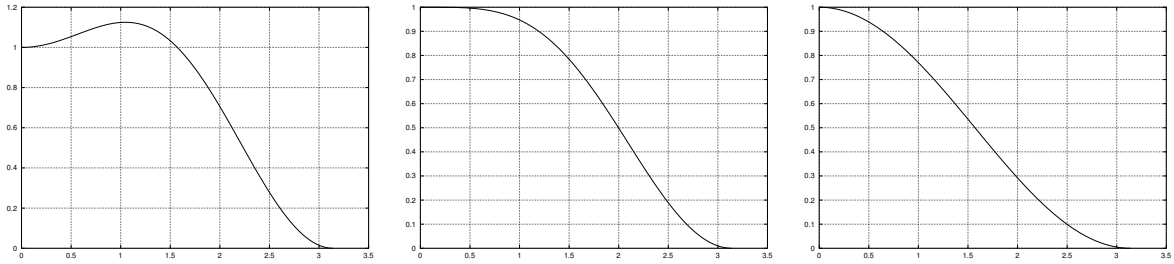


Abbildung 1.6: Die drei Filter zu den zusätzlichen Forderungen: Erhaltung der Frequenz $\frac{\pi}{2}$ (links), "Flachheit" an 0 (mitte) und Balanciertheit (rechts).

Diese drei Filter sind in Abb. 1.6 dargestellt.

Auch wenn dieses Beispiel sehr einfach ist, zeigt es bereits, wie beim Filterdesign vorgegangen wird: Man gibt sich normalerweise das Verhalten des Filters im Frequenzbereich vor und bestimmt dann die Koeffizienten des Filters so, daß diese Verfahren realisiert oder zumindest approximiert wird. Und das zu ist einiges zu sagen.

Bemerkung 1.22 (Filterdesign/Transferfunktion)

1. *Die Transferfunktion eines Filters, insbesondere eines nichttrivialen kausalen Filters, ist normalerweise eine* komplexwertige *Funktion. Deren* Realteil *liefert den* Gewichtungsfaktor *für die jeweilige Frequenz während ihr* Imaginärteil *die* Phasenverschiebung *angibt. Da letztere eigentlich nicht hörbar ist, gibt man oftmals bei Bandpassfiltern nur den Realteil des Filters vor.*

2. *Es mag zuerst etwas seltsam anmuten, daß der Frequenzgang des Filters immer nur von $-\pi$ bis π läuft, schließlich hätte man doch eigentlich gerne Filter, die beispielsweise hörbare Frequenzen von $10-20000$ Hz verarbeiten können. Und hier ist der Punkt, an dem der Abtastsatz 1.16 ins Spiel kommt: Der Filter operiert auf dem* diskreten *Signal, das eben durch hinreichend feines Abtasten gewonnen sein muß, und zwar so fein, daß die Fouriertransformierte des Signals auf das* relevante *Band $[-\pi,\pi]$ beschränkt wird.*

Das viel größere Problem besteht aber darin, daß die Transferfunktionen von FIR–Filtern nur eine sehr "kleine" Familie von Funktionen aus $L_2(\mathbb{T})$, oder was auch immer man als Filtermenge nehmen möchte, darstellen und man so bei weitem nicht jeden Filter, den man gerne hätte, auch wirlich exakt als FIR–Filter bekommt.

Definition 1.23 *Eine Funktion $f \in C^\infty(\mathbb{T})$ der Form*

$$f(x) = \sum_{|k|\le n} f_k\, e^{ikx}, \qquad f_k \in \mathbb{C}, \quad x \in \mathbb{T},$$

heißt trigonometrisches Polynom *der Ordnung n.*

Übung 1.8 Zeigen Sie: Jedes trigonometrische Polynom der Ordnung n läßt sich auch als

$$f(x) = a_0 + \sum_{k=1}^{n} \left[a_k\, \cos^k x + b_k \sin^k x\right], \qquad a_0, \ldots, a_n,\, b_1, \ldots, b_n \in \mathbb{C}$$

schreiben. ♢

Es ist nun klar, daß ein LTI–Filter F, dargstellt durch $f \in \ell(\mathbb{Z})$, genau dann ein FIR–Filter ist, wenn $f \in \ell_{00}(\mathbb{Z})$, also seine Fouriertransformation ein trigonometrisches Polynom ist. Eigentlich nicht so schlimm, denn die trigonometrischen Polynome sind *dicht* in $L_2(\mathbb{T})$ und wir wissen sogar, wie man die *beste Approximation* zu einer Transferfunktion $g = \widehat{f}$ berechnet: Man nimmt die Fourierreihe

$$\mathscr{F} g = \sum_{k\in\mathbb{Z}} g_k\, e^{ik\cdot}, \qquad g_k = \frac{1}{2\pi} \int_{\mathbb{T}} g(t)\, e^{-ikt}$$

und bildet deren n–te *Partialsumme*

$$\mathscr{F}_n g := \sum_{|k|\le N} g_k\, e^{ik\cdot} =: \widehat{h},$$

was uns auch schon die Koeffizienten unseres Filters h liefert. Und $\widehat{h}$ ist sogar dasjenige trigonometrische Polynom der Ordnung n, das in der Norm von $L_2(\mathbb{T})$ die Transferfunktion am besten annähert, also eine *Bestapproximation*, siehe z.B. [46]. Mehr zu Fourierreihen findet man beispielsweise in [26, 33, 61]. Leider ist das aber auch nicht so einfach.

Beispiel 1.24 (Tiefpassfilter) *Wir wollen jetzt einen Filter F designen, der nur die tiefen Frequenzen durchläßt, sagen wir die untere Hälfte. Ideal wäre dieser Filter also durch*

$$\widehat{f} = \chi_{[-\pi/2,\pi/2]}$$

definiert – das ist natürlich kein trigonomentrisches Polynom und somit auch kein FIR–Filter, also auch nicht praktisch realisierbar. Um also näherungsweise Tiefpassfilter zu bekommen, bestimmen wir die Fourierkoeffizienten von $g = \widehat{f}$ und erhalten, daß

$$g_0 = \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} dt = \frac{1}{2}$$

und, für $k \neq 0$*,*

$$\begin{aligned} g_k &= \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} e^{-ikt}dt = \frac{1}{2\pi}\,\frac{e^{-ikt}}{-ik}\Big|_{t=-\pi/2}^{\pi/2} = \frac{1}{2\pi}\,\frac{e^{-ik\pi/2}-e^{ik\pi/2}}{-ik} = \frac{1}{k\pi}\sin\frac{k}{2}\pi \\ &= \begin{cases} 0, & k \in 2\mathbb{Z}, \\ \frac{(-1)^{(k-1)/2}}{k\pi}, & k \in 2\mathbb{Z}+1, \end{cases} \end{aligned}$$

das heißt,

$$g_{2k+1} = \frac{(-1)^k}{k\pi}, \qquad g_{2k+2} = 0, \qquad k \in \mathbb{N}_0.$$

Die Partialsumme der Ordnung $n = 2m+1$ *ist also*

$$\begin{aligned} h_n(\xi) &= \frac{1}{2} + \sum_{k=0}^{m} \frac{(-1)^k}{(2k+1)\pi}\,\underbrace{\left[e^{i(2k+1)\xi} + e^{-i(2k+1)x}\right]}_{2\cos(2k+1)\xi} \\ &= \frac{1}{2} + \sum_{k=0}^{m} (-1)^k \frac{2}{(2k+1)\pi}\,\cos(2k+1)\xi, \end{aligned}$$

was uns den näherungsweisen FIR–Filter liefert. Nur ist es mit der Approximationsqualität nicht so weit her, wie uns Abb. 1.7 zeigt.

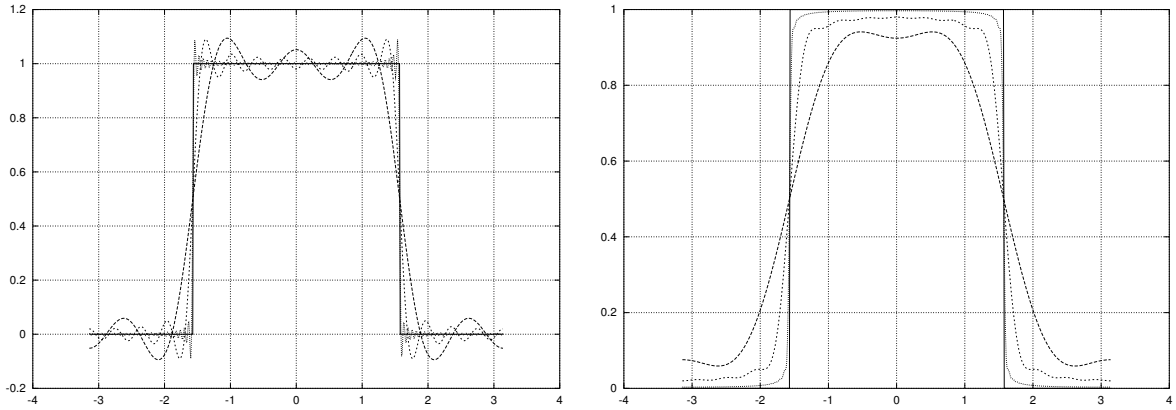


Abbildung 1.7: *Links:* Approximation des Bandpassfilters durch Partialsummen für $n = 5, 15, 100$ (Werte eher zufällig). Man beachte, daß die "Überschiesser" nur schmaler, nicht aber kleiner werden.

Rechts: Approximation durch ein anderes Approximationsverfahren, nämlich die *Fejérschen Mittel*. Diese haben zwar eine größere Abweichung vom Bandpass als die Partialsummen, verzichten dafür aber auf wilde Oszillationen.

Abb. 1.7 zeigt das sogenannte *Gibbs*[48] *–Phänomen*: Die Partialsummen zu "steilflankigen" Filtern liefern immer Überschwingphänomene, die auch mit steigender Approximationsqualität nicht verschwinden. Dies macht FIR–Filter für die Konstruktion von Bandpassfiltern recht ungeeignet. Außerdem ist die Qualität, mit der man nichtglatte Funktionen durch trigonometrische Polynome approximieren kann, die sogenannte *Approximationsordnung*, ebenfalls stark eingeschränkt, siehe z.B. [11, 35, 39, 46]. Man kann, wie das rechte Bild in Abb. 1.7 zeigt, dem Gibbs–Phänomen durch Wahl eines geeigneten Approximationsverfahrens entgehen[49], das "gestalterhaltende[50]" Eigenschaften besitzt, dafür aber mit langsamerer Konvergenz bezahlt – für Details siehe z.B. nochmals [11, 35, 39, 46].

[48] Josiah Willard Gibbs, 1839–1903, Professor für mathematische Physik in Yale; das Gibbs–Phänomen wurde aber angeblich nicht von ihm entdeckt.

[49] Anstelle der n–ten Partialsumme betrachten man hier das *arithmetische Mittel* der Partialsummen der Ordnungen $0, 1, \ldots, n$.

[50] "Shape preserving".

Contrariwise, [. . .] if it was so, it might be; and if it were so, it would be; but as it isn't, it ain't. That's logic.

L. Carroll, *Through the looking glass*

Ein größeres Repertoire an Filtern

2

Wie Beispiel 1.24 gezeigt hat weisen die FIR–Filter oder äquivalent die trigonometrischen Polynome gewisse Approximationsprobleme auf und zwingen beim Filterdesign eine Entscheidung zwischen

- *Effizienz:* Eine hohe *Approximationsordnung* gewährleistet eine möglichst gute Annäherung des Wunschverhaltens unter Verwendung von möglichst wenig Schaltgliedern. Denn natürlich kostet bei einer Implementierung jedes Schaltglied bzw. jeder *Tap* des Filters einen gewissen Aufwand.
- *Verzögerung:* Das viel größere Problem bei "langen" Filtern ist aber die Tatsache, daß in jedem Taktschritt die Daten im Filter nur eine Stelle "weiterrücken", je größer also der Träger des Filters ist, desto länger dauert es, bis die Daten wirklich verarbeitet sind und das Ergebnis wirklich am hinteren Ende des Filters ankommt. Ist nun diese sogenannte *Latenzzeit* relativ groß, kann es, gerade in Echtzeitanwendungen, zu Problemen kommen. Beispielsweise können Filter mit unterschiedlichen Latenzzeiten, wenn sie parallel verwendet werden, in der Soundverarbeitung zu Phasenverschiebungen führen, die als ein Jaulen wahrgenommen werden können.
- *Gestalterhaltung:* Der Filter weicht zwar weiter vom Wunschfilter ab, ähnelt diesem aber dafür optisch.

zu treffen. Und in beiden Fällen muss man deutliche Abstriche machen.
Aus diesem Grund ist es sicherlich vernünftig, sich zu fragen, ob es außer den FIR–Filtern[51] vielleicht noch andere Filtertypen gibt, die sich mit den drei Schaltgliedern realisieren lassen und die andererseits ein besseres Approximationsverhalten zeigen können.

2.1 Die z–Transformation

Die Transformation, die wir jetzt einführen, ist eigentlich nur eine Variante der Fouriertransformation, aber dann auch wieder nicht so wirklich. Wir werden nämlich jetzt Folgen in formale Potenzreihen transformieren.

[51] Wir erinnern uns: Kausal/nichtkausal war nicht wirklich das Problem!

Definition 2.1 (Laurentreihen und z–Transformation)

1. *Mit* $\Pi[\mathbb{C}]$ *bezeichnen wir den Ring aller* Polynome *mit komplexen Koeffizienten und mit* $\Lambda[\mathbb{C}]$ *den Ring der* Laurentpolynome*, also*

$$\Pi[\mathbb{C}] := \left\{ f(z) = \sum_{j=0}^{n} f_j \, z^j \ : \ f_j \in \mathbb{C},\, n \in \mathbb{N}_0 \right\}$$

und

$$\Lambda[\mathbb{C}] := \left\{ f(z) = \sum_{j=-n}^{n} f_j \, z^j \ : \ f_j \in \mathbb{C},\, n \in \mathbb{N}_0 \right\}$$

2. *Mit* $\Pi(\mathbb{C})$ *und* $\Lambda(\mathbb{C})$ *bezeichnen wir die* formalen[52] Potenz- *bzw.* Laurentreihen *der Form*[53]

$$f(z) = \sum_{k\in\mathbb{N}_0} f_k \, z^k \quad \textit{bzw.} \quad f(z) = \sum_{k\in\mathbb{Z}} f_k \, z^k, \qquad z \in \mathbb{C}^\times := \mathbb{C} \setminus \{0\}.$$

3. *Die* z–Transformierte $c^* \in \Lambda(\mathbb{C})$ *eines Signals* $c \in \ell(\mathbb{Z})$ *ist definiert als*

$$c^*(z) := \sum_{k\in\mathbb{Z}} c(k) \, z^{-k}, \qquad z \in \mathbb{C}^\times. \tag{2.1}$$

Übung 2.1 Zeigen Sie:

1. Jedes $f \in \Lambda[\mathbb{C}]$ kann als $f(z) = z^k \, p(z)$, $k \in \mathbb{Z}$, $p \in \Pi[\mathbb{C}]$ geschrieben werden.

2. $f \in \Lambda[\mathbb{C}]$ ist genau dann invertierbar, d.h. $1/f \in \Lambda[\mathbb{C}]$, wenn $f(z) = c\, z^k$ für ein $c \in \mathbb{C}^\times$ und ein $k \in \mathbb{Z}$ ist.

3. Für $c \in \mathbb{C}^\times$ gehört die Funktion
$$f(z) = \frac{1}{z - c}$$
zu $\Pi(\mathbb{C})$. Geben Sie die formale Potenzreihe zu f an.

4. Für $c \in \ell(\mathbb{Z})$ gilt
$$c^* \in \Lambda[\mathbb{C}] \qquad \Longleftrightarrow \qquad c \in \ell_{00}(\mathbb{Z}).$$

[52] Das heißt, bei so einer Reihe interessiert es uns zuerst einmal nicht, ob oder wo sie konvergiert!

[53] Laurentreihen sind an der Stelle $z = 0$ nicht definiert, sobald sie keine Potenzreihe sind. Daher muß man sich auf $\mathbb{C}^\times = \mathbb{C} \setminus \{0\}$, die Menge der *Einheiten*, d.h. der invertierbaren Elemente in $\mathbb{C}$ beschränken.

♢

Daß die z–Transformierte gar nichts so wirklich neues ist, sieht man schon daran, daß sie auf dem komplexen Einheitskreis[54] $\mathbb{T} = \{z \in \mathbb{C} \,:\, |z| = 1\}$ mit der Fouriertransformierten übereinstimmt:

$$c^* \left(e^{i\xi}\right) = \widehat{c}(\xi), \qquad \xi \in \mathbb{T}.$$

Wenn man nun noch bedenkt, daß jedes Laurentpolynom wie auch jede konvergente Laurentreihe durch ihr Verhalten auf $\mathbb{T}$ eindeutig definiert sind[55], dann ist die Erweiterung von $\mathbb{T}$ auf ganz $\mathbb{C}^\times$ so wild eigentlich nicht.

Bemerkung 2.2 (Eigenschaften der z–Transformation) *Zuerst einmal ein paar Bemerkungen über die z–Transformation:*

1. *Anstelle von (2.1) verwendet man oftmals auch gerne das* Symbol

$$c^\star(z) := c^* \left(z^{-1}\right) = \sum_{k\in\mathbb{Z}} c(k)\, z^k,$$

was eigentlich keinen richtigen, systematischen Unterschied macht. Trotzdem ist die z–Transformierte in der Literatur, insbesondere in der Signalverarbeitungsliteratur, recht konsistent wie in (2.1) definiert, siehe z.B. [15] – und dann sollte man sich auch dran halten.

2. *Die z–Transformation benimmt sich zur Faltung genau wie die Fouriertransformation:*

$$\begin{aligned}
(c * d)^* (z) &= \sum_{k\in\mathbb{Z}} (c * d)(k)\, z^{-k} = \sum_{k\in\mathbb{Z}} \left(\sum_{j\in\mathbb{Z}} c(j)\, d(k-j) \right) z^{-j} z^{j-k} \\
&= \sum_{j\in\mathbb{Z}} c(j)\, z^{-j} \sum_{k\in\mathbb{Z}} d(k-j)\, z^{-(k-j)} = \left(\sum_{j\in\mathbb{Z}} c(j)\, z^{-j} \right) \left(\sum_{k\in\mathbb{Z}} c(k)\, z^{-k} \right) \\
&= c^*(z)\, d^*(z).
\end{aligned}$$

3. *Es gibt auch eine* inverse z–Transformation*, und zwar*

$$f \in \Lambda(\mathbb{C}) \;\mapsto\; f_*(k) := \frac{1}{2\pi i} \oint_\Gamma f(z)\, z^{k-1}\, dz, \tag{2.2}$$

wobei Γ eine geschlossene Kurve in $\mathbb{C}$ ist, beispielsweise ein Kreis um den Ursprung, also $r\mathbb{T}$ mit $r > 0$, in deren Innerem der Nullpunkt und alle Singularitäten[56] von $f(z)$ liegen. Dies ist eine Folgerung aus dem Cauchyschen[57] Integralsatz*, siehe z.B. [30, Theorem 7.2.1].*

[54] Denn wir wegen des kanonischen Isomorphismus $z = e^{i\xi}$ ebenfalls mit $\mathbb{T}$ bezeichnen werden.

[55] Anschaulich klar: Die Fourierreihe legt ja alle Koeffizienten der Laurentreihe fest.

[56] Also Pole.

[57] Augustin Louis Cauchy, 1789–1857, wurde von den "Freunden der Familie" Laplace und Lagrange, insbesondere letzterem, zur Mathematik gebracht, studierte unter anderem bei Ampére. Ein Journalist sagte angeblich einmal über ihn: *"... it is certain a curious thing to see an academician who seemed to fulfil the respectable functions of a missionary preaching to the heathens."*; Zitat aus [37].

2.2 Rationale Filter und ihre Realisierung

Die z–Transformation ist nun der Schlüssel zu realisierbaren Filtern mit wesentlich mehr Freiheitsgraden, den sogenannten *rationalen Filtern.* Ein Filter F, dargestellt durch eine Folge[58] f heißt *rational*, wenn es (Laurent–)Polynome $p, q \in \Lambda[\mathbb{C}]$ gibt, so daß

$$f^*(z) = \frac{p(z)}{q(z)}, \qquad z \in \mathbb{C}^\times \setminus Z(q), \tag{2.3}$$

wobei $Z(q) := \{z \in \mathbb{C} : q(z) = 0\}$ die (endliche) Nullstellenmenge von q bezeichnet, von der man sich tunlichst fernhalten sollte, denn ansonsten landet man in einem Pol von f^*. Über die Konvergenz der zu f^* gehörigen Laurentreihe und somit der Frage, wann und wo $f^*(z)$ als Funktion überhaupt wohldefiniert ist, schweigen wir uns noch ein bisschen aus, das kommt dann im Kapitel 2.3.
Was uns zuerst einmal interessiert ist, daß p und q natürlich in keinster Weise eindeutig sind, denn schließlich kann man den Bruch beliebig mit $c\,z^k$, $c \in \mathbb{C}^\times$, $k \in \mathbb{Z}$, erweitern und so sogar dafür sorgen, daß p und q beispielsweise beide Polynome sind. Da p und q auf alle Fälle Laurentpolynome sind, können wir sie auch als z–Transformierte von zwei Folgen $p, q \in \ell_{00}(\mathbb{Z})$ auffassen und anstelle von $f^* = p/q$ auch $f^* = p^*/q^*$ schreiben.

Betrachten wir nun die Identität $y = Fx$ in der z–Transformierten, dann erhalten wir, daß

$$y^* = (Fx)^* = f^*\,x^* = \frac{p^*}{q^*}\,x^*, \qquad \text{also} \qquad y^*\,q^* = p^*\,x^*,$$

Jetzt machen wir Gebrauch von der Normierungsfreiheit und fordern, daß

$$q\left(z^{-1}\right) \in \Pi[\mathbb{C}] \qquad \text{und} \qquad q(z) = 1 + q(1)\,z^{-1} + \cdots + q(m)z^{-m}$$

ist[59]. Diese Normierung läßt sich für $q^* \neq 0$ immer erreichen. Damit ist dann aber

$$y^* = p^*\,x^* + \underbrace{(1 - q^*)}_{=:r}\,y^*, \qquad \text{also} \qquad y = p * x + r * y. \tag{2.4}$$

Und was haben wir hier gewonnen? Ganz einfach: Der Filter r hat die Form

$$r = \begin{array}{c} \\ \overset{0}{\downarrow} \\ \hline \ldots 0 \mid 0 \mid q(1)\ldots q(m) \mid 0\ldots \\ \hline \end{array}$$

und ist somit ein *kausaler* Filter, das heißt, er verwendet zum Zeitpunkt k nur Werte $y(j)$, $j < k$. Und die sind aber schon bekannt! Mit anderen Worten: Man muß "nur" das Ergebnissignal y nach einem Verzögerungsschritt[60] in den "Nennerfilter" schicken und die Ergebnisse der beiden Filter schließlich aufsummieren. Dies ist in Abb. 2.1 schematisch dargestellt. Insbesondere benötigt man also für einen rationalen Filter mit Zählerpolynom[61] p vom Grad m und Nennerpolynom q vom Grad n

[58] Also ist F, um genau zu sein, wieder einmal ein LTI–Filter.
[59] Die zweite Bedingung schreibt man in der Literatur zur Signalverarbeitung gerne als $q(\infty) = 1$.
[60] Da ja $r(0) = 0$ ist und der Filter r so erst mit $q(1)$ beginnt!
[61] Damit lassen sich Zähler und Nenner kausal realisieren und es sind keine weiteren Zeitverzögerer nötig.

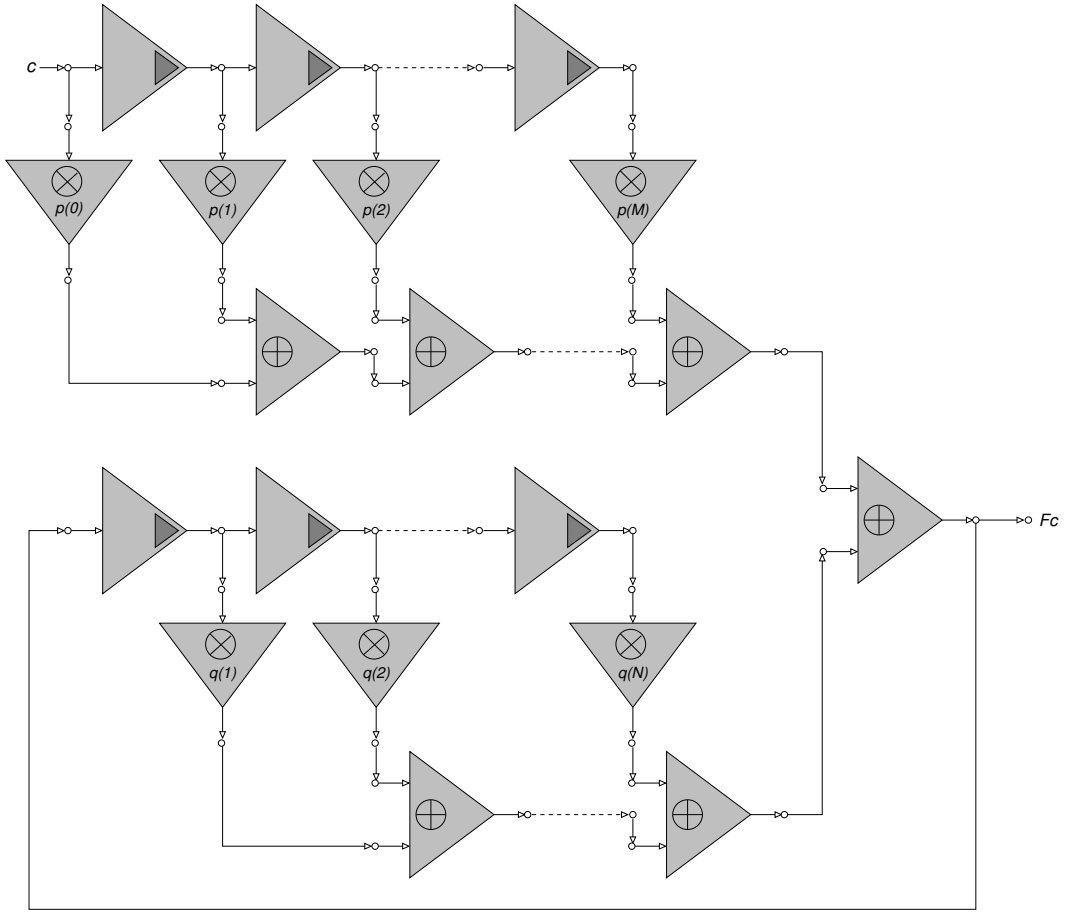


Abbildung 2.1: Ein rationaler Filter, realisiert mittels *delayed feedback*: Das mit p gefilterte Signal wird verzögert in den Filter q geschickt und die beiden Ergebnisse summiert.

- $n + m + 1$ Multiplizierer,
- $n + m$ Addierer und
- $n + m$ Verzögerer,

was einen durchaus vertretbaren Aufwand darstellt.

Beispiel 2.3 (Summationsfilter) *Das einfachste Beispiel für einen rationalen Filter ist der kausale* Summationsfilter S*, definiert durch*

$$(Sc)(k) = \sum_{j=-\infty}^{k} c(j).$$

Das Signal Sc ist eine Art "Stammfunktion" zu c. Man sieht nun leicht, daß

$$Sc\,(k) = \sum_{j=0}^{\infty} c(k-j) = \sum_{j\in\mathbb{Z}} s(j)\,c(k-j) = s * c, \qquad s(j) = \begin{cases} 1, & j \geq 0, \\ 0, & j < 0, \end{cases}$$

weswegen

$$s^*(z) = \sum_{k=0}^{\infty} z^{-k} = \frac{1}{1-z^{-1}}, \qquad |z| > 1,$$

ist; die Forderung $|z| < 1$ sorgt dafür, daß die Potenzreihe absolut konvergiert. Der Filter S ist ein IIR–Filter*, denn die Impulsantwort s ist ja nun für alle nichtnegativen Zeitpunkte ungleich Null.*
Die Realisierung ist nun aber denkbar einfach: Wir erhalten die Zähler- und Nennerpolynome sofort als

$$p(z) = 1, \qquad q(z) = 1 - z^{-1} \qquad \Longrightarrow \qquad r(z) = z^{-1},$$

und somit ist p der triviale Filter und r besteht gerade aus einem Verzögerer, wie man in Abb. 2.2 sehen kann.

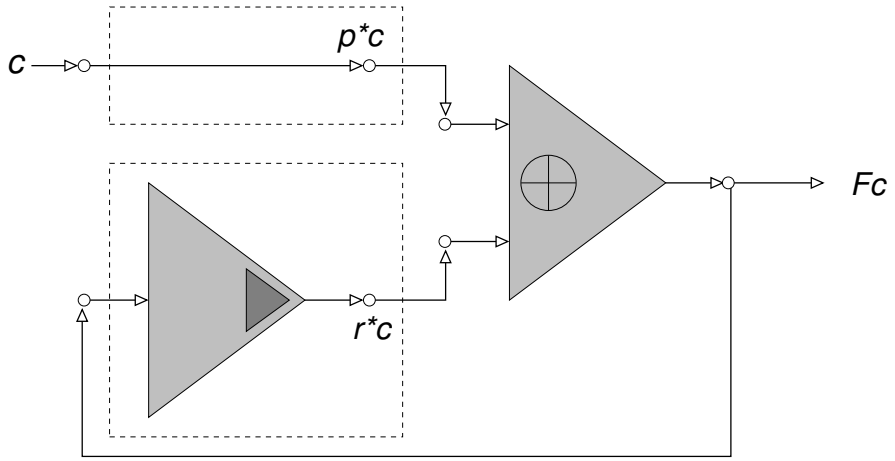


Abbildung 2.2: Das Schaltbild des Summationsfilters, eine richtig einfache Angelegenheit. Die Filter für Zähler und Nenner sind eingerahmt.

Aus dem Summationsfilter von Beispiel 2.3 kann man sich dann auch einen *Integrierer* basteln. Dazu tastet man eine Funktion $f \in L_1(\mathbb{R})$ mittels des Sampling–Operators S_h ab und erhält als Näherung für das Riemann–Integral

$$\int_0^x f(t)\,dt = h \sum_{k=0}^{\lfloor x/h \rfloor} f\,(kh) = [h\,S\,S_h f]\,(\lfloor x/h \rfloor)\,, \tag{2.5}$$

"technisch" müssen wir also nur einen Multiplikator nachschalten.

Allerdings ist die Riemannsumme numerisch eine eher "dumme" Methode, um ein Integral zu berechnen, sie ist nur genau für konstante Funktionen. Aber unsere rekursiven Filter können schon noch mehr. Dazu noch ein Beispiel aus [24, Kapitel 3.10].

Beispiel 2.4 (Integrationsfilter) *Eine bessere Intgrationsregel oder* Quadraturformel *ist die sogenannte* Trapezregel. *Dabei wird das Integral von kh bis $(k+1)h$ durch den Mittelwert $\frac{1}{2}\left[f(kh)+f\left((k+1)h\right)\right]$ angenähert*[62] *und liefert uns so die Regel*

$$\begin{aligned} y(k) &= \sum_{j=-\infty}^{k} \frac{1}{2}\left[x(j-1)+x(j)\right] = \sum_{j=-\infty}^{k-1} \frac{1}{2}\left[x(j-1)+x(j)\right] + \frac{1}{2}\left[x(k-1)+x(k)\right] \\ &= y(k-1) + \frac{1}{2}\left[x(k-1)+x(k)\right]. \end{aligned}$$

Übergang zur z–Transformierten liefert dann

$$\begin{aligned} y^*(z) &= \sum_{k\in\mathbb{Z}} y(k)\, z^{-k} = \sum_{k\in\mathbb{Z}} \left[y(k-1) + \frac{1}{2}\left(x(k-1)+x(k)\right)\right] z^{-k} \\ &= \sum_{k\in\mathbb{Z}} y(k-1)\, z^{-k} + \frac{1}{2}\sum_{k\in\mathbb{Z}} x(k-1)\, z^{-k} + \frac{1}{2}\sum_{k\in\mathbb{Z}} x(k)\, z^{-k} \\ &= \sum_{k\in\mathbb{Z}} y(k)\, z^{-k-1} + \frac{1}{2}\sum_{k\in\mathbb{Z}} x(k)\, z^{-k-1} + \frac{1}{2}\sum_{k\in\mathbb{Z}} x(k)\, z^{-k} \\ &= z^{-1}\, y^*(z) + \frac{1}{2}\left(1+z^{-1}\right)\, x^*(z), \end{aligned}$$

also

$$f_T^*(z)\, x^*(z) = y^*(z) = \frac{1}{2}\,\frac{1+z^{-1}}{1-z^{-1}}\, x^*(z), \qquad z\in\mathbb{C}^\times. \tag{2.6}$$

Analog transformiert man die Simpson–Regel[63]

$$y(k) = y(k-2) + \frac{1}{6}\left[x(k-2) + 4\,x(k-1) + x(k)\right], \qquad k\in\mathbb{Z},$$

in

$$f_S^*(z)\, x^*(z) = y^*(z) = \frac{1}{6}\,\frac{1+4z^{-1}+z^{-2}}{1-z^{-2}}\, x^*(z) = \frac{1}{6}\,\frac{z+4+z^{-1}}{z-z^{-1}}\, x^*(z) \qquad z\in\mathbb{C}^\times. \tag{2.7}$$

Die beiden Filter aus Beispiel 2.4 kann man sich nun im "Frequenzbereich" ansehen, wobei man

$$\widehat{f_T}\,(\xi) = f_T^*\left(e^{i\xi}\right) = \frac{1}{2}\frac{1+e^{-i\xi}}{1-e^{-i\xi}} = \frac{1}{2}\frac{e^{i\xi/2}+e^{-i\xi/2}}{e^{i\xi/2}-e^{-i\xi/2}} = \frac{1}{2i}\frac{\cos\xi/2}{\sin\xi/2}$$

[62] Weswegen die Quadraturformel den *Exaktheitsgrad* 1 hat, also alle *linearen* Polynome exakt integriert, siehe beispielsweise [44].

[63] Integration des quadratischen Polynoms, das an drei aufeinanderfolgenden Punkten interpoliert.

und analog

$$\widehat{f_S}(\xi) = \frac{1}{6}\,\frac{4+\left(e^{i\xi}+e^{-i\xi}\right)}{e^{i\xi}-e^{-i\xi}} = \frac{1}{6}\frac{4+2\,\cos\xi}{2i\,\sin\xi} = \frac{1}{6i}\frac{2+\cos\xi}{\sin\xi}.$$

Beide Filter haben nun eine Singularität an der Frequenz $\xi = 0$, und das muß auch so sein! Da der Filter ja versucht, eine näherungsweise Stammfunktion zu berechnen, müsste ja für eine Funktion g

$$g(x) = (Fg)'(x) = (f * g)(x)$$

gelten, was nach Übergang zur Fouriertransformierten unter Verwendung von (1.10) und (1.12) gerade

$$\widehat{g}(\xi) = i\xi\,\widehat{f}(\xi)\,\widehat{g}(\xi) \qquad \Longrightarrow \qquad \widehat{f}(\xi) = \frac{1}{i\xi}$$

für den *perfekten Integrationsfilter* liefert. Die Singularität an 0 ist nun genau der Indikator dafür, daß unsere Quadraturfilter Funktionen mit Frequenz 0, also konstante Funktionen, exakt integrieren. Aus diesem Grund betrachten wir die *relativen* Abweichungen der Filter

$$\frac{\widehat{f_T}(\xi)}{\widehat{f}(\xi)} = \cos\xi/2\,\frac{\xi/2}{\sin\xi/2}, \tag{2.8}$$

$$\frac{\widehat{f_S}(\xi)}{\widehat{f}(\xi)} = \frac{1}{3}\,(2+\cos\xi)\,\frac{\xi}{\sin\xi}; \tag{2.9}$$

für $\xi = 0$ sind diese Quotienten offensichtlich 1 und da außerdem beide Funktionen gerade sind, genügt es, sie auf $[0,\pi]$ zu betrachten. Die Funktionen sind in Abb. 2.3 zu sehen. Man beachte,

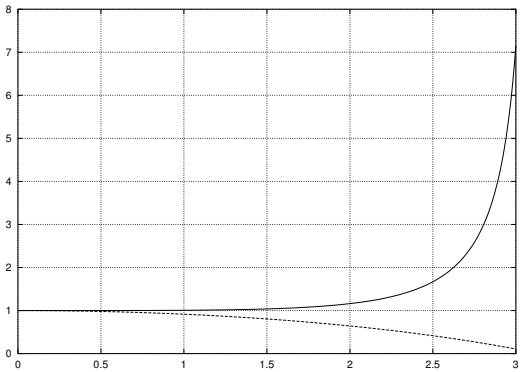


Abbildung 2.3: Relativer Frequenzgang des Trapez- und des Simpson–Filters. Der Simpson–Filter ist im "niederfrequenten" Bereich deutlich besser, bezahlt aber dafür bei den höheren Frequenzen.

daß der relative Fehler von (2.8) mit zunehmender Frequenz zwar mehr und mehr abnimmt, aber wenigstens beschränkt bleibt, während für $\xi \to \pi$ in (2.9) eine Singularität auftaucht – ein weiterer Effekt des Abtastsatzes, da sich da alle hohen Frequenzen aufsummieren, die ja zu beliebig großen Integralen führen können.

Bemerkung 2.5 *Wie die Integrationsfilter zeigen, besitzen rationale Filter eine durchaus wünschenswerte Fähigkeit, die den FIR–Filtern fehlt: Sie können Singularitäten modellieren, in diesem Falle die Singularität der Fouriertransformierten an $\xi = 0$, die bei der Integration aus prinzipiellen Gründen vorhanden sein muß. Das ist ein weiterer Grund, sich mit rationalen Filtern zu befassen.*

2.3 Stabilität

Die Beispiele für rationale Filter aus dem vorangegangenen Kapitel zeigen, daß rationale Filter zwar alle realisierbar sind. Leider macht es aber einen gewaltigen Unterschied, wo nun die Nullstellen der Nenner, also die Pole der rationalen Funktion liegen. Dies wollen wir uns zuerst einmal anhand eines Beispiels aus [23, Beispiel 2, S. 64] veranschaulichenn.

Beispiel 2.6 (Entwicklung eines rekursiven Filters) *Gegeben die z–Transformierte*

$$f^*(z) = \frac{c}{(1 - az^{-1})\,(1 - bz^{-1})}, \qquad c \in \mathbb{C}^\times, \qquad z \in \mathbb{C} \setminus \{0, a, b\}\,.$$

Wie lauten die Koeffizienten $f(k)$, $k \in \mathbb{Z}$? Man könnte natürlich nun "blind" die inverse z–Transformation aus (2.2) verwenden, aber viel geschickter ist es, für f^* die Partialbruchzerlegung

$$f^*(z) = \frac{A}{1 - az^{-1}} + \frac{B}{1 - bz^{-1}} = \frac{A\,(1 - bz^{-1}) + B\,(1 - az^{-1})}{(1 - az^{-1})\,(1 - bz^{-1})}$$

anzusetzen, was das Gleichungssystem

$$\begin{bmatrix} 1 & 1 \\ b & a \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} c \\ 0 \end{bmatrix} \qquad \Longrightarrow \qquad A = \frac{ac}{a - b}, \quad B = \frac{-bc}{a - b}$$

liefert – zumindest solange $a \neq b$ ist. Bleibt also nur noch die Entwicklung

$$\frac{1}{1 - az^{-1}} = \sum_{k=0}^{\infty} a^k\, z^{-k},$$

um die zumindest formale Darstellung

$$f^*(z) = A \sum_{k=0}^{\infty} a^k\, z^{-k} + B \sum_{k=0}^{\infty} b^k\, z^{-k} = \sum_{k=0}^{\infty} \frac{\left(a^{k+1} - b^{k+1}\right)\, c}{a - b}\, z^{-k}, \qquad z \in \mathbb{C}^\times,$$

und damit

$$f(k) = \begin{cases} 0, & k < 0, \\ \dfrac{\left(a^{k+1} - b^{k+1}\right)\, c}{a - b}, & k \geq 0, \end{cases} \tag{2.10}$$

liefert. Dieser Filter hat keine *endliche Impulsantwort, dafür aber eine andere Eigenschaft, die sich je nach Wahl von a und b unterscheidet:*

- *Sind $|a|, |b| < 1$, dann fallen die Koeffizienten der Impulsantwort f aus (2.10) für $k \to \infty$ ab und man spricht von einem* Dämpfungsverhalten *des Filters.*

- *Ist hingegen $|a| < 1 < |b|$, dann ist natürlich*

$$|f(k)| \simeq \left|b^{k+1}\right|$$

und daher wächst die Impulsantwort über alle Grenzen, wenn $b < 0$ ist sogar in oszillierender Manier. Aus Symmetriegründen passiert für $|a| > 1 > |b|$ zudem auch nichts anderes.

- *Ist schließlich $1 < |a|, |b|$, dann liefert unter der Annahme*[64] *$1 < a \leq b$ die Identität*

$$\frac{\left(a^{k+1} - b^{k+1}\right)\, c}{a-b} = \sum_{j=0}^{k} a^j b^{k-j} \geq \sum_{j=0}^{k} a^j = \frac{a^{k+1}-1}{a-1},$$

was ebenfalls ein eher divergentes Verhalten zeigt.

Es mag zuerst etwas seltsam erscheinen, daß wir die Potenzreihe eines rationalen Filters bestimmen, wo wir doch gerade die Tatsache so gepriesen haben, daß dieser Filter mit endlich vielen Schaltgliedern realisierbar ist, indem man die rationale Struktur ausnutzt. Trotzdem ist diese Entwicklung wichtig, da sie uns die Impulsantwort des Filters liefert und so dessen Verhalten beschreibt. Und es macht eben einen ganz gewaltigen Unterschied, ob die Impulsantwort abklingt, oder ob sie immer heftiger und mit immer größerer Amplitude oszilliert.

Definition 2.7 (Dämpfung) *Ein Filter F heißt*

1. dämpfend*, wenn die Impulsantwort für $k \to \pm\infty$ verschwindet, d.h., wenn*

$$f \in \ell_0(\mathbb{Z}) := \left\{ c \in \ell(\mathbb{Z}) \;:\; \lim_{k\to\infty} c(k) = \lim_{k\to-\infty} c(k) = 0 \right\}$$

2. beschränkt*, wenn die Impulsantwort beschränkt ist, d.h. $f \in \ell_\infty(\mathbb{Z})$ ist.*

Ein FIR–Filter mit $f \in \ell_{00}(\mathbb{Z})$ hat natürlich alle diese "schönen" Eigenschaften, d.h., er ist dämpfend und beschränkt, ja es ist sogar $f\, g \in \ell_{00}(\mathbb{Z})$ für jedes beliebige Signal $g \in \ell(\mathbb{Z})$.
Warum sind nun gerade beschränkte oder noch besser dämpfende Filter so sehr von Interesse? Dazu sehen wir uns nur einmal den Filter mit der Impulsantwort $f(k) = k$, $k \in \mathbb{N}$, und $f(k) = 0$ für $k < 0$, an. Schickt man nun in diesen Filter zwei Diracpulse mit gewisser Wartezeit, also das Signal $c = \delta + \tau_k \delta$, dann ist

$$Fc(j) = \begin{cases} 0, & j \leq 0, \\ j, & 0 < j \leq k, \\ j + (j-k), & j > k. \end{cases}$$

[64] Die keine echte Einschränkung darstellt!

Theoretisch ist das nicht sonderlich aufregend. Ist nun aber k relativ groß, dann kann es passieren, daß in einer Fließpunktarithmetik der Term j die hinzukommenden Terme $j - k$ so stark dominiert, daß letztere praktisch unter den Tisch fallen und das zweite Signal einfach ignoriert wird. Das wäre besonders fatal, wenn man das Signal $\delta - \tau_k \, \delta$ in das System füttert, denn dann könnte das System nach dem Zeitpunkt k einfach weiterwachsen, anstatt in den konstanten Zustand $Fc(j) = k$, $j \geq k$, zu fallen.

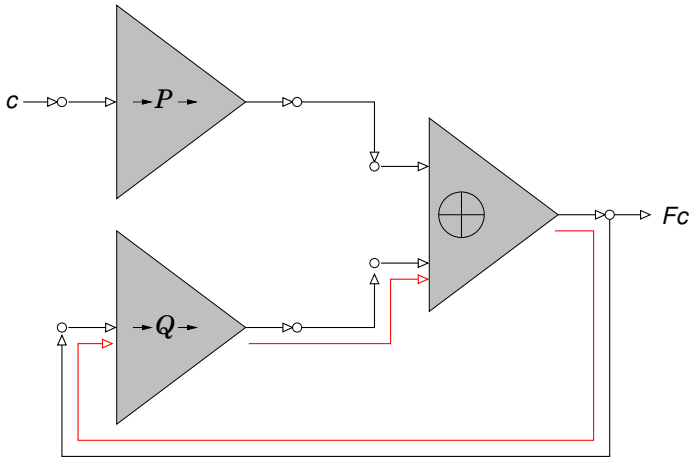


Abbildung 2.4: Der "Teufelskreis" in einem rekursiven System ohne Dämpfung. Einmal angestoßen entwickelt der Filter Q, der zum Nenner(Laurent-)polynom $q = q^*$ gehört ein Eigenleben und regt sich möglicherweise immer mehr auf – bis zu einem Punkt, an dem das System eingehende Signale als "zu klein" erachtet und ignoriert.

Man kann sich die Problematik auch anders veranschaulichen, indem man den rationalen Filter wie in Abb. 2.4 als Blockschaltbild aus den beiden Filtern P und Q darstellt. Durch die Rückführung der Information in Q kann das System, wenn es über keine Dämpfung verfügt, in einem Anregungszustand gefangen sein, aus dem es letzen Endes nicht mehr herauskommt.

Deswegen ist es also eine wichtige Eigenschaft rekursiver bzw. rationaler Filter *stabil* zu sein, das heißt, ein Dämpfungsverhalten aufzuweisen, denn andernfalls würde das einmal angeregte System irgendwann einmal jede neu ankommende Information so stark dominieren, daß sie schlichtweg irrelevant würde. Glücklicherweise kann man Stabilität rationaler Filter recht einfach charakterisieren.

Satz 2.8 (Stabilität rationaler Filter) *Ein rationaler*[65] *Filter F mit*

$$f^*(z) = \frac{p(z)}{q(z)}, \qquad z \in \mathbb{C}^\times$$

ist genau dann stabil, wenn alle Nullstellen von q im Inneren des Einheitskreises liegen.

[65] Rationale Funktionen werden immer in *gekürzter* Form angenommenm, also ohne gemeinsame Nullstellen von Zähler- und Nennerpolynom.

Um Satz 2.8 zu beweisen, sammeln wir erst einmal ein bißchen Information über rationale Filter. Zuerst einmal der einfachste Fall.

Proposition 2.9 *Ein elementarer rationaler Filter der Form*

$$f^*(z) = \frac{a}{1-\zeta\, z^{-1}}, \qquad z \in \mathbb{C}^\times \setminus \{\zeta\}, \quad \zeta \in \mathbb{C}, \tag{2.11}$$

ist genau dann dämpfend, wenn $|\zeta| < 1$ *ist.*

Beweis: Wir verwenden die Entwicklungsformel

$$\frac{1}{1-\zeta\, z^{-1}} = \sum_{k=0}^{\infty} \zeta^k\, z^{-k}, \tag{2.12}$$

um

$$f(k) = \begin{cases} 0, & k < 0, \\ a\,\zeta^k, & k \geq 0, \end{cases}$$

zu erhalten, was für $k \to -\infty$ *trivialerweise* immer gegen Null konvergiert und für $k \to +\infty$ genau dann, wenn $|\zeta| < 1$ ist. □

Der Beweis zeigt sogar eine noch schönere Eigenschaft: Wenn so ein einfacher Filter wie in (2.11) die Dämpfungseigenschaft hat, dann fällt er nicht etwa "irgendwie" ab, sondern *exponentiell*, das heißt, es gibt eine Konstante $\rho < 1$, nämlich $\rho = |\zeta|$, so daß $|f(k)| \leq \rho^{|k|}$ ist. Diese Beobachtung erweitern wir nun noch etwas.

Proposition 2.10 *Für einen rationaler Filter der Form*

$$f^*(z) = \frac{a}{(1-\zeta\, z^{-1})^\alpha}, \qquad z \in \mathbb{C}^\times \setminus \{\zeta\}, \quad \zeta \in \mathbb{C}, \quad \alpha \in \mathbb{N}, \tag{2.13}$$

sind äquivalent:

1. *F besitzt ein Dämpfungsverhalten.*
2. *F besitzt ein* exponentielles *Dämpfungsverhalten.*
3. $|\zeta| < 1$.

Beweis: Wir defieren die polynomialen Folgen $\pi_\alpha \in \ell(\mathbb{Z})$, $\alpha \in \mathbb{N}_0$ durch

$$\pi_0 \equiv 1 \qquad \text{und} \qquad \pi_{\alpha+1}(k) = \sum_{j=0}^{k} \pi_\alpha(j).$$

Nach [22, S. 283, (6.78)][66] ist π_α ein Polynom vom Grad α der Form

$$\pi_\alpha(k) = \frac{k^\alpha}{\alpha!} + \cdots$$

[66] Da wird auch gezeigt wie man auf die Formel kommt, bei der die sogenannten *Bernoulli–Zahlen* involviert sind.

und bestimmt auch das Verhalten in (2.13): Es gilt nämlich, daß

$$\frac{1}{(1-\zeta\,z^{-1})^{\alpha}} = \sum_{k=0}^{\infty} \pi_{\alpha-1}(k)\,\zeta^k\,z^{-k}, \qquad |z| > |\zeta|. \tag{2.14}$$

Die zeigt man durch Induktion über α, wobei der Fall $\alpha = 1$ gerade (2.12) ist und wir ansonsten

$$\begin{aligned}
\frac{1}{(1-\zeta\,z^{-1})^{\alpha+1}} &= \frac{1}{(1-\zeta\,z^{-1})}\,\frac{1}{(1-\zeta\,z^{-1})^{\alpha}} = \left(\sum_{k=0}^{\infty} \pi_0(k)\,\zeta^k\,z^{-k}\right)\left(\sum_{k=0}^{\infty} \pi_{\alpha-1}(k)\,\zeta^k\,z^{-k}\right) \\
&= \sum_{j,k=0}^{\infty} \zeta^j\,\pi_{\alpha-1}(k)\,\zeta^k\,z^{-(j+k)} = \sum_{j,k=0}^{\infty} \pi_{\alpha-1}(k)\,\zeta^{j+k}\,z^{-(j+k)} \\
&= \sum_{k=0}^{\infty} \pi_{\alpha-1}(k) \sum_{j=k}^{\infty} \zeta^j\,z^j = \sum_{j=0}^{\infty} \zeta^j\,z^j \left(\sum_{k=0}^{j} \pi_{\alpha-1}(k)\right) = \sum_{j=0}^{\infty} \pi_{\alpha}(j)\,\zeta^j\,z^j,
\end{aligned}$$

womit (2.14) bewiesen ist.
Damit sind wir aber fertig, denn die Folge $\pi_\alpha(j)\,\zeta^j$ konvergiert für $j \to \infty$ genau dann gegen Null wenn $|\zeta| < 1$ ist. Die exponentielle Abklingordnung erhalten wir, indem wir ein beliebiges $|\zeta| < \rho < 1$ wählen, denn dann haben wir, daß

$$\left|\pi_\alpha(j)\,\zeta^j\right| \le |\pi_\alpha(j)| \underbrace{\left(\frac{|\zeta|}{|\rho|}\right)^j}_{\to 0} \rho^j$$

und da $\pi_\alpha(j)\ (\zeta/\rho)^j \to 0$ für $j \to \infty$, gibt es eine Konstante $C > 0$, die diesen Ausdruck für alle $j \in \mathbb{N}$ majorisiert. Somit ist

$$\left|\pi_\alpha(j)\,\zeta^j\right| \le C\,\rho^j$$

und kling daher exponentiell ab. □

Bemerkung 2.11 *Der Beweis von Proposition 2.10 zeigt, daß man die exponentielle Abklingordnung ρ beliebig knapp oberhalb von $|\zeta|$ wählen kann, allerdings natürlich um den Preis einer immer übleren Konstante. Der Wert $|\zeta|$ hingegen wird, zumindest für $\alpha > 1$,* nicht *funktionieren.*

Um den Beweis von Satz 2.8 zu vervollständigen, müssen wir schließlich noch die Partialbruchzerlegung aus Beispiel 2.6 auf allgemeine theoretische Füße stellen. Dazu verwenden wir einen algebraischen Klassiker, nämlich die oftbenutzte *Bézout–Identität*.

Satz 2.12 (Bézout–Identität) *Seien $f, g \in \Lambda[\mathbb{C}]$ zwei Laurentpolynome ohne gemeinsame Nullstellen. Dann gibt es $p, q \in \Lambda[\mathbb{C}]$, so daß*

$$f\,p + g\,q = 1 \tag{2.15}$$

ist.

Korollar 2.13 *Seien $f, g \in \Lambda[\mathbb{C}]$ zwei Laurentpolynome ohne gemeinsame Nullstellen. Dann gibt es zu jedem $h \in \Lambda[\mathbb{C}]$ Laurentpolynome $p, q \in \Lambda[\mathbb{C}]$, so daß*

$$f\,p + g\,q = h \tag{2.16}$$

ist.

Beweis: Nach Satz 2.12 gibt es $\widetilde{p}$ und $\widetilde{q}$, so daß $f\,\widetilde{p} + g\,\widetilde{q} = 1$ und

$$p = h\,\widetilde{p}, \qquad q = h\,\widetilde{q}$$

erfüllen (2.16). □

Bemerkung 2.14 *Kein vernünftiger Mensch wird eine Bézout–Indetität der Form (2.16) so auflösen wie im Beweis von Korollar 2.13, denn dann sind die Grade der Koeffizienten p, q einfach zu groß. Tatsächlich bestimmt man diese Polynome über den* erweiterten euklidischen Algorithmus*, eines der grundlegenden Verfahren in der Computeralgbra, siehe z.B. [18, 45].*

Übung 2.2 Zeigen Sie, daß die Polynome in (2.15) nicht eindeutig sind und geben Sie an, wie man ausgehend von *einer* speziellen Lösung p^*, q^* *alle* Lösungen p, q von (2.15) finden kann. Dieses Verfahren ist in der Signalverarbeitungs-/Waveletgemeinde unter dem Namen *"Lifting scheme"* populär, siehe [60]. ♢

Beweis von Satz 2.12: Ohne Einschränkung können wir annehmen, daß $f, g \in \Pi[\mathbb{C}]$ Polynome sind, denn andernfalls bräuchten wir beide nur mit einer Potenz von z zu multiplizieren, die wir später als Potenz von z^{-1} auf p bzw. q schieben können. Seien $m = \deg f$ und $n = \deg g$, dann können wir die Polynome p, q sogar vom Grad $n-1$ bzw. $m-1$ wählen. Mit $\zeta_1, \ldots, \zeta_m$ und $\eta_1, \ldots, \eta_n$ bezeichnen wir die Nullstellen von f bzw. g, d.h.,

$$f = f_m\,(x-\zeta_1)\cdots(x-\zeta_m) \qquad \text{und} \qquad g = g_n\,(x-\eta_1)\cdots(x-\eta_n)\,.$$

Nun wählen wir $p \in \Pi_{n-1}[\mathbb{C}]$ als Lösung des Interpolationsproblems[67]

$$p\,(\eta_j) = \frac{1}{f\,(\eta_j)}, \qquad j = 1, \ldots, n,$$

und $q \in \Pi_{m-1}[\mathbb{C}]$ als Lösung von

$$q\,(\zeta_j) = \frac{1}{g\,(\zeta_j)}, \qquad j = 1, \ldots, m.$$

Diese Größen sind wohldefiniert, da f und g keine gemeinsamen Nullstellen besitzen. Das Polynom $h = f\,p + g\,q$ hat dann Grad $n+m-1$ und erfüllt

$$h\,(\zeta_j) = \underbrace{f\,(\zeta_j)}_{=0}\,p\,(\zeta_j) + g\,(\zeta_j)\,\underbrace{q\,(\zeta_j)}_{=1/g(\zeta_j)} = 1, \qquad j = 1, \ldots, m,$$

[67] Sind die Nullstellen von f und g alle einfach, dann handelt es sich um ein *Lagrange–Interpolationsproblem* (also Interpolation von Funktionswerten), ansonsten um ein *Hermite–Interpolationsproblem*, bei dem auch Ableitungen zu interpolieren sind – auf letzteren Fall wollen wir hier nicht im Detail eigehen, siehe aber z.B. [10].

sowie

$$h(\eta_j) = f(\eta_j) \underbrace{p(\eta_j)}_{=1/f(\eta_j)} + \underbrace{g(\eta_j)}_{=0} q(\eta_j) = 1, \qquad j = 1, \ldots, n,$$

und damit stimmt $h \in \Pi_{n+m-1}[\mathbb{C}]$ an den $n+m$ Punkten $\zeta_1, \ldots, \zeta_m, \eta_1, \ldots, \eta_n$ mit der konstanten Funktion mit Wert 1 überein, weswegen $h = 1$ sein muß. Das vervollständigt den Beweis. □

Was jetzt noch bleibt ist die "Partialbruchzerlegung" von rationalen Funktionen und die läuft nun wieder via Bézout.

Lemma 2.15 *Sei $f = p/q$, $p, q \in \Pi[\mathbb{C}]$ eine rationale Funktion und seien $\zeta_1, \ldots \zeta_m$ die Nullstellen von q mit Vielfachheiten $\mu_1, \ldots, \mu_m$, also*

$$q = \prod_{j=1}^{m} (\cdot - \zeta_j)^{\mu_j}.$$

Dann gibt es Polynome $p_1, \ldots, p_m$, so daß

$$f(z) = \sum_{j=1}^{m} \frac{p_j(z)}{(z-\zeta_j)^{\mu_j}}, \qquad z \in \mathbb{C} \setminus Z(q). \tag{2.17}$$

Natürlich gilt Lemma 2.15 auch für Quotienten von Laurentpolynomen, also für $p, q \in \Lambda[\mathbb{C}]$: Man muß nur zuerst so normalisieren, daß $q \in \Pi[\mathbb{C}]$ ist und kann dann einen eventuellen Faktor der Form z^{-k} aus p herausziehen, den man danach einfach wieder auf die Polynome p_j verteilt. Durch Umnormalisierung der Summanden erhält man dann die folgende Variation von Lemma 2.15.

Korollar 2.16 *Sei $f = p/q$, $p, q \in \Lambda[\mathbb{C}]$, und seien $\zeta_1, \ldots, \zeta_m \in \mathbb{C}^\times$ die Nullstellen[68] von q mit Vielfachheiten $\mu_1, \ldots, \mu_m$. Dann gibt es Laurentpolynome $p_1, \ldots, p_m \in \Lambda[\mathbb{C}]$, so daß*

$$f(z) = \sum_{j=1}^{m} \frac{p_j(z)}{(1-\zeta_j z^{-1})^{\mu_j}}, \qquad z \in \mathbb{C}^\times \setminus Z(q). \tag{2.18}$$

Beweis von Lemma 2.15: Induktion über m, für $m = 1$ ist nichts zu tun, denn dann haben wir ja nur einen Faktor. Sei also die Aussage für ein $m \geq 1$ bewiesen, dann nehmen wir an, daß q von der Form $q(z) = z^{\mu_1+\cdots+\mu_m} + \cdots$ ist, was sich durch Normalisierung ja immer erreichen läßt und zerlegen es in

$$q(z) = \underbrace{(z-\zeta_1)^{\mu_1} \cdots (z-\zeta_{m-1})^{\mu_{m-1}}}_{=:q_0(z)} \underbrace{(z-\zeta_m)^{\mu_m}}_{=:q_m(z)}, \qquad Z(q_0) \cap Z(q_1) = \emptyset.$$

[68] Hier ist eine kleine aber feine Fußangel verborgen: Das Hin- und Herspringen zwischen Polynomen und Laurentpolynomen ist nur dann "verlustfrei" möglich, wenn diese Nullstellen alle von Null verschieden sind. Ein Laurentpolynom kann keine Nullstelle an $z = 0$ haben (es ist da ja noch nicht einmal definiert), ein Polynom sehr wohl!

Nach Korollar 2.13 gibt es daher Polynome p_0, p_m, so daß

$$p_m\, q_0 + p_0\, q_m = p$$

und damit ist

$$\frac{p_0(z)}{q_0(z)} + \frac{p_m(z)}{(z-\zeta_m)^{\mu_m}} = \frac{p_0(z)}{q_0(z)} + \frac{p_m(z)}{q_m(z)} = \frac{p_0(z)\, q_m(z) + p_m(z)\, q_0(z)}{q_0(z)\, q_m(z)} = \frac{p(z)}{q(z)} = f(z).$$

Nach der Iduktionshypothese können wir nun die rationale Funktion p_0/q_0, die nur die $m-1$ verschiedenen Nullstellen $\zeta_1, \ldots, \zeta_{m-1}$ besitzt, weiterzerlegen und landen so bei (2.15). □

Und damit sind wir am Ziel unseres Stabilitätssatzes für rationale Filter angelangt.

Beweis von Satz 2.8: Seien $\zeta_1, \ldots, \zeta_m$ die Nullstellen des Nenners von f^*, also die Pole von f^*, mit den zugehörigen Vielfachheiten $\mu_1, \ldots, \mu_m$. Unter Verwendung der Partialbruchzerlegung (2.18) aus Korollar 2.16 und der Identität (2.14) erhalten wir dann, daß[69]

$$\begin{aligned} f^*(z) &= \sum_{j=1}^{m} \frac{p_j^*(z)}{(1-\zeta_j\, z^{-1})^{\mu_j}} = \sum_{j=1}^{m} \left[\sum_{k=-N}^{N} p_j(k)\, z^{-k}\right] \left[\sum_{\ell=0}^{\infty} \pi_{\mu_j-1}(\ell)\, \zeta_j^{\ell}\, z^{-\ell}\right] \\ &= \sum_{j=1}^{m} \sum_{k=-N}^{N} \sum_{\ell=0}^{\infty} p_j(k)\, \pi_{\mu_j-1}(\ell)\, \zeta_j^{\ell}\, z^{-(\ell+k)} = \sum_{j=1}^{m} \sum_{k=-N}^{N} \sum_{\ell\geq k} p_j(k)\, \pi_{\mu_j-1}(\ell-k)\, \zeta^{-k}\, \zeta_j^{\ell}\, z^{-\ell} \\ &= \sum_{j=1}^{m} \sum_{\ell=-N}^{\infty} \underbrace{\left[\sum_{k=-N}^{\ell} p_j(k)\, \pi_{\mu_j-1}(\ell-k)\, \zeta^{-k}\right]}_{=:r_j(\ell,\zeta_j)} \zeta_j^{\ell}\, z^{-\ell} \end{aligned}$$

wobei $r_j\,(\ell, \zeta_j)$ ein Polynom in ℓ und ζ_j ist. Damit ist also

$$f(k) = \sum_{j=1}^{m} r_j\,(k, \zeta_j)\; \zeta_j^k, \qquad k \geq -N. \tag{2.19}$$

Sind nun alle $|\zeta_j| < 1$, dann konvergiert jeder individuelle Summand in (2.19) für $k \to \infty$ gegen Null und der Filter ist (exponentiell) dämpfend[70], für $|\zeta_j| > 1$ hingegen divergiert der entsprechende Term in (2.19), und zwar so schnell wie $|\zeta_j|^k$ – auch hier ist in der Grenze der polynomiale Term vernachlässigbar. Gilt also beispielsweise, daß[71]

$$|\zeta_1| > |\zeta_2| \geq \cdots \geq |\zeta_m|$$

und $|\zeta_1|$, dann kann der Filter nicht dämpfend sein. Den Fall mehrerer "fernster" Pole kann man auch, allerdings mit etwas mehr Schwierigkeiten, erledigen, siehe Übung 2.3 für einen ersten Schritt. □

[69] Wir schreiben jetzt wieder die Laurentpolynome p_j aus (2.18) als z–Transformierte.

[70] Das Argument ist genau das gleiche wie im Beweis von Proposition 2.10.

[71] Bis auf das strikte "$>$" am Anfang lassen sich die Pole immer so anordnen.

Übung 2.3 Zeigen Sie: Für $\zeta, \eta \in \mathbb{C}$ mit $|\zeta| = |\eta| > 1$ sind die Koeffizienten der formalen Potenzreihe

$$\frac{p(z)}{1-\zeta\, z^{-1}} + \frac{q(z)}{1-\eta\, z^{-1}}$$

genau dann beschränkt, wenn sie alle verschwinden. ♢

Damit haben wir also die vollständige Beschreibung von rationalen Filtern, die das gewünschte Dämpfungsverhalten aufweisen und somit stabil sind. Man kann allerdings für endliche Signale noch ein bißchen mehr erhalten, indem man vom Mittel der "Zeitumkehr" Gebrauch macht. Dazu nehmen wir an,

$$f^*(z) = \frac{p(z)}{q(z)}, \qquad z \in \mathbb{C}^\times \setminus Z(q),$$

wäre die z–Transformation der Impulsantwort eines Filters F, der Pole *innerhalb und außerhalb* des Einheitskreises hat, aber *keinen* Pol auf dem Einheitskreis, d.h., $Z(q) \cap \mathbb{T} = \emptyset$. Nun zerlegen wir q in das Produkt seiner Nullstellen und zwar, bis auf eine Konstante c, die wir genauso in das Zählerpolynom p schieben können, als

$$q(z) = c\, \underbrace{\left(1-z^{-1}\zeta_1\right)\cdots\left(1-z^{-1}\zeta_m\right)}_{=:q_<(z)}\, \underbrace{\left(1-z^{-1}\eta_1\right)\cdots\left(1-z^{-1}\eta_n\right)}_{=:q_>(z)}, \qquad |\zeta_j|<1,\ |\eta_j|>1.$$

Offensichtlich haben $q_<$ und $q_>$ keine gemeinsamen Nullstellen und daher gibt es Laurentpolynome $p_<$ und $p_>$, so daß

$$f^*(z) = \frac{p_<(z)}{q_<(z)} + \frac{p_>(z)}{q_>(z)} =: f^*_<(z) + f^*_>(z).$$

Dann sind aber die Pole von $f^*_>\left(z^{-1}\right)$ alle im Inneren des Einheitskreises und wir können den Filter mit Hilfe der "Zeitspiegelung"[72] als

$$\left(F_> x\right)^*(z) = \left(\sigma_{-1}\left[F_> x\right]\right)^*\left(z^{-1}\right)$$

schreiben, wobei

$$\begin{aligned}
\left(\sigma_{-1}\left[F_> x\right]\right)^*(z) &= \sum_{j\in\mathbb{Z}} \left[F_> x\right](-j) z^{-j} = \sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} f_>(k)\; x(-j-k)\, z^{-j} \\
&= \sum_{k\in\mathbb{Z}} f_>(k)\; z^k \sum_{j\in\mathbb{Z}} x(-j-k)\, z^{-j-k} = \sum_{k\in\mathbb{Z}} f_>(k)\; z^k \sum_{j\in\mathbb{Z}} x(-j)\, z^{-j} \\
&= f^*_>\left(z^{-1}\right)\; (\sigma_{-1}x)^*(z).
\end{aligned}$$

Die z–Transformation $f^*_>\left(z^{-1}\right)$, die zum Filter $\sigma_{-1}F_>$ gehört, hat nun aber alle Pole im Inneren des Einheitskreises, denn die Operation $z \mapsto z^{-1}$ entspricht ja gerade einer Spiegelung am Einheitskreis. Das liefert eine gedämpften und somit *numerisch stabil* berechenbaren Filter und somit das stabilere Verfahren

$$F_> x = \sigma_{-1}\left[\left(\sigma_{-1}F_>\right) x\right],$$

[72] Denn nichts anderes ist die Skalierung σ_{-1}, definiert durch $\sigma_{-1}c = c(-\cdot)$.

das mit Spiegelung von Filter und Signal auskommt und eigentlich nur auf der Identität $\sigma_{-1}^2 = I$ basiert. Die numerisch stabile Realisierung von F ist also

$$Fx = F_< x + \sigma_{-1}\left[(\sigma_{-1}F_>)\, x\right],$$

was *formal* keinen, aber *numerisch* einen ganz gewaltigen Unterschied bei der Realisierung rationaler Filter macht, siehe [7].

Damit haben wir im Moment erst einmal genug von rationalen Filtern gesehen. Wie man mit ihnen Filterdesign betreibt, das werden wir später noch sehen.

. . . eine brilliante Lösung eines mathematischen Problems: so brilliant, daß sie erregend wirkt, aber zugleich so mathematisch, daß sie keine Illusionen über ihre blendende Logik zuläßt.

Roberto Cotroneo, *Presto con fuoco/Die verlorene Partitur*

Fourier – schnell und diskret

3

In diesem Kapitel wollen wir uns mit **dem** fundamentalen Algorithmus der Signalverarbeitung beschäftigen, nämlich mit der *schnellen Fouriertransformation* oder *FFT* ("*F*ast *F*ourier *T*ransform"). Als vor einiger Zeit die Liste der 10 bedeutendsten und einflußreichsten Verfahren aufgestellt wurde, war die FFT unangefochtener und eindeutiger Sieger. Und dabei handelt es sich eigentlich bei der FFT um eine unglaublich einfache Idee. Doch da die FFT eigentlich "nur" eine schnelle Berechnungsmethode der *diskreten Fouriertranformation* oder *DFT*[73] darstellen, ist es vernünftig, sich diese zuerst einmal anzusehen.

3.1 Die diskrete Fouriertransformation

Eigentlich ist die Fouriertransformation einer Folge eine seltsame Operation, bildet sie doch, im Gegensatz zur "normalen" Fouriertransformation, eine Folge $c \in \ell(\mathbb{Z})$ auf die 2π–periodische Funktion $\widehat{c} \in C(\mathbb{T})$ ab, was man auch daran sieht, daß die inverse Fouriertransformation (1.20) einer Folge eine völlig andere Struktur hat als die Fouriertransformation selbst. Ganz abgesehen davon ist es sowieso nicht möglich, *kontinuierliche* Frequenzinformation zu verarbeiten, so daß man auch im Frequenzbereich abtasten muß. Wegen der 2π–Periodizität der Fouriertransformierten $\widehat{c}$ empfiehlt es sich natürlich, diese Abtastgenauigkeit von der Form $h = \frac{2\pi}{n}$ für ein $n \in \mathbb{N}$ zu wählen. Dann ergibt sich die Folge

$$\widehat{c}_n = \mathrm{DFT}_n := S_{2\pi/n}\widehat{c} = \sum_{k\in\mathbb{Z}} c(k)\, e^{-2\pi i k\,\cdot/n}. \tag{3.1}$$

Definition 3.1 *Die Folge $\widehat{c}_n$ aus (3.1) bezeichnet man als* diskrete Fouriertransformierte *oder* DFT *der Folge $c \in \ell(\mathbb{Z})$ von der Ordnung n.*

Daß die DFT ein durchaus wichtiges Konzept ist, kann man schon daran erkennen, daß man sie nicht nur in allen Büchern über digitale Signalverarbeitung findet, sondern daß es sogar

[73] Glücklicherweise haben *"diskret"* und *"discrete"* denselben Anfangsbuchstaben.

eigene Bücher über die DFT gibt, z.B. [3]. Wegen

$$\widehat{c}_n\,(\cdot + n) = \sum_{k\in\mathbb{Z}} c(k)\, e^{-2\pi i\, k\,(\cdot/n+1)} = \widehat{c}_n$$

ist auch die DFT periodisch und es genügt, lediglich einen Block von n Einträgen zu speichern, das heißt, die DFT ist durch die Werte

$$\widehat{c}_n(k), \qquad k \in \mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z} \simeq \{0,\dots,n-1\}\,,$$

festgelegt.

Bemerkung 3.2 *Unter $\mathbb{Z}_n$ ist normalerweise* mehr *zu verstehen, als "nur" die Menge $\{0,\dots,n-1\}$, zum Beispiel sind alle Operationen auf $\mathbb{Z}_n$ sind immer modulo n aufzufassen. Wir werden aber hier nicht so pingelig auf diesen Details herumreiten und $\mathbb{Z}_n$ auch manchmal nur für die Menge verwenden – zumindest solange die exakte Bedeutung des Symbols aus dem Kontext ohne allzuviel Aufwand ersichtlich wird.*

Auf periodischen oder *periodisierten* Folgen[74] $c \in \ell(\mathbb{Z}_m)$, also mit Periodenlänge m kann man die DFT sogar als Matrix darstellen, nämlich als

$$\widehat{c}_n = V_{n,m}\, c, \qquad V_{n,m} := \left[e^{-2i\pi jk/n} \;:\; j \in \mathbb{Z}_n,\, k \in \mathbb{Z}_m\right].$$

Ist $n = m$, dann schreiben wir einfach V_n. Das ist auch die "Standardversion" der DFT, bei der Signale in Signale gleicher Länge oder gleichen Informationsgehalts transformiert werden. Noch ein Wort zur Periodisierung: Ist $c \in \ell_{00}(\mathbb{Z})$, dann kann man c ja so schieben, daß $\operatorname{supp} c \in \mathbb{Z}_n$ und dann kann man c mit V_n diskret Fourier–transformieren.

Beispiel 3.3 *Wir bestimmen die diskrete Fouriertransformierte der Folge $S_{2\pi/512}\cos$ auf $\mathbb{Z}_{512}$. Realteil und Imagnärteil sind in Abb 3.1 dargestellt.*

Generell liefern Sinus- und Kosinusschwingungen mit Frequenzen, die Teiler der Abtastrate sind, scharfe Spitzen in Real- bzw. Imaginärteil der DFT während entsprechende Funktionen mit "unpassenden" Frequenzen "verwaschen" werden.

Beispiel 3.4 *Wir betrachten die Funktion*

$$f(x) = \cos x - \cos 80x + \cos 130.7x + \sin 16x - \sin 277.8x$$

und bestimmen $DFT_{512} S_{2\pi/512} f$. Das Ergebnis ist in Abb. 3.2 zu sehen.

Im weiteren gehen wir nun davon aus, daß sowohl c also auch seine diskrete Fouriertransformierte zu $\ell(\mathbb{Z}_n)$ gehören, daß wir es also mit der Matrix V_n zu tun haben. Sehen wir uns

[74] Was nicht passt wird passend gemacht.

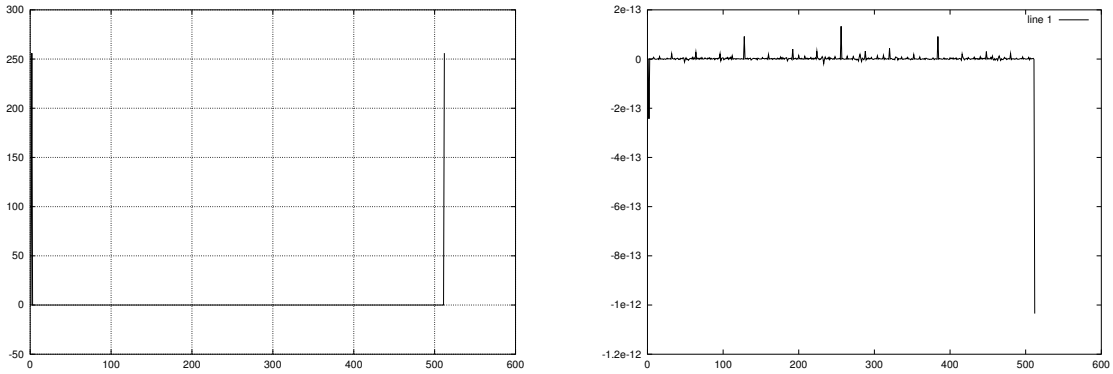


Abbildung 3.1: Die diskrete Fouriertransformierte

$$\mathrm{DFT}_{512}\, S_{2\pi/512}\cos\,.$$

Im linken Bild der Realteil, der starke Ausschläge an 1 und $511 \simeq -1$ hat – was ja auch passt, denn da

$$\cos x = \frac{e^{ix} + e^{-ix}}{2}$$

ist, tauchen in ihm gerade die beiden Frequenzen ± 1 auf. Und natürlich ist der Imagiärteil rechts praktisch Null und besteht eigentlich nur aus numerischem Müll, auch wenn dieser mit 10^{-13} durchaus in einer nicht so begeisternded Größenordnung liegt.

---

die Matrix mal genauer an; dazu ist es vernünftig $\omega = e^{-2\pi i/n}$ zu definieren und uns daran zu erinnern, daß ω eine (primitive) *n–te Einheitswurzel* ist, daß also $\omega^n = 1$ gilt. Dann ist

$$V_n = \left[\omega^{jk} \;:\; j,k \in \mathbb{Z}_n\right] = \begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ 1 & \omega^1 & \dots & \omega^{n-2} & \omega^{n-1} \\ 1 & \omega^2 & \dots & \omega^{2(n-2)} & \omega^{2(n-1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \omega^{n-2} & \dots & \omega^4 & \omega^2 \\ 1 & \omega^{n-1} & \dots & \omega^2 & \omega^1 \end{bmatrix}$$

Diese Matrix hat eine sehr einfache Inverse.

Lemma 3.5 (Inverse DFT) *Für $n \in \mathbb{N}$ ist*

$$V_n^{-1} = \frac{1}{n}\left[e^{2\pi ijk/n} \;:\; j,k \in \mathbb{Z}_n\right] = \frac{1}{n}\left[\omega^{-jk} \;:\; j,k \in \mathbb{Z}_n\right] \tag{3.2}$$

Beweis: Wir bezeichnen die Matrix auf der rechten Seite von (3.2) mit W_n, dann ist

$$\left(V_n\, W_n\right)_{jk} = \frac{1}{n}\sum_{\ell\in\mathbb{Z}_n} \omega^{j\ell}\,\omega^{-\ell k} = \frac{1}{n}\sum_{\ell\in\mathbb{Z}_n} \left(\omega^{j-k}\right)^{\ell}.$$

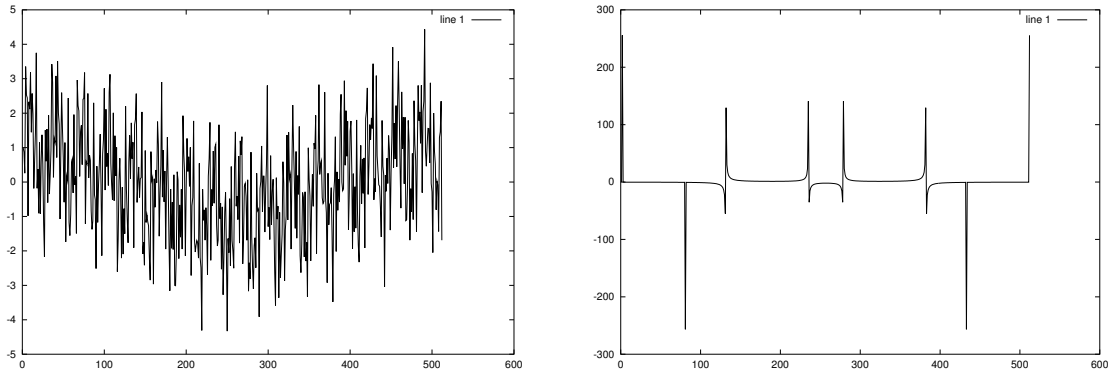


Abbildung 3.2: Die Funktion aus Beispiel 3.4 (links – sieht mit etwas Phantasie fast wie ein Sprachsignal aus) und Real- und Imaginärteil ihrer DFT (rechts). Die ganzzahligen Frequenzen liefern scharfe Zacken und zwar *entweder* im Real- *oder* im Imaginärteil, wohingegen die nichtganzzahligen Frequenzen zu "verschmierten" Ausschlägen in beiden Teilen des Spektrums führen.

Damit ist

$$(V_n W_n)_{jj} = \frac{1}{n} \sum_{\ell \in \mathbb{Z}_n} \left(\omega^0\right)^\ell = \frac{n}{n} = 1$$

und ansonsten

$$(V_n W_n)_{jk} = \frac{1}{n} \sum_{\ell=0}^{n-1} \left(\omega^{j-k}\right)^\ell = \frac{1}{n} \, \frac{\omega^0 - \left(\omega^{j-k}\right)^n}{1 - \omega^{j-k}} = \frac{1}{n} \, \frac{1 - (\omega^n)^{j-k}}{1 - \omega^{j-k}}$$

und da $\omega^n = 1$ und $-n < j - k < n$, also auch $\omega^{j-k} \neq 1$ ist, erhalten wir, daß

$$(V_n W_n)_{jk} = \frac{1}{n} \, \frac{1 - 1^{j-k}}{1 - \omega^{j-k}} = 0$$

ist, womit der Beweis vollständig ist. □

Nun ist aber $\omega^{-1} = e^{2\pi i/n}$ bei genauem Hinsehen nichts anderes als $\overline{\omega}$, das heißt, wir erhalten wegen der Symmetrie von V_n, daß

$$V_n^{-1} = \frac{1}{n}\overline{V_n} = \frac{1}{n}\overline{V_n^T} = \frac{1}{n} V_n^H,$$

was wir auch wie folgt formulieren können.

Korollar 3.6 *Die Matrix $n^{-1/2} V_n$ ist unitär.*

Bemerkung 3.7 *Die Verteilung des Faktors $\frac{1}{n}$ zwischen V_n und V_n^{-1} erscheint willkürlich und unsymmetrisch und tatsächlich gibt es auch Leute, die die DFT mit einem Vorfaktor $\sqrt{n}^{-1}$*

einführen. Das macht zwar die Konstante "schöner", liefert, wie Korollar 3.6 zeigt, auch eine unitäre Matrix, zerstört aber die Interpretation als Abtastung des trigonometrischen Polynoms und führt eine zusätzliche irrationale[75] *Größe ein. Außerdem spielt die Konstante* $1/\sqrt{n}$ *eine Rolle ganz analog zur Konstante* $1/2\pi$ *bei der Fouriertransformierten und deren Inverser. Und je mehr Analogien, desto besser – allerdings hängt bei der DFT die Normierungsgröße von der Länge des Vektors ab. Und ob die Matrix nun unitär ist oder nur* $V_n V_n^H = V_n^H V_n = nI$ *gilt, das ist genauso wesentlich wie der Unterschied zwischen orthogonal und orthonormal.*

Offenbar ist für $c \in \ell(\mathbb{Z}_n)$ die DFT $\widehat{c}_n \in \ell(\mathbb{Z}_n)$ eine, wenn nicht sogar *die* "natürliche" Operation. Wir stellen nun ein paar Eigenschaften von $\mathrm{DFT}_n : \ell(\mathbb{Z}_n) \to \ell(\mathbb{Z}_n)$ zusammen, und zwar in Analogie zu Satz 1.8. Dazu brauchen wir auch den zur DFT gehörigen Faltungsbegriff und das ist die *zyklische Faltung*, die die "periodische" Struktur von $\mathbb{Z}_n$ ausnutzt[76].

Definition 3.8 (Zyklische Faltung) *Zu* $c, d \in \ell(\mathbb{Z}_n)$ *ist die* zyklische Faltung $c * d = c *_n d \in \ell(\mathbb{Z}_n)$ *definiert als*

$$(c * d)(j) = \sum_{k \in \mathbb{Z}_n} c(k)\, d(j-k), \qquad j \in \mathbb{Z}_n,$$

wobei $j - n$ *enstprechend den Rechenregeln in* $\mathbb{Z}_n$*, also modulo* n *zu verstehen ist.*

Satz 3.9 (Eigenschaften der DFT) *Für* $n \in \mathbb{N}$ *gilt:*

1. DFT_n *ist eine invertierbare lineare Abbildung von* $\ell(\mathbb{Z}_n)$ *in sich mit*[77]

$$\|\widehat{c}_n\|_2 = \sqrt{n}\|c\|_2, \qquad c \in \ell(\mathbb{Z}_n). \tag{3.3}$$

2. *Für* $j \in \mathbb{Z}_n$ *ist*

$$(c(\cdot + j))_n^\wedge(k) = \omega^{-jk}\,\widehat{c}_n(k) \qquad \textit{und} \qquad \left(\omega^{j\cdot}\, c\right)_n^\wedge(k) = \widehat{c}_n(k+j), \qquad k \in \mathbb{Z}_n. \tag{3.4}$$

3. *Für* $c, d \in \ell(\mathbb{Z}_n)$ *ist*

$$(c *_n d)_n^\wedge = \widehat{c}_n\, \widehat{d}_n. \tag{3.5}$$

Beweis: *1.)* Linearität ist klar und Invertierbarkeit folgt aus Lemma 3.5 – da ist die inverse DFT ja explizit angegeben. Für (3.3) wenden wir die unitäre Invarianz der 2–Norm[78] an und erhalten

$$\|\widehat{c}_n\|_2 = \|V_n c\|_2 = \sqrt{n}\,\left\|\left(n^{-1/2}V_n\right)c\right\|_2 = \sqrt{n}\|c\|_2.$$

2.) Durch einfaches Nachrechnen erhält man

$$(c(\cdot + j))_n^\wedge(k) = \sum_{\ell \in \mathbb{Z}_n} c(j+\ell)\ \omega^{k\ell} = \sum_{\ell \in \mathbb{Z}_n} c(\ell)\,\omega^{k(\ell-j)} = \omega^{-kj} \sum_{\ell \in \mathbb{Z}_n} c(\ell)\,\omega^{k\ell} = \omega^{-kj}\,\widehat{c}_n(k)$$

[75] Naja, ganz so schlimm ist es auch wieder nicht, es ist ja "nur" eine Wurzel und die kann man auch algorithmisch effektiv zum Grundkörper $\mathbb{Q}$ adjungieren, siehe z.B. [45].

[76] Hier rechnen wir jetzt *wirklich* modulo n.

[77] Natürlich sind die p–Normen zu $c \in \ell(\mathbb{Z}_n)$ als $\left(\sum_{k\in\mathbb{Z}_n} |c(k)|^p\right)^{1/p}$ bzw. $\max_{k\in\mathbb{Z}_n} |c(k)|$ definiert.

[78] Zur Erinnerung: Für unitäres U, d.h. $U^H U = I$ und beliebiges x ist $\|x\|_2^2 = x^H x = x^H U^H U x = \|Ux\|_2^2$.

und

$$\left(\omega^{j\cdot}\, c\right)_n^{\wedge}(k) = \sum_{\ell\in\mathbb{Z}_n} \omega^{j\ell}\, c(\ell)\; \omega^{k\ell} = \sum_{\ell\in\mathbb{Z}_n} c(\ell)\; \omega^{(k+j)\ell} = \widehat{c}_n(k+j)\,.$$

3.) Da ω eine n–te Einheitswurzel ist, also $\omega^n = 1$ und somit auch $\omega^{k+n} = \omega^k$ gilt, ist $k \mapsto \omega^k$ eine Folge in $\ell(\mathbb{Z}_n)$. Daher erhalten wir für $j \in \mathbb{Z}_n$, daß

$$(c *_n d)_n^{\wedge}(j) = \sum_{\ell\in\mathbb{Z}_n} \left(\sum_{k\in\mathbb{Z}_n} c(k)\, d(\ell-k) \right) \omega^{j\ell} = \sum_{k,\ell\in\mathbb{Z}_n} c(k)\, d(\ell-k)\, \omega^{jk}\omega^{j(\ell-k)} = \widehat{c}_n(j)\, \widehat{d}_n(j).$$

Die andere Hälfte von (3.5) funktioniert analog. □

3.2 Diskret versus diskretisiert

In den allermeiste Praxisfällen, beispielsweise bei der Verarbeitung von Tönen oder auch Hirnstrommessungen, resultieren die zu verarbeitenden, diskreten Daten, aus der *endlichen* Abtastung eines kontinuierlichen Signals, also[79]

$$c(k) = (S_h f)(k), \qquad k \in \mathbb{Z}_N.$$

Wenn wir nun die DFT $\widehat{c}_N$ zu diesem Signal c berechnen, dann berechnen wir eine Diskretisierung des zugehörigen trigonometrischen Polynoms

$$\widehat{c}(\xi) = \sum_{k\in\mathbb{Z}} c(k)\, e^{-ik\xi} = \sum_{k\in\mathbb{Z}_N} f(hk)\, e^{-ik\xi},$$

auf dem Gitten $2\pi\mathbb{Z}_N/N$, also

$$\widehat{c}_n(j) = \sum_{k\in\mathbb{Z}_N} f(hk)\, e^{-2i\pi jk/N}, \qquad j \in \mathbb{Z}_N.$$

Dieser Vektor hat aber erst einmal keine direkte Verbindung zu dem, was wir eigentlich berechnen wollen, nämlich eine Diskretisierung der Fouriertransformation von f. Und das kann eben auch wieder zu Artefakten führen.

Beispiel 3.10 *Wir betrachten die DFT einer Abtastung der wohlbekannten* sinc *–Funktion, deren Fouriertransformierte ja eine charakteristische Funktion ist, und tasten sie, in* `Octave`*–Notation mittels*[80]

```
ocatve> N = 512; c = sinc( 100*pi*(0:N-1)/N );
```

ab. Das Ergebnis einer Fouriertransformation mit anschliessendem `fftshift`[81] *ist in Abb. 3.3 gezeigt – die Frequenzen am Rand des Bandpassfilters sind, wie man sieht, deutlich überhöht.*

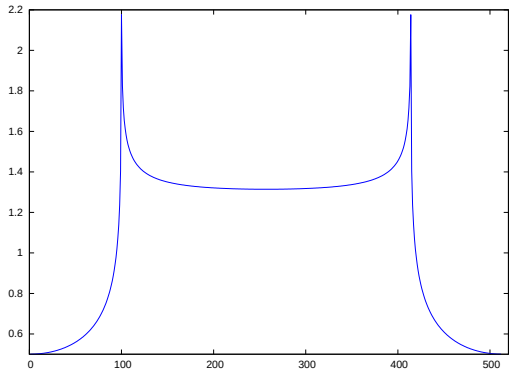


Abbildung 3.3: Die DFT der sinc –Funktion aus Beispiel 3.10. Man sieht, daß am Rand des Frequenzintervalls recht böse Artefakte auftreten, die mit Diskretisierung allein nicht erklärbar sind.

Um eine bessere Annäherung an die eigentliche Funktion zu erhalten, verwenden wir einen sogenannten *Quasiinterpolanten* als Approximation an f auf der Basis der Abtastungen. Zu einer Funktion $\phi \in L_{00}(\mathbb{R})$ ist der Quasiinterpolant recht einfach als skalierte Faltung

$$Q_{h,\phi}c := \phi * c\left(h^{-1}\cdot\right) = \sum_{k\in\mathbb{Z}_N} c(k)\phi\left(h^{-1}\cdot - k\right) = \sum_{k\in\mathbb{Z}_N} f\left(hk\right)\,\phi\left(h^{-1}\cdot - k\right)$$

definiert. Ist ϕ sogar eine *kardinale Funktion*, das heißt, gilt $\phi|_{\mathbb{Z}} = \delta$, dann ist $Q_{h,\phi}(hk) = S_h f(k) = f(hk)$, $k \in \mathbb{Z}_N$, es werden also die abgetasteten Daten interpoliert[82]. Andernfalls hofft man, so zumindest eine Approximation zu erhalten.

Beispiel 3.11 *Die gebräuchlichsten Funktionen für derartige Quasiinterpolanten sind die* kardinalen Splines*, also die Splines, deren unendliche Knotenmenge gerade $\mathbb{Z}$ ist. Darunter befinden sich* interpolatorische*, nämlich die Splines zu den Ordnungen* 0 *und* 1*, und* nichtinterpolatorische*, nämlich der ganze Rest. Die so resultierenden Approximationsoperatoren, die sogenannten* Schoenbergoperatoren *, sind beispielsweise in [46, 48, 50] beschrieben.*

Wenn wir einmal davon ausgehen, daß $Q_{h,\phi}c$ die Funktion f halbwegs approximiert, daß also $\|f - Q_{h,\phi}c\|_1$ klein ist, dann ist die Fouriertransformierte von $Q_{h,\phi}c$ auch eine gute Approximation der Fouriertransformierten von f und wir können letztere aus den abgetasteten Daten als

$$\left(Q_{h,\phi}c\right)^\wedge(\xi) \;=\; \left(\sigma_{h^{-1}}\left(\phi * c\right)\right)^\wedge(\xi) = h\left(\phi * c\right)^\wedge(h\xi) = h\,\widehat{\phi}(h\xi)\,\widehat{c}(h\xi).$$

[79] Wir verwenden hier N für die Anzahl der Abtastungen, um zum Ausdruck zu bringen, daß es sich dabei um eine sehr große Anzahl handelt.

[80] Man sollte die sinc –Funktion nicht auf dem ganzzahligen Gitter abtasten, da bekäme man eine δ–Folge,

[81] Diese `Octave`–Funktion sorgt dafür, daß die Nullfrquenz in die Mitte des Vektors geschoben wird.

[82] Was in gewissem Sinne das “Quasi” erklärt.

berechnen. Ersetzen wir in dieser Gleichung noch ξ durch $h^{-1}\xi$, und diskretieren ξ auf dem *diskreten Torus* $\mathbb{T}_N := 2\pi\mathbb{Z}_N/N$, so erhalten wir, daß

$$\widehat{f}\left(\frac{2k\pi}{Nh}\right) \simeq (Q_{h,\phi}c)^\wedge\left(\frac{2k\pi}{Nh}\right) = h\,\widehat{\phi}\left(\frac{2k\pi}{N}\right)\widehat{c}_N(k), \qquad k \in \mathbb{Z}_N. \tag{3.6}$$

Diese einfache Formel verknüpft nun die diskrete Fouriertransformation $\widehat{c}_N(k) = (S_h f)^\wedge_N$ der Abtastung mit einer *näherungsweisen* Diskretisierung der Fourietransformierten von f und erklärt auch sehr schön die Zusammenhänge:

1. Die *Abtastrate* bzw. *Samplingrate*[83] h bestimmt, welche Frequenzen von f wirklich in der DFT $\widehat{c}_N$ codiert sind und je kleiner h ist, desto größer wird dieser Frequenzbereich[84].

2. Die *Frequenzauflösung*, also die Anzahl der wirklich berechneten Spektrumseinträge, hängt hingegen von der gewählten Diskretisierungszahl N ab - je größer N ist, desto genauer wird das Spektrum dargestellt, je kleiner N ist, desto mehr Frquenzen werden zu einem Block zusammengefasst. Natürlich steigt mit wachsendem N auch der Rechenaufwand, doch dazu gleich mehr.

3. So einfach entkoppeln kann man diese beiden Größen nicht! Normalerweise resultieren die abgetasteten Daten ja aus Messungen über einen gewissen "nicht zu kurzen" Bereich bzw. Zeitraum[85], so daß Nh normalerweise von signifikanter Größe sein wird, was dazu führt, daß eine hohe Abtastrate in der Praxis auch mit einer hohen Frquenzauflösung verbunden sein dürfte.

4. Der Normierungsfaktor h in (3.6) ist nur dann wichtig, wenn wir uns wirklich für die konkreten Werte im Spektrum interessieren, geht es uns nur um die *Spektralverteilung*, dann können wir ihn berücksichtigen oder nicht.

5. Der Abstand zwischen $\widehat{f}$ und dem wirklich berechneten $(Q_{j,\phi}f)^\wedge$ beeinflusst natürlich ganz entscheidend, wie genau die berechneten Werte wirklich die diskretisierte Fouriertransfer beschreibt. Da $\|\widehat{g}\|_\infty \leq \|f\|_1$ gilt, und da wir nur auf dem Intervall $[0, Nh]$ abtasten, ist die L_1–Approximationsgüte

$$\|f - Q_{h,\phi}f\|_1 := \int_0^{Nh} |f(t) - Q_{h,\phi}f(t)|\,dt \leq Nh \max_{0\leq t\leq Nh} |f(t) - Q_{h,\phi}f(t)|$$

entscheidend für die Qualität unserer Näherung. Bei kardinalen Splinefunktionen gibt es Abschätzungen hierfür[86], siehe [48], die normalerweise von der Größenordnung Ch^2 sind.

[83] Der internationale Terminus Technicus.

[84] Wer hätte das gedacht?

[85] Beispielsweise eine Aufnahme eines Musikstücks oder die Erfassung der Hirnstromdaten während eines psychologischen Experiments.

[86] Das Zauberwort hierfür heisst *Schoenbergoperator*.

6. Verwendet man Splinefunktionen als ϕ, dann ist der Korrekturfilter $\widehat{\phi}$ besonders einfach zu berechnen, nämlich eine Potenz der sinc –Funktion, aber, wegen der Abtastung nur der ersten "Berg" dieser Funktione. Für hohe Frquenzen fällt so eine Potenz für steigende Ordnung der Spline natürlich auch immer schneller ab, sorgt also für eine stärkere Dämpfung unerwünschter hoher Frequenzen.

7. Lässt man ϕ in (3.6) weg, dann wählt man $\widehat{\phi}$ als $\chi_{[0,1]}$, also ϕ als sinc –Funktion und anstelle von $\widehat{f}$ diskretisiert man die Fouriertransformierte der interpolatorischen Rekonstruktion aus dem Abtastsatz. Die ist aber eben keine sonderlich gute Approximation, es sei denn, sie rekonstruiert exakt, das heisst, f müsste bandbeschränkt und die Abtastrate passend gewählt sein. Nur gerade dafür gibt es in den wenigsten Fällen wirklich eine Garantie.

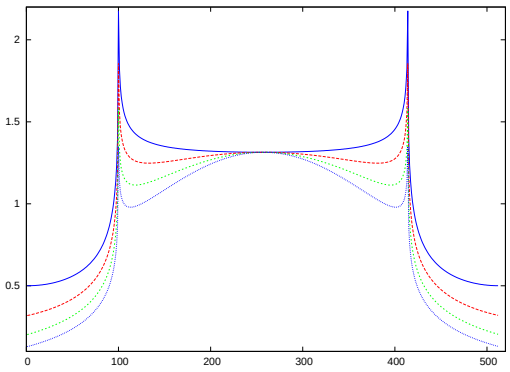


Abbildung 3.4: Filterung der Funktion aus Abb. 3.3 mit Splines der Ordnungen $0, 1, 2, 3$. Man sieht sehr schön, daß der Ausreisser deutlich kleiner wird, das allerdings um den Preis einer "Delle" am Rand. So ganz gut approximieren die Splines die sinc –Funktion also leider nicht.

3.3 Die schnelle Fouriertransformation

Nun liegt der besondere Charme der DFT aber nicht nur in der Tatsache, daß sie eine konsistente Erweiterung der Fouriertransformierten für periodische oder periodisierte Folgen ist, sondern vor allem daran, daß man sie besonders schnell durchführen kann. Das führt zur *schnellen Fouriertransformation* oder *Fast Fourier Transform*, kurz als *FFT* bezeichnet. Genaugenommen ist die FFT also eine FDFT, eine schnelle *diskrete* Fouriertransformation. Dieses Verfahren wurde 1965 in [6] von Cooley und Tukey (wieder)entdeckt, siehe dazu [4, 5], und funktioniert nicht nur für die diskrete Fouriertransformation, sondern auf recht beliebigen Ringen – alles, was man wirklich braucht ist eine *primitive n–te Einheitswurzel* wie unser ω. Dabei ist die Idee hinter der FFT auch noch sehr einfach: Nehmen wir einmal an, daß $n = 2m$ eine gerade Zahl wäre

und bemerken wir, daß

$$\omega^2 = e^{-2\pi i 2/n} = e^{-2\pi i m} =: \omega_m, \qquad \omega_n := \omega,$$

dann ist für $c \in \ell(\mathbb{Z}_n)$ und $j \in \mathbb{Z}_n$

$$\begin{aligned}\widehat{c}_n(j) &= \sum_{k\in\mathbb{Z}_n} c(k)\,\omega^{jk} = \sum_{k\in\mathbb{Z}_m} c(2k)\,\omega^{2jk} + \sum_{k\in\mathbb{Z}_m} c(2k+1)\,\omega^{j(2k+1)} \\ &= \sum_{k\in\mathbb{Z}_m} c(2k)\,\omega_m^{jk} + \omega^j \sum_{k\in\mathbb{Z}_m} c(2k+1)\,\omega_m^{jk} = (c(2\cdot))^\wedge_m(j) + \omega^j\ (c(2\cdot+1))^\wedge_m(j),\end{aligned}$$

also

$$\widehat{c}_n = (c(2\cdot))^\wedge_m + \omega^{\cdot}\,(c(2\cdot+1))^\wedge_m \tag{3.7}$$

Worin liegt nun der Wert dieser Darstellung? Nun, wenn wir einmal annehmen, daß die Werte $\omega, \ldots, \omega^{n-1}$ in tabellierter Form vorliegen und vorberechnet sind[87], dann benötigt die "naive" Realisierung der DFT als Multiplikation einer $n \times n$–Matrix mit einem n–Vektor $O(n^2)$ Rechenoperationen. Nehmen wir an, die Berechnung über (3.7) würde $F(n)$ Rechenoperationen benötigen. Dann sagt uns (3.7), daß wir zur Berechnung von $\widehat{c}_n$ die beiden DFTs der Länge $m = n/2$ berechnen müssen (Aufwand $2F(n/2)$), den zweiten komonentenweise mit dem Vektor $[\omega^j : j \in \mathbb{Z}_n]$ multiplizieren (Aufwand n) und die beiden komponentenweise addieren[88] (Aufwand n). Insgesamt müssen wir also einen Aufwand von $2\,(F(n/2)+n)$ betreiben und erhalten so die Beziehung

$$F(n) = 2\,(F(n/2)+n)\,, \tag{3.8}$$

für den unbekannten Aufwand n. Nehmen wir mal an, daß $n = 2^\ell$ für $\ell \in \mathbb{N}$ ist, dann gilt die Beziehung

$$F(n) = 2^k F\left(2^{\ell-k}\right) + k\,2^{\ell+1}, \qquad k = 1, \ldots, \ell, \tag{3.9}$$

was für $k = 1$ gerade (3.8) ist und sich sonst induktiv aus

$$\begin{aligned}F(n) &= 2^k F\left(2^{\ell-k}\right) + k\,2^{\ell+1} = 2^k\ 2\left[F\left(2^{\ell-k-1}\right) + 2^{\ell-k}\right] + k\,2^{\ell+1} \\ &= 2^{k+1} F\left(2^{\ell-k-1}\right) + 2^{\ell+1} + k\,2^{\ell+1} = 2^{k+1} F\left(2^{\ell-k-1}\right) + (k+1)\,2^{\ell+1}\end{aligned}$$

ergibt. Betrachten wir nun speziell (3.9) für $k = \ell = \log_2 n$, dann ist

$$F(n) = \underbrace{2^\ell}_{=n} F(1) + \underbrace{\ell\,2^{\ell+1}}_{=2n\,\log_2 n} = n\,(2\,\log_2 n + F(1)) = O(n\,\log_2 n)\,,$$

was *deutlich* besser ist als $O(n^2)$. Tatsächlich ist $O(n\,\log_2 n)$ eine typische asymptotische Komplexität für derartige Methoden, die auf dem Prinzip "Halbieren und Rekursion" basieren,

[87] Diese Werte sind ja für alle Vektoren $c \in \ell(\mathbb{Z}_n)$ dieselben und können daher beispielsweise in einem Cachespeicher vorgehalten werden. Und selbst wenn sie nicht vorberechnet sind, dann kann man sie immer noch mit einem Aufwand von "nur" $O(n)$ bestimmen.

[88] Diese beiden Vektoren sind m–periodisch, werden also einfach fortgesetzt

diese Tatsache wird bei diskreten Komplexitätsbetrachtungen auch gerne als *"Master Theorem"* bezeichnet, siehe z.B. [58].
Und diese Komplexitätsaussage gilt nicht nur für Zweierpotenzen! Ist nämlich $2^{\ell-1} < n \leq 2^\ell$, dann ersetzen wir einfach n durch 2^ℓ indem wir die Vektoren beispielsweise durch Nullen ergänzen[89] und so eine Komplexität von

$$\begin{aligned} 2^\ell F(1) + 2\ell\, 2^\ell &\leq 2n\, F(1) + 2\log_2(2n)\, 2n = 2n\, F(1) + 4n\,(\log_2 n + 1) \\ &= 2n\,(2\log_2 n + F(1) + 2)\,, \end{aligned}$$

also im wesentlichen nur einen Faktor 2 erhalten – asymptotisch ist das immer noch $O\,(n\,\log_2 n)$.

Beispiel 3.12 *Natürlich ist das mit dem "Auffüllen" nur ein reines Komplexitätsargument und nicht das, was man in der Realität machen sollte. Bildet man beispielsweise $S_{2\pi/500}\cos(50\cdot)$ also eine Folge in $\ell\,(\mathbb{Z}_{500})$ und füllt diese Folge mit 12 Nullen zu einem Element von $\ell\,(\mathbb{Z}_{512})$ auf, dann verschmieren sich die Frequenzen ganz gewaltig und zwar reell ebenso wie imaginär[90], siehe Abb. 3.5*

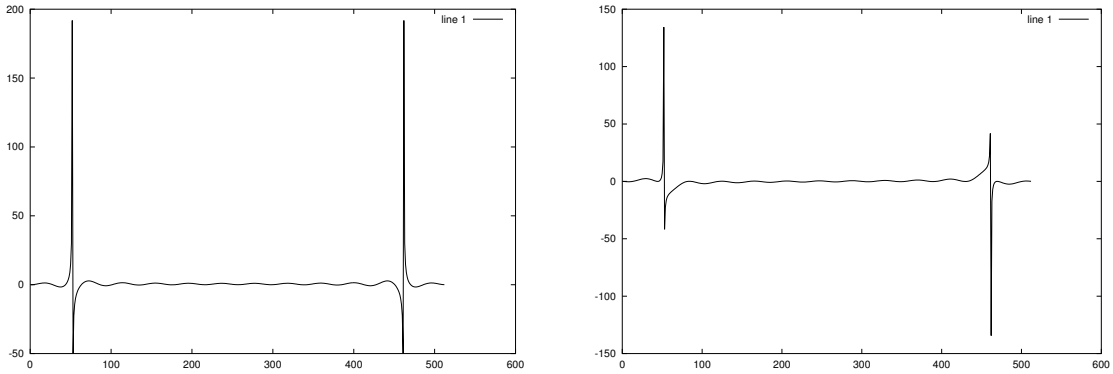


Abbildung 3.5: Real- (links) und Imaginärteil (rechts) der FFT von $S_{2\pi/500}\cos(50\cdot)$, aufgefüllt auf 512 Einträge. Zwar sind die beiden "Frequenzen" 50 und 450 immer noch deutlich sichtbar, aber die sehr großen Imaginärteile sind schon irreführend.

Nun könnte man sagen, das Problem mit Abb. 3.5 läge daran, daß jedwede Form von Periodizität kaputtgemacht wird und man das Signal besser periodisch fortsetzen sollte – aber dann passen halt die beiden Perioden 500 und 512 auch wieder nicht zusammen, außer in dem glücklichen Fall, daß das Signal so hochfrequent ist, daß die Periodisierung gerade eine volle Signalperiode erwischt.

Die Form der FFT, die wir hier betrachtet haben, ist die sogenannte *Radix–2* FFT, da die Zerlegungen auf der Basis 2, also auf Halbierung des Datenbestands beruhen. Man kann das

[89] Was zwar, wie wir sofort sehen werden, die Komplexität nicht signifikant verschlechtert, aber unsere schöne Periodizität zerstören wird!

[90] Zur Erinnerung: imaginäre Frequenzen dürften in der DFT gar nicht auftreten!

aber auch mit jeder anderen Zahl, beispielsweise mit der Basis $p \in \mathbb{N}$ und erhält dann für $m = n/p$ die analoge Zerlegung "modulo p"

$$\begin{aligned}\widehat{c}_n(j) &= \sum_{k\in\mathbb{Z}_m}\sum_{\ell\in\mathbb{Z}_p} c\,(pk+\ell)\;\omega^{j(pk+\ell)} = \sum_{\ell\in\mathbb{Z}_p}\omega^{j\ell}\left[\sum_{k\in\mathbb{Z}_m} c\,(pk+\ell)\;\omega_p^{jk}\right]\\ &= \sum_{\ell\in\mathbb{Z}_p}\omega^{j\ell}\;\left(c\,(p\cdot+\ell)\right)_m^{\wedge}(j),\end{aligned}$$

wir mußsen als *mehr* DFTs aber für *kürzere* Segmente berechnen. Der Aufwand hierbei ist dann $O\left(n\,\log_p n\right)$, bleibt also bis auf eine Konstante unverändert.

Übung 3.1 Zeigen Sie, daß der Rechenaufwand bei der Radix–p–FFT von der Größenordnung $O\left(n\,\log_p n\right)$ ist. ♢

Worin liegen nun die Vorteile der Radix–p–FFT?

- Das Parallelisierungspotential ist höher: Man kann die p Summanden von $\widehat{c}_n(j)$ ja *unabhängig* voneinander berechnen, sofern man nur die Eingangsdaten passend aufspaltet. Hat man beispielsweise p FFT–Prozessoren zur Verfügung, so macht man zuerst einen Radix–p–Schritt und läßt diese dann unabhängig voneinander operieren, was die Laufzeit sofort auf den Faktor $1/p$ reduziert, ganz unabängig davon, bezüglich welcher Basis die einzelnen FFTs arbeiten – das Problem ist so auf ziemlich einfache Weise *linear skalierbar*.

- Man braucht keine Zweierpotenzen sondern könnte eine *Primfaktorzerlegung* von n verwenden, indem man auf die entsprechenden Radix–p_j–Verfahren zurückgreift, $n = \prod p_j$. Insbesondere erspart man sich auch auf diese Art und Weise das fehleranfällige "Auffüllen" auf eine Zweierpotenz. Der Nachteil ist allerdings auch klar: Die Bestimmung von Primfaktoren ist nicht so einfach und schnell durchzuführen wie eine Division durch 2, die auf einem binären Digitalrechner[91] durch einen einfachen Schiebeprozess realisiert wird.

3.4 Anwendungen der FFT

Daß die FFT so ein bedeutsames Verfahren ist, liegt daran, daß sie nicht nur die DFT schnell berechnet, sondern dazu benutzt werden kann, auch viele andere Berechnungen schnell durchzuführen. Das Stichwort hierbei heißt diskrete Faltung! Nach (3) aus Satz 3.9 kann man nämlich die zyklische Faltung $c *_n d$ zweier Folgen aus $\ell\,(\mathbb{Z}_n)$, eine typische $O\,(n^2)$–Geschichte wenn man es naiv macht, auch dadurch bestimmen, daß man die FFT von c und d berechnet (Aufwand $C\,n\log_2 n$), diese dann *komponentenweise* multipliziert (Aufwand $C\,n$) und das Ergebnis wieder rücktransformiert. Damit ist der Aufwand auch schon auf $O\,(n\,\log_2 n)$ gedrückt. Und mit zyklischen Faltungen kann man schon eine Menge anfangen und die FFT beispielsweise für die folgenden Zwecke nutzen:

[91] "Binär sollte man durchaus dazusagen! Auch wenn im Moment alle "handelsüblichen" Prozessoren auf der Basis 2 operieren, ist nicht garantiert, daß das immer so bleiben wird. Der Fließpunkt–Standard IEEE 854 erlaubt laut [29] beispielsweise Arithmetik zu den Basen 2 und 10.

1. *Schnelle Filterung:* Man kann die Faltung zweier Folgen, also insbesondere die Anwendung eines FIR–Filters schneller über die FFT berechnen. Allerdings ist etwas Vorsicht geboten, denn die "normale" Faltung ist *nicht* periodisch. Das kann man aber dadurch korrigieren, daß man die zu faltenden Folgen vorne und hinten mit Nullen auffüllt, so daß die Überlappeffekte der zyklischen Faltung vermieden werden.

2. *Schnelle Matrixmultiplikation:* Eine Matrix $A \in \mathbb{R}^{\mathbb{Z}_n \times \mathbb{Z}_n}$ heißt *Toeplitz–Matrix*, wenn sie von der Form[92]
$$A = [a_{jk} = a_{j-k} \ : \ j,k \in \mathbb{Z}_n]$$
ist. Dann gilt für die Matrix–Vektor–Multiplikation
$$Ax = \left[\sum_{k \in \mathbb{Z}_n} a_{j-k}\, x_k \ : \ j \in \mathbb{Z}_n \right] = a *_n x, \qquad a = a_{0\cdot},$$
und entsprechend
$$AB = [a *_n b_{\cdot k}(j) \ : \ j,k \in \mathbb{Z}_n],$$
so daß man die Linksmultiplikation einer Toeplitz–Matrix an einen Vektor mit $O(n \log_2 n)$ anstelle von $O(n^2)$ und an eine Matrix mit $O(n^3 \log_2 n)$ anstelle von $O(n^3)$ bewältigen kann. Besonders interessant wird das, wenn man eine Toeplitz–Matrix von links an eine große *dünnbesetzte* Matrix multipliziert, denn dann kann man mit etwas Glück sogar noch deutlich bessere Ergebnisse erzielen. Solche Methoden spielen eine Rolle bei der *Vorkonditionierung* von linearen Gleichungssystemen für die Verwendung iterativer Methoden.

3. *Schnelle Polynom- und Ganzzahlmultiplikation:* Bei der Multiplikation von zwei Polynomen
$$f(z)\, g(z) = \left(\sum_{j=0}^{m} f_j\, z^j \right) \left(\sum_{k=0}^{n} g_k\, z^k \right) = \sum_{j=0}^{n+m} \underbrace{\left(\sum_{k+\ell=j} f_k\, g_\ell \right)}_{\sim f *_j g} z^j$$
spielen zyklische Faltungen offenbar eine bedeutende Rolle bei der Bestimmung der Koeffizienten des Produkts, weswegen man auch hier durch die FFT signifikante Beschleunigungsgewinne erzielen kann, siehe z.B. [52]

   Zur Ganzzahlmultiplikation kommt man, indem man eine ganze Zahl in einem Stellenwertsystem zur Basis $B > 0$ als Polynom $f(B)$ auffaßt und dann die Multiplikation der Zahlen auf eine Polynommultiplikation mit geeigneter Behandlung des Übertrags "zurückführt". Diese Idee führt zum "klassischen" Algorithmus von Schönhage und Strassen [51, 53], allerdings sind da noch jede Menge von Detailproblemen zu behandeln, siehe auch [18, 45].

Für praktische Anwendungen können wir uns das Leben aber leichtmachen: Sowohl `Matlab` als auch `Matlab` verfügen über die eingebaute Funktion `fft`, die einen Vektor schnell und diskret in seine DFT gleicher Länge überführt – so wurden ja auch die Beispiele berechnet.

[92] Ist unserere $\mathbb{Z}_n$–Notation nicht hilfreich?

3.5 Realisierung der FFT

Jetzt, wo wir wissen, daß die DFT so eine tolle Sache ist, sollten wir uns natürlich auch ansehen, wie FFT in der Praxis realisiert wird, auch wenn wir hier nur einen Bruchteil dessen betrachten können, was bekannt und in "realistischen" Implementierungen auch eingebaut ist, siehe [13]. Trotzdem wollen wir die grundlegende Implementierungsidee, das "Butterfly"–Element einmal betrachten. Dazu beginnen wir mit dem einfachsten Fall, nämlich $\ell = 1$, also $n = 2^\ell = 2$ und damit $c = [c(0), c(1)] \in \ell(\mathbb{Z}_2)$. Die FFT bestimmt hier die DFT der Ordnung 0 von $c(0)$ und $c(1)$, also[93]

$$\sum_{k=0}^{0} c(0)\,\omega^k = c(0) \qquad \text{sowie} \qquad \sum_{k=0}^{0} c(1)\,\omega^k = c(1)$$

und kombiniert diese mittels

$$\widehat{c}_1(j) = \begin{cases} c(0) + \omega^0\, c(1) = c(0) + c(1), & j = 0, \\ c(0) + \omega^1\, c(1) = c(0) + \omega\, c(1), & j = 1. \end{cases}$$

Man kann das aber auch anders sehen: Die beiden "transformierten Signale" $c(0)$ und $c(1)$ werden übereinandergeschrieben und "über Kreuz" verknüpft, um das Ergebnis zu liefern, siehe Abb. 3.6. Bei diesem "über–Kreuz–Prozess" wird die untere Hälfte natürlich entsprechend gewichtet.

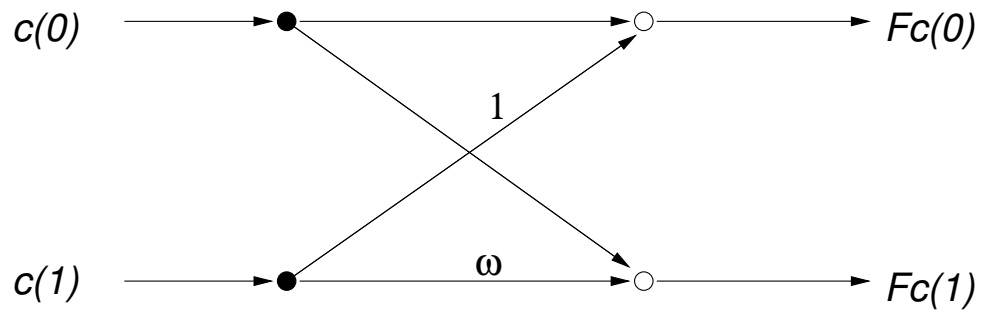


Abbildung 3.6: Das "Butterfly–Schema" zur Berechnung der FFT in der einfachsten Form $n = 2$; hier steht natürlich Fc für $\widehat{c}$.

Im "allgemeinen" Fall können wir die FFT zuerst einmal als Blockschaltbild von FFTs halber Länge darstellen und deren Ergebnisse gemäß (3.7) zusammensetzen, siehe Abb. 3.7. Natürlich ist dabei in beiden Fällen "FFT"derselbe "Block" von Operationen.
Man beachte, daß bei dieser blockweisen Verarbeitung die Einträge des Signals c umgestellt werden müssen, nämlich getrennt in gerade und ungerade Einträge, wobei die geraden zuerst kommen. Im Verarbeitungsblock "FFT" passiert nun wieder dasselbe und bei jedem rekursiven Aufruf von "FFT" während der Rekursion auch wieder.

[93] Da trivialerweise $\mathbb{Z}_1 = \{0\}$ ist.

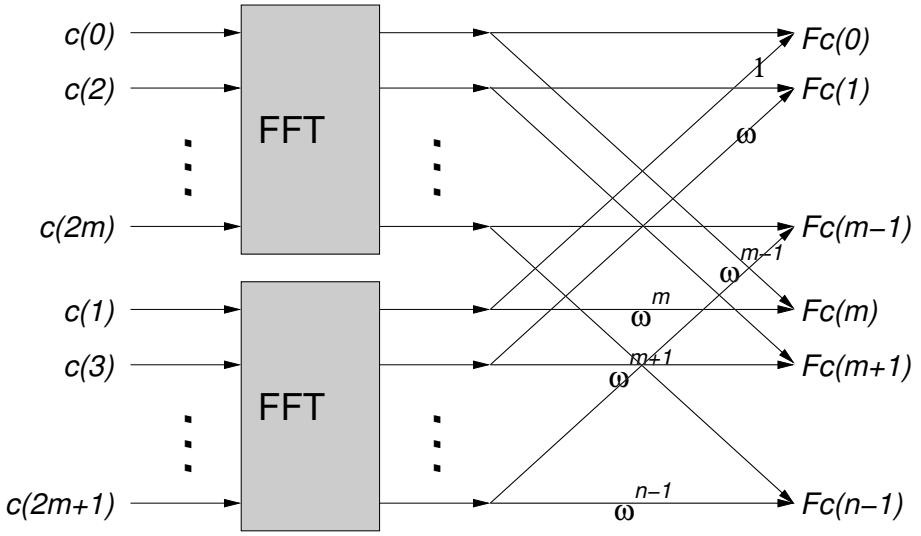


Abbildung 3.7: Das Butterfly–Schema als Blockschaltbild mit $m = \frac{n}{2} - 1$. Die Hälfte der Daten, nämlich die mit den ungeraden Indizes wird ebenso in die FFT halber Länge eingefüttert wie die Daten mit den ungeraden Indizes. Die Ergebnisse werden dann kombiniert.

Durch diese Rekursion und das Aufspalten in einen geraden und einen ungeraden Anteil werden die Elemente von c offenbar "gemischt", also umgeordnet, und da stellt sich natürlich die Frage, ob und inwieweit man diesen Prozess auch vernünftig beschreiben kann. Dazu ein Beispiel.

Beispiel 3.13 *Wie sieht dieser Datenzugriff für* $n = 4$ *aus? Hier wird* $c(0), \ldots, c(3)$ *in*

$$c(0), c(2), c(1), c(3)$$

umsortiert. Etwas interessanter wird es dann schon für $n = 8$*, wo wir die Sortierung*

$c(0)$	$c(0)$	$c(0)$	000	000
$c(1)$	$c(2)$	$c(4)$	100	001
$c(2)$	$c(4)$	$c(2)$	010	010
$c(3)$	$c(6)$	$c(6)$	110	011
$c(4)$	$c(1)$	$c(1)$	001	100
$c(5)$	$c(3)$	$c(5)$	101	101
$c(6)$	$c(5)$	$c(3)$	011	110
$c(7)$	$c(7)$	$c(7)$	111	111

erhalten, aus der wir dann auch das Schema ablesen können, das sich hinter dieser Ordnung verbirgt: Die Indizes werden in Binärschreibweise dargestellt, rückwarts gelesen und die Daten bezüglich dieser Reihenfolge angeordnet, siehe Abb. 3.8. Diesen Vorgang bezeichnet man als bit reversal.

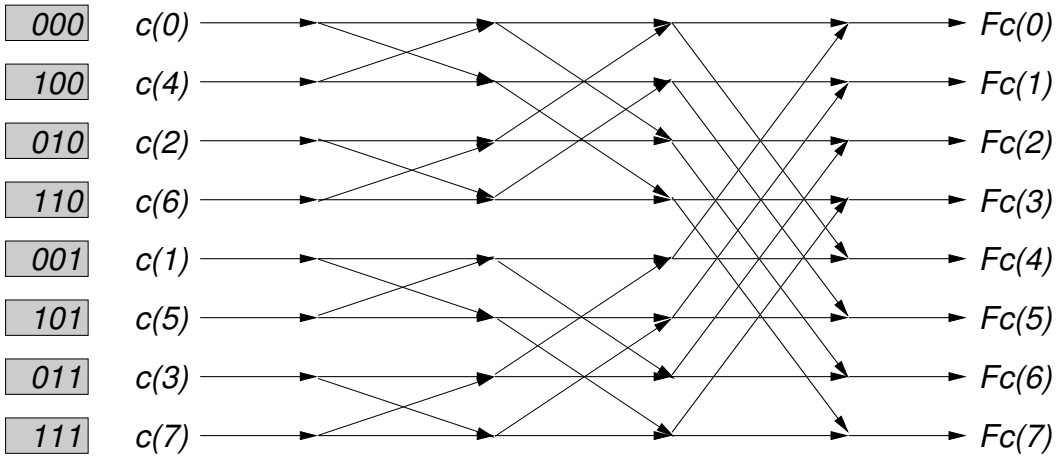


Abbildung 3.8: Die vollständige Butterfly–Struktur der FFT für $n = 8$. Durch Auflösung der Rekursion kommen zuerst die "kleinen" FFT–Blöcke der Länge 2, dann die der Länge 4 und schließlich die Kombination der Länge 8. Die Potenzen von ω zu ergänzen bleibt als Übungsaufgabe. Würde man hingegen die Eingabedaten links in ihrer "natürlichen" Reihenfolge auflisten, dann würde man ein ziemlich chaotisches Bild erhalten, siehe z.B. [32, Bild 6.4.6, S. 239].

Die Prozedur des *"bit reversal"*, die in Beispiel 3.13 exemplarisch dargestellt wurde, ist einer der wesentlichen Aspekte bei einer wirklich effizienten Implementierung der FFT – alle Prozesse wie Subsampling[94], Sortieren und Mischen können so, zumindest was die Indizes betrifft, auf *Bitebene* und somit recht effektiv durchgeführt werden.

Weitere algorithmische Details über die Implementierung und Realisierung der FFT – und davon gibt es jede Menge, was allerdings aufgrund der Tatsache, daß es sich bei der FFT um **das** numerische Verfahren schlechthin handelt, nicht weiter überraschen sollte, finden sich beispielsweise und vor allem in [62].

3.6 Undichte Fenster

Es ist natürlich schwer, ein Signal "unendlicher Dauer" oder auch nur eine sehr lange Folge $c \in \ell(\mathbb{Z}_n)$ für ein sehr großes n der Fouriertransformation zu unterziehen[95], denn selbst wenn $n \log_2 n$ als ein eher langsames Wachstum gilt überfordert die Transformation eines Musikstücks[96] auf "einen Durchgang" immer noch die Kapazitäten verfügbarer Rechner, ganz zu schweigen von kleinen Signalprozessoren, die beispielsweise in einen portablen CD- oder

[94] Also das Isolieren der geradzahligen und ungeradzahligen Teilfolgen.

[95] Wie bitte schreibt man das: "Fourierzutransformieren" oder "zu Fouriertransformieren" – "Fourier zu transformieren" ist jedenfalls etwas anderes.

[96] Die durchschnittliche Länge einer `wav`–Datei beträgt etwa 40 MB und selbst wenn genug Speicher zur Verfügung stehen sollte ist das Einlesen und Wegschreiben derartiger Datenmengen immer noch mit ganz immensem Aufwand verbunden!

MP3–Player eingebaut werden sollen. Deswegen verarbeitet man nicht das gesamte Signal, sondern eben nur Stücke des Signals, die man durch einen Filterungsprozess erhält. Um eine FFT der Länge $n = 2^\ell$ zu berechnen betrachtet man dazu die *"Fenster"*

$$\ell(\mathbb{Z}_n) \ni \; c_k = (Fc)(\cdot + kn) = (f * c)(\cdot + kn), \qquad k \in \mathbb{Z},$$

wobei der einfachste Filter natürlich $f = \delta$ ist, was lediglich das Signal in Stücke der Länge n hackt. Aber das kann jetzt zu richtigen Schwierigkeiten führen.

Beispiel 3.14 *Wir betrachten die Funktion* $\cos(64x)$*, abgetastet an den* $n = 768 = 3 * 256$ *Punkten aus* $2\pi/768 * \mathbb{Z}_n$ *und transformiert auf den* 6 *Intervallen der Länge* 128*. In Abb. 3.9 sieht man, daß es mit der Frequenzauflösung nun dahingeht.*

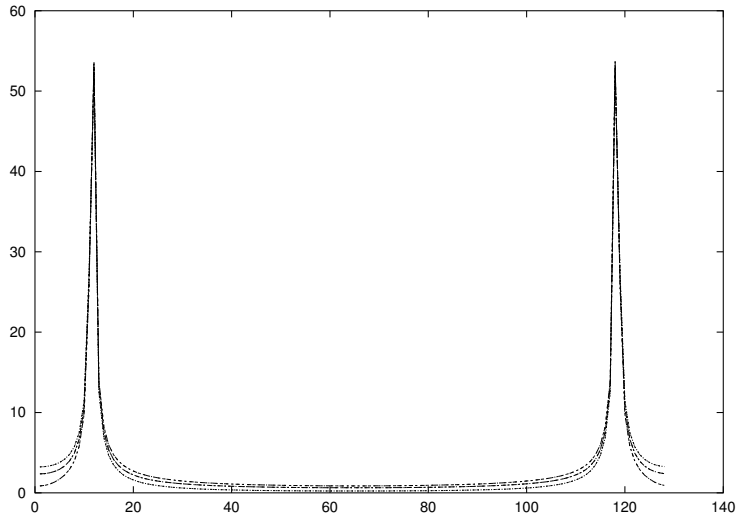


Abbildung 3.9: Absolutbeträge der Fouriertransformierten der Fenster der Länge 128. Man sieht deutlich, daß die Frequenzen "verschmiert" sind, obwohl ja eigentlich nur eine einzige Frequenz (und deren gerade Fortsetzung) auftauchen dürften. Außerdem variiert dieser Effekt mit dem jeweiligen Intervall – die Variationen sind nicht gerade dramatisch aber doch deutlich sichtbar.

Der in Abb. 3.9 dargestellte Effekt ist ein typisches Phänomen für die gefensterte Fouriertransformation: Die eigentlich scharf lokalisierten Frequenzen "laufen aus", wenn Abtastfrequenz und Fensterbreite nicht nicht mit der Periodizität der zugrundeliegenden Funktion kompatibel sind. Dieses "undichten" Frequenzen bezeichnet man auch als "Leck–Effekt", was vom englischen *"leakage phenenomenon"* stammt und nichts mit [20] zu tun hat. Aber sehen wir uns doch erst einmal an, wo dieser Effekt herstammt. Bei den "naiven" Fenstern berechnen wir ja

die Blöcke

$$\begin{aligned}
[c\,(\cdot+2kn)]_n^\wedge &= \left[\tau_{2kn}\left(c\times\sum_{j\in\mathbb{Z}_n}\tau_j\delta\right)\right]_n^\wedge = \underbrace{\omega^{2kn\cdot}}_{=1}\left(c\times\sum_{j\in\mathbb{Z}_n}\tau_j\delta\right)_n^\wedge \\
&= S_{2\pi/n}\left(c\times\sum_{j\in\mathbb{Z}_n}\tau_j\delta\right)^\wedge = S_{2\pi/n}\left[\widehat{c}\,*\left(\sum_{j\in\mathbb{Z}_n}e^{i\,j\cdot}\right)\right] \\
&= \left[S_{2\pi/n}\widehat{c}\right]*\left[\sum_{j\in\mathbb{Z}_n}e^{2\pi ij\cdot/n}\right],
\end{aligned}$$

und selbst wenn $\widehat{c}$ perfekt lokalisiert wäre, sorgt die Faltung mit den Exponentialfolgen für ein Verschmieren oder "Auslaufen" der Frequenzen.

Analog sieht man das Phänomen auch für die kontinuierliche Fouriertransformierte, denn da man für "hinreichend brave" f, g wegen

$$\begin{aligned}
&(fg)^\wedge(\xi) = 2\pi\ (fg)^\vee(-\xi) = 2\pi\ \left[(f^\vee)^\wedge\ (g^\vee)^\wedge\right]^\vee(-\xi) = 2\pi\ \left[(f^\vee * g^\vee)^\wedge\right]^\vee(-\xi) \\
&= \ 2\pi\ (f^\vee * g^\vee)(-\xi) = 2\pi\int_{\mathbb{R}} f^\vee(\theta)\, g^\vee(-\xi-\theta)\ d\theta \\
&= \ \int_{\mathbb{R}} f^\vee(-\theta)\ [2\pi g^\vee(-\xi+\theta)]\ d\theta = \int_{\mathbb{R}}\frac{1}{2\pi}\widehat{f}(\theta)\,\widehat{g}\,(\xi-\theta)\ d\theta = \frac{1}{2\pi}\,\widehat{f}*\widehat{g}(\xi),
\end{aligned}$$

auch das Produkt als Faltung darstellen kann, hat die "zentrierte Fensterung" an $x+[-h,h]$, also $F_{x,h} := f\,\chi_{x+[-h,h]}$ die Fouriertransformierte

$$F_{x,h}^\wedge(\xi) = \frac{1}{2\pi}\,\left(\widehat{f}*e^{-ix\xi}\,\widehat{\chi}_{[-h,h]}\right)(\xi)$$

und da

$$\widehat{\chi}_{[-h,h]}(\xi) = \int_{-h}^{h} e^{-i\xi t}\,dt = \frac{e^{-ih\xi}-e^{ih\xi}}{-i\xi} = 2h\,\mathrm{sinc}\,(\xi/\pi),$$

erhält man auch hier eine Faltung mit einer Funktion, die zwar abklingt, aber sehr wohl ihren Beitrag zur Frequenzverschmierung leistet. Was man auch sehr gut sieht ist, daß zu kleine Fenster nicht viel bringen – ist ja auch klar, denn ist f stetig oder zumindest global beschränkt[97], dann ist

$$\|F_{x,h}\|_1 \le h\,\|f\|_\infty\,, \qquad x\in\mathbb{R}, h>0,$$

was für $h\to 0$ natürlich gegen Null konvergiert, wobei sich die Fouriertransformierte gern anschließt.
Wir fassen zusammen:

> *Der "Leck–Effekt" ist eine Konsequenz der Multiplikation mit einer abschneidenden Fensterfunktion.*

[97] Was ja auch in unseren Annahmen an "brave" Funktionen steckt, siehe Übung 3.2.

Übung 3.2 Zeigen Sie: Ist $f \in L_1(\mathbb{R})$ die inverse Fouriertransformierte einer Funktion g, dann ist f gleichmäßig stetig und gleichmäßig beschränkt. ♢

Das ideale Fenster wäre damit also eine Funktion w mit $\widehat{w}(\xi) = 2\pi\,\delta$, wobei jetzt δ tatsächlich einmal die *Dirac–Distribution*[98] definiert durch

$$\int_{\mathbb{R}} f(x)\delta(x) := f(x), \qquad f \in C_{00}^{\infty}(\mathbb{R})$$

bezeichnet. Nur ist dann leider[99]

$$w(x) = 2\pi\,\delta^{\vee}(x) = \int_{\mathbb{R}} e^{i\theta x}\delta(\theta)\,d\theta = e^{i0} = 1,$$

und das ideale Fenster ist damit als "kein Fenster" identifiziert. Ein Kompromiß ist also unvermeidbar! Und jetzt schlägt auch noch die Heisenbergsche Unschärferelation, Satz 1.18, die uns nicht nur sagt, daß es ein ideales Fenster mit guter Zeit- und Frequenzlokalisierung gar nicht geben kann, sondern, daß

$$\operatorname{supp} w = [-S, S]$$

automatisch die "Verschmierung"

$$\operatorname{supp} \widehat{w} \supseteq \left[-\frac{1}{4\pi T}, \frac{1}{4\pi T}\right]$$

mit sich bringt – jedes im Zeitbereich wirksame Fenster **muß** im Frequenzbereich auslaufen. Alles, was man also tun kann, ist einen Kompromiß einzugehen, der neben der Zeitlokalisierung wenigstens eine bessere Frequenzlokalisierung als die sinc –Funktion liefert.

3.7 Fensterfunktionen

Das Problem mit der Restriktion auf einen Zeitbereich, also dem sogenannten *Rechtecksfenster*, besteht offenbar darin, daß die sehr gute Qualität der Einschränkung im Zeitbereich – die durch Multiplikation mit dem Rechtecksfenster erhaltene Funktion ist *exakt* die Einschränkung der Funktion auf den vorgegebenen Zeitbereich – mit einem Qualitätsverlust im Frequenzbereich erkauft werden muß, die Faltung mit der sinc –Funktion sorgt für relativ starke Verschmierungen.
In Abb. 3.10 sind das Rechtecksfenster und der Absolutbetrag seiner normierten Fouriertransformierten, berechnet mit dem Kommando "`fft( f,512 );`", dargestellt. Wie wir sehen werden, ist das Abklingverhalten im Frequenzbereich schlichtweg zu langsam, um die Lecks halbwegs in Grenzen zu halten.

Die Fenster, die wir uns nun ansehen wollen, sind allesamt auf einen diskreten Bereich $0, 1, \ldots, N$ normiert, werden also immer nur an den *ganzzahligen* Stellen ausgewertet, geplottet und Fourier–transformiert werden, obwohl es sich bei ihnen um kontinuierliche Funktionen handelt.

[98] Was Distributionen angeht, ist [65] wieder einmal eine gute Adresse, zum einfachen (und relativ billigen) Einstieg ist aber auch [16] gar nicht schlecht.

[99] Wenn man mal geflissentlich ignoriert, daß $e^{i\cdot}$ nicht zu $C_{00}^{\infty}(\mathbb{R})$ gehört, weil es am kompakten Träger mangelt.

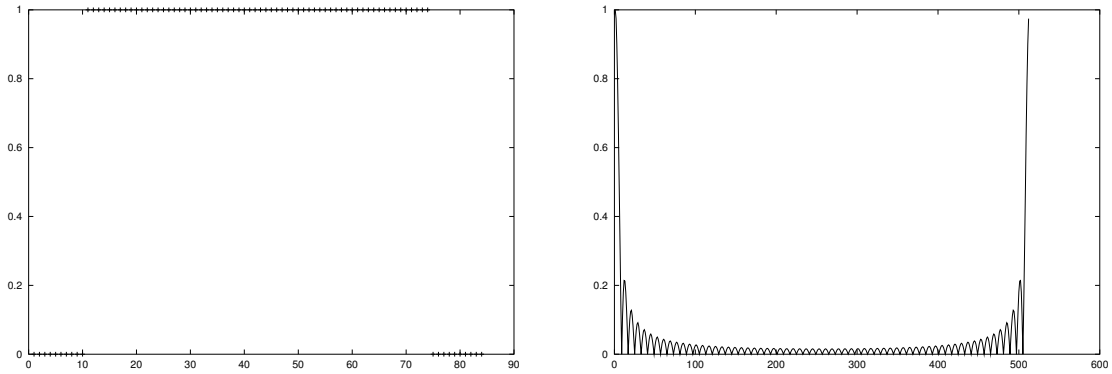


Abbildung 3.10: Das Rechtecksfenster der Länge 64 und seine der Absolutbetrag der Fouriertransformierten, also die sinc –Funktion.

Die vielleicht einfachste Wahl des Fensters besteht darin, vom Rechteck zum Dreieck überzugehen und die Fensterfuktion als

$$f_{Ba}(x) := \chi_{[0,N]} \ \min\left\{\frac{x}{2N}, 1 - \frac{x}{2N}\right\} \tag{3.10}$$

zu wählen. Dies bezeichnet man gerne auch als *Bartlett–Fenster*, siehe Abb. 3.11. Natürlich verschmiert dieses Fenster wesentlich stärker im Zeitbereich, ist dafür aber im Frequenzbereicht wesentlich angenehmer. Was die Abbildung allerdings nahelegt ist, daß die Fensterfunktion[100] zwar gleichmäßig, aber auch sehr flach ansteigt und deswegen vielleicht "steilflankigere" Fensterfunktionen besser wären.
Das *Hann–Fenster* in Abb. 3.12, oftmals auch *Hanning–Fenster* genannt[101], folgt genau dieser Idee und ist definiert als

$$f_{Hn}(x) = \chi_{[0,N]} \ \frac{1}{2}\left(1 - \cos\frac{2\pi}{N}x\right). \tag{3.11}$$

Durch die Kosinusfunktion ist es "breiter" und "steilflankiger" als das Dreiecksfenster. Eng damit verwandt ist das *Hamming–Fenster*

$$f_{Hm}(x) = \chi_{[0,N]} \ \left(0.54 - 0.46\,\cos\frac{2\pi}{N}x\right), \tag{3.12}$$

das eine etwas "breitere" Kosinusfunktion verwendet, dafür aber an den Enden 0 und N des "interessanten" Intervalls nicht einmal mehr stetig ist. Man kann sich vorstellen, daß man für

[100] Im Zeitbereich

[101] Laut [32, S. 182] ist das Fenster nach dem österreichischen Meteorologen Julius von Hann, 23.3. 1839 – 1.10.1921, benannt, der gab 1866 erstmals die richtige Begründung für Föhnwetter angab, suchte seine Ferienorte oftmals nach Häufigkeit der dort auftretenden Gewitter aus Wettervohersagen äußerst skeptisch gegenüberstand. In der englischsprachigen Literatur und insbesondere in `Matlab` und `Matlab` wird oft der Begriff *"Hanning–Fenster"* bzw. die Funktion `hanning` verwendet.

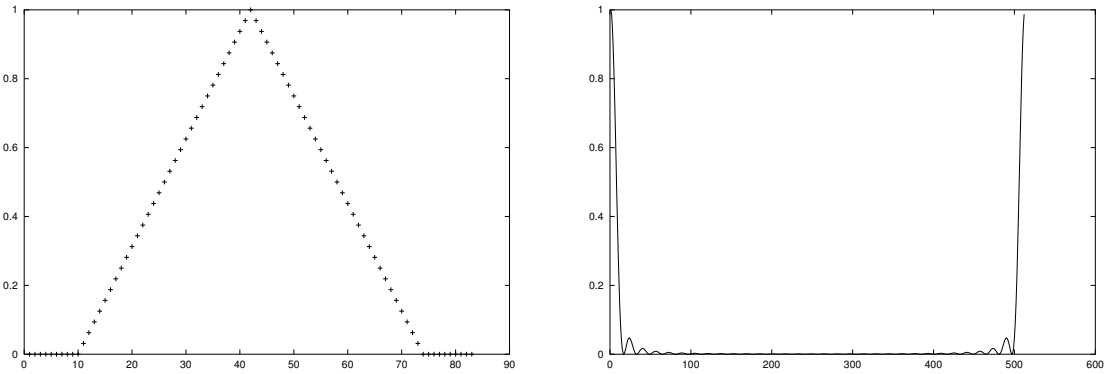


Abbildung 3.11: Das Dreiecks- oder Bartlett–Fenster und sein spektrales Verhalten. Wie man sieht, ist der Abfall “nach innen” wesentlich schneller.

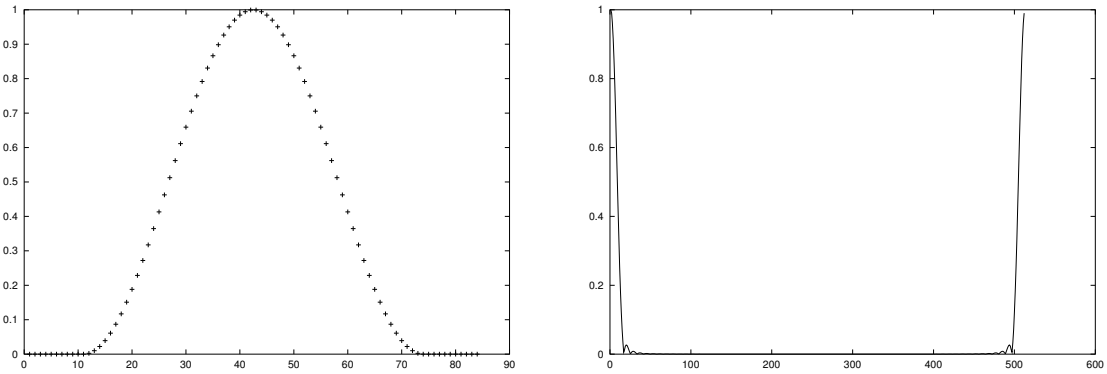


Abbildung 3.12: Das Hann–Fenster und seine Fouriertransformierte

dieses etwas schlechtere Verhalten im Zeitbereich tatsächlich im Frequenzbereich gewinnt. Um dies zu sehen, stellen wir die Spektra nicht in der normalen Skala dar, sondern in einer logarithmischen und zwar als

$$10 \log_{10} \frac{\left|\widehat{f}\,(2\pi \cdot /N)\right|^2}{\left|\widehat{f}(0)\right|^2}$$

in *Dezibel*[102], kurz *dB* an und vergleichen dort einmal die vier Frequenzgänge. Der Vorteil ist klar: In einer logarithmischen Skala wird der Abfall wesentlich deutlicher. Diese Plots sind in Abb. 3.13 dargestellt. Allerdings hat ja beispielsweise die Fouriertransformierte des Rechtecksfensters, also die sinc –Funktion, jede Menge von Nullstellen, weswegen der Logarithmus

[102] Trotz des einen “l” ist die Skala wohl nach Alexander Graham Bell benannt.

ziemlich schlecht aussehen würde. Daher betrachten wir die leicht modifizierte Funktion

$$10\log_{10}\left(t+\frac{\left|\widehat{f}\,(2\pi\cdot/N)\right|^2}{\left|\widehat{f}(0)\right|^2}\right),\qquad t>0. \tag{3.13}$$

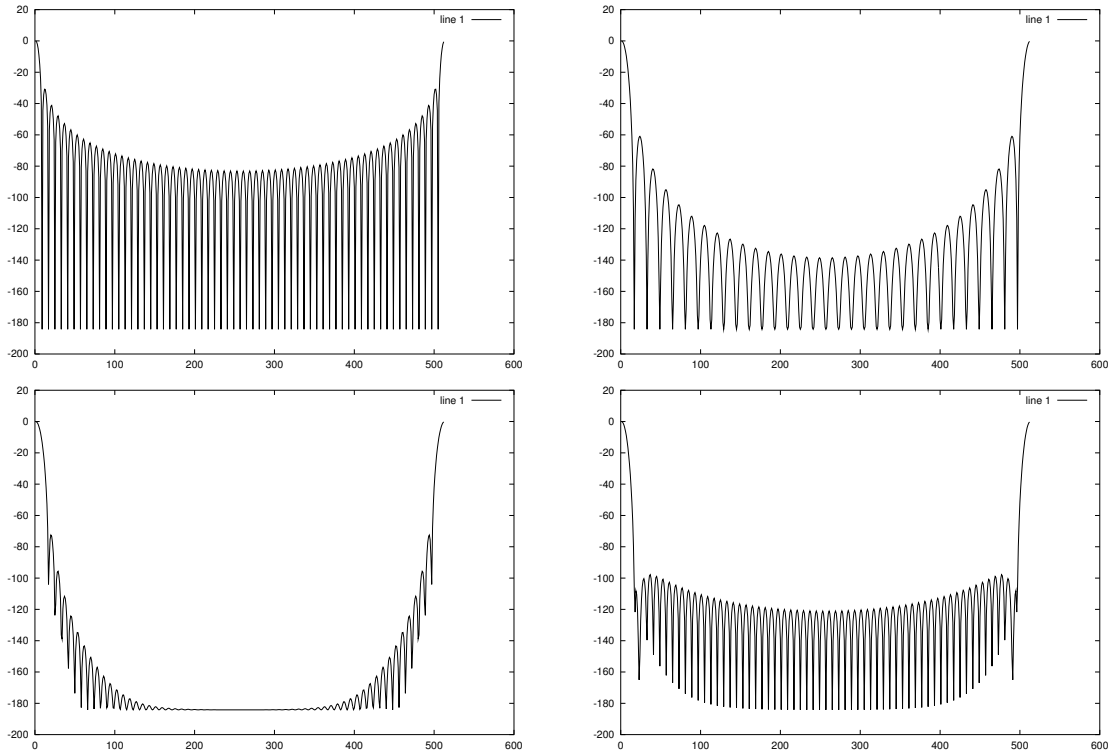


Abbildung 3.13: Die Spektra in dB, allerdings mit einem "Threshhold"–Faktor von $t = 10^{-8}$ in (3.13).

Besonders interessant ist offenbar der Unterschied zwischen Hann- und Hamming–Fenster: Während das Hann–Fenster im Inneren deutlich stärker abfällt[103], fehlt dem Hamming–Fenster das ausgeprägte erste innere Maximum mit Wert -80, sondern das Spektrum bleibt bei -100. Somit ist das Hamming–Fenster etwas vorteilhafter im niederfrequenten Bereich.
Hier sind schließlich noch ein paar weitere Fensterfunktionen:

Gauß–Fenster: Man verwendet die "Gauß–Glocke"

$$f_{Ga}(x)=\chi_{[0,N]}\,e^{-18(x/N-1/2)^2}, \tag{3.14}$$

siehe [38], wobei die Konstante 18 wohl eher willkürlich gewählt ist.

[103] Die Fouriertransformierte hat in Wirklichkeit nur *eine* Nullstelle, und zwar an $\pi/2$ bzw. der entsprechenden Stelle.

Blackman–Fenster: Diese Erweiterung des Hann- und Hamming–Fensters verwendet die Funktion

$$f_{Bl}(x) = \chi_{[0,N]} \left(0.42 - 0.5 \cos \frac{2\pi}{N} x + 0.08 \cos \frac{4\pi}{N} x \right), \tag{3.15}$$

fügt also einen höherfrequenten Cosinusterm hinzu.

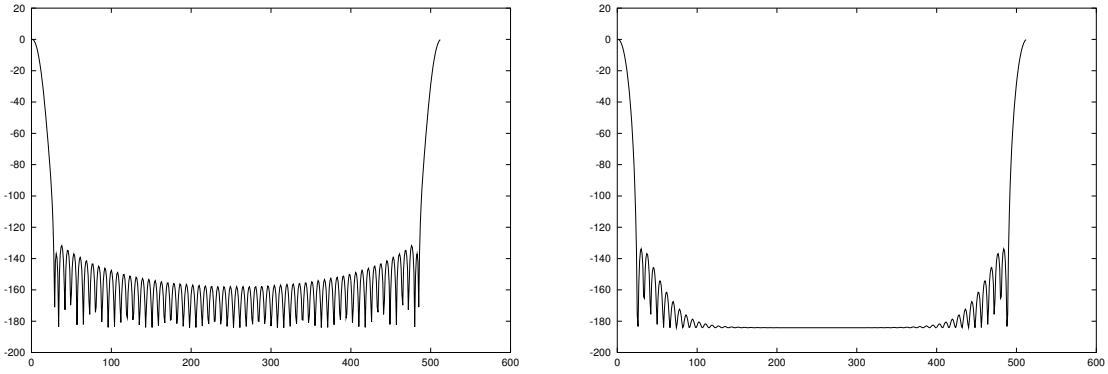


Abbildung 3.14: Die Spektren des Gauß- und des Blackman–Fensters, wieder in dB und wieder mit Hinzunahme des Wertes 10^{-8}.

Das ist aber noch lange nicht alles, es gibt noch eine ganze Reihe von weiteren Fensterfunktionen, beispielsweise das Kaiser–Fenster, siehe [32, S. 185], das auf *modifizierten Bessel–Funktionen erster Art* basiert – aber eine exakte Definition dieser Funktionen würde hier zu weit führen.

Es ist ein große Stärkung beim Studieren, wenigstens für mich, alles, was man liest, so deutlich zu fassen, daß man eigne Anwendungen davon oder gar Zusätze dazu machen kann. Man wird am Ende dann geneigt zu glauben, man habe alles selbst erfinden können, und so was macht Mut. So wie nichts mehr abschreckt als Gefühl von Superiorität im Buch.

G. Chr. Lichtenberg

4 Subband–Coding und Wavelets

In diesem Kapitel wollen wir uns schließlich noch mit einem eher "modernen" Aspekt der Signalverarbeitung befassen, nämlich mit *diskreten* Wavelettransformationen und den zugehörigen Filterungsprozessen. Dafür aber zuerst einmal ein vorbereitendes Kapitel über Filter, die das genaue Gegenteil eines Bandpass–Filters sind.

4.1 Allpass–Filter

Besonders einfache – oder soll man sagen "gutartige" – Filter sind diejenigen, die alle Frequenzen gleichbehandeln und diese mit Amplitude 1 wiedergeben. Anders gesagt, es sind Filter, die für alle Signale c die Eigenschaft

$$|\widehat{c}(\xi)| = \left|(Fc)^{\wedge}(\xi)\right| = \left|(f * c)^{\wedge}(\xi)\right| = \left|\widehat{f}(\xi)\right| \, |\widehat{c}(\xi)|$$

haben, deren Transferfunktion also

$$\left|\widehat{f}(\xi)\right| = 1 \tag{4.1}$$

erfüllt. Ein solcher Filter heißt *Allpass–Filter*, da er alle Frequenzen gleichmäßig passieren läßt. Wir wollen uns einmal ansehen, wie so ein Filter als *rationaler* Filter aussieht. Dazu zuerst eine kleine Bemerkung, die zeigt, daß eine rationale Funktion durch ihr Verhalten auf dem Rand $\mathbb{T} = \{z \in \mathbb{C} \, : \, |z| = 1\}$ des Einheitskreises eindeutig festgelegt wird.

Proposition 4.1 *Es seien $f = p/q$ und $g = r/s$ rationale Funktionen in z mit*

$$f(z) = g(z), \qquad z \in \mathbb{T}.$$

Dann ist $f(z) = g(z)$ für alle $z \in \mathbb{C}$.

Beweis: Seien zunächst $f, g \in \Pi[\mathbb{C}]$, also Polynome, das heißt, $q = s = 1$. Dann ist $f - g$ ein Polynom vom Grad, sagen wir einmal n, und damit durch sein Verhalten an $n + 1$ beliebigen Punkten festgelegt[104], die wir ohne weiteres auch auf den Einheitskreis legen können. Da dort aber f und g übereinstimmen, müssen sie überall übereinstimmen.
Für beliebige *rationale* Funktionen bemerken wir, daß

$$\frac{p(z)}{q(z)} = f(z) = g(z) = \frac{r(z)}{s(z)}, \quad z \in \mathbb{T}, \qquad \Longleftrightarrow \qquad p(z)\, s(z) = r(z)\, q(z), \qquad z \in \mathbb{T},$$

was nach unserer Aussage über Polynome aber wieder zu

$$p(z)\, s(z) = r(z)\, q(z), \qquad z \in \mathbb{C}, \qquad \Longleftrightarrow \qquad f(z) = \frac{p(z)}{q(z)} = \frac{r(z)}{s(z)} = g(z), \quad z \in \mathbb{C},$$

äquivalent ist. □

Als nächstes benötigen wir die *Spiegelung* eines Polynoms.

Definition 4.2 (Spiegelung) *Es sei $p \in \Pi_n$ ein Polynom mit $p^{(n)} \neq 0$, also ein Polynom vom Grad* exakt n. *Die* Spiegelung $p^\sharp$ *von p ist definiert als*

$$p^\sharp(z) = z^n\, \overline{p\,(\overline{z}^{-1})}, \qquad z \in C^\times. \tag{4.2}$$

Warum die Spiegelung "Spiegelung" heißt, wird klar, wenn man sich mal ansieht, was bei diesem Prozess passiert. Dazu sei $p(z) = p_n z^n + \cdots + p_0$, $p_n \neq 0$. Dann ist

$$\sum_{j=0}^{n} p_j^\sharp\, z^j := p^\sharp(z) = z^n \overline{\sum_{j=0}^{n} p_j\, \overline{z}^{-j}} = z^n \sum_{j=0}^{n} \overline{p_j}\, z^{-j} = \sum_{j=0}^{n} \overline{p_j}\, z^{n-j} = \sum_{j=0}^{n} \overline{p_{n-j}}\, z^j,$$

also $p_j^\sharp = \overline{p_{n-j}}$, $j = 0, \ldots, n$. Mit anderen Worten: Die Koeffizienten des Spiegelpolynoms sind die Koeffizienten des Ausgangspolynoms konjugiert komplex und in umgekehrter Reihenfolge gelesen. Schreibt man außerdem p in *faktorisierter* Form als $p(z) = p_n\,(z - \zeta_1) \cdots (z - \zeta_n)$, dann ist

$$p^\sharp(z) = z^n\, \overline{p_n} \prod_{j=1}^{n} \overline{(\overline{z}^{-1} - \zeta_j)} = z^n\, \overline{p_n} \prod_{j=1}^{n} \left(z^{-1} - \overline{\zeta_j}\right) = \overline{p_n} \prod_{j=1}^{n} \left(1 - \overline{\zeta_j}\, z\right). \tag{4.3}$$

Hat also p die Nullstellen ζ_j, dann hat $p^\sharp$ die am Einheitskreis gespiegelten Nullstellen $\overline{\zeta_j}^{-1}$.

Satz 4.3 (Charakterisierung von Allpass–Filtern) *Ein rationaler Filter F mit z–Transformierter $f^*(z) = \frac{p(z)}{q(z)}$ ist genau dann ein Allpass–Filter, wenn $q = p^\sharp$ ist.*

[104] Bekanntlich ist auch die Interpolation mit *komplexen* Polynomen vom Grad n an $n + 1$ *komplexen* Punkten eindeutig – der Beweis ist genau wie im Rellen, tatsächlich funktioniert er über *jedem* Körper; so kann man übrigens den *Chinesischen Restsatz* beweisen, siehe [45].

Korollar 4.4 *Es gibt genau einen FIR–Allpass–Filter, nämlich* $f^* = 1$.

Unter Verwendung von (4.3) erhalten wir dann eine weitere Charakterisierung von Allpass–Filtern in *faktorisierter* Form.

Korollar 4.5 *Ein rationaler Filter* F *ist genau dann ein Allpass–Filter, wenn*

$$f^*(z) = c \prod_{j=1}^{n} \frac{z - \zeta_j}{1 - \overline{\zeta_j}\, z}, \qquad |c| = 1, \tag{4.4}$$

wobei $c = p_n/\overline{p_n}$ *ist.*

Definition 4.6 (Blaschke–Produkt) *Eine Funktion der Form (4.4) bezeichnet man als* Blaschke–Produkt *der Ordnung* n.

Beweis von Satz 4.3: Daß F ein Allpass–Filter ist liefert, durch Quadrierung von (4.1), daß

$$1 = \left|\widehat{f}(\xi)\right|^2 = \left|f^*\left(e^{i\xi}\right)\right|^2, \qquad \xi \in \mathbb{R},$$

und somit

$$1 = |f^*(z)|^2 = \frac{|p(z)|^2}{|q(z)|^2}, \qquad z \in \mathbb{T},$$

also

$$p(z)\,\overline{p(z)} = |p(z)|^2 = |q(z)|^2 = q(z)\,\overline{q(z)}, \tag{4.5}$$

insbesondere ist also $\deg p = \deg q =: n$. Da für $z \in \mathbb{T}$, also $z = e^{i\theta}$, $\theta \in [-\pi, \pi]$ die Identität $\overline{z} = e^{-i\theta} = 1/z$ gilt, ist

$$\overline{p(z)} = \sum_{j=0}^{n} \overline{p_j}\,\overline{z}^j = \sum_{j=0}^{n} \overline{p_j} z^{-j} = z^{-n} p^\sharp(z).$$

Setzen wir das in (4.5) ein und multplizieren wir die resultierende Gleichung mit z^n, dann ist

$$p(z)\, p^\sharp(z) = q(z)\, q^\sharp(z), \qquad z \in \mathbb{T},$$

und da wir p und q als teilerfremd annehmen dürfen[105], muß auf $\mathbb{T}$ die Gleichheit $p = q^\sharp$ und $q = p^\sharp$ gelten, was wir nach Proposition 4.1 auf ganz $\mathbb{C}$ fortsetzen können.
Für die Umkehrung brauchen wie nur zu beachten, daß

$$\frac{|p(z)|^2}{|q(z)|^2} = \frac{p(z)\ z^{-n}\, p^\sharp(z)}{q(z)\ z^{-n}\, q^\sharp(z)} = \frac{p(z)}{q^\sharp(z)}\,\frac{p^\sharp(z)}{q(z)}$$

ist. □

[105] Ansonsten ist die rationale Funktion nicht wirklich sinnvoll.

4.2 Upsampling, Downsampling und Filterbänke

Die Idee des *Subband Coding*, das insbesondere bei der Datenkompression gerne verwendet wird, besteht darin, ein Signal in mehrere Teilsignale zu zerlegen und jedes dieser Teilsignale separat zu codieren – am besten natürlich so, daß sich die Subbänder nach Möglichkeit ergänzen.

Beispiel 4.7 *Die naheliegendste Idee wäre natürlich, für die Subband–Zerlegung verschiedene Bandpass–Filter*

$$\chi^{\vee}_{[t_j,t_{j+1}]}, \qquad 0 = t_0 < t_1 < \cdots < t_{n-1} < t_n = 2\pi$$

zu verwenden und so das vollständige Frequenzband mittels exakter Bandpass–Filter aufzuspalten und diese Teilsignale weiterzuverarbeiten[106]*. Dieser Ansatz erfreut zwar mit Sicherheit den Phono–Freak, ist aber ziemlich aufwendig. Daher wollen wir uns eine einfachere Zerlegung ansehen, die sehr leicht zu realisieren ist und uns obendrein zu Wavelets führen wird.*

Definition 4.8 *Sei $n \in \mathbb{N}$; um Trivialitäten auszuschließen, nehmen wir außerdem an, daß $n \geq 2$ ist. Der Operator $\downarrow_n$, der $c \in \ell(\mathbb{Z})$ das Signal*

$$\downarrow_n c = c(n\cdot)$$

zuordnet, heißt Downsampling–Operator *der Ordnung n, der Operator $\uparrow_n$ mit*

$$\uparrow_n c(j) = \begin{cases} c(j/n), & j \in n\mathbb{Z}, \\ 0, & j \in \mathbb{Z} \setminus n\mathbb{Z}, \end{cases}$$

heißt Upsampling–Operator *der Ordnung n.*

Offensichtlich gilt $\downarrow_n \uparrow_n = I \neq \uparrow_n \downarrow_n$. Mit Hilfe des Downsampling–Operators können wir nun ein Signal c in die "Teilbänder"

$$c_j = \downarrow_n \tau_j c, \qquad j \in \mathbb{Z}_n,$$

zerlegen, die man über die Formel

$$c = \sum_{j \in \mathbb{Z}_n} \tau_j \uparrow_n c_j$$

wieder zu c kombinieren kann. Der Zerlegungsprozess liefert hierbei

$$\begin{array}{ccccccccccc} \ldots & c(-1) & \boxed{c(0)} & c(1) & \ldots & c(n-1) & \boxed{c(n)} & c(n+1) & \ldots & \to c_0 \\ \ldots & c(0) & c(1) & c(2) & \ldots & c(n) & c(n+1) & c(n+2) & \ldots & \to c_1 \\ & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \\ \ldots & c(n-2) & c(n-1) & c(n) & \ldots & c(2n-2) & c(2n-1) & c(2n) & \ldots & \to c_{n-1} \end{array}$$

so daß

$$c_j = c\,(n\cdot + j)\,, \qquad j \in \mathbb{Z}_n,$$

[106] Das ist das, was ein "idealer" Equalizer in einer Stereoanlage wohl machen würde.

ist, was sich mit Upsampling und Translation zu

$$\begin{array}{rccccccccl} c_0 \to & 0 & c(0) & 0 & \dots & 0 & c(n) & 0 & \dots & \\ c_1 \to & 0 & 0 & c(1) & \dots & 0 & 0 & c(n+1) & \dots & \\ & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \\ c_{n-1} \to & c(-1) & 0 & 0 & \dots & c(n-1) & 0 & 0 & \dots & \\ \hline & c(-1) & c(0) & c(1) & \dots & c(n-1) & c(n) & c(n+1) & \dots & \to c \end{array}$$

kombiniert wie behauptet. Beim "Subband Coding" filtern wir nun zuerst die Daten mit n Filtern F_j, $j \in \mathbb{Z}_n$, und wenden dann auf jedes der Resultate den Downsampling–Operator an, also

$$c_j = \downarrow_n \; F_j \, c, \qquad j \in \mathbb{Z}_n. \tag{4.6}$$

Als schematisches Bild sieht das dann wie folgt aus:

$$\begin{array}{rcccccc} & & \nearrow & \boxed{F_0} & \to & \boxed{\downarrow_n} & \to \; c_0 \\ c \; \to & \odot & & \vdots & & \vdots & \vdots \\ & & \searrow & \boxed{F_{n-1}} & \to & \boxed{\downarrow_n} & \to \; c_{n-1} \end{array} \tag{4.7}$$

Die Zerlegung aus (4.7) bezeichnet man als *Analyse–Filterbank* $F = [F_j \; : \; j \in \mathbb{Z}_n]$. Durch das Downsampling enthält jede Komponente c_j von c nur den n–ten Teil der Information des Signals $F_j c$, also auch nur den n–ten Teil der Information von c, vorausgesetzt, die Filter sind vernünftig.

Wir kennen ja bereits das einfachste Beispiel einer Filterbank, nämlich den Fall, daß $F_j = \tau_j$, $j \in \mathbb{Z}_n$, ist, dann entspricht die Filterbank tatsächlich der Zerlegung des Signals modulo n.

Als nächstes wollen wir eine weitere Beschreibung der Filterbank geben, wobei wir wieder annehmen, daß jeder Filter F_j ein LTI– Filter mit Impulsantwort f_j ist, daß also $F_j c = f_j * c$ gilt. Das führt zum folgenden Begriff.

Definition 4.9 (Modulationsmatrix) *Die* Modulationsmatrix $M(z)$ *zu der Filterbank*

$$F = [F_j \; : \; j \in \mathbb{Z}_n]$$

ist definiert als

$$\begin{aligned} M(z) \; &:= \; \frac{1}{n} \left[f_j^* \left(e^{2\pi i k/n} \, z \right) \; : \; j,k \in \mathbb{Z}_n \right] \\ &= \; \frac{1}{n} \begin{bmatrix} f_0^*(z) & f_0^* \left(e^{2\pi i/n} z \right) & \dots & f_0^* \left(e^{2\pi i (n-1)/n} z \right) \\ \vdots & \vdots & \ddots & \vdots \\ f_{n-1}^*(z) & f_{n-1}^* \left(e^{2\pi i/n} z \right) & \dots & f_{n-1}^* \left(e^{2\pi i (n-1)/n} z \right) \end{bmatrix}. \end{aligned} \tag{4.8}$$

Eine besonders einfache Gestalt hat die Modulationsmatrix wieder einmal für $n = 2$, wo man

$$M(z) = \begin{bmatrix} f_0^*(z) & f_0^*(-z) \\ f_1^*(z) & f_1^*(-z) \end{bmatrix}$$

erhält. Die Bedeutung der Modulationsmatrix liegt aber darin, daß sie die Aktion der Filterbank beschreibt.

Satz 4.10 *Für die Filterbank aus (4.7) gilt*

$$\left[c_j^*\left(z^n\right)\ :\ j\in\mathbb{Z}_n\right]=M(z)\ \left[c^*\left(e^{2\pi ij/n}z\right)\ :\ j\in\mathbb{Z}_n\right],\tag{4.9}$$

das heißt,

$$\begin{bmatrix} c_0^*\left(z^n\right)\\ \vdots\\ c_{n-1}^*\left(z^n\right)\end{bmatrix}=M(z)\begin{bmatrix} c^*(z)\\ c^*\left(e^{2\pi i/n}z\right)\\ \vdots\\ c^*\left(e^{2\pi i(n-1)/n}z\right)\end{bmatrix}\tag{4.10}$$

Übung 4.1 Bestimmen Sie die Modulationsmatrix für die Translationsfilter $F_j=\tau_j$, $j\in\mathbb{Z}_n$. ♢

Im Fall $n=2$ wird (4.10) dann zu

$$\begin{bmatrix} c_0^*\left(z^2\right)\\ c_1^*\left(z^2\right)\end{bmatrix}=\begin{bmatrix} f_0^*(z) & f_0^*(-z)\\ f_1^*(z) & f_1^*(-z)\end{bmatrix}\begin{bmatrix} c^*(z)\\ c^*(-z)\end{bmatrix}.$$

Jetzt aber an den Beweis von Satz 4.10. Dazu benötigen wir zuerst etwas Information, wie sich die Up- und Downsampling–Operatoren auf ein Signal auswirken.

Lemma 4.11 *Für $n\in\mathbb{N}$ ist*

$$\left(\downarrow_n c\right)^*\left(z^n\right)=\frac{1}{n}\sum_{k\in\mathbb{Z}_n}c^*\left(e^{2\pi ik/n}z\right),\qquad\left(\uparrow_n c\right)^*(z)=c^*\left(z^n\right),\tag{4.11}$$

Beweis: Da

$$\frac{1}{n}\sum_{k\in\mathbb{Z}_n}e^{-2\pi ijk/n}=\begin{cases}1, & j\in n\mathbb{Z},\\ 0, & j\notin n\mathbb{Z},\end{cases}$$

siehe Lemma 3.5, ergibt sich die linke Indentität in (4.11) aus

$$\begin{aligned}\left(\downarrow_n c\right)^*\left(z^n\right) &= \sum_{j\in\mathbb{Z}}\left(\downarrow_n c\right)(j)\,z^{-nj}=\sum_{j\in\mathbb{Z}}c(nj)\,z^{-nj}=\sum_{j\in\mathbb{Z}}c(j)\,z^{-j}\left[\frac{1}{n}\sum_{k\in\mathbb{Z}_n}e^{-2\pi ijk/n}\right]\\ &= \frac{1}{n}\sum_{j\in\mathbb{Z}}c(j)\sum_{k\in\mathbb{Z}_n}\left(e^{2\pi ik/n}z\right)^{-j}=\frac{1}{n}\sum_{k\in\mathbb{Z}_n}c^*\left(e^{2\pi ik/n}z\right),\end{aligned}$$

Die rechte Seite von (4.11) erhält man hingegen aus der einfachen Rechnung

$$\left(\uparrow_n c\right)^*(z)=\sum_{j\in\mathbb{Z}}c(j)\,z^{-nj}=c^*\left(z^n\right).$$

□

Besonders einfach wird (4.11) natürlich wieder im Fall $n=2$. Da $e^{i\pi}=-1$ ist, ergibt sich dann nämlich

$$\left(\downarrow_2 c\right)^*\left(z^2\right)=\frac{1}{2}\left(c^*(z)+c^*(-z)\right),\qquad\left(\uparrow_2 c\right)^*(z)=c^*\left(z^2\right).\tag{4.12}$$

Es gibt aber noch eine andere Interpretation von (4.11). Dazu schreiben wir $c^*(z)$ als

$$\begin{aligned} c^*(z) &= \sum_{j\in\mathbb{Z}} c(j)z^{-j} = \sum_{k\in\mathbb{Z}_n}\sum_{j\in\mathbb{Z}} c(nj+k)\,z^{-nj-k} = \sum_{k\in\mathbb{Z}_n} z^{-k}\sum_{j\in\mathbb{Z}} \left(\uparrow_n \tau_k c\right)(j)\, z^{-nj} \\ &= \sum_{k\in\mathbb{Z}_n} z^{-k}\ \left(\uparrow_n \tau_k c\right)^*(z^n) =: \sum_{k\in\mathbb{Z}_n} z^{-k}\,\widetilde{c}_k^*\,(z^n)\,, \end{aligned}$$

wobei $\widetilde{c}_k$ dasjenige Signal ist, das wir durch die einfachste Filterbank erhalten, die nur Anteile modulo n bestimmt. Setzen wir jetzt $z = e^{2\pi ij/n}z$ ein, dann ist

$$c^*\left(e^{2\pi ij/n}z\right) = \sum_{k\in\mathbb{Z}_n} e^{-2\pi ijk/n} z^{-k}\,\widetilde{c}_k^*\,(z^n)$$

und somit

$$\begin{aligned} \frac{1}{n}\sum_{j\in\mathbb{Z}_n} c^*\left(e^{2\pi ij/n}z\right) &= \frac{1}{n}\sum_{j\in\mathbb{Z}_n}\sum_{k\in\mathbb{Z}_n} e^{-2\pi ijk/n} z^{-k}\,\widetilde{c}_k^*\,(z^n) \\ &= \sum_{k\in\mathbb{Z}_n} \underbrace{\left(\frac{1}{n}\sum_{j\in\mathbb{Z}_n} e^{-2\pi ijk/n}\right)}_{=\delta_{0k}} z^{-k}\,\widetilde{c}_k^*\,(z^n) = \widetilde{c}_0^*\,(z^n)\,. \end{aligned}$$

Multipliziert man nun mit $e^{2\pi ik/n}$, dann erhält man, daß

$$z^{-k}\,\widetilde{c}_k^*\,(z^n) = e^{2\pi ik/n}\,\frac{1}{n}\sum_{j\in\mathbb{Z}_n} c^*\left(e^{2\pi ij/n}z\right), \qquad k\in\mathbb{Z}_n. \tag{4.13}$$

Übung 4.2 Wie sieht (4.13) im Fall $n = 2$ aus? ♢

Beweis von Satz 4.10: Da $c_j = F_j c = f_j * c$ gesetzt wurde, erhalten wir für $j\in\mathbb{Z}_n$

$$\begin{aligned} c_j^*\,(z^n) &= \left[\downarrow_n (f_j * c)\right]^*(z^n) = \frac{1}{n}\sum_{k\in\mathbb{Z}_n} (f_j * c)^*\left(e^{2\pi ik/n}z\right) \\ &= \frac{1}{n}\sum_{k\in\mathbb{Z}_n} f_j^*\left(e^{2\pi ik/n}z\right)\ c^*\left(e^{2\pi ik/n}z\right). \end{aligned}$$

Daraus folgt (4.10) sofort durch Übergang zur Matrix–Vektor–Schreibweise. □

Da es ja keinen Unterschied macht, ob man die Sachen auf ganz $\mathbb{C}$ oder nur auf dem Einheitskreis betrachtet, können wir die Modulationsmatrix auch für die Fouriertransformierte der Signale betrachten. Und in der Tat: Setzen wir $z = e^{i\xi/n}$ in (4.10) ein, dann folgt, daß

$$\begin{bmatrix} \widehat{c}_0(\xi) \\ \vdots \\ \widehat{c}_{n-1}(\xi) \end{bmatrix} = \begin{bmatrix} c_0^*\,(z^n) \\ \vdots \\ c_{n-1}^*\,(z^n) \end{bmatrix} = M(z)\ \left[c^*\left(e^{2\pi ij/n}\,z\right)\ :\ j\in\mathbb{Z}_n\right]$$

$$
\begin{aligned}
&= M\left(e^{i\xi/n}\right)\left[c^*\left(e^{2\pi ij/n}\,e^{i\xi/n}\right)\ :\ j\in\mathbb{Z}_n\right] = M\left(e^{i\xi/n}\right)\left[c^*\left(e^{i(\xi+2j\pi)/n}\right)\ :\ j\in\mathbb{Z}_n\right]\\
&= \begin{bmatrix} f_0^*\left(e^{i\xi/n}\right) & f_0^*\left(e^{i(\xi+2\pi)/n}\right) & \dots & f_0^*\left(e^{i(\xi+2(n-1)\pi)/n}\right)\\ \vdots & \vdots & \ddots & \vdots\\ f_{n-1}^*\left(e^{i\xi/n}\right) & f_{n-1}^*\left(e^{i(\xi+2\pi)/n}\right) & \dots & f_{n-1}^*\left(e^{i(\xi+2(n-1)\pi)/n}\right)\end{bmatrix}\begin{bmatrix} c^*\left(e^{i\xi/n}\right)\\ c^*\left(e^{i(\xi+2\pi)/n}\right)\\ \vdots\\ c^*\left(e^{i(\xi+2(n-1)\pi)/n}\right)\end{bmatrix}\\
&= \begin{bmatrix} \widehat{f}_0\left(\frac{\xi}{n}\right) & \widehat{f}_0\left(\frac{\xi}{n}+2\pi\frac{1}{n}\right) & \dots & \widehat{f}_0\left(\frac{\xi}{n}+2\pi\frac{n-1}{n}\right)\\ \vdots & \vdots & \ddots & \vdots\\ \widehat{f}_{n-1}\left(\frac{\xi}{n}\right) & \widehat{f}_{n-1}\left(\frac{\xi}{n}+2\pi\frac{1}{n}\right) & \dots & \widehat{f}_{n-1}\left(\frac{\xi}{n}+2\pi\frac{n-1}{n}\right)\end{bmatrix}\begin{bmatrix} \widehat{c}\left(\frac{\xi}{n}\right)\\ \widehat{c}\left(\frac{\xi}{n}+2\pi\frac{1}{n}\right)\\ \vdots\\ \widehat{c}\left(\frac{\xi}{n}+2\pi\frac{n-1}{n}\right)\end{bmatrix},
\end{aligned}
$$

also

$$
\left[\widehat{c}_j(\xi)\ :\ j\in\mathbb{Z}_n\right] = \left[\widehat{f}_j\left(\frac{\xi+2k\pi}{n}\right)\ :\ j,k\in\mathbb{Z}_n\right]\ \left[\widehat{c}\left(\frac{\xi+2j\pi}{n}\right)\ :\ j\in\mathbb{Z}_n\right] \tag{4.14}
$$

Die Matrix

$$
\left[\widehat{f}_j\left(\frac{\xi+2\pi k}{n}\right)\ :\ j,k\in\mathbb{Z}_n\right]
$$

bezeichnet man übrigens als *Polyphase Matrix* zur Filterbank. Nun sollte natürlich auch wieder zusammenwachsen, was wir gerade eben so mühsam getrennt haben – wir möchten also den Analyseprozess umkehren und aus den Subband–Daten c_j, $j\in\mathbb{Z}_n$, wieder *ein* Signal c basteln[107]. Dazu gehen wir genau den umgekehrten Weg, indem wir erst ein Upsampling der Daten durchführen, dann diese Daten mit Filtern G_j, $j\in\mathbb{Z}_n$, filtern und schließlich die Ergebnisse aufsummieren, also

$$
c' = \sum_{j\in\mathbb{Z}_n} G_j\ \uparrow_n c_j, \tag{4.15}
$$

bzw. die *Synthese–Filterbank*

$$
\begin{array}{ccccccc}
c_0 & \to & \boxed{\uparrow_n} & \to & \boxed{G_0} & \searrow & \\
\vdots & & \vdots & & \vdots & & \oplus\ \to c\\
c_{n-1} & \to & \boxed{\uparrow_n} & \to & \boxed{G_{n-1}} & \nearrow &
\end{array} \tag{4.16}
$$

verwenden. Damit ist also

$$
c^*(z) = \sum_{j\in\mathbb{Z}_n}\left(G_j\ \uparrow_n c_j\right)^*(z) = \sum_{j\in\mathbb{Z}_n} g_j^*(z)\left(\uparrow_n c_j\right)^*(z) = \sum_{j\in\mathbb{Z}_n} g_j^*(z)\,c_j^*\left(z^n\right). \tag{4.17}
$$

So weit, so gut, so einfach. Um aber zurück zur Modulationsmatrix zu kommen, erinnern wir uns daran, daß die Eingangsdaten für diese aus dem Vektor $\left[c^*\left(e^{2\pi ij/n}z\right)\ :\ j\in\mathbb{Z}_n\right]$ bestanden

[107] Wobei wir nicht unbedingt annehmen, daß die c_j jemals durch eine Analyse–Filterbank aus diesem c enstanden sein sollen.

haben, und den erhalten wir, indem wir z in (4.17) durch $e^{2\pi ij/n}z$, $j \in \mathbb{Z}_n$, ersetzen, berücksichtigen, daß $\left(e^{2\pi ij/n}z\right)^n = z^n$ ist, und schließlich das Ganze wieder in Matrix–Vektor–Form als

$$\begin{bmatrix} c^*(z) \\ \vdots \\ c^*\left(e^{2\pi i(n-1)/n}z\right) \end{bmatrix} = \underbrace{\begin{bmatrix} g_0^*(z) & \dots & g_{n-1}^*(z) \\ \vdots & \ddots & \vdots \\ g_0^*\left(e^{2\pi i(n-1)/n}z\right) & \dots & g_{n-1}^*\left(e^{2\pi i(n-1)/n}z\right) \end{bmatrix}}_{=:\widetilde{M}(z)} \begin{bmatrix} c_0^*(z^n) \\ \vdots \\ c_{n-1}^*(z^n) \end{bmatrix} \tag{4.18}$$

bzw.

$$\left[c^*\left(e^{2\pi ij/n}z\right) \;:\; j \in \mathbb{Z}_n\right] = \widetilde{M}(z)\,\left[c_j^*(z^n) \;:\; j \in \mathbb{Z}_n\right] \tag{4.19}$$

schreiben. Was wir natürlich betrachten wollen, ist nicht das Analyse- oder das Synthesesystem allein, sondern das Gesamtsystem, die *Filterbank*

$$c \;\to\; \odot \begin{array}{ccccccccc} \nearrow & \boxed{F_0} & \to & \boxed{\downarrow_n} & \to & c_0 & \to & \boxed{\uparrow_n} & \to & \boxed{G_0} & \searrow \\ & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & \\ \searrow & \boxed{F_{n-1}} & \to & \boxed{\downarrow_n} & \to & c_{n-1} & \to & \boxed{\uparrow_n} & \to & \boxed{G_{n-1}} & \nearrow \end{array} \oplus \;\to c' \tag{4.20}$$

und eine natürliche "Minimalforderung" ist sicherlich, daß bei dem System das, was man vorne reinsteckt auch hinten wieder rauskommt.

Definition 4.12 (Perfect Reconstruction) *Die Filterbank (F, G) hat die Fähigkeit zur* perfekten Rekonstruktion[108]*, wenn in (4.20) $c' = c$ für alle Eingabedaten c gilt.*

Bemerkung 4.13

1. *Unter Verwendung der Modulationsmatrizen kann man die perfekte Rekonstruktion besonders nett beschreiben. Setzt man nämlich (4.9) in (4.19) ein, so erhält man, daß perfekte Rekonstruktion äquivalent zu*

$$\left[c^*\left(e^{2\pi ij/n}z\right) \;:\; j \in \mathbb{Z}_n\right] = \widetilde{M}(z)\,M(z)\,\left[c^*\left(e^{2\pi ij/n}z\right) \;:\; j \in \mathbb{Z}_n\right] \tag{4.21}$$

   ist. Eine Filterbank erlaubt also mit Sicherheit perfekte Rekonstruktion, wenn

$$\widetilde{M}(z)\,M(z) = I$$

   ist.

2. *Manchmal ist man etwas großzügiger und erlaubt, daß die Filterbank zwar die Eingabedaten rekonstruiert, erlaubt aber* Zeitverzögerungen*, d.h., $c' = \tau_j c$ für ein $j \in \mathbb{Z}$. Nachdem*

$$(\tau_k c)^*(z) = z^{-k}\,c^*(z)$$

[108] In Englisch *"perfect reconstruction (property)"*.

ist, ergibt sich somit, daß in diesem Fall

$$\left[e^{-2\pi ijk/n} z^{-k} c^*\left(e^{2\pi ij/n} z\right) \ : \ j \in \mathbb{Z}_n\right] = \widetilde{M}(z)\, M(z)\ \left[c^*\left(e^{2\pi ij/n} z\right) \ : \ j \in \mathbb{Z}_n\right]$$

was für

$$\widetilde{M}(z)\, M(z) = \text{diag}\ \left[e^{-2\pi ijk/n} z^{-k} \ : \ j \in \mathbb{Z}_n\right]$$

erfüllt ist.

3. *Man sieht leicht, daß*

$$\widetilde{M}(z)\, M(z) = \left[\sum_{\ell\in\mathbb{Z}} g_\ell^*\left(e^{2\pi ij/n}\, z\right) f_\ell^*\left(e^{2\pi ik/n}\, z\right) \ : \ j,k \in \mathbb{Z}_n\right]$$

ist.

Es ist klar, daß die perfekte Rekonstruktion folgt, wenn $\widetilde{M}\, M = I$ ist – man muß das ja nur in (4.21) einsetzen. Was nicht ganz so offensichtlich ist, ist die Tatsache, daß auch die Umkehrung gilt.

Satz 4.14 (Perfect Reconstruction) *Eine rationale Filterbank*[109] (F, G) *besitzt die Fähigkeit zur perfekten Rekonstruktion genau dann, wenn* $\widetilde{M}(z)\, M(z) = I$, $z \in \mathbb{C}$.

Beweis: Die Richtung $\Leftarrow$ ist klar. Für die Umkehrung betrachten wir die äquivalente Form

$$0 = \left(I - \widetilde{M}(z)\, M(z)\right)\ \left[c^*\left(e^{2\pi ij/n} z\right) \ : \ j \in \mathbb{Z}_n\right]$$

von (4.21) und setzen $c = \tau_k \delta$, $k \in \mathbb{Z}_n$. Da in diesem Fall $c^*(z) = z^k$, also

$$\left[c^*\left(e^{2\pi ij/n} z\right) \ : \ j \in \mathbb{Z}_n\right] = z^k\ \left[e^{2\pi ijk/n} \ : \ j \in \mathbb{Z}_n\right]$$

ist und da $z^k \neq 0$ auf $\mathbb{T}$ ist, erhalten wir, daß

$$0 = \left(I - \widetilde{M}(z)\, M(z)\right)\ \left[e^{2\pi ijk/n} \ : \ j \in \mathbb{Z}_n\right].$$

Nun sind aber die Vektoren $\left[e^{2\pi ijk/n} \ : \ j \in \mathbb{Z}_n\right]$, $k \in \mathbb{Z}_n$, *linear unabhängig*, da sie zusammen die Inverse der DFT[110] aufspannen. Und daher muß $I - \widetilde{M}(z)\, M(z)$ für alle $z \in \mathbb{T}$, also auch für alle $z \in \mathbb{C}$ gleich Null sein. □

Mit anderen Worten: Man kann ein Analysesystem genau dann wieder perfekt rekonstruieren, wenn die entsprechende Modulationsmatrix invertierbar ist. Nur stellt sich jetzt die Frage, in welchem Ring $M(z)$ invertierbar ist, denn das führt zu unterschiedlichen Voraussetzungen an $\det M(z)$.

[109] Also eine Filterbank, bei der alle beteiligten Filter rationale Filter sind.

[110] Bis auf den Normalisierungsfaktor $\frac{1}{n}$ natürlich.

1. Invertierbarkeit in $\Pi^{n\times n}$: Es muß für alle $z \in \mathbb{C}$ die Beziehung $\det M(z) = c$, $c \in \mathbb{C}^\times$, gelten, die Determinante muß also eine Konstante sein.

2. Invertierbarkeit in $\Lambda^{n\times n}$: Hier genügt es, wenn es $k \in \mathbb{Z}$ und $c \in \mathbb{C}^\times$ gibt, so daß $\det M(z) = c\, z^k$, $z \in \mathbb{C}$, ist.

3. Invertierbarkeit in $(\Pi/\Pi)^{n\times n}$, dem Raum der rationalen Funktionen: Hier reicht $\det M(z) \not\equiv 0$.

Das ist eine allgemeine algebraische Tatsache: Eine Matrix $A \in R^{n\times n}$, wobei R ein beliebiger Ring ist, ist genau dann invertierbar, wenn $\det A \in R^\times$ ist. Solche Matrizen heißen *unimodular*.

Und schließlich noch eine Klasse von besonders "einfachen" Filterbänken, nämlich die, bei denen die einzelnen Kanäle der Filterbank sauber voneinander getrennt sind – eine ideale Kombination von Hoch- und Tiefpassfilter hätte diese Eigenschaft. Die Forderung ist, daß der Syntheseanteil $\uparrow_n G_j$ nur den dazugehörigen Signalanteil c_j durchläßt, aber alle anderen c_k, $k \in \mathbb{Z}_n \setminus \{j\}$, vollständig *sperrt*. Das bedeutet also[111], daß

$$0 = (G_j \uparrow_n \downarrow_n F_k)^*(z) = g_j^*(z)\,\frac{1}{n}\sum_{\ell\in\mathbb{Z}_n} f_k^*\left(e^{2\pi i\ell/n} z\right), \qquad z \in \mathbb{T},$$

ist. Ersetzen wir z durch $e^{2\pi i m/n} z$ für $m \in \mathbb{Z}_n$, dann ist auch[112]

$$0 = g_j^*\left(e^{2\pi i m/n}\, z\right)\,\frac{1}{n}\sum_{\ell\in\mathbb{Z}_n} f_k^*\left(e^{2\pi i(\ell+m)/n} z\right) = g_j^*\left(e^{2\pi i m/n}\, z\right)\,\frac{1}{n}\sum_{\ell\in\mathbb{Z}_n} f_k^*\left(e^{2\pi i\ell/n} z\right),$$

was sich auch als

$$0 = \left(\widetilde{M}^T(z)\, M(z)\right)_{jk}, \qquad j,k \in \mathbb{Z}_n, \quad j \neq k, \qquad z \in \mathbb{T},$$

schreiben läßt. Das heißt, daß bis auf Normalisierung der resultierenden Diagonalmatrix

$$\widetilde{M}^T(z) = M^{-1}(z), \qquad z \in \mathbb{T}, \tag{4.22}$$

gelten muß. Deswegen ist es auch nicht verwunderlich, daß man in diesem Fall von einer *biorthogonalen* Filterbank spricht. Eine *orthogonale* Filterbank liegt hingegen vor, wenn zusätzlich $\widetilde{M} = M$ ist – dann hätte man $M^T = M^{-1}$ und die Matrix $M(z)$ muß für jedes $z \in \mathbb{T}$ orthogonal sein.

4.3 Zweikanal–Filterbänke

So viel Allgemeinheit reicht für's erste. Jetzt wenden wir uns wieder dem ach so einfachen Fall $n = 2$ zu. Dabei bemerken wir zuerst, daß hier eine Hälfte der Filterbank zusammen mit der

[111] Hier betrachten wir nur den "Fourieranteil", das heißt, wir wählen $z \in \mathbb{T}$.

[112] Der Trick besteht in der Ersetzung des Summationsindex ℓ durch $\ell - m$.

perfekten Rekonstruktion bereits die andere Hälfte der Filterbank festlegt. Das folgt aus der Formel

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = (ad - bc) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

die für jede 2×2–Matrix gilt, also

$$\underbrace{\begin{bmatrix} a & b \\ c & d \end{bmatrix}}_{=:A} = \frac{1}{\det A} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix},$$

was mit

$$M(z) = \begin{bmatrix} f_0^*(z) & f_0^*(-z) \\ f_1^*(z) & f_1^*(-z) \end{bmatrix}$$

die Identität

$$\widetilde{M}(z) = M^{-1}(z) = \frac{1}{f_0^*(z)\, f_1^*(-z) - f_0^*(-z)\, f_1^*(z)} \begin{bmatrix} f_1^*(-z) & -f_0^*(-z) \\ -f_1^*(z) & f_0^*(z) \end{bmatrix}$$

liefert, wobei $\widetilde{M}$ tatsächlich die gewünschte Struktur

$$\widetilde{M}(z) = \begin{bmatrix} g_0^*(z) & g_1^*(z) \\ g_0^*(-z) & g_1^*(-z) \end{bmatrix},$$

mit

$$g_0^*(z) = \frac{f_1^*(-z)}{f_0^*(z) f_1^*(-z) - f_0^*(-z) f_1^*(z)}, \qquad g_1^*(z) = \frac{-f_0^*(-z)}{f_0^*(z) f_1^*(-z) - f_0^*(-z) f_1^*(z)},$$

hat.

Die weitere Darstellung in diesem Kapitel folgt [27] und basiert auf den Konzept der *Spiegelfilter*, das wir bereits von den Allpass–Filtern her kennen. Einen solchen Spiegelfilter zu einem Filter F mit z–Transformierter f^* erhalten wir, indem wir den Filter G mit $g^*(z) := f^*\left(z^{-1}\right)$ betrachten. Dabei ist zu beachten:

1. Die Verwendung von Spiegelfiltern ist besonders einfach! Man muß nur die Koeffizienten eines Filters speichern.

2. Der Name "Spiegelfilter" ist klar, da

$$f*(z) = \sum_{k \in \mathbb{Z}} f(k)\, z^{-k} \qquad \Longrightarrow \quad g^*(z) = f^*\left(z^{-1}\right) = \sum_{k \in \mathbb{Z}} f(k)\, z^k = \sum_{k \in \mathbb{Z}} f(-k)\, z^{-k},$$

   also $g(k) = f(-k)$ ist – die Filterkoeffizienten werden am Ursprung gespiegelt.

3. Da wir immer von *reellen* Filtern ausgegangen sind, ist

$$g^*(z) = \overline{f^*\left(\overline{z}^{-1}\right)},$$

bis auf einen Faktor der Form z^n ist also g^* nichts anderes als $(f^*)^\sharp$. Allerdings macht dieser Faktor, der ja bei der Definition von $p^\sharp$ vom Grad des Polynoms p abhing, ja nur dann Sinn, wenn die so "transformierte" Funktion auch wirklich ein Poynom war. Und wenn man also mit rationalen Funktionen arbeiten möchte, dann kann man ihn auch gleich weglassen.

4. Für $z \in \mathbb{T}$ ist ja $z^{-1} = \overline{z}$, also ist

$$g^*(z) = f^*\left(z^{-1}\right) = f^*\left(\overline{z}\right) = \overline{f^*(z)} \tag{4.23}$$

Nun übertragen wir das auf unsere Filterbank, indem wir

$$g_0^*(z) = f_0^*\left(z^{-1}\right) \qquad \text{und} \qquad g_1^*(z) = f_1^*\left(z^{-1}\right) \tag{4.24}$$

setzen, womit sich die Perfekte Rekonstruktion unter Verwendung von (4.17) als

$$\begin{aligned} c^*(z) &= [g_0^*(z)\ g_1^*(z)] \begin{bmatrix} c_0^*\left(z^2\right) \\ c_0^*\left(z^2\right) \end{bmatrix} = [g_0^*(z)\ g_1^*(z)]\ \frac{1}{2} \begin{bmatrix} f_0^*(z) & f_0^*(-z) \\ f_1^*(z) & f_1^*(-z) \end{bmatrix} \begin{bmatrix} c^*(z) \\ c^*(-z) \end{bmatrix} \\ &= \frac{1}{2}\left[g_0^*(z)\, f_0^*(z) + g_1^*(z)\, f_1^*(z)\right]\ c^*(z) + \frac{1}{2}\left[g_0^*(z)\, f_0^*(-z) + g_1^*(z)\, f_1^*(-z)\right]\ c^*(-z) \\ &= \frac{1}{2}\left[f_0^*\left(z^{-1}\right)\ f_0^*(z) + f_1^*\left(z^{-1}\right)\ f_1^*(z)\right]\ c^*(z) \\ &\quad + \frac{1}{2}\left[f_0^*\left(z^{-1}\right)\ f_0^*(-z) + f_1^*\left(z^{-1}\right)\ f_1^*(-z)\right]\ c^*(-z), \end{aligned}$$

also

$$\begin{aligned} 2 &= f_0^*\left(z^{-1}\right)\ f_0^*(z) + f_1^*\left(z^{-1}\right)\ f_1^*(z) \\ 0 &= f_0^*\left(z^{-1}\right)\ f_0^*(-z) + f_1^*\left(z^{-1}\right)\ f_1^*(-z) \end{aligned} \tag{4.25}$$

liefert. Die Herleitung zeigt aber auch, daß unter der Annahme (4.24), daß Analyse und Synthese Spiegelungen voneinander sind, die perfekte Rekonstruktion *äquivalent* zu (4.25) ist. Es gibt aber noch eine andere Art, diese Bedingungen zu sehen: Unter Verwendung von (4.24) ist

$$\widetilde{M}(z) = \begin{bmatrix} g_0^*(z) & g_1^*(z) \\ g_0^*(-z) & g_1^*(-z) \end{bmatrix} = \begin{bmatrix} f_0^*\left(z^{-1}\right) & f_1^*\left(z^{-1}\right) \\ f_0^*\left(-z^{-1}\right) & f_1^*\left(-z^{-1}\right) \end{bmatrix} = 2M^T\left(z^{-1}\right),$$

so daß die perfekte Rekonstruktion und damit auch (4.25) auch äquivalent zu

$$I = \widetilde{M}(z)\, M(z) = 2M^T\left(z^{-1}\right)\ M(z), \qquad z \in \mathbb{C}^\times, \tag{4.26}$$

ist – die Matrix $2^{-1/2}M(z)$ ist *unitär* auf $\mathbb{T}$. Die Gleichungen (4.25) stellen aber besondere Forderungen an den Filter F_0.

Lemma 4.15 *Sind f_0^* und f_1^* rationale Funktionen mit reellen Koeffizienten, dann ist*

$$0 = f_0^*\left(z^{-1}\right)\, f_0^*(-z) + f_1^*\left(z^{-1}\right)\, f_1^*(-z) \tag{4.27}$$

genau dann erfüllt, wenn es $k \in \mathbb{Z}$ und einen rationalen Allpass–Filter h gibt, so daß

$$f_1^*(z) = z^{2k-1}\, f_0^*\left(-z^{-1}\right)\, h\left(z^2\right) \tag{4.28}$$

ist.

Beweis: Setzt man (4.28) in die rechte Seite von (4.27) ein, so erhält man für $z \in \mathbb{T}$, daß

$$\begin{aligned} & f_0^*\left(z^{-1}\right)\, f_0^*(-z) + z^{1-2k}\, f_0^*\left(-z\right)\, h\left(z^{-2}\right)\, (-z)^{2k-1}\, f_0^*\left(z^{-1}\right)\, h\left(z^2\right) \\ &= \; f_0^*\left(z^{-1}\right)\, f_0^*(-z)\left[1 - h\left(z^{-2}\right)\, h\left(z^2\right)\right] = f_0^*\left(z^{-1}\right)\, f_0^*(-z)\left[1 - \left|h\left(z^2\right)\right|^2\right] = 0, \end{aligned}$$

was damit nach Proposition 4.1 auch auf ganz $\mathbb{C}$ gelten muß. Damit ist "$\Leftarrow$" bewiesen.
Für die Umkehrung definieren wir

$$g(z) := \frac{f_1^*(z)}{f_0^*\left(-z^{-1}\right)}, \qquad z \in \mathbb{C}^\times,$$

und erhalten, indem wir z in (4.27) durch z^{-1} ersetzen und durchdividieren, daß

$$0 = \frac{f_0^*(z)}{f_1^*\left(-z^{-1}\right)} + \frac{f_1^*(z)}{f_0^*\left(-z^{-1}\right)} = \frac{1}{g\left(-z^{-1}\right)} + g(z) = \frac{1 + g\left(-z^{-1}\right)\, g(z)}{g\left(-z^{-1}\right)},$$

also

$$-1 = g\left(-z^{-1}\right)\, g(z).$$

Schreiben wir nun

$$g(z) = \sum_{k\in\mathbb{Z}} g(k)\, z^k = \sum_{k\in\mathbb{Z}} g(2k) z^{2k} + z \sum_{k\in\mathbb{Z}} g(2k+1) z^{2k} = g_0\left(z^2\right) + z\, g_1\left(z^2\right),$$

dann ist für $z \in \mathbb{T}$

$$\begin{aligned} g\left(-z^{-1}\right)\, g(z) &= \left[g_0\left(z^{-2}\right) - z^{-1}\, g_1\left(z^{-2}\right)\right]\, \left[g_0\left(z^2\right) + z\, g_1\left(z^2\right)\right] \\ &= \; g_0\left(z^{-2}\right)\, g_0\left(z^2\right) - z^{-1}\, g_1\left(z^{-2}\right)\, g_0\left(z^2\right) + z\, g_0\left(z^{-2}\right)\, g_1\left(z^2\right) - g_1\left(z^{-2}\right)\, g_1\left(z^2\right) \\ &= \; \left|g_0\left(z^2\right)\right|^2 - z^{-1}\, g_0\left(z^2\right)\, g_1\left(z^{-2}\right) + z\, g_0\left(z^{-2}\right)\, g_1\left(z^2\right) - \left|g_1\left(z^2\right)\right|^2. \end{aligned}$$

Da $g_0\left(z^2\right)\, g_1\left(z^{-2}\right)$ und $g_0\left(z^{-2}\right)\, g_1\left(z^2\right)$ Laurentreihen in z^2 sind, kann diese Reihe nur dann konstant sein, wenn

$$0 = \underbrace{z^{-1}\, g_0\left(z^2\right)\, g_1\left(z^{-2}\right)}_{=:h(z)} - \underbrace{z\, g_0\left(z^{-2}\right)\, g_1\left(z^2\right)}_{=h(z^{-1})=\overline{h(z)}} = 2\Re h(z) = 2\Re\left(z^{-1}\, g_0\left(z^2\right)\, g_1\left(z^{-2}\right)\right)$$

für alle $z \in \mathbb{T}$ gilt, das heißt, wenn $g_0\left(z^2\right)\ g_1\left(z^{-2}\right) = 0$ ist, also wenn entweder $g_0 = 0$ ist oder $g_1 = 0$ ist oder beide. Wäre aber $g_1 = 0$, dann wäre

$$g\left(-z^{-1}\right)\ g(z) = \left|g_0\left(z^2\right)\right|^2 \geq 0$$

und damit niemals $= -1$. Also muß $g_1 \neq 0$ und damit $g_0 = 0$ sein und es ist

$$g(z) = z\, g_1\left(z^2\right), \quad z \in \mathbb{C}, \qquad \text{und} \qquad |g_1(z)| = 1, \quad z \in \mathbb{T}.$$

Damit ist $g_1\left(z^2\right)$ ein gerader[113] Allpass–Filter und damit kann g als

$$z\, h\left(z^2\right) = g(z) = \frac{f_1^*(z)}{f_0^*\left(-z^{-1}\right)}$$

geschrieben werden, woraus

$$f_1^*(z) = z\, f_0^*\left(-z^{-1}\right)\ h\left(z^2\right)$$

folgt. Beliebige ungerade Potenzen von z wie in (4.27) können dann durch "Verschiebung" von h erhalten werden. □

Bemerkung 4.16 *Ist man nur an FIR–Filtern interessiert und ist $f_0^* \in \Pi$, dann ist lediglich $h(z) = z^k$ für beliebiges $k \in \mathbb{Z}$, also nur ein "Verzögerer" zulässig. Die rationale Filtertheorie erweitert also die zulässigen Filterklassen ganz enorm.*

Das nächste Konzept ist das der Autokorrelation eines Filters und die zugehörige Eigenschaft der z–Transformation.

Definition 4.17 (Korrelation) *Die* Korrelation *zweier Folgen $f, g \in \ell(\mathbb{Z})$ ist definiert als*

$$f \star g := \sum_{k\in\mathbb{Z}} f(k)\, g\left(\cdot + k\right),$$

und die Autokorrelation *von $f \in \ell(\mathbb{Z})$ als $f \star f$.*

Es kann nie schaden, die Korrelation bezüglich der z–Transformation zu betrachten, wobei man

$$\begin{aligned}
(f \star g)^*(z) &= \sum_{j\in\mathbb{Z}} (f \star g)(j)\, z^{-j} = \sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} f(k)\, g(j+k)\, z^{-j} \\
&= \sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} f(k)\, z^k\, g(j+k)\ z^{-j-k} = \sum_{k\in\mathbb{Z}} f(k)\, z^k \sum_{j\in\mathbb{Z}} g(j)\ z^{-j} \\
&= f^*\left(z^{-1}\right)\ g^*(z)
\end{aligned}$$

[113] Also eine Funktion, die bezüglich der Vertauschung $z \leftrightarrow -z$ invariant ist.

erhält. Insbesondere hat also die Autokorrelation die z–Transformation

$$(f \star f)^* (z) = f^*(z)\, f^* \left(z^{-1}\right).$$

Mit Hilfe der Autokorrelation können wir die erste Bedingung von (4.25) auch anders schreiben! Ist nämlich g die Autokorrelation von f_0^*, dann bedeutet diese Identität unter Verwendung von (4.28) nämlich nichts anderes als

$$\begin{aligned} 2 &= f_0^* \left(z^{-1}\right)\, f_0^*(z) + f_1^* \left(z^{-1}\right)\, f_1^*(z) \\ &= f_0^* \left(z^{-1}\right)\, f_0^*(z) + z^{1-2k}\, f_0^* (-z)\, h\left(z^2\right)\, z^{2k-1}\, f_0^* \left(-z^{-1}\right)\, h\left(z^{-2}\right) \\ &= f_0^* \left(z^{-1}\right)\, f_0^*(z) + f_0^*(-z)\, f_0^* \left(-z^{-1}\right) \underbrace{\left|h\left(z^2\right)\right|^2}_{=1} \\ &= g^*(z) + g^*(-z) = \sum_{k\in\mathbb{Z}} g(k) z^{-k} + \sum_{k\in\mathbb{Z}} g(k)(-z)^{-k} = \sum_{k\in\mathbb{Z}} g(2k)\, z^{-2k}, \end{aligned}$$

also $g(2\cdot) = \delta$ und daher $g(z) = 1 + z\,\widetilde{g}\,(z^2)$. Offensichtlich ist nicht jede rationale Funktion die z–Transformierte einer Autokorrelation – aufgrund ihrer besonderen Struktur müssen die Nullstellen symmetrisch am Einheitskreis liegen: ist ζ eine Nullstelle einer solchen Funktion, dann auch ζ^{-1}.
Wir fassen jetzt einmal schnell die bisher hergeleiteten Beschreibungen *orthogonaler Filterbänke* zusammen.

Satz 4.18 (Orthogonale Filterbänke) *Seien F_0, F_1 Filter mit zugehörigen rationalen z–Transformationen f_0^*, f_1^*. Dann sind äquivalent:*

1. $M^T\left(z^{-1}\right)\, M(z) = \frac{1}{2} I$, $z \in \mathbb{C}^\times$.
2. *Es gibt $k \in \mathbb{Z}$ und einen rationalen Allpass–Filter h, so daß*

   $$f_1^*(z) = z^{2k-1}\, f_0^* \left(-z^{-1}\right)\, h\left(z^2\right) \qquad \textit{und} \qquad f_0^*(z)\, f_0^* \left(z^{-1}\right) + f_0^*(-z)\, f_0^* \left(-z^{-1}\right) = 2.$$

3. *Die z–Transformierte g der Autokorrelation von f_0^* hat die Form $g(z) = 1 + z\,\widetilde{g}\,(z^2)$ und es gibt $k \in \mathbb{Z}$ und einen rationalen Allpass–Filter h, so daß*

   $$f_1^*(z) = z^{2k-1}\, f_0^* \left(-z^{-1}\right)\, h\left(z^2\right)$$

   ist.

Mit Hilfe dieses Satzes können wir nun – und das Resultat stammt aus [27] – *alle* orthogonalen Filterbänke vollständig beschreiben beziehungsweise parametrisieren.

Satz 4.19 (Parametrisierung orthogonaler Filterbänke) *Alle orthogonalen rationalen Zweikanal–Filterbänke lassen sich auf die folgende Art und Weise konstruieren:*

1. *Wähle ein* beliebiges *Polynom* $0 \neq p \in \Pi$ *und setze*

$$g(z) = 2\,\frac{p(z)\,p\left(z^{-1}\right)}{p(z)\,p\left(z^{-1}\right) + p(-z)\,p\left(-z^{-1}\right)}, \qquad z \in \mathbb{C}^\times. \tag{4.29}$$

2. *Faktorisiere* $z^k g(z)$ *für passendes* k *als* $f(z)\,f\left(z^{-1}\right)$.

3. *Setze*

$$f_0^*(z) = h_0(z)\,f(z) \qquad \textit{und} \qquad f_1^*(z) = z^{2k-1}\,f_0^*\left(-z^{-1}\right)\,h_1\left(z^2\right), \qquad z \in \mathbb{C}^\times, \tag{4.30}$$

wobei h_0 *und* h_1 *beliebige Allpass–Filter sind.*

4. *Setze*

$$g_0^*(z) = f_0^*\left(z^{-1}\right) \qquad \textit{und} \qquad g_1^*(z) = f_1^*\left(z^{-1}\right), \qquad z \in \mathbb{C}^\times. \tag{4.31}$$

Bemerkung 4.20

1. *Das Konstruktionsverfahren hat im Prinzip vier freie Parameter: Das "Startpolynom"* p*, die Allpass–Filter* h_0 *und* h_1 *und den Verzögerungsparameter* k*.*

2. *Die Autokorrelationsfunktion* $g(z)$ *hat auf dem Einheitskreis, also für* $z \in \mathbb{T}$*, die Form*

$$g(z) = 2\frac{|p(z)|^2}{|p(z)|^2 + |p(-z)|^2}$$

und hat daher genau dann einen Pol auf dem Einheitskreis, wenn p *und* $p(-\cdot)$ *eine eine gemeinsame Nullstelle haben. Insbesondere ist* g *für alle* $p \neq 0$ *wohldefiniert, weil dann der Nenner nicht konstant verschwindet.*

3. *Die Faktorisierung* $g(z) \to f(z)\,f\left(z^{-1}\right)$ *kann man natürlich geschickt durchführen: Man schiebt alle "guten" Pole* innerhalb *des Einheitskreises in* $f(z)$ *und alle "bösen" Pole* außerhalb *des Einheitskreises in* $f\left(z^{-1}\right)$*. Damit ist der resultierende Filter sogar* stabil *und zwar* automatisch*. Diese Idee ist nicht so neu – es ist im wesentlichen die Idee, die auch den klassischen* Butterworth–Filtern *zugrundeliegt, siehe z.B. [24].*

Beweis von Satz 4.19: Wir zeigen zuerst, daß jeder Schritt der Konstruktion durchführbar ist; daß sie uns eine orthogonale Filterbank liefert, haben wir schon in Satz 4.18 gesehen. Da Zähler und Nenner von g symmetrisch in z und z^{-1} sind, ist mit jeder Nullstelle und jedem Pol ζ von g auch ζ^{-1} eine Nullstelle bzw. ein Pol von g. Daher lassen sich Zähler und Nenner in der Form

$$c\prod_{j=1}^{n}(z-\zeta_j)\left(z-\zeta_j^{-1}\right) = (-1)^n z^n \zeta_1^{-1}\cdots\zeta_n^{-1}\prod_{j=1}^{n}(z-\zeta_j)\left(z^{-1}-\zeta_j\right)$$

schreiben, was

$$f(z) = \sqrt{(-1)^n\zeta_1^{-1}\cdots\zeta_n^{-1}}\prod_{j=1}^{n}(z-\zeta_j)$$

liefert; insbesondere ist also g die z–Transformierte einer Autokorrelation, was sich auch nicht ändert, wenn wir f mit einem Allpass–Filter h_0 multiplizieren, der ja die Eigenschaft

$$h_0(z)\, h_0\left(z^{-1}\right) = |h_0(z)|^2 = 1, \qquad z \in \mathbb{T}$$

hat, die sich wieder einmal auf $\mathbb{C}^\times$ fortsetzt. Darüberhinaus ist

$$g(z) + g(-z) = 2\,\frac{p(z)\,p\left(z^{-1}\right) + p(-z)\,p\left(-z^{-1}\right)}{p(z)\,p\left(z^{-1}\right) + p(-z)\,p\left(-z^{-1}\right)} = 2,$$

was uns zusammen mit Satz 4.18 zeigt, daß wir es tatsächlich mit einer orthogonalen Filterbank zu tun haben.
Für die Umkehrung sei g die z–Transformierte der Autokorrelation von f_0, das eine orthogonale Filterbank definiert. Dann ist g von der Form[114]

$$g(z) = \frac{p(z)\,p\left(z^{-1}\right)}{q(z)\,q\left(z^{-1}\right)}$$

für passende Polynome p, q und, wegen der Orthogonalität,

$$g(z) = 1 + z\,\widetilde{g}\left(z^2\right) = 1 + z\,\frac{\widetilde{p}\,(z^2)}{\widetilde{q}\,(z^2)} = \frac{\widetilde{q}\,(z^2) + z\,\widetilde{p}\,(z^2)}{\widetilde{q}\,(z^2)}$$

Sehen wir uns nun den Zähler an und spalten in gerade und ungerade Potenzen von z auf, dann erhalten wir durch Koeffizientenvergleich, daß

$$\begin{aligned} \widetilde{q}\left(z^2\right) &= \frac{1}{2}\left[p(z)\,p\left(z^{-1}\right) + p(-z)\,p\left(-z^{-1}\right)\right], \\ z\,\widetilde{p}\left(z^2\right) &= \frac{1}{2}\left[p(z)\,p\left(z^{-1}\right) - p(-z)\,p\left(-z^{-1}\right)\right], \end{aligned}$$

und daher

$$g(z) = \frac{p(z)\,p\left(z^{-1}\right)}{\widetilde{q}\,(z^2)} = 2\,\frac{p(z)\,p\left(z^{-1}\right)}{p(z)\,p\left(z^{-1}\right) + p(-z)\,p\left(-z^{-1}\right)}.$$

Damit ist der Beweis komplett. □

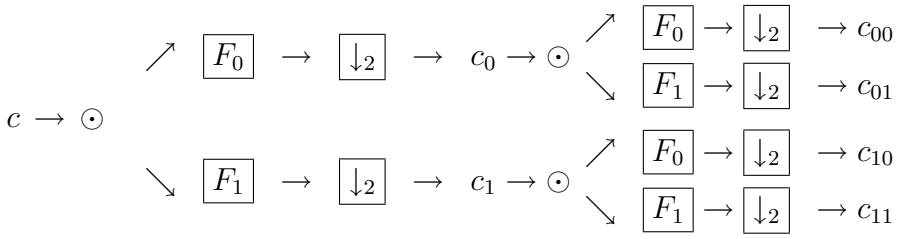
4.4 Subband–Kaskaden, Subdivision und Wavelets

Nachdem wir jetzt unsere orthogonalen Filter vollständig charakterisiert haben, wird es wieder Zeit, sich damit zu beschäftigen, wie man diese Filter verwendet. Wir erinnern uns dabei, daß unser Analysefilter hier immer von der Form

$$c \;\to\; \odot \;\begin{array}{lllll} \nearrow & \boxed{F_0} & \to & \boxed{\downarrow_2} & \to \; c_0 \\ \searrow & \boxed{F_1} & \to & \boxed{\downarrow_2} & \to \; c_1 \end{array}$$

[114] Das folgt direkt aus der Definition der Autokorrelation

war. Und jetzt können wir natürlich c_0 und/oder c_1 wieder in dieselbe Analyse–Filterbank stecken und die Daten so weiter zerlegen, also



und so weiter – auf diese Weise erhält man eine *Baumstruktur* von Signalen. In der "normalen" Waveletanalysis zerlegt man nur die Daten weiter, die aus dem "Tiefpassfilter" herauskommen, aber nachdem wir bisher ja noch nicht einmal spezifiziert haben, ob wir F_0 oder F_1 als Tiefpass ansehen, bleibt uns erst einmal nichts anderes übrig, als symmetrisch zu zerlegen und so bei der Baumstruktur anzukommen. Diesen Ansatz bezeichnet man auch als *Wavelet Packages*.

Nach r derartigen Kaskadenschritten erhalten wir somit die Ausgabesignale

$$c_j^r \in \ell(\mathbb{Z}), \qquad j = 0, \ldots, 2^r - 1,$$

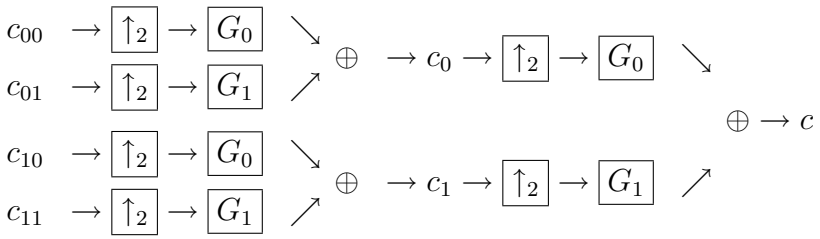
wobei die *Binärdarstellung* des Index j genau die Filterungskaskade angibt, der unser Signal unerworfen wurde. Genauer: Ist

$$j = \sum_{k=0}^{r-1} \epsilon_k \, 2^k =: \epsilon_{r-1} \cdots \epsilon_0, \qquad \epsilon_k \in \{0, 1\},$$

dann ist

$$c_j^r = \downarrow_2 \; F_{\epsilon_0} \cdots \downarrow_2 \; F_{\epsilon_{r-1}} \, c.$$

Umgekehrt wird natürlich auch die Synthese wieder durch Kaskaden realisiert, diesmal durch Kaskaden der Synthese–Filterbank:



Diese Synthese–Kaskade erlaubt es uns nun, das Problem von einer ganz anderen Warte aus zu sehen, indem wir die Analyse mal für einen Moment vergessen und die Synthese als Methode ansehen, Funktionen zu generieren – das sind dann auch schon die sogenannten *Subdivision–Schemata*.

Und zwar geben wir uns jetzt ein Signal x vor, speisen das in c_0^r ein und sehen uns das Signal $c = c[x, r]$ an, das auf diese Art und Weise herauskommt. Offensichtlich ist dann

$$c = (G_0 \uparrow_2)^r \; x,$$

also, nach Lemma 4.11,

$$c^*(z) = \left[(G_0 \uparrow_2)^r\ x\right]^*(z) = g_0^*(z)\ \left[\uparrow_2\ (G_0 \uparrow_2)^{r-1}\ x\right]^*(z) = g_0^*(z)\ \left[(G_0 \uparrow_2)^{r-1}\ x\right]^*\left(z^2\right)$$

und somit

$$c^*(z) = g_0^*(z)\cdots g_0^*\left(z^{2^{r-1}}\right)\ x^*\left(z^{2^r}\right) = \left[\prod_{j=0}^{r-1} g_0^*\left(z^{2^j}\right)\right]\ x^*\left(z^{2^r}\right), \tag{4.32}$$

was uns unter Verwendung der Spiegelbeziehung $g_0^* = f_0^*\left(z^{-1}\right)$ schließlich

$$c^*(z) = \left[\prod_{j=0}^{r-1} f_0^*\left(z^{-2^j}\right)\right]\ x^*\left(z^{2^r}\right) \tag{4.33}$$

liefert. Wegen des Upsamplings enthält die Folge $c = (G_0 \uparrow_2)^r\, x$ deutlich mehr Information als x, nämlich das 2^r–fache. Außerdem ist

$$S_G x := G_0 \uparrow_2 x = g_0 * (\uparrow_2 x) = \sum_{k\in\mathbb{Z}} g_0\left(\cdot - 2k\right)\ x(k)$$

und somit $\tau_2 S_G = S_G \tau$, weswegen man die Folge $S_G^r x$ als diskrete Funktion an den Abszissen $2^{-r}k$, $k \in \mathbb{Z}$, auffassen sollte: Hat beispielsweise x zumindest lokal ein k–periodisches Verhalten[115], dann hat $S_G^r x$ lokal ein $2^r k$–periodisches Verhalten. Und was auch noch sofort auffällt: Da S_G ein linearer Operator ist und wir jede Folge x formal als

$$x = \sum_{k\in\mathbb{Z}} x(k)\,\tau_k\,\delta, \qquad \delta(j) = \delta_{j0}, \quad j \in \mathbb{Z},$$

schreiben können, ist

$$c^*(z) = \sum_{k\in\mathbb{Z}} z^k\, x(k)\ \left[\prod_{j=0}^{r-1} f_0^*\left(z^{-2^j}\right)\right]\ \underbrace{\delta^*\left(z^{2^r}\right)}_{=1},$$

weswegen es genügt, sich auf $x = \delta$ zu beschränken – in diesem Fall ist bereits alle Information enthalten.

Diese diskrete Funktion c, genauer, die diskrete Funktion $c\left(2^{-r}\cdot\right) \in \ell\left(2^{-r}\mathbb{Z}\right)$, die an den dyadischen Punkten der Ordnung r definiert ist, wollen wir uns nun mal in der Fouriertransformation ansehen, das heißt, wir betrachten die Funktionen

$$\varphi_r(\xi) = \left[c\left(2^{-r}\cdot\right)\right]^\wedge(\xi), \qquad r \in \mathbb{N}_0.$$

Natürlich ist diese Funktion gar nicht wirklich definiert. Ersetzt man aber in (4.33) z durch $e^{i\xi/2^r}$, so erhält man eine Folge φ_r von trigonometrischen Polynomen[116] auch formal korrekt als

$$\varphi_r(\xi) = 2^{-r}\,\widehat{c}\,(\xi/2^r) = 2^{-r} c^*\left(e^{i\xi/2^r}\right) = \prod_{j=0}^{r-1} \frac{1}{2}\, f_0^*\left(e^{-i\xi/2^{r-j}}\right) = \prod_{j=1}^{r} \frac{1}{2}\, f_0^*\left(e^{-i\xi/2^j}\right),$$

[115] Und dazu gehört insbesondere ein kompakter Träger!

[116] Wir bleiben jetzt bei $x = \delta$!

und wir können die Grenzfunktion

$$\varphi(\xi) := \varphi_\infty(\xi) := \lim_{r\to\infty} \varphi_r(\xi) = \prod_{j=1}^{\infty} \frac{1}{2}\, f_0^*\left(e^{-i2^{-j}\xi}\right) \tag{4.34}$$

definieren – vorausgesetzt natürlich, das unendliche Produkt konvergiert. Dann hat φ eine sehr nette Eigenschaft, nämlich

$$\begin{aligned}\varphi(\xi) &= \frac{1}{2}\, f_0^*\left(e^{-i\xi/2}\right) \prod_{j=2}^{\infty} \frac{1}{2}\, f_0^*\left(e^{-i2^{-j}\xi}\right) = \frac{1}{2}\, f_0^*\left(e^{-i\xi/2}\right) \underbrace{\prod_{j=1}^{\infty} \frac{1}{2}\, f_0^*\left(e^{-i2^{-j}(\xi/2)}\right)}_{=\varphi(\xi/2)} \\ &= \frac{1}{2}\, f_0^*\left(e^{-i\xi/2}\right)\, \varphi\left(\frac{\xi}{2}\right). \end{aligned} \tag{4.35}$$

Das können wir nun nochmals ein bißchen anders sehen, indem wir auch noch annehmen, daß $\varphi \in L_1(\mathbb{R})$ ist[117], denn dann können wir invers fouriertransformieren und erhalten so eine gleichmäßig stetige Funktion $\phi = \varphi^\vee$. Mit $\varphi = \widehat{\phi}$ wird (4.35) dann[118] zu

$$\widehat{\phi}(\xi) = \frac{1}{2}\, f_0^*\left(e^{-i\xi/2}\right)\, \widehat{\phi}\left(\frac{\xi}{2}\right) = \frac{1}{2} \underbrace{\widehat{f_0}\left(-\frac{\xi}{2}\right)}_{=\widehat{g_0}(\xi/2)} \widehat{\phi}\left(\frac{\xi}{2}\right) = \left[(g_0 * \phi)\,(2\cdot)\right]^\wedge(\xi),$$

also

$$\phi = (g_0 * \phi)\,(2\cdot) = \sum_{k\in\mathbb{Z}} g_0(k)\, \phi\,(2\cdot - k)\,. \tag{4.36}$$

Diese Gleichung, als *Verfeinerungsgleichung*, *Refinement Equation* oder *Zweiskalenbeziehung* bezeichnet, bedeutet, daß man die Funktion ϕ als Überlagerung von Translaten ihrer "gestauchten" Version darstellen kann, siehe Abb. 4.1 – dies schafft nicht jede Funktion, sondern es ist eine *Forderung* an die Funktion. Und tatsächlich erhält man Funktionen, die (4.36) erfüllen, eigentlich auch nur als Ergebnis des Subdivision–Prozesses. Aber zu diesem Zweck braucht man natürlich Kriterien für die Konvergenz des unendlichen Produkts in (4.34). Klar ist natürlich, daß die Konvergenz eines unendlichen Produkts erzwingt, daß die einzelnen Glieder des Produkts gegen 1 konvergieren, so daß[119] ohne

$$1 = \frac{1}{2} \lim_{r\to\infty} f_0^*\left(e^{-i2^{-r}\xi}\right) = \frac{1}{2}\, f_0^*(1) = \frac{1}{2}\, \widehat{f_0}(0) \tag{4.37}$$

gar nichts läuft. Eine *hinreichende* Bedingung für die Existenz einer stetigen Funktion ϕ, die die Zweiskalenbeziehung (4.36) liefert das folgende Resultat, das auf Daubechies [8] zurückgeht und das mitsamt Beweis aus [63] entnommen ist.

[117] Hier ist $\mathbb{R}$ anstelle von $\mathbb{T}$ wichtig, denn φ_r ensteht ja durch *Dilatation* des trigonometrischen Polynoms $\widehat{c}$.

[118] Wir erinnern uns: $g_0^*(z) = f_0^*\left(z^{-1}\right)$, also ist $\widehat{g_0}(\xi) = g_0^*\left(e^{i\xi}\right) = f_0^*\left(e^{-i\xi}\right) = \widehat{f_0}(-\xi)$.

[119] Zumindest für rationale Filter f_0^* ohne Pol an $z = 1$, und nur solche interessieren uns ja hier eigentlich.

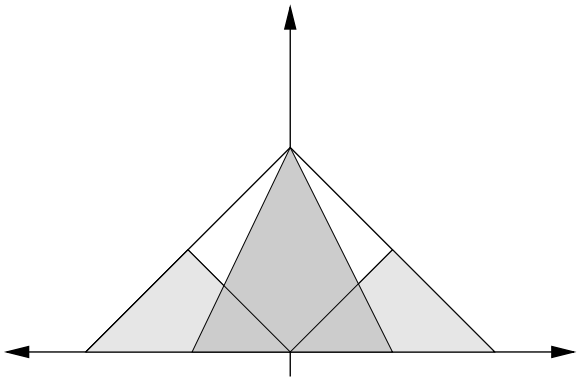


Abbildung 4.1: Wie man die stückweise lineare “Dachfunktion” als Zweiskalenbeziehung darstellt.

Proposition 4.21 (Existenz stetiger verfeinerbarer Funktionen) *Läßt sich f_0^* als*

$$f_0^*(z) = \left(\frac{1+z}{2}\right)^k q(z) \qquad \textit{mit} \qquad \max_{z\in\mathbb{T}} |q(z)| < 2^k \quad \textit{und} \quad q(1) = 2, \tag{4.38}$$

schreiben, dann existiert eine stetige Lösung von (4.36).

Beweis: Die Idee besteht darin, zu zeigen, daß unter der Voraussetzung (4.38) das unenedliche Produkt (4.34) konvergiert und eine L_1–Funktion liefert, deren inverse Fouriertransformation dann existieren muß und als ϕ gewählt werden kann.
Dazu zerlegen wir das Produkt in

$$\prod_{j=1}^{\infty} \frac{1}{2}\widehat{f}_0\left(2^{-r}\xi\right) = \prod_{j=1}^{\infty}\left(\frac{1+e^{-i2^{-r}\xi}}{2}\right)^k \prod_{j=1}^{\infty}\frac{1}{2}\widehat{q}\left(2^{-r}\xi\right) \tag{4.39}$$

und behandeln die beiden Faktoren auf der rechten Seite separat. Da

$$\begin{aligned}\frac{1-e^{-i\xi}}{\xi} &= \frac{1+e^{-i\xi/2}}{2}\,\frac{1-e^{-i\xi/2}}{\xi/2} = \frac{1+e^{-i\xi/2}}{2}\,\frac{1+e^{-i\xi/4}}{2}\,\frac{1-e^{-i\xi/4}}{\xi/4} \\ &= \cdots = \underbrace{\frac{1-e^{-i2^{-r}\xi}}{2^{-r}\xi}}_{\to i}\prod_{r=1}^{N}\frac{1+e^{-i2^{-r}\xi}}{2},\end{aligned}$$

also

$$\prod_{r=1}^{\infty}\frac{1+e^{-i2^{-r}\xi}}{2} = \frac{1-e^{-i\xi}}{i\,\xi}$$

ist, hat das erste Produkt den Wert[120]

$$\left(\frac{1-e^{-i\xi}}{i\,\xi}\right)^k = \left(e^{-i\xi/2}\,\frac{e^{i\xi/2}-e^{-i\xi/2}}{\xi}\right)^k = e^{-ik\xi/2}\left(\frac{\sin\xi/2}{\xi/2}\right)^k,$$

was im Absolutbetrag $\leq C_1\,(1+|\xi|)^{-k}$ für eine passende Konstante $C_1>0$ ist.
Jetzt zum zweiten Faktor in (4.39), wo wir die Abkürzung $h=\frac{1}{2}q$ verwenden wollen. Da $h(1)=1$ ist, gibt es eine Konstante $C_2>0$, so daß für $|\xi|\leq 1$ die Abschätzung[121] $\left|h\left(e^{-i\xi}\right)\right| \leq 1+C_2|\xi| \leq e^{C_2|\xi|}$ erfüllt ist, also gilt für $|\xi|\leq 1$ die Abschätzung

$$\prod_{j=1}^{\infty}\left|h\left(e^{-i2^{-r}\xi}\right)\right| \leq \prod_{j=1}^{\infty} e^{C_2 2^{-r}|\xi|} = \exp\left(\sum_{j=1}^{\infty}\frac{C_2|\xi|}{2^r}\right) = e^{C_2|\xi|} \leq e^{C_2}. \tag{4.40}$$

Für beliebiges $\xi\in\mathbb{R}$ wählen wir nun $n\in\mathbb{N}$ so, daß $2^{n-1}\leq|\xi|<2^n$, und verwenden die folgende Aufspaltung zusammen mit (4.40):

$$\begin{aligned}\prod_{j=1}^{\infty}\left|h\left(e^{-i2^{-r}\xi}\right)\right| &= \prod_{j=1}^{n}\left|h\left(e^{-i2^{-r}\xi}\right)\right| \prod_{j=n+1}^{\infty}\left|h\left(e^{-i2^{-r}\xi}\right)\right| \\ &= \prod_{j=1}^{n}\left|h\left(e^{-i2^{-r}\xi}\right)\right| \prod_{j=1}^{\infty}\left|h\left(e^{-i2^{-r}(\xi/2^n)}\right)\right| \leq \prod_{j=1}^{n}\left|h\left(e^{-i2^{-r}\xi}\right)\right|\; e^{C_2} \leq B^n\, e^{C_2}\end{aligned}$$

mit

$$B := \max_{z\in\mathbb{T}}|h(z)| \leq 2^{k-1-\varepsilon} \qquad \text{für} \varepsilon>0.$$

Damit ist

$$B^n \leq 2^{n(k-1-\varepsilon)} \leq \underbrace{\left(2\,2^{n-1}\right)}_{\leq 2\,|\xi|\leq 1+|\xi|}{}^{k-1-\varepsilon} \leq 2^k\;(1+|\xi|)^{k-1-\varepsilon}$$

und somit

$$\prod_{j=1}^{\infty}\left|f_0^*\left(e^{-i2^{-r}\xi}\right)\right| \leq C_1\;(1+|\xi|)^{-k}\;e^{C_2 k}\,2^k\;(1+|\xi|)^{k-1-\varepsilon} \leq C_3\,(1+|\xi|)^{-1-\varepsilon}.$$

Also ist das Produkt überall absolut konvergent und[122] die resultierende Funktion gehört zu $L_1(\mathbb{R})$ und läßt daher eine inverse Fouriertransformierte zu. □

Übung 4.3 Zeigen Sie: Ist ϕ eine (nichttriviale) Lösung der Zwei–Skalen–Gleichung (4.36), dann ist $\widehat{\phi}(0)\neq 0$ und $\widehat{g}(0)=2$. ♢

[120] Wenn man genau hinschaut, erkennt man hier die Fouriertransformierte des zentrierten kardinalen B–Splines. Diese tauchen also wieder einmal an ganz zentraler Stelle auf.

[121] Für wen diese Abschätzung etwas suspekt erscheint: Die Funktionen $1+cx$ und e^{cx} haben für $x=0$ denselben Wert aber die Ableitungen c und $c\,e^{cx}\leq c$, also wächst die zweite Funktion schneller.

[122] Wer's genau haben will, braucht jetzt "dominated convergence"

Bemerkung 4.22 *Die Forderung (4.21) betrifft ja eigentlich f_0 und nicht die für die Zwei–Skalen–Gleichung (4.36) relevante Folge g_0. Da wir aber in (4.21) ja ohne weiteres z durch z^{-1} ersetzen können, ist es völlig irrelevant, ob die hinreichende Bedingung für f_0 oder g_0 formuliert wird.*

Sei nun also ϕ die[123] Lösung der Zwei–Skalen–Gleichung (4.36), dann gibt es drei Möglichkeiten, wie wir diese Funktion konstruieren können:

1. Über die Fourier–Transformierte:
$$\phi = \left[\prod_{j=1}^{\infty} \frac{1}{2}\, \widehat{f}_0\left(-\frac{\cdot}{2}\right)\right]^{\vee}; \tag{4.41}$$
diese Werte könnte man an ganzzahligen Werten ausrechnen und dann die inverse FFT zur Bestimmung von ϕ verwenden.

2. Über das *Kaskaden–Schema*:
$$\phi = \lim_{j\to\infty} T_G^j\, \psi, \qquad T_G\psi := (g_0 * \psi)\,(2\cdot)\,, \tag{4.42}$$
mit einer beliebigen[124] Startfunktion ψ. Hier hat man es mit einem Verfahren zu tun, daß gegen den *Fixpunkt* ψ des *Transferoperators* T_G konvergiert.

3. Über das Subdivision–Schema:
$$\phi = \lim_{j\to\infty} S_G^j \delta, \qquad \lim_{j\to\infty} \sup_{k\in\mathbb{Z}} \left|\phi\left(2^{-r}k\right) - S_G\delta(k)\right| = 0, \tag{4.43}$$
wobei die Grenzfunktion an einem immer dichteren Gitter von dyadischen Punkten bestimmt wird.

Und wo sind nun die Wavelets? Nur noch eine Interpretation weit weg. Wir betrachten jetzt, wo wir unsere schöne Funktion ϕ haben, ein Signal c nicht mehr als ein diskrete Funktion, sondern als Koeffizienten der Funktion
$$f_c := c * \phi = \sum_{k\in\mathbb{Z}} c(k)\, \phi(\cdot - k).$$
Das einfachste Beispiel hierfür sind die stückweise konstanten oder stückweise linearen Funktionen mit $\phi = \chi_{[0,1]}$ bzw. $\phi = \chi_{[0,1]} * \chi_{[0,1]}$ – letzteres ist die "Dachfunktion" aus Abb. 4.1. Da $\phi = T_G\phi$ ist
$$\begin{aligned} f_c &= \sum_{j\in\mathbb{Z}} c(j)\, \phi(\cdot - j) = \sum_{j\in\mathbb{Z}} c(j)\ (T_G\phi)\,(\cdot - j) = \sum_{j\in\mathbb{Z}} c(j)\ \sum_{k\in\mathbb{Z}} g_0(k)\, \phi\,(2\cdot - 2j - k) \\ &= \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} c(j)\, g_0(k)\, \phi\,(2\cdot - 2j - k) = \sum_{k\in\mathbb{Z}} \underbrace{\sum_{j\in\mathbb{Z}} g_0\,(k-2j)\ c(j)}_{=S_G c(k)}\, \phi\,(2\cdot - k) \\ &= (S_G c * \phi)\,(2\cdot), \end{aligned}$$

[123] Das mit "die" oder "eine" ist in Wirklichkeit gar nicht so trivial!

[124] Na gut, mehr oder weniger beliebigen.

und damit entspricht $S_G c$ den Koeffizienten bezüglich der "gestauchten" Funktion $\phi(2\cdot)$. Und das genau bringt uns nun zu den Wavelets: Unser Ein- und damit auch Ausgangssignal c der "perfect reconstruction" Filterbank

$$c \to \odot \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{ccccccc} \boxed{F_0} & \to & \boxed{\downarrow_2} & \to & c_0 & \to & \boxed{\uparrow_2} & \to & \boxed{G_0} \\ \boxed{F_1} & \to & \boxed{\downarrow_2} & \to & c_1 & \to & \boxed{\uparrow_2} & \to & \boxed{G_1} \end{array} \begin{array}{c} \searrow \\ \nearrow \end{array} \oplus \to c \tag{4.44}$$

interpretieren wir als Koeffizienten einer Funktion $f = c * \phi(2\cdot)$, die zum Raum

$$V_1 = \text{span } \{\phi(2 \cdot -k) \ : \ k \in \mathbb{Z}\}$$

gehört. Wegen der Zweiskalenbeziehung (4.36) ist auch $\phi \in V_1$ und da V_1 obendrein *translationsinvariant*[125] ist, erhalten wir, daß

$$V_1 \supseteq V_0 := \text{span } \{\phi(\cdot - k) \ : \ k \in \mathbb{Z}\}. \tag{4.45}$$

Setzen wir nun

$$V_j = \text{span } \left\{\phi\left(2^j \cdot -k\right) \ : \ k \in \mathbb{Z}\right\},$$

dann ist natürlich

$$V_0 \subseteq V_1 \subseteq V_2 \subseteq \cdots$$

Damit bilden die Räume V_j, $j \in \mathbb{N}$ oder $j \in \mathbb{Z}$, eine *Multiresolution Analysis* oder *MRA*, wie sie von Mallat eingeführt wurde, siehe beispielsweise [9, 36, 38, 63] und noch viele weitere mehr. Die (minimalen) Eigenschaften einer MRA sind:

1. Eine aufsteigende Skala von Räumen $V_0 \subseteq V_1 \subseteq \cdots$
2. (Ganzzahlige) Translationsinvarianz der Räume V_j.
3. Skalenbeziehung: $f \in V_j \Longrightarrow f(2\cdot) \in V_{j+1}$.

Und wo bleibt nur das Wavelet? Ganz einfach: Wir haben in der Filterbank (4.44) unser Signal c oder alternativ die Funktion $c * \phi(2\cdot)$ in zwei Signale c_0 und c_1 zerlegt. Nach unserer Konstruktion entspricht nun die Filterung $c_0 = \downarrow_2 \ F_0\, c$, das heißt die Bestimmung der Funktion $c_0 * \phi$, einer *Projektion* von V_1 auf V_0, das heißt,

$$c * \phi(2\cdot) \in V_0 \qquad \Longleftrightarrow \qquad c_1 = 0. \tag{4.46}$$

Dies folgt aus der Tatsache, daß die Modulationsmatrix M orthogonal und insbesondere invertierbar ist! Gäbe es nämlich zwei Darstellungen c_0, c_1 und $\widetilde{c}_0, \widetilde{c}_1$, so daß

$$G_0 \ \uparrow_2 \ c_0 + G_1 \ \uparrow_2 \ c_1 = G_0 \ \uparrow_2 \ \widetilde{c}_0 + G_1 \ \uparrow_2 \ \widetilde{c}_1$$

ist, dann ist, nach Übergang zur z–Transform

$$0 = M^T\left(z^{-1}\right) \left(\begin{bmatrix} c_0^*(z^2) \\ c_1^*(z^2) \end{bmatrix} - \begin{bmatrix} \widetilde{c}_0^*(z^2) \\ \widetilde{c}_1^*(z^2) \end{bmatrix} \right) \qquad \Longrightarrow \qquad \begin{bmatrix} c_0^*(z^2) \\ c_1^*(z^2) \end{bmatrix} = \begin{bmatrix} \widetilde{c}_0^*(z^2) \\ \widetilde{c}_1^*(z^2) \end{bmatrix},$$

[125] Ist $f \in V_1$, so ist auch $f(\cdot - k) \in V_1$ für alle $k \in \mathbb{Z}$.

denn $M(z)$ ist invertierbar. Das liefert uns (4.46).

Damit gehört dann aber die Folge c_1 zur Darstellung des Projektionsfehlers, also dem Anteil in $W_0 := V_1 \ominus V_0$, korrekter gesagt, $V_1 = V_0 \oplus W_0$, den wir nun als

$$\begin{aligned} c * \phi(2\cdot) &= \sum_{k\in\mathbb{Z}} c(k)\phi\left(2\cdot - k\right) = \sum_{k\in\mathbb{Z}}\sum_{j\in\mathbb{Z}} \left[g_0\left(k-2j\right)\, c_0(j) + g_1\left(k-2j\right)\, c_1(j)\right]\phi\left(2\cdot - k\right) \\ &= \sum_{k\in\mathbb{Z}} \left(S_G c_0\right)(k)\phi(2\cdot - k) + \sum_{j\in\mathbb{Z}} c_1(j) \underbrace{\left[\sum_{k\in\mathbb{Z}} g_1(k)\,\phi\left(2(\cdot - j) - k\right)\right]}_{=:\psi(\cdot - j)} \\ &= c_0 * \phi + c_1 * \psi \end{aligned}$$

bestimmen können.

Definition 4.23 *Die Funktion*

$$\psi := g_1 * \phi(2\cdot) = \sum_{k\in\mathbb{Z}} g_1(k)\,\phi\left(2\cdot - k\right) \tag{4.47}$$

heißt Wavelet *zur* Skalierungsfunktion ϕ.

Damit können wir dann durch Kaskadenschaltung unserer Analyse–Filterbank die *Waveletzerlegung* einer Funktion

$$f = c * \phi\left(2^n\cdot\right) = \sum_{k\in\mathbb{Z}} c(k)\,\phi\left(2^n\cdot - k\right) \quad \in \quad V_n = V_0 \oplus \bigoplus_{j=1}^{n-1} W_j$$

bestimmen, indem wir das "abgeschnittene" Pyramidenschema

$$c \to \odot \begin{array}{ll} \nearrow & \boxed{F_0} \to \boxed{\downarrow_2} \to \odot \\ \searrow & \boxed{F_0} \to \boxed{\downarrow_2} \to c_1 \end{array} \begin{array}{ll} \nearrow & \boxed{F_0} \to \boxed{\downarrow_2} \to c_{00} \to \cdots \\ \searrow & \boxed{F_1} \to \boxed{\downarrow_2} \to c_{01} \end{array}$$

zur Zerlegung verwenden und dann f als

$$f = c_1 * \psi\left(2^{n-1}\cdot\right) + c_{01} * \psi\left(2^{n-2}\cdot\right) + \cdots + c_{0\cdots01} * \psi + c_{0\cdots0} * \phi \tag{4.48}$$

darstellen. Diese Darstellung (4.48) bezeichnet man dann als die *Waveletzerlegung* von f, die Koeffizientenfolgen $c_1, c_{01}, \ldots, c_{0\ldots01}$ als *Waveletkoeffizienten*.

Bemerkung 4.24 (Orthogonalität) *Oftmals wird bei einer MRA, zumeist in L_2, verlangt, daß die Skalierungsfunktion* orthogonale *Ganzzahltranslate besitzt, das heißt, daß*

$$\int_{\mathbb{R}} \phi(x)\,\phi(x-k)\,dx = \delta_{k0}\,\|\phi\|_2^2, \qquad k \in \mathbb{Z},$$

ist. Eine notwendige *Bedingung hierfür ist, daß die Modulationsmatrix orthogonal ist oder daß, äquivalent, die Bedingung (4.25) erfüllt ist. Allerdings ist (4.25) nicht ganz hinreichend für Orthogonalität, man braucht noch zusätzliche Bedingungen an die Funktion ϕ.*

Zum Abschluß noch ein Beispiel für eine MRA und ein Wavelet, die man nicht ganz so oft sieht, nämlich die MRA, die die sinc –Funktion als Skalierungsfunktion hat. Da mit $\phi(x) = \operatorname{sinc} x = \frac{\sin x}{x}$ wir $\widehat{\phi} = \frac{1}{2}\chi_{[-1,1]}$ haben, was man durch die *inverse* Fouriertransformierte

$$\frac{1}{2}\,\chi^{\vee}_{[-1,1]}(x) = \frac{1}{2}\int_{-1}^{1} e^{ixt}dt = \frac{e^{ix}-e^{-ix}}{2ix} = \frac{\sin x}{x}$$

leicht nachprüft, erhalten wir die Verfeinerungsgleichung durch

$$\widehat{\phi}(\xi) = \widehat{g}_0\left(\frac{\xi}{2}\right)\,\widehat{\phi}\left(\frac{\xi}{2}\right) = \widehat{g}_0\left(\frac{\xi}{2}\right)\,\frac{1}{2}\,\chi_{[-2,2]}(\xi),$$

was uns

$$\chi_{[-1,1]} = \widehat{g}_0(\xi) = \sum_{k\in\mathbb{Z}} g_0(k)\,e^{-ik\xi}, \qquad g_0(k) = \frac{1}{2\pi}\int_{-1}^{1} e^{-ikt}\,dt = \frac{\sin k}{k\pi}$$

liefert, siehe auch Beispiel 1.24. Nach Satz 4.18 ist dann

$$g_1^*(z) = z\,g_0^*\left(-z^{-1}\right), \qquad \Longrightarrow \qquad g_1(k) = (-1)^k\,g_0\left(1-k\right), \quad k\in\mathbb{Z},$$

womit wir das sinc –Wavelet als

$$\psi(x) = g_1 * \phi\,(2\cdot) = \sum_{k\in\mathbb{Z}}(-1)^k\,\frac{\sin(1-k)}{(1-k)\pi}\,\frac{\sin x - k}{x-k}$$

erhalten. Laut [63, S. 221] ist dieses Wavelet auch als *Littlewood–Paley–Wavelet* bekannt, signaltheoretisch entspricht es einer *Octave–Band–Zerlegung*.

Weia!
Waga!
Woge, du Welle,
walle zur Wiege!
Wagalaweia!
Wallala weiala weia!

R. Wagner, *Das Rheingold*

5 Ecken, Kanten, Wavelets

Gut, wir haben jetzt also Subband–Coding und seine Verwandten kennengelernt und können die Wavelettransformierte einer (diskreten) Funktion mit unserer Filterbank bestimmen. Die "Magie" dieser Zerlegung besteht nun darin, daß wir sie unter gewissen Voraussetzungen[126] für

- Kompression
- Eckenerkennung
- Entrauschen

benutzen können. Der Grund ist eigentlich ganz einfach: *Waveletkoeffizienten* sind dort "groß", wo die Funktion Defizite bei der Differenzierbarkeit hat. Um das ein wenig klarer zu machen, brauchen wir aber erst noch etwas Theorie.

5.1 Polynome und Strang–Fix

Um es nochmals klarzustellen: Eigentlich haben wir nur die Filter zur Verfügung, aber über Subdivision konnten wir[127] dem Filter die Funktion ϕ zuordnen, die nun wieder zu einem Signal c die Funktion $\phi * c$ liefert, sozusagen durch "virtuelle Iteration" des Tiefpass–Synthese–Filters.

Die erste Frage, mit der wir uns befassen ist, unter welchen Voraussetzungen wir *Polynome* durch $\phi * c$ generieren können, also wann

$$\Pi_n \subseteq \mathbb{S}(\phi) := \{\phi * c \ : \ c \in \ell(\mathbb{Z})\} \tag{5.1}$$

ist. Der Grund ist relativ einfach: ϕ hat kompakten Träger und wenn nun eine Funktion f glatt, genauer gesagt, $n+1$–mal differenzierbar ist, dann können wir sie lokal durch ihr Taylorpolynom annähern, das Taylorpolynom durch $\phi * c$ und so approximiert ϕ also f lokal. Die kleine

[126] Aber was geht in der Mathematik schon ohne Voraussetzungen?
[127] Natürlich wieder nur unter gewissen Voraussetzungen.

Abweichung zwischen f und $\phi * c$ stellen wir durch das Wavelet dar und weil es nur ein geringer Fehler ist, müssen die Waveletkoeffizienten klein sein. Das ist, grob gesagt, das Programm für dieses Kapitel.

Definition 5.1 *Ein Menge F von Funktionen heißt* translationsinvariant[128]*, wenn*

$$f \in F \qquad \Leftrightarrow \qquad f(\cdot + k) \in F, \quad k \in \mathbb{Z}.$$

Der von ϕ erzeugte translationsinvariante *Raum wird geschrieben als*

$$\mathbb{S}(\phi) = \{\phi * c \ : \ c \in \ell(\mathbb{Z})\}.$$

Übung 5.1 Zeigen Sie:

1. $\mathbb{S}(\phi)$ ist translationsinvariant.

2. Die Polynomräume Π_n sind translationsinvariant für $n \in \mathbb{N}_0$.

♢

Wenn wir uns also nun mit der Frage befassen wollen, unter welchen Voraussetzungen $\Pi_n \subseteq \mathbb{S}(\phi)$ gilt, dann kann es nicht schaden, sich zuerst einmal zu überlegen, was das wohl für Signale sein werden, die Polynome erzeugen. Und warum sollten wir es nicht einfach mal mit Polynomen versuchen – das folgende Resultat von de Boor zeigt uns, daß wir damit gar nicht so falsch liegen.

Proposition 5.2 *Sei $\phi \in C_{00}(\mathbb{R})$ eine stetige Funktion mit kompaktem Träger*[129]*. Ist $\Pi_n \subset \mathbb{S}(\phi)$, dann ist für jedes $p \in \Pi_n$ auch*[130]

$$\phi * p = \sum_{j \in \mathbb{Z}} \phi(\cdot - j)\ p(j) \quad \in \Pi_n \qquad \textit{und} \qquad \deg \varphi * p \le \deg p. \tag{5.2}$$

Mit anderen Worten: "Produktion" und "Reproduktion" von Polynomen sind beinahe äquivalent.

[128] Im englischen Sprachgebrauch hat sich der Begriff *shift invariant* eingebürgert.

[129] Auch wenn wir beispielsweise in der Norm von L_2 approximieren wollen, werden die Funktionen, mit denen wir das tun wollen, doch zumeist zu $C_{00}(\mathbb{R})$ gehören; da diese Funktionen normalerweise inverse Fouriertransformationen sind, siehe Proposition 4.21, dann bleibt ihnen ja sowieso nichts anderes übrig, als gleichmäßig stetig zu sein und wenn man "richtig" mit ihnen rechnen will, dann muß man entweder kompakten Träger oder ziemlich flottes Abklingen fordern.

[130] Die Faltung in (5.2) ist zuerst einmal mehrdeutig, aber hier wollen wir Polynome immer nur im "diskreten Sinne" mit Funktionen falten, also Polynome p hier als die Folgen $(p(j) \ : \ j \in \mathbb{Z}) \in \ell(\mathbb{Z})$ auffassen.

Beweis: Sei $\Pi_n \ni p = \phi * c$ für ein $c \in \ell(\mathbb{Z})$, das ja existieren muß, weil $\Pi_n \subset \mathbb{S}(\phi)$. Dann ist

$$\begin{aligned}
\phi * p &= \sum_{j\in\mathbb{Z}} \phi(\cdot - j)\; p(j) = \sum_{j\in\mathbb{Z}} \phi(\cdot - j) \sum_{k\in\mathbb{Z}} \phi(j-k)\, c(k) \\
&= \sum_{j,k\in\mathbb{Z}} \phi(\cdot - j - k)\; \phi(j)\, c(k) = \sum_{j\in\mathbb{Z}} \phi(j) \sum_{k\in\mathbb{Z}} \underbrace{\phi(\cdot - j - k)\; c(k)}_{\phi * c(\cdot - j) = p(\cdot - j)} \\
&= \sum_{j\in\mathbb{Z}} \phi(j)\, p(\cdot - j). \qquad (5.3)
\end{aligned}$$

Ist nun $p \in \Pi_n$, so ist auch $p(\cdot - j)$ ein Polynom[131] vom selben Grad wie p in Π_n und da φ kompakten Träger hat, ist die Summe auf der rechten Seite eine *endliche* Linearkombination von Polynomen vom Grad $\leq \deg p$, also wieder in ein Polynom vom Grad $\leq \deg p$. □

Was natürlich schön wäre, das wäre $\deg \varphi * p = \deg p$ in (5.2), nur leider können wir das nicht erwarten, wie das folgende Beispiel zeigt.

Beispiel 5.3 *Sei*

$$\phi := \begin{cases} -1, & x \in [-1, 0), \\ 1, & x \in [0, 1), \\ 0, & sonst, \end{cases}$$

dann ist

$$2 = \sum_{j\in\mathbb{Z}} (-1)^j\, \phi(\cdot - j),$$

also $\Pi_0 \subset \mathbb{S}(\phi)$*, aber*

$$\phi * 1 = \sum_{j\in\mathbb{Z}} \phi(\cdot - j) = 0,$$

der Grad wird also echt "kleiner"! Nun gut, dieses Funktion φ *ist ja auch nicht stetig, aber erstens wurden im obigen Beweis ja eigentlich nur kompakter Träger und* gleichmäßige Beschränktheit *von* φ *verwendet und zweitens kann man natürlich auch Beispiele höherer Ordnung angeben. Die sind halt dann bloß nicht mehr so einfach.*

Korollar 5.4 *Erfüllt* ϕ *neben den Voraussetzungen von Proposition 5.2 auch noch*

$$0 \neq (\phi * 1)\,(0) = \sum_{j\in\mathbb{Z}} \phi(j), \qquad (5.4)$$

so ist $\deg \phi * p = \deg p$ *für alle* $p \in \Pi_n$*.*

Beweis: Nehmen wir der Einfachheit halber an, daß $p(x) = x^k + \cdots$, $k \leq n$, was man durch geeignete Normierung ja immer erreichen kann. Daß $\deg \varphi * p \leq \deg p = k$ ist, das wissen wir

[131] Ja, die Polynome von einem bestimmten Höchstgrad bilden einen *endlichdimensionalen* translationsinvarianten Raum – wer hätte gedacht, daß es so was gibt?

ja schon aus Proposition 5.2. Würde aber die strikte Ungleichung "$<$" gelten, so ergibt (5.3) daß

$$0 = (\phi * p)^{(k)} = \sum_{j\in\mathbb{Z}} \phi(j)\ \underbrace{p^{(k)}(\cdot - j)}_{=k!} = k! \sum_{j\in\mathbb{Z}} \phi(j) \neq 0,$$

was natürlich einen Widerspruch darstellt. □

Jetzt aber wird es Zeit, die Polynomreproduktion durch Eigenschaften der Funktion ϕ zu beschreiben. Dazu definieren wir diese magischen Eigenschaften am besten erst einmal.

Definition 5.5 (Strang–Fix–Bedingungen) *Eine Funktion $f \in L_1(\mathbb{R})$ erfüllt die* Strang–Fix–Bedingungen *der Ordnung $r \geq 0$ wenn*

1. $\widehat{f}(0) \neq 0$.
2. *für $j = 0, \ldots, r$ ist*
$$\widehat{f}^{(j)}(2k\pi) = 0, \qquad k \in \mathbb{Z} \setminus \{0\}. \tag{5.5}$$

Bemerkung 5.6 *1. Diese Bedingungen an die Fouriertransforierte einer Funktion wurden erstmals von Schoenberg*[132] *in [49] aufgestellt und untersucht, ihren Namen haben sie jedoch von der Arbeit [59]*[133].

2. Die erste der Strang–Fix–Bedingungen, also $\widehat{f}(0) \neq 0$, haben wir schon einmal gesehen, nämlich in Übung 4.3. Da sorgte es einfach dafür, daß

$$\widehat{\phi}(\xi) = \left[\prod_{j=1}^{\infty} \widehat{g}\left(2^{-j}\xi\right)\right] \widehat{\phi}(0)$$

zu einer vernünftigen Funktion[134] *führt.*

Satz 5.7 (Strang–Fix) *Erfüllt die Funktion $\phi \in C_{00}(\mathbb{R})$ die Strang–Fix–Bedingungen der Ordnung n, dann ist $\Pi_n \subseteq \mathbb{S}(\phi)$.*

Wir werden Satz 5.7 sogar in einer etwas allgemeineren Form beweisen, indem wir Erhaltung polynomialer Räume für solche translationsinvarianten Räume beweisen, die (5.4) erfüllen.

Satz 5.8 *Sei $\phi \in C_{00}(\mathbb{R})$ und $\phi * 1(0) \neq 0$. Dann ist*

$$\Pi_n \subseteq \mathbb{S}(\phi) \qquad \Longleftrightarrow \qquad \widehat{\phi}^{(j)}(2k\pi) = 0, \quad j = 0, \ldots, n,\ k \in \mathbb{Z} \setminus \{0\}. \tag{5.6}$$

[132] Isaac J. Schoenberg, 1903–1990, Studium der Mathematik in Berlin und Göttingen, befasste sich unter anderem mit analytischer Zahlentheorie, totaler Positivität, isometrischen Einbettungen metrischer Räume in Hilberträume. In der Numerik als "Vater der Splines" am bekanntesten. Schoenberg war mit Landaus Tochter Charlotte verheiratet, seine Schwester mit Hans Rademacher.

[133] Es ist bemerkenswert, daß dies eine der wenigen mathematischen Arbeiten ist, deren Autoren *nicht* in alphabetischer Reihenfolge aufgeführt werden.

[134] Ist $\widehat{\phi} = 0$, dann kann auch ϕ nur die Nullfunktion sein und das ist etwas dürftig.

Satz 5.7 folgt nun aus Satz 5.8 als unmittelbare Anwendung der Poissonschen Summenformel (1.21), denn erfüllt ϕ auch nur die Strang–Fix–Bedingungen der Ordnung 0, so ist

$$\phi * 1 = \sum_{k\in\mathbb{Z}} \phi(k) = \sum_{k\in\mathbb{Z}} \widehat{\phi}(2k\pi) = \underbrace{\widehat{\phi}(0)}_{\neq 0} + \sum_{k\in\mathbb{Z}\setminus\{0\}} \underbrace{\widehat{\phi}(2k\pi)}_{=0} \neq 0 \tag{5.7}$$

und Satz 5.7 ist gerade die Richtung "$\Longleftarrow$" von Satz 5.8.

Beweis von Satz 5.8: Beginnen wir mit der Richtung "$\Rightarrow$", für die wir die Voraussetzung $\phi * 1 \neq 0$ noch nicht einmal brauchen werden. Dabei betrachten wir für $j = 0, \ldots, n$ und festes $x \in \mathbb{R}$ die Funktion $\psi(t) = (-t)^j\, \phi(x+t)$ und erhalten, da $\psi \in C_{00}(\mathbb{R})$, über die Poissonsche Summenformel (1.21), (1.8) und (1.13), daß

$$\begin{aligned} p(x) \;&:=\; \sum_{k\in\mathbb{Z}} k^j \phi(x-k) = \sum_{k\in\mathbb{Z}} \psi(k) = \sum_{k\in\mathbb{Z}} \widehat{\psi}(2k\pi) = \sum_{k\in\mathbb{Z}} \left((-\cdot)^j\, \phi(\cdot + x)\right)^{\wedge}(2k\pi) \\ &=\; \sum_{k\in\mathbb{Z}} i^j\, \frac{d^j}{d\xi^j} \left(\widehat{\phi}(\xi) e^{ix\xi}\right)(2k\pi) = \sum_{k\in\mathbb{Z}} i^j \sum_{\ell=0}^{j} \binom{j}{\ell} \widehat{\phi}^{(\ell)}(2k\pi)\; (ix)^{j-\ell}\; e^{2i\pi kx}, \end{aligned}$$

also

$$\sum_{k\in\mathbb{Z}} k^j \phi(x-k) = \sum_{k\in\mathbb{Z}} i^j \sum_{\ell=0}^{j} \binom{j}{\ell} \widehat{\phi}^{(\ell)}(2k\pi)\; (ix)^{j-\ell}\; e^{2i\pi kx}, \qquad x \in \mathbb{R}. \tag{5.8}$$

Per Induktion über j können wir annehmen, daß alle Terme mit $\ell < j$ in der Summe auf der rechten Seite verschwinden[135], und so erhalten wir, daß für jedes $x \in \mathbb{R}$

$$p(x) = \sum_{k\in\mathbb{Z}} i^j\, \widehat{\phi}^{(j)}(2k\pi)\, e^{2i\pi kx}$$

ist. Die Funktion auf der linken Seite ist ein Polynom[136] in x nach Proposition 5.2, was auf der rechten Seite steht hingegen periodisch mit Periode 1, da $e^{2ik\pi} = 1$ für alle $k \in \mathbb{Z}$ und deswegen bleibt p nichts anderes übrig, als konstant zu sein, was aber gerade $\widehat{\phi}^{(j)}(2k\pi) = 0$ für $k \neq 0$ impliziert.

Für die andere Richtung, "$\Leftarrow$", verwenden wir nochmals (5.8) und erhalten, nach Einsetzen der Strang–Fix–Bedingungen $\widehat{\phi}^{(j)}(2k\pi) = 0$, $k \in \mathbb{Z} \setminus \{0\}$, daß[137]

$$\begin{aligned} &\sum_{k\in\mathbb{Z}} k^j \phi(x-k) \\ &\quad= \; i^j \sum_{\ell=0}^{j} \binom{j}{\ell} \widehat{\phi}^{(\ell)}(0)\; (ix)^{j-\ell} + \sum_{k\in\mathbb{Z}\setminus\{0\}} i^j \sum_{\ell=0}^{j} \binom{j}{\ell} \underbrace{\widehat{\phi}^{(\ell)}(2k\pi)}_{=0}\; (ix)^{j-\ell}\; e^{2i\pi kx} \\ &\quad= \; \sum_{\ell=0}^{j} \underbrace{i^{j+\ell} \binom{j}{\ell} \widehat{\phi}^{(j-\ell)}(0)}_{=:c_\ell}\; x^\ell = \sum_{\ell=0}^{j} c_\ell\, x^\ell \quad \in \Pi_j, \end{aligned}$$

[135] Der Induktionsanfang, $j = 0$, ist trivialerweise erfüllt, denn dann haben wir einfach keine Bedingung.

[136] Vom Grad $\leq n$, aber das ist eher sekundär.

[137] Das ist auch die wesentliche Idee hinter den Strang–Fix–Bedingungen: Beim Übergang zur Fouriertransformierten korrespondieren Multiplikation mit Polynomen und Ableitungen, siehe Satz 1.8.

und da $c_j = (-1)^j \widehat{\phi}(0) \neq 0$ ist[138], ist $\deg p = j$. Damit sind aber die Polynome

$$p_j := \sum_{k\in\mathbb{Z}} k^j \phi(\cdot - k), \qquad j = 0, \ldots, n,$$

linear unabhängig und bilden demzufolge eine Basis von Π_n. Insbesondere gilt dann aber auch $\Pi_n \subset \mathbb{S}(\phi)$. □

Beispiel 5.9 *Hier ein paar Beispiele für Funktionen, die die Strang–Fix–Bedingungen erfüllen (oder auch nicht):*

1. *Der (kardinale) B–Spline N_j erfüllt eine Strang–Fix–Bedingung der Ordnung j. Da*

$$\widehat{N_j}(\xi) = \left(\frac{1 - e^{-i\xi}}{i\xi}\right)^{j+1} =: \psi^{j+1}(\xi), \qquad \textit{also} \qquad \widehat{N_j}(0) = 1$$

mit $\psi(2k\pi) = 0$, hat $\widehat{N_j}$ an $2k\pi$, $k \in \mathbb{Z}\setminus\{0\}$ eine Nullstelle der Ordnung j, also

$$\widehat{N_j}^{(\ell)}(2k\pi) = 0, \qquad \ell = 0, \ldots, j, \quad k \in \mathbb{Z}\setminus\{0\}.$$

Damit erfüllen die B–Splines die Strang–Fix–Bedingungen und besitzen die Fähigkeit zur Polynomreproduktion. Was nicht verwunderlich ist, sind doch alle Polynome trivialerweise auch stückweise Polynome.

2. *Wie sieht es nun mit der Funktion ϕ aus Beispiel 5.3 aus? Da wir in L_1 auch $\phi = \chi - \chi(\cdot + 1)$ schreiben können[139], erhalten wir, daß*

$$\widehat{\phi}(\xi) = \widehat{\chi}(\xi) - e^{i\xi}\widehat{\chi}(\xi) = \left(1 - e^{i\xi}\right)\left(\frac{1 - e^{-i\xi}}{i\xi}\right).$$

An den Stellen $2k\pi$ verschwinden also tatsächlich $\widehat{\phi}$, was aber für die Probleme sorgt, ist die Tatsache, daß hier auch $\widehat{\phi}(0) = 0$ ist.

3. *Das ändert sich auch nicht groß, wenn wir die immer glatteren Funktionen*

$$\psi_j := \underbrace{\chi * \cdots * \chi}_{j} * \phi = N_{j-1} * \phi, \qquad j \in \mathbb{N}, \qquad \psi_0 = \phi, \tag{5.9}$$

einführen, also

$$\begin{aligned}\widehat{\psi_j}(\xi) &= \widehat{\phi}(\xi)\,\widehat{N_{j-1}}(\xi) = \left(1 - e^{i\xi}\right)\left(\frac{1 - e^{-i\xi}}{i\xi}\right)\left(\frac{1 - e^{-i\xi}}{i\xi}\right)^j \\ &= \left(1 - e^{i\xi}\right)\left(\frac{1 - e^{-i\xi}}{i\xi}\right)^{j+1} = \left(1 - e^{i\xi}\right)\widehat{N_j}(\xi),\end{aligned}$$

[138] Hier gehen wieder die Annahme $\phi * 1 \neq 0$ und die Poissonsche Summationsformel ein, genauer, Gleichung (5.7).

[139] Funktionen in L_p–Räumen sind nur bis auf Menge vom Maß Null eindeutig bestimmt!

die sogar

$$\widehat{\psi_j}^{(\ell)}(2k\pi) = 0, \qquad \ell = 0, \ldots, j+1, \quad k \in \mathbb{Z} \setminus \{0\}$$

erfüllen, aber wegen $\widehat{\psi_j}(0) = 0$ *immer noch knapp an den Strang–Fix–Bedingungen scheitern.*

4. *Trotzdem: Der letzte Teil der Bemerkung zeigt uns, wie wir die "Strang–Fix–Ordnung" einer Funktion erhöhen können, nämlich indem wir sie mit charakteristischen Funktionen oder eben gleich mit B–Splines falten.*

Nun ist das mit den Strang–Fix–Bedingungen ja alles schön und gut, aber bisher haben wir es mit Eigenschaften von *Funktionen* zu tun und haben unseren Filter eigentlich völlig ignoriert! Andererseits ist ja ϕ aber nicht irgendeine Funktion[140], sondern ist durch den Filter $g = g_0$ definiert:

$$\widehat{\phi}(\xi) = \frac{1}{2}\,\widehat{g}_0\left(\frac{\xi}{2}\right)\,\widehat{\phi}\left(\frac{\xi}{2}\right), \qquad \xi \in \mathbb{R}. \tag{5.10}$$

Setzen wir in (5.10) $\xi = 0$, dann erhalten wir, daß

$$\widehat{\phi}(0) = \frac{1}{2}\,\widehat{g}_0(0)\,\widehat{\phi}(0)$$

und die erste Strang–Fix–Bedingung, $\widehat{\phi}(0) \neq 0$, liefert sogar, daß $\widehat{g}_0(0) = 2$ sein muß. Mit $\xi = 2k\pi$ liefert die Strang–Fix–Bedingung der Ordnung 0, daß

$$0 = \widehat{\phi}(2k\pi) = \frac{1}{2}\,\widehat{g}_0\,(k\pi)\;\widehat{\phi}(k\pi) = \frac{1}{2}\,\widehat{\phi}(k\pi)\;\begin{cases} 2, & k \in 2\mathbb{Z}, \\ \widehat{g}_0\,(\pi)\,, & k \in 2\mathbb{Z}+1, \end{cases}$$

denn schließlich ist $\widehat{g}_0$ ja 2π–periodisch. Wäre also $\widehat{g}_0(\pi) \neq 0$, dann müsste $\widehat{\phi}\,(\pi + 2k\pi) = 0$ sein für alle $k \in \mathbb{Z}$. Das schließen wir einfach mal aus und erhalten die folgende Beobachtung.

Lemma 5.10 *Ist* $\left[\widehat{\phi}\,(\pi + 2k\pi)\;:\;k \in \mathbb{Z}\right] \neq 0$*, dann ist* $\widehat{g}_0(\pi) = 0$*.*

Als nächstes differenzieren wir (5.10) und erhalten, daß

$$\widehat{\phi'}(\xi) = \frac{d}{d\xi}\left[\frac{1}{2}\,\widehat{g}_0\left(\frac{\xi}{2}\right)\,\widehat{\phi}\left(\frac{\xi}{2}\right)\right] = \frac{1}{4}\left[\widehat{g_0'}\left(\frac{\xi}{2}\right)\,\widehat{\phi}\left(\frac{\xi}{2}\right) + \widehat{g}_0\left(\frac{\xi}{2}\right)\,\widehat{\phi'}\left(\frac{\xi}{2}\right)\right] \tag{5.11}$$

Mit $\xi = 0$ und unter Berücksichtigung von $\widehat{g}_0(0) = 2$ erhalten wir somit, daß

$$\widehat{\phi'}(0) = \frac{1}{4}\,\widehat{g_0'}(0)\,\widehat{\phi}(0) + \frac{1}{2}\,\widehat{\phi'}(0) \qquad \Rightarrow \qquad \widehat{g_0'}(0) = 2\frac{\widehat{\phi'}(0)}{\widehat{\phi}(0)},$$

[140] Für den Fall, daß es niemand gemerkt hat: Die Strang–Fix–Bedingungen gelten für *beliebige* Funktionen, das einzige, worum es dabei wirklich geht, ist die Tatsache daß wir es mit ganzzahligen Translaten einer Funktion zu tun haben. Insofern ist es kein Wunder, daß die Bedingungen in [49] erstmals für Splines aufgestellt wurden.

was uns nicht wirklich weiterhilft, während $\xi = 2k\pi$, $k \neq 0$ wieder unter Verwendung der Annahme $\left[\widehat{\phi}\,(\pi + 2k\pi) \;:\; k \in \mathbb{Z}\right] \neq 0$, uns

$$\begin{aligned} 0 &= \widehat{\phi}'(2k\pi) = \frac{1}{4}\widehat{g}_0'\,(k\pi)\;\widehat{\phi}\,(k\pi) + \frac{1}{4}\underbrace{\widehat{g}_0\,(k\pi)}_{=0,\;k\in 2\mathbb{Z}+1}\underbrace{\widehat{\phi}'\,(k\pi)}_{=0,\;k\in 2\mathbb{Z}} = \frac{1}{4}\widehat{g}_0'\,(k\pi)\;\widehat{\phi}\,(k\pi) \\ &= \begin{cases} 0, & k \in 2\mathbb{Z}, \\ \widehat{g}_0'(\pi)\,\widehat{\phi}(k\pi), & k \in 2\mathbb{Z}+1, \end{cases} \end{aligned}$$

und unsere Annahme liefert nun auch, daß $\widehat{g}_0'(\pi) = 0$ sein muss. Wenn wir ein wenig systematischer an die Sache herangehen, dann erhalten wir schließlich die folgende Bedingung an unseren Filter.

Satz 5.11 *Erfüllt $\phi \neq 0$ die Zwei–Skalen–Gleichung (5.10) und*

$$\left[\widehat{\phi}\,(\pi + 2k\pi) \;:\; k \in \mathbb{Z}\right] \neq 0, \tag{5.12}$$

Dann erfüllt ϕ die Strang–Fix–Bedingungen der Ordnung $r \geq 0$ genau dann, wenn

$$\widehat{g}_0(0) = 2, \qquad \widehat{g}_0^{(k)}(\pi) = 0, \quad k = 0, \ldots, r. \tag{5.13}$$

Beweis: Für $k = 0, \ldots, r$ betrachten wir die k–te Ableitung von (5.10) und erhalten mit der Leibniz–Regel, daß

$$\widehat{\phi}^{(k)}(\xi) = \frac{1}{2^{k+1}} \sum_{j=0}^{k} \binom{k}{j} \widehat{g}_0^{(j)}\left(\frac{\xi}{2}\right)\,\widehat{\phi}^{(k-j)}\left(\frac{\xi}{2}\right). \tag{5.14}$$

Nehmen wir nun an, daß die Strang–Fix–Bedingungen erfüllt sind, und daß wir bereits für ein $k \leq r$ gezeigt haben, daß $\widehat{g}_0^{(j)}(\pi) = 0$ für $j < k$ erfüllt sein muss. Dann liefert (5.14), daß für $\ell \in \mathbb{Z}$

$$0 = \widehat{\phi}^{(k)}(2\ell\pi) = \frac{1}{2^{k+1}}\left[\widehat{g}_0^{(k)}\,(\ell\pi)\;\widehat{\phi}\,(\ell\pi) + \sum_{j=0}^{k-1}\binom{k}{j}\widehat{g}_0^{(j)}\,(\ell\pi)\;\widehat{\phi}^{(k-j)}\,(\ell\pi)\right]$$

und die Summe hat den Wert Null, da für ungerade k die g_0–Terme, für gerade k hingegen die ϕ–Terme verschwinden – und das war dann auch schon der Induktionsschritt.

Auch die Umkehrung erfolgt induktiv – wir nehmen jetzt an, daß wir für ein k bereits $\widehat{\phi}^{(j)}(2\ell\pi) = 0$ gezeigt haben, dann erhalten wir für $\ell \neq 0$ aus (5.14), daß

$$\begin{aligned} \widehat{\phi}^{(k)}(2\ell\pi) &= \frac{1}{2^{k+1}}\left[\widehat{g}_0\,(\ell\pi)\;\widehat{\phi}^{(k)}\,(\ell\pi) + \sum_{j=1}^{k}\binom{k}{j}\widehat{g}_0^{(j)}\,(\ell\pi)\;\widehat{\phi}^{(k-j)}\,(\ell\pi)\right] \\ &= \frac{1}{2^{k+1}}\,\widehat{g}_0\,(\ell\pi)\;\widehat{\phi}^{(k)}\,(\ell\pi) = \begin{cases} 2^{-k}\widehat{\phi}^{(k)}(\ell\pi), & \ell \in 2\mathbb{Z}, \\ 0, & \ell \in 2\mathbb{Z}+1. \end{cases} \end{aligned}$$

Schreiben wir also $\ell = 2^m \ell'$, $\ell' \in 2\mathbb{Z}+1$, dann ist schließlich

$$\widehat{\phi}^{(k)}(2\ell\pi) = 2^{-km}\widehat{\phi}^{(k)}\left(2\ell'\pi\right) = 0,$$

was den Beweis komplettiert. □

Durch Übergang vom trigonometrischen Polynom zum Laurentpolynom und Ausnutzung der Tatsache, daß bei (Laurent-)Polynomen[141] Nullstellen abfaktorisiert werden können, erhalten wir die folgenden Beschreibungen der Strang–Fix–Bedingungen.

Korollar 5.12 *Für eine Funktion die die Zweiskalengleichung (4.36) und (5.12) erfüllt, sind die folgenden Bedingungen äquivalent:*

1. *ϕ erfüllt die Strang–Fix–Bedingungen der Ordnung r.*

2. *Es gilt*
$$g_0^*(1) = 2, \qquad g_0^{*(k)}(-1) = 0, \quad k = 0, \ldots, r. \tag{5.15}$$

3. *g_0^* ist faktorisierbar als*[142]
$$g_0^*(z) = \left(\frac{z+1}{2}\right)^{r+1} h(z), \qquad h(1) = 2. \tag{5.16}$$

Beweis: Wir betrachten einfach die Ableitungen

$$\widehat{g}_0'(\xi) = \frac{d}{d\xi} g_0^*\left(e^{i\xi}\right) = \left.\frac{d}{dz} g_0^*(z)\right|_{z=e^{i\xi}} \frac{d}{d\xi} e^{i\xi} = i\, e^{i\xi}\, \left(g_0^*\right)'\left(e^{i\xi}\right),$$

was sich per Induktion zu

$$\widehat{g}_0^{(k)}(\xi) = \sum_{j=1}^{k} \binom{k}{j} e^{i(k-j+1)\xi}\, \left(g_0^*\right)^{(j)}\left(e^{i\xi}\right), \tag{5.17}$$

erweitern lässt und uns die Darstellung

$$\left[\widehat{g}_0^{(k)} \;:\; k = 1, \ldots, r\right](\xi) = L(\xi)\, \left[\left(g_0^*\right)^{(k)} \;:\; k = 1, \ldots, r\right]\left(e^{i\xi}\right) \tag{5.18}$$

liefert, wobei $L(\xi)$ eine untere Dreiecksmatrix mit von Null verschiedenen Diagonalementen ist. Insbesondere ist $L(\xi)$ invertierbar und daher sind die Ableitungsbedingungen in (5.13) und (5.15) äquivalent. □

[141] Im Gegensatz zu trigonometrischen Polynomen.

[142] Man beachte: Genau so eine Zerlegung wird in Proposition 4.21 verwendet und das ist natürlich **kein** Zufall.

Bemerkung 5.13 *Hinter der Bedingung (5.12) aus Satz 5.11 steckt wesentlich mehr! Sie ist ein Spezialfall der Forderung*

$$\left[\widehat{\phi}\,(\xi + 2k\pi) \;:\; k \in \mathbb{Z}\right] \neq 0, \qquad \xi \in [0, 2\pi], \tag{5.19}$$

die mittels der Fouriertransformation die Stabilität *von ϕ beschreibt, also die Existenz von Konstanten $0 < A, B < \infty$, so daß*

$$A \left\|c\right\| \leq \left\|\phi * c\right\| \leq B \left\|c\right\| \tag{5.20}$$

gilt. Und das ist wichtig, denn es bedeutet, daß man beliebig zwischen der Funktion und den Koeffizienten hin- und herspringen darf und daß dieser Übergang stetig ist – man kann durchaus sagen, daß Stabilität eigentlich eine Minimalvoraussetzung für einen "vernünftigen" numerischen *Umgang mit ϕ ist. Insbesondere gilt für jede stabile Funktion ϕ, daß*

$$\phi * c = 0 \quad \Leftrightarrow \quad c = 0, \qquad \textit{bzw.} \qquad \phi * c = \phi * c' \quad \Leftrightarrow \quad c = c'$$

ist.

Beispiel 5.14 *Daß die Bedingung (5.12) im allgemeinen auch nicht fallengelassen werden kann, sieht man sehr schön am Beispiel der Funktion $\phi = \chi_{[0,2)}$. Die ist verfeinerbar:*

$$\phi = \chi_{[0,2]} = \chi_{[0,1]} + \chi_{[0,2]} = \phi\,(2\cdot) + \phi\,(2 \cdot -2)$$

also ist $g_0(0) = g_0(2) = 1$ und somit $g_0^*(z) = 1 + z^2$, also $g_0^*(1) = g_0^*(-1) = 2$ und damit sind (5.15) und (5.16) natürlich verletzt.

Die Fouriertransformation $\widehat{\phi}$ lässt sich andererseits sehr einfach als

$$\widehat{\phi}(\xi) = \int_0^2 e^{-i\xi t} dt = \frac{1 - e^{-2\xi}}{i\xi} = 2e^{-i\xi} \frac{e^{i\xi} - e^{-i\xi}}{2i\,\xi} = 2e^{-i\xi} \operatorname{sinc} \xi$$

berechnen, was bekanntlich an $k\pi$, $k \in \mathbb{Z} \setminus \{0\}$, verschwindet, siehe Abb. 1.1, und somit insbesondere (5.12) verletzt.

Zum Abschluß dieses Abschnitts gönnen wir uns noch eine *explizite* Formel für die Koeffizienten, die man braucht, um Polynome darzustellen.

Lemma 5.15 *Erfüllt ϕ die Strang–Fix–Bedingungen der Ordnung n, und sind p_j die Polynome, für die $\phi * p_j = (\cdot)^j$, $j = 0, \ldots, n$, gilt dann erhält man für die Funktionale*

$$\lambda(f)(k) := \sum_{j=0}^{n} \frac{f^{(j)}(k)}{j!}\, p_j(0), \qquad f \in C^n\left(\mathbb{R}\right), \quad k \in \mathbb{Z}, \tag{5.21}$$

daß[143]

$$p = Q_h p := \sigma_{h^{-1}}\left(\phi * \lambda\left(\sigma_h p\right)\right) = \left(\phi * \lambda\left(p(h\cdot)\right)\right)\left(h^{-1}\cdot\right), \qquad p \in \Pi_n. \tag{5.22}$$

[143] Man bezeichnet einen Operator der Form $f \mapsto \phi * \lambda(f) = \sum_k \lambda_k(f)\, \phi(\cdot - k)$ auch als *Quasiinterpolanten* – allerdings gibt es Leute, die schon bei der Erwähnung dieses Namens Ausschlag bekommen, denn die wenigsten Quasiinterpolanten interpolieren und sind somit also ziemlich "quasi".

Beweis: Es genügt, denn Fall $h = 1$ zu betrachten: Ist nämlich $Q_1 p = \phi * \lambda(p) = p$, so ist natürlich auch

$$Q_h\, \sigma_{h^{-1}} p = (\phi * \lambda\,(\sigma_h \sigma_{h^{-1}} p))\,\left(h^{-1}\cdot\right) = (Q_1 p)\,\left(h^{-1}\cdot\right) = \sigma_{h^{-1}} p,$$

und ersetzt man x durch hx in dieser Gleichung, so folgt $Q_h p = p$.
Für $\ell \in \mathbb{Z}$ und $j = 0, \ldots, n$ ist

$$(x-\ell)^j = (\phi * p_j)\,(\cdot - \ell) = \sum_{k\in\mathbb{Z}} \phi\,(\cdot - \ell - k)\; p_j(k) = \sum_{k\in\mathbb{Z}} \phi\,(\cdot - k)\; p_j\,(k-\ell) \tag{5.23}$$

Nun hat aber jedes $p \in \Pi_n$ an der Stelle ℓ die (endliche) Taylorentwicklung

$$p(x) = \sum_{j=0}^{n} \frac{p^{(j)}(\ell)}{j!}\; (x-\ell)^j \tag{5.24}$$

Setzen wir nun (5.23) in (5.24) ein, dann erhalten wir, daß

$$\begin{aligned} p(x) &= \sum_{j=0}^{n} \frac{p^{(j)}(\ell)}{j!} \sum_{k\in\mathbb{Z}} \phi\,(\cdot - k)\; p_j\,(k-\ell) = \sum_{k\in\mathbb{Z}} \phi\,(\cdot - k) \underbrace{\sum_{j=0}^{n} \frac{p^{(j)}(\ell)}{j!}\, p_j\,(k-\ell)}_{=:q_\ell(k)} \\ &= \phi * q_\ell(x), \end{aligned}$$

wobei $q_\ell \in \Pi_n$. Nach Übung 5.2 gilt dann für beliebige $\ell, \ell' \in \mathbb{Z}$, daß $q_\ell = q_{\ell'}$, also insbesondere

$$q_0(k) = q_k(k) = \sum_{j=0}^{n} \frac{p^{(j)}(k)}{j!}\, p_j(0), \qquad k \in \mathbb{Z},$$

woraus unmittelbar $p = \phi * q_0 = Q_1 p$, also (5.22) folgt. □

Übung 5.2 Zeigen Sie: Erfüllt φ die Strang–Fix–Bedingungen der Ordnung n, dann gilt für $p \in \Pi_n$

$$\phi * p \equiv 0 \qquad \Longleftrightarrow \qquad p = 0.$$

♢

5.2 Waveletkoeffizienten glatter Funktionen

Nun kommen wir zum eigentlichen Thema dieses Kapitels, nämlich der Approximationsgüte von Funktionen, die die Strang–Fix–Bedingung erfüllen, und welche Konsequenzen das für die Koeffizienten der Wavelet–Darstellung hat. Und wieder: **Eigentlich** kennen wir die Funktionen ja gar nicht, sondern haben nur die Filterbank zur Verfügung, aber das ist nicht so schlimm, denn

- die "magische" Funktion ϕ ergibt sich aus den Filtern[144] durch den bereits erwähnten Subdivision–Prozess,
- die Berechnung der Waveletkoeffizienten ist wieder nur ein Filterungsprozess.

Anders gesagt: Auch wenn wir Resultate für Funktionen beweisen, so sind diese Funktionen trotzdem nur ein *implizites* Hilfsmittel, aber eben ein hilfreiches. Wir beginnen mit einer Aussage über die *Approximationsgüte* von ϕ.

Satz 5.16 *Erfüllt die Funktion $\phi \in C_{00}(\mathbb{R})$ Strang–Fix–Bedingungen der Ordnung n, dann gibt es zu jedem $f \in C^{n+1}(\mathbb{R})$ und $h > 0$ ein $c_h \in \ell(\mathbb{Z})$, so daß*

$$\left\| f - \sigma_{1/h} \left(\phi * c_h \right) \right\|_p \leq C\, h^{n+1} \left\| f^{(n+1)} \right\|_p . \tag{5.25}$$

Wir können das auch noch anders sagen: Erfüllt ϕ die Strang–Fix–Bedingungen, dann erlaubt $\mathbb{S}(\phi)$ sehr gute Approximation an zwar nicht alle Funktionen[145], aber doch an differenzierbare Funktionen. In diesem Sinne ist Satz 5.16 ein "translationsinvariantes" Gegenstück zu den *Jackson–Sätzen* der Approximationstheorie, siehe z.B. [11, 35, 46], die für differenzierbare Funktionen die Approximationsordnung, also die Konvergenzordnung eine Approximationsoperators, mit der Glattheit der zu approximierenden Funktion in Verbindung bringen.

Die zweite Aussage ist für uns fast noch wichtiger, denn sie sagt uns, daß wir *lokal* glatte Funktionen daran erkennen können, daß ihre Waveletkoeffizienten für hohe Auflösungen schnell abfallen. Dazu schreiben wir eine Funktion f in ihrer unendlichen Waveletzerlegung

$$f = \phi * c + \sum_{k=0}^{\infty} \psi * d_k \left(2^k \cdot \right) ,$$

deren Existenz wir dank der guten Approximation von ϕ aus Satz 5.16 immer annehmen können, indem wir f durch $\mathbb{S}\left(\sigma_{2^k}\phi\right)$ approximieren und dafür dann jeweils die Waveletzerlegungen bestimmt.

Satz 5.17 *Erfüllt $\phi \in C_{00}(\mathbb{R})$ Strang–Fix–Bedingungen der Ordnung n und ist f in einer Umgebung U von $x^* \in \mathbb{R}$ differenzierbar von der Ordnung $n+1$, dann gilt für $k \in \mathbb{N}$ und $\ell \in 2^k x^* - \operatorname{supp}\phi$, daß*

$$|d_k(\ell)| \leq C 2^{-kn} \left\| f^{(n+1)} \right\|_{\infty,U} .$$

Beweis von Satz 5.16: Es genügt, zu fordern, daß $f \in C_{00}^{n+1}(\mathbb{R})$ ist – das folgt aus dem kompakten Träger von ϕ, denn wenn wir f an irgendeiner Stelle x approximieren, dann sind nur endlich viele Werte von $\lambda(f)(\cdot)$ "aktiv". Und in den L_p–Funktionen sind die stetigen Funktionen mit kompaktem Träger dicht. Außerdem nehmen wir noch an, daß $\operatorname{supp}\phi \subseteq [0, N]$, was wir für hinreichend großes N immer durch eine ganzzahlige Verschiebung von ϕ erreichen können, und die Verschiebung ist für den Raum $\mathbb{S}(\phi)$ irrelevant.

[144] Genauer: Aus dem Low–Pass–Synthesefilter.

[145] Wer würde das auch erwarten?

Für $x \in \mathbb{R}$ und $h > 0$ sei

$$T_n f := \sum_{j=0}^{n} \frac{f^{(j)}(x)}{j!} \, (\cdot - x)^j$$

das Taylor–Polynom der Ordnung n an f bezüglich der Stelle x. Dann ist, da $f(x) = T_n f(x) = Q_h \left(T_n f\right)(x)$

$$\begin{aligned} |f(x) - Q_h f(x)| &= \left| Q_h (\underbrace{f - T_n f}_{=:g})(x) \right| = |Q_h g(x)| = \left| \sum_{k \in \mathbb{Z}} \lambda \left(g(h\cdot)\right)(k) \, \phi\left(h^{-1}x - k\right) \right| \\ &= \left| \sum_{k \in x/h + (-N,0)} \lambda \left(g(h\cdot)\right)(k) \, \phi\left(h^{-1}x \cdot - k\right) \right| \end{aligned}$$

Für beliebiges $1 \le p < \infty$ und $q = (p-1)/p$, also $1/p + 1/q = 1$, ist dann, mit Übung 5.3,

$$|f(x) - Q_h f(x)|^p \le N^{p-1} \sum_{k \in x/h + (-N,0)} \left| \lambda \left(g(h\cdot)\right)(k) \, \phi\left(h^{-1}x \cdot - k\right) \right|^p \tag{5.26}$$

Als nächstes schauen wir uns mal den Ausdruck $\lambda \left(g(h\cdot)\right)(k)$, $k \in \mathbb{Z}$, an. Dazu bemerken wir zuerst einmal, daß für $y \in \mathbb{R}$ und $\ell = 0, \ldots, n$ die Gleichung[146]

$$\left(T_n f\right)^{(\ell)}(y) = \sum_{j=0}^{n} \frac{f^{(j)}(x)}{j!} \underbrace{\frac{d^\ell}{dy^\ell} (y-x)^j}_{j!/(j-\ell)!(y-x)^{j-\ell}} = \sum_{j=\ell}^{n} \frac{f^{(j)}(x)}{(j-\ell)!} (y-x)^{j-\ell} = \sum_{j=0}^{n-\ell} \frac{f^{(j+\ell)}}{j!} (y-x)^j,$$

also

$$\left(T_n f\right)^{(\ell)} = T_{n-\ell} f^{(\ell)}, \qquad \ell = 0, \ldots, n, \tag{5.27}$$

gilt. Da $g = f - T_n f$ ist, ergibt sich also nach (5.21)

$$\begin{aligned} \lambda \left(g(h\cdot)\right)(k) &= \sum_{j=0}^{n} \frac{p_j(0)}{j!} h^j g^{(j)}(hk) = \sum_{j=0}^{n} \frac{h^j \, p_j(0)}{j!} \left(f - T_n f\right)^{(j)}(hk) \\ &= \sum_{j=0}^{n} \frac{h^j \, p_j(0)}{j!} \left(f^{(j)} - T_{n-j} f^{(j)}\right)(hk). \end{aligned}$$

Nach der *Taylor–Formel mit Integralrestglied*,

$$\left(f - T_n f\right)(y) = \frac{1}{n!} \int_x^y f^{(n+1)}(t) \, (y-t)^n \, dt, \qquad f \in C^{n+1}(\mathbb{R}), \quad y \in \mathbb{R}, \tag{5.28}$$

[146] Unter Verwendung der Konvention, daß $j! = \infty$ für $j < 0$, also insbesondere $1/j! = 0$.

siehe z.B. [28, S. 285][147] oder Übung 5.4, ist somit

$$\lambda\left(g(h\cdot)\right)(k) = \sum_{j=0}^{n} \frac{h^j\, p_j(0)}{j!(n-j)!} \int_x^{hk} f^{(n+1)}(t)\,(hk-t)^{n-j}\,dt. \tag{5.29}$$

Jetzt können wir unsere Bausteinchen zusammensetzen! Da die Summe nur über solche Kombinationen x, k läuft, für die $x/h - k \in [0, N]$, also $x - hk \in h[0, N]$, also $|x - hk| \le Nh$ gilt, können wir nun (5.26) nach x integrieren, (5.29) einsetzen, um so

$$\begin{aligned}
\|f - Q_h f\|_p^p &= \int_{\mathbb{R}} |f(x) - Q_h f(x)|^p\,dx \\
&\le N^{p-1} \int_{\mathbb{R}} \sum_{k\in x/h+(-N,0)} \left|\lambda\left(g(h\cdot)\right)(k)\,\phi\left(h^{-1}x-k\right)\right|^p dx \\
&\le ((n+1)N)^{p-1} \int_{\mathbb{R}} \sum_{k\in x/h+(-N,0)} \sum_{j=0}^{n} \underbrace{\left|\frac{p_j(0)}{j!(n-j)!}\right|^p}_{\le C^p} \times \\
&\qquad \times \left|h^j \int_0^{hk-x} f^{(n+1)}(x+t)\,(hk-x-t)^{n-j}\,dt\;\phi\left(h^{-1}x-k\right)\right|^p dx \\
&\le ((n+1)N)^{p-1} C^p \int_{\mathbb{R}} \sum_{k\in x/h+[-N,0]} \sum_{j=0}^{n} h^{jp} \underbrace{|x-hk|^{p-1}}_{\le (Nh)^{p-1}} \int_0^{kh-x} \underbrace{\left|(hk-x-t)^{n-j}\right|^p}_{\le (2Nh)^{(n-j)p} \le (2N)^{np} h^{(n-j)p}} \times \\
&\qquad \times \left|f^{(n+1)}(x+t)\right|^p \underbrace{\left|\phi\left(h^{-1}x-k\right)\right|^p}_{\le \|\phi\|_\infty^p}\,dt\,dx \\
&\le \underbrace{(n+1)^p N^p C^p N^{p-1} N^{np}\, 2^{np}\, \|\phi\|_\infty^p}_{=:C_1^p}\, h^{np}\, h^{p-1} \int_{\mathbb{R}} \int_0^{hN} \left|f^{(n+1)}(x+t)\right|^p dt\,dx \\
&= C_1^p h^{np}\, h^{p-1} \int_0^{hN} \int_{\mathbb{R}} \left|f^{(n+1)}(x)\right|^p dx\,dt = N C_1^p\, h^{(n+1)p} \left\|f^{(n+1)}\right\|_p^p
\end{aligned}$$

zu erhalten, was (5.25) mit der Konstanten

$$C := 2^n N^{n+2}(n+1)\ \|\phi\|_\infty \tag{5.30}$$

liefert. □

Übung 5.3 Zeigen Sie, daß für jede Folge $a \in \ell(\mathbb{Z})$, jedes $1 \le p < \infty$ und $N \in \mathbb{N}$ die Abschätzung

$$\left|\sum_{j=1}^{N} a_j\right|^p \le N^{p-1} \sum_{j=1}^{N} |a_j|^p$$

[147] Wer hofft, dort die Lösung von Übung 5.4 zu finden, muß leider enttäuscht werden, denn es ist dort auch nur als Übungsaufgabe aufgelistet.

gilt. ♢

Übung 5.4 Beweisen Sie die Taylor–Formel (5.28).
Hinweis: Partielle Integration. ♢

Übung 5.5 Beweisen Sie Satz 5.16 für $p = \infty$, also

$$\sup_{x\in\mathbb{R}} \left| f(x) - \sigma_{1/h}\left(\phi * c_h\right)\right| \leq C\, h^{n+1} \sup_{x\in\mathbb{R}} \left| f^{(n+1)}(x)\right|.$$

♢

Lemma 5.18 *Ist ϕ stabil und erfüllt die Strang–Fix–Bedingungen der Ordnung n, dann gibt es Polynome $p_j \in \Pi_j$, $j = 0, \ldots, n$, so daß $S_G p_j = 2^j p_j$.*

Beweis: Wir normieren ϕ so, daß $\phi * 1 = 1$ und beginnen mit $n = 0$, wo wir $p_0 = 1$ wählen können, da

$$\phi * p_0 = 1 = \phi * p_0(2\cdot) = \phi * S_G p_0,$$

also wegen der Stabilität[148] $p_0 = S_G p_0$ ist.

Für allgemeines n wählen wir p_n so, daß $\phi * p_n(x) = x^n$ und erhalten daß

$$\phi * (2^n p_n) = 2^n \phi * p_n = (2\cdot)^n = \phi * p_n\,(2\cdot) = \phi * S_G p_n,$$

und das war's dann auch schon. □

Lemma 5.19 *Ist ϕ stabil und erfüllt die Strang–Fix–Bedingungen der Ordnung n, dann ist $0 = F_1 p =\downarrow_2 f_1 * p$ für alle $p \in \Pi_n$.*

Definition 5.20 *Erfüllt ein Filter F die Bedingung $F\Pi_n = 0$, dann sagt man, der Filter habe $n+1$* verschwindende Momente[149].

Anders gesagt bedeutet Lemma 5.19: Erfüllt die Skalierungsfunktion ϕ die Strang–Fix–Bedingungen der Ordnung n, dann hat F_1, der *duale Filter*, $n+1$ verschwindende Momente.

Beweis: Das Lemma ist eine ziemlich unmittelbare Konsequenz aus (4.46), denn da $\Pi_n \subset V_0 \subset V_1$ ist, liefert die Projektion eines Polynoms von V_1 auf V_0, daß

$$0 = c_1 =\downarrow_2 F_1 p_k, \qquad k = 0, \ldots, n,$$

sein muss. □

Beweis von Satz 5.17: Da ϕ endlichen Träger hat, gibt es $N \in \mathbb{N}$, so daß $\operatorname{supp}\phi \subseteq [-N, N]$ ist[150] und nehmen wir an, daß $f \in C^{n+1}(U)$ für eine Umgebung U von x^* ist. Für $k \in \mathbb{N}$ und $x \in \mathbb{R}$ ist nun

$$\phi * S_{2^{-k}} f\left(2^k x\right) = \sum_{\ell\in\mathbb{Z}} \phi\left(2^k x - \ell\right)\, f\left(2^{-k}\ell\right) = \sum_{\left|2^k x - \ell\right| \leq N} \phi\left(2^k x - \ell\right)\, f\left(2^{-k}\ell\right),$$

[148] Siehe Bemerkung 5.13.

[149] Zur Erinnerung: *Momente* sind das Ergebnis der Anwendung eines Operators (z.B. Integration) auf Monome und beschreiben das Verhalten des Operators auf den Polynomen.

[150] Hier ist der symmetrische Träger "angenehmer" als der asymmetrische $[0, N]$.

es gilt also

$$\ell \in 2^k x + [-N, N] \qquad \Rightarrow \qquad 2^{-k}\ell \in x + 2^{-k}\,[-N, N]\,,$$

und es geht also an jeder Stelle x nur *lokale* Information über f ein. Wählen wir k so groß, daß $x^* + 2^{-k}\,[-N, N] \subset U$ und verwenden wir eine Taylor–Entwicklung von f um x^* mit

$$|f(x) - T_n f(x)| \le \left(2^{-k}N\right)^{n+1} \left\|f^{n+1}\right\|_{\infty,U}, \qquad x \in U,$$

wobei $\|f\|_{\infty,U} = \sup_{x\in U} |f(x)|$, dann ist für jedes x mit $x + 2^{-k}\,[-N, N] \in U$

$$\begin{aligned}
&\phi * S_{2^{-k}} f\left(2^k x\right) \\
&= \sum_{\left|2^k x - \ell\right| \le N} \phi\left(2^k x - \ell\right)\, T_n f\left(2^{-k}\ell\right) + \sum_{\left|2^k x - \ell\right| \le N} \phi\left(2^k x - \ell\right)\, \left(f - T_n f\right)\left(2^{-k}\ell\right) \\
&= \phi * q_n\left(2^k x\right) + \phi * c\left(2^k x\right) = \phi * q_n\left(2^{k-1} x\right) + \phi * c_0\left(2^{k-1} x\right) + \psi * d_0\left(2^{k-1} x\right).
\end{aligned}$$

Die relevanten Waveletkoeffizienten in c_1 haben aber nun die Eigenschaft, daß für $\ell \in 2^k x + [-N, N]$ die *lokale* Abschätzung

$$\left|c\left(2^{-k}\ell\right)\right| \le \|\downarrow_2 F_1\| \left(2^{-k}N\right)^{n+1} \left\|f^{n+1}\right\|_{\infty,U} =: C 2^{-k(n+1)} \left\|f^{n+1}\right\|_{\infty,U}$$

gilt, wobei die Konstante C von k und f unabhängig ist und damit nur durch Charakteristiken der Filterbank, nämlich Polynomreproduktion und Norm des Waveletfilters, bestimmt wird. □

5.3 Anwendungen

Jetzt können wir uns endlich an die Anwendung unserer Waveletzerlegung und Filterung machen, um damit Signale zu verarbeiten. Die Idee ist einfach: Wir interpretieren ein Signal $c \in \ell(\mathbb{Z})$ als Funktion $\phi * c\left(2^k \cdot\right)$ für ein hinreichend großes $k \in \mathbb{N}$ und zerlegen

$$\begin{array}{ccccccccc}
c =: c^{(0)} & \to & c^{(1)} & \to & c^{(2)} & \to & \dots & \to & c^{(k)} \\
& \searrow & & \searrow & & \searrow & & \searrow & \\
& & d^{(1)} & & d^{(2)} & & \dots & & d^{(k)}
\end{array}$$

Ist nun f ein glattes Signal der Differenzierbarkeitsordnung $m \le n$, dann müsste

$$d^{(j+1)}\left(\lfloor 2^{k-j} x \rfloor\right) \sim 2^{(j-k)m} \left|f^{m+1}(x)\right|, \qquad j = 0, \dots, k-1,$$

sein und wir können so die lokale Glattheit der Funktion abschätzen. Insbesondere ist aber damit zu rechnen, daß Ecken und Kanten einer Funktion sich als große Waveletkoeffizienten zeigen, zumindest wenn man ordentlich mit einer Zweierpotenz multipliziert.

Beispiel 5.21 *Wir sehen uns einmal die Ecken einer einfachen "Dachfunktion" an, siehe Abb. 5.1, und erkennen hier sehr schön, daß die normalisierten Waveletkoeffizienten alle auf die Singularität "zeigen". Wenn man genau hinsieht, dann kann man sogar erkennen, daß die Größe der Koeffizienten bei "spitzeren" Winkeln zunimmt – nicht so überraschend, denn je spitzer der Winkel, desto größer ist die Krümmung, also die zweite Ableitung*[151].

[151] Achtung: Das ist keine bewiesene Aussage, sondern lediglich pure Heuristik. Trotzdem ist sowas für die Intuition hilfreich.

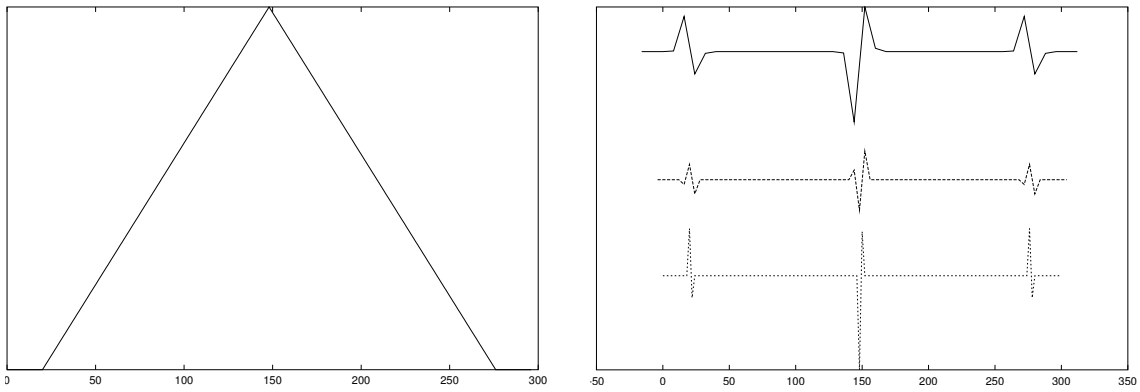


Abbildung 5.1: Eine stückweise lineare Funktion (links) und ihre Waveletkoeffizienten (rechts) bezüglich eines Wavelets mit hinreichender Strang–Fix–Ordnung.

Beispiel 5.21 ist recht eindrucksvoll, aber natürlich rein akademisch und künstlich. In der Realität sind Signale natürlich nicht so einfach und brav, sondern haben ein bißchen von allem.

Beispiel 5.22 *Das Signal in Abb 5.2 zeigt einen Ausschnitt aus einem EEG–Signal, siehe [56]*

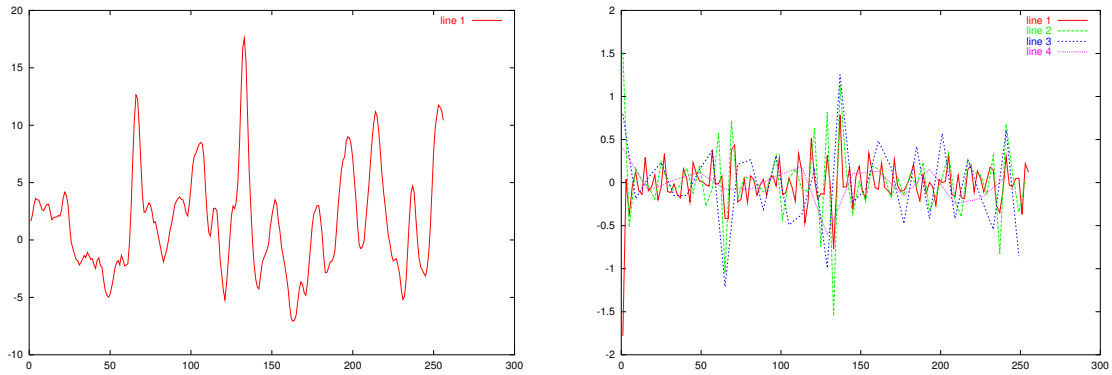


Abbildung 5.2: Ausschnitt aus einer Hirnstrommessung (links) und die zugehörigen Waveletkoeffizienten (rechts).

für Informationen über das Experiment, bei dem sie erhoben wurden, und die zugehörigen Waveletkoeffizienten. Man sieht doch, daß in diesem Fall die "Spitzen" wesentlich unschärfer sind und die "Ecken" und "Kanten" nicht so scharf charakterisieren.

Machen wir uns also systematischer an die Sache: Das "Credo" ist, daß große Waveletkoeffizienten "wichtige" Stellen der Funktion beschreiben, während kleine Waveletkoeffizienten zu Stellen gehören, an denen die Funktion auch in niedrigerer Auflösung gut beschrieben werden kann.

Beginnen wir also mit einem Signal $c \in \ell(\mathbb{Z})$, dann interpretieren wir dieses Signal als eine Funktion

$$f_c = \phi * c\,(2^n \cdot) \tag{5.31}$$

für ein passend gewähltes $n \in \mathbb{N}$, wobei die *Skalierungsfunktion* ϕ wieder nur *implizit* durch die Impulsantwort g_0 unserer Filterbank gegeben ist. Für dieses Signal bestimmen wir nun die Waveletzerlegung

$$f_c = \phi * c_n + \sum_{j=0}^{n-1} \psi * d_{n-j}\left(2^j \cdot\right), \tag{5.32}$$

wobei $c_0 = c$ und die Koeffizientenberechnungen

$$\begin{array}{rcll} c_{j+1} & = & \downarrow_2 g_0 * c_j, & \\ d_{j+1} & = & \downarrow_2 g_1 * c_j, & j = 0, \ldots, n-1, \end{array} \tag{5.33}$$

sich jetzt ausschließlich in der Filterbank abspielen. Durch das Downsampling enthalten sowohl c_{j+1} als auch d_{j+1} in etwa halb so viel Information wie c_j, so daß wir für Signale $c \in \ell_{00}(\mathbb{Z})$ mit endlichem Träger ganz automatisch die Einschränkung

$$n \leq \log_2 \#\operatorname{supp} c, \qquad \operatorname{supp} c = [j \ : \ c(j) \neq 0]\,,$$

haben[152].

Bemerkung 5.23 (Filterlänge) *Wenn wir eine Faltung $a*b$, $a, b \in \ell_{00}(\mathbb{Z})$ betrachten, dann hat das Ergebnis immer einen größeren Träger als b, da es ja eine "Überlappung" zwischen a und b gibt. Formal: Ist* $\operatorname{supp} a = [m, n]$ *und* $\operatorname{supp} b = [r, s]$*, dann ist*

$$a * b(j) = \sum_{k \in \mathbb{Z}} a(j-k)\, b(k) = \sum_{j=r}^{s} a(j-k)\, b(k) \neq 0,$$

wenn

$$j - [r, s] \cap [m, n] \neq \emptyset \qquad \Leftrightarrow \qquad j \in [m + r, n + s]$$

gilt, und wenn wir annehmen, daß $a(m)$, $a(n)$, $b(r)$ und $b(s)$ alle von Null verschieden sind, dann ist $\operatorname{supp} a * b = [m + r, n + s]$.

Für unsere Filterung eines Signals bedeutet das, daß die Länge des gefilterten Signals sich um die Länge des Filters erhöht und daß damit die Ausgaben der Analyse–Filterbank eben nur ungefähr *halbe Länge bzw. halben Informationsgehalt haben.*

Bemerkung 5.24 (Artefakte) *Die Sache mit der Filterlänge hat aber noch eine wesentlich schwerwiegendere Konsequenz! Normalerweise kann man ja nur endliche Signale messen bzw. aufzeichnen*[153] *oder die Information ist von Natur aus endlich, z.B. bei Bildern. Das heißt aber nicht, daß man das Signal außerhalb des Messbereichs einfach als 0 oder periodisch fortsetzen dürfte – man weiß es einfach* ***nicht****! Nun braucht aber die Filterung am "Rand" des Signals Informationen "außerhalb" des Signals und diese kann man ja nur raten, was dann aber natürlich zu massiven Artefakten führen kann.*

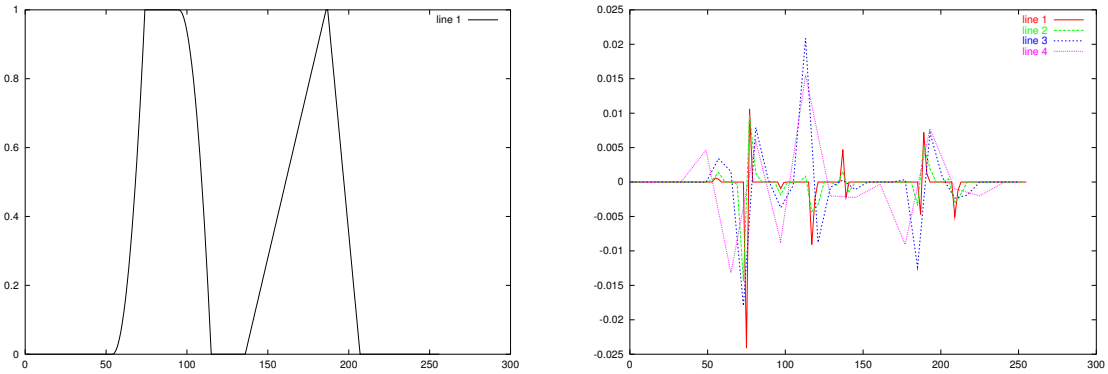


Abbildung 5.3: Ein einfaches Testsignal und die zugehörigen Waveletkoeffizienten. Wie man sieht, ist die Lokalisierung der Singularitäten weit davon entfernt, perfekt zu sein.

Beispiel 5.25 (Artefakte) *Hier zwei Beispiele für Artefakte.*

1. *In Abb. 5.3 ist ein "Testsignal" mit Übergängen verschiedener Glattheit zu sehen. Die Waveletkoeffizienten dazu deuten auch immer noch auf die Singularitäten, aber man sieht doch, daß die Lokalisierung nun wesentlich "unschärfer" ist. Außerdem tauchen am Rand jetzt signifikante Waveletkoeffizienten auf, die reine Artefakte sind, siehe Abb 5.4.*

2. *Auch Periodisierung ist nicht wirklich die Lösung, denn da kommt es darauf an,wie man periodisiert. Abb 5.5 zeigt dies ziemlich drastisch.*

Jetzt aber endlich zu der Art und Weise, wie man mit Wavelets Ecken erkennt, Entrauschen betreibt oder komprimiert – ganz egal, was man davon angeht, das Stichwort heißt[154] *Thresholding*.

Definition 5.26 *Für eine Konstante $\theta \in \mathbb{R}_+$ setzt ein* Thresholding–Operator $T_\theta : \ell(\mathbb{Z}) \to \ell(\mathbb{Z})$ *kleine Konstanten auf Null, und zwar vermittels*

1. (Hard thresholding)

$$T_\theta^h c(k) = \begin{cases} c(k), & |c(k)| \geq \theta, \\ 0, & |c(k)| < \theta, \end{cases} \qquad k \in \mathbb{Z}. \tag{5.34}$$

2. (Soft thresholding)

$$T_\theta^s c(k) = \operatorname{sgn} c(k)\ (|c(k)| - \theta)_+ = \begin{cases} c(k) - \theta, & c(k) \geq \theta, \\ 0, & |c(k)| \leq \theta, \\ c(k) + \theta, & c(k) < -\theta, \end{cases} \qquad k \in \mathbb{Z}. \tag{5.35}$$

[152] In dieser Formel bezeichnet $[\cdots]$ zur Abwechslung die *konvexe Hülle*.
[153] Unsere Möglichkeiten sind halt immer nur endlich, leider oder glücklicherweise.
[154] Auf gut Deutsch ...

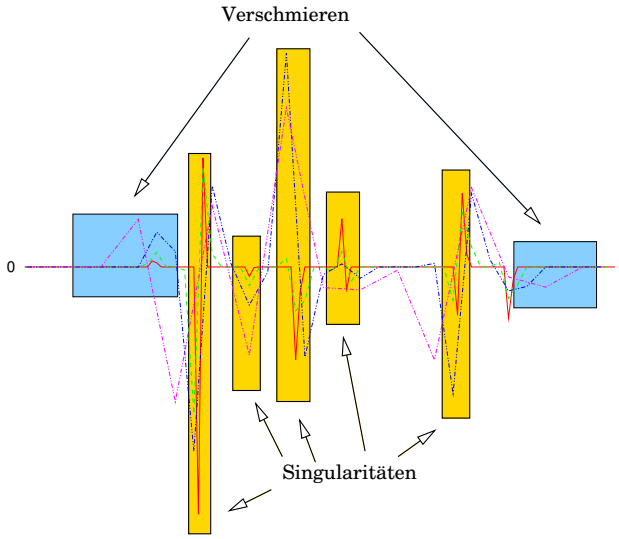


Abbildung 5.4: Erklärung der Waveletkoeffizienten in Abb 5.3: Die inneren Waveletkoeffizienten sind verschmiert, am Rand hingegen erscheinen Fortsetzungsartefakte.

Der Vorteil des "soft Thresholding" im Gegensatz zur "harten" Variante ist, daß die Abbildung

$$c \mapsto \operatorname{sgn} c \; (|c| - \theta)_+$$

stetig ist – allerdings wird auch das Signal c bei dieser Methode im "relevanten" Teil um die Konstante θ verringert. So ganz fällt allerdings auch das Soft Thresholding nicht vom Himmel, es ist nämlich Lösung eines Minimierungsproblems.

Proposition 5.27 *Für $c \in \ell_{00}(\mathbb{Z})$ ist*

$$T_\theta^s c = \operatorname{argmin} \; \|c - x\|_2^2 + 2\theta \; \|x\|_1 \tag{5.36}$$

Beweis: Nehmen wir an, daß $\operatorname{supp} c = [0, N]$, dann muß natürlich auch die Minimallösung c^* von (5.36) Träger $[0, N]$ haben[155], so daß wir das zu minimierende Funktional aus (5.36) in

$$\lambda_c(x) = \sum_{j=0}^{N} \left(c(j) - x(j)\right)^2 + 2\theta \sum_{j=0}^{N} |x(j)|$$

umschreiben können. Ist nun $c(j) \in \{0\} \cup [\theta, \infty)$, dann können wir $x(j) = c(j)$ wählen und besser geht es nicht. Andernfalls nehmen wir an, daß $c = c(j) \in (0, \theta)$ ist und suchen[156]

$$\min_x (x - c)^2 + 2\theta|x|, \qquad \Rightarrow \qquad 0 = 2(x - c) + 2\theta \operatorname{sgn} x$$

[155] Sonst käment ja nur unnötige Terme in (5.36) vor ...

[156] So ganz sauber ist der Beweis hier nicht, aber dafür einfach und anschaulich und er erspart uns weitere Fallunterscheidungen.

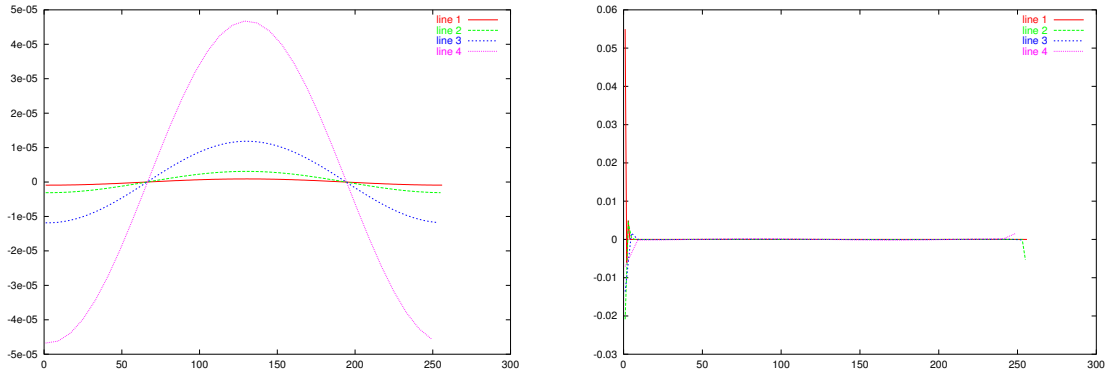


Abbildung 5.5: Waveletkoeffizienten zu periodischen Fortsetzungen der Sinunsfunktion, einmal mit Periode passend zu der des Sinus (links) und einma mit unpassender Periode (rechts). Man beachte die Skalierung der Achsen!

und da der erste Term für positives x sicherlich kleiner wird, erhalten wir, daß $x = c - \theta$. Dasselbe Argument können wir auch auf $c < 0$ anwenden, um auf $x = c + \theta$ zu kommen. Kombiniert man das, so ergibt sich gerade (5.35). □

Ein weiterer wesentlicher Freiheitsgrad besteht natürlich auch in der Wahl des Parameters θ, den man entweder fest und unabhängig von c wählen kann, aber besser in Abhängigkeit von c bestimmen sollte, beispielsweise[157]

- als $\lambda \|c\|_\infty$, $0 < \lambda < 1$, also als Bruchteil des größten auftretenden Wertes. Man wirft bei diesem Verfahren alles weg, was klein im Vergleich zu dieser Norm ist.

- als Bruchteil des Mittelwerts:

$$\theta = \lambda \frac{\|c\|_1}{\#c} = \frac{\lambda}{\#c} \sum_{k\in\mathbb{Z}} |c(k)|\,, \qquad \#c := \#\{k \ : \ c(k) \neq 0\}\,.$$

- als relativen Median[158] für $0 < \lambda < 1$:

$$\theta = |c(k)| \qquad \text{so daß} \qquad \#\{j \ : \ |c(j)| \leq |c(k)|\} \sim \lambda \#c.$$

All diese Ansätze haben ihre Vor- und Nachteile, der größte Vorteil des absoluten θ besteht aber sicherlich darin, daß dabei natürlich kein weiterer Rechenaufwand anfällt, die Bestimmung des Medians ist im wesentlichen so schwer wie Sortieren der Folge, die Berechnung der Mittelwerte hingegen eine Summation.

[157] Bei diesen Ansätzen muss man natürlich annehmen, daß $c \in \ell_{00}(\mathbb{Z})$ ist.
[158] Den eigentlichen Median erhält man für $\lambda = \frac{1}{2}$.

Unter Verwendung dieses Operators können wir nun die Anwendungen skizzieren. Zuerst berechnet man immer eine Waveletzerlegung[159] wie in (5.33)

$$c \mapsto (d_0, \ldots, d_{n-1}, c_n)$$

auf die wir nun folgendermaßen Thresholding anwenden:

Eckenerkennung: Je nachdem welcher Ordnung[160] die Ecken sein sollen, bestimmen wir $d_k^* = T_\theta 2^{kn} d_k$ für ein relativ großes θ und sehen uns die Lokalisierung dieser Koeffizienten an. Häufen sie sich irgendwo, dann ist das ein Indikator für eine Singularität.

Kompression: Hier bilden wir $d_k^* = T_\theta d_k$ sowie $c_n^* = T_\theta c_n$, wobei θ die Kompressionsrate und auch die Qualität des Ergebnisses[161] kontrolliert, und speichern dann nur die von Null verschiedenen Koeffizienten von $\left(d_0^*, \ldots, d_{n-1}^*, c_n^*\right)$.

Entrauschen: Hier geht man davon aus, daß Rauschen hochfrequent und von relativ kleiner Amplitude ist und sich somit nur in kleinen Störungen der Waveletkoeffizienten zeigt. Also bilden wir wieder $d_k^* = T_\theta d_k$ und bestimmen dann das neue, entrauschte c^* mit Hilfe des Synthesefilters[162] aus $\left(d_0^*, \ldots, d_{n-1}^*, c_n\right)$.

Klassifizierung: Das ist schon ein etwas subtileres Thema. Hier bestimmt man die Waveletzerlegung und extrahiert wenige *maximale* Waveletkoeffizienten, was zu einem Muster aus Werten (Absolutbetrag) und Indizes (Lage der "Ecke") führt. Mit anderen Worten: es ergibt sich ein relativ kleiner Vektor von Zahlen und dieser Vektor wird dann in ein "normales" Klassifizierungssystem[163] gefüttert.

[159] Und wir sind jetzt zurück bei den Filtern ...

[160] Sprünge, Ecken, Krümmungsdefekte (= zweite Ableitung) ...

[161] Wie man ja von JPEG her weiß, laufen diese beiden Interessen einander entgegen – hohe Kompression ist halt leider nicht ohne Qualitätsverlust zu haben.

[162] Also der rechten Hälfte der Filterbank, für irgendas muss die ja auch gut sein.

[163] Neuronales Netzwerk oder Lerntheorie heißen hier die Stichworte.

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*

6 Bilder

In der digitalen Signalverarbeitung hat man ein sehr einfaches Modell eines Bildes, nämlich als Matrix von *Pixeln*, also einzelnen Bildpunkten. Anders gesagt, ein Bild ist gegeben als eine Matrix $\boldsymbol{A} \in \mathbb{R}^{M\times N}$ mit M Zeilen und N Spalten, als eine doppeltunendliche Matrix[164] $\boldsymbol{A} = a(j,k)$, $j,k \in \mathbb{Z}$, oder als eine Funktion $a(x,y)$, $x,y \in \mathbb{R}$. Die Einträge von $\boldsymbol{A}$ sind entweder Zahlen, die dann als Graustufen interpretiert werden[165] oder Tripel von Zahlen, die dann als RGB–Werte interpretiert werden, also Rot-, Gelb-, und Blauanteil der Farbe. Mehr über Farbenlehre, Farbdarstellungen und wie man von einem Modell ins andere umrechnet findet sich in [14]. Wenn wir Quantisierung vernachlässigen[166] und jeden Farbkanal für sich betrachten, dann können wir die Einträge aber auch wieder als reelle Zahlen ansehen.

6.1 Ein paar Grundlagen

Jetzt ist es also so weit, wir betrachten *zweidimensionale* Signale, also Funktionen von $\mathbb{R}^2 \to \mathbb{R}$ bzw. $\mathbb{Z}^2 \to \mathbb{R}$. Formalisieren wir kurz, womit wir es hier zu tun haben.

Definition 6.1 (Signalklassen) *Mit $\ell(\mathbb{Z}^2)$ bezeichnen wir die Menge aller Signale der Form*[167]

$$c = \big(c(\alpha) \ : \ \alpha \in \mathbb{Z}^2\big) = (c(j,k) \ : \ j,k \in \mathbb{Z})$$

[164] Jedes endliche Bild kann in dieses Konzept eingebettet werden, nur haben wir natürlich dann wieder das Problem mit den Randartefakten.

[165] Normalerweise variieren diese von $0,\ldots,255$, aber mit den Quantisierungsdetails, die in der Praxis extrem relevant sind, wollen wir uns hier nicht beschäftigen.

[166] Und das wollen wir hier auch tun.

[167] Den Index $\alpha = (\alpha_1,\alpha_2) \in \mathbb{Z}^2$ bezeichnet man auch als *Multiindex*.

unter Verwendung der Normen

$$\|c\|_p := \left(\sum_{\alpha \in \mathbb{Z}^2} |c(\alpha)|^p \right)^{1/p}, \qquad \|c\|_\infty = \sup_{\alpha \in \mathbb{Z}^2} |c(\alpha)| .$$

Analog sind für Funktionen $f : \mathbb{R}^2 \to \mathbb{R}$ die Normen als

$$\|f\|_p = \left(\int_{\mathbb{R}^2} |f(t)|^p \, dt \right)^{1/p}, \qquad \|f\|_\infty = \sup_{x \in \mathbb{R}^2} |f(x)|$$

definiert.

Auch eine Fouriertransformation gibt es in $\mathbb{R}^s$, und zwar, für $\xi \in \mathbb{R}^2$,

$$\widehat{f}(\xi) = \int_{\mathbb{R}^2} f(t) e^{-i\xi^T t} \, dt = \int_{\mathbb{R}^2} f(t) e^{-i(\xi_1 t_1 + \xi_2 t_2)} \, dt = \int_{\mathbb{R}^2} f(t) \, e^{-i\xi_1 t_1} \, e^{-i\xi_2 t_2} \, dt.$$

Und wo es eine Fouriertransformation gibt, da ist auch eine Faltung nicht weit, und die sieht auch noch praktisch ganz genauso aus wie die Faltung in einer Variablen:

$$f * g = \int_{\mathbb{R}^2} f(\cdot - t) \, g(t) \, dt.$$

Übung 6.1 Zeigen Sie, daß auch in zwei Variablen die Identität

$$(f * g)^\wedge (\xi) = \widehat{f}(\xi) \, \widehat{g}(\xi)$$

gilt. ♢

Übung 6.2 Bestimmen Sie für Fouriertransformationen in zwei Variablen

- die inverse Fouriertransformation,
- die Perseval/Plancherel–Formel,
- die Poissonsche Summenformel.

♢

Auch LTI–Filter sind damit kein Problem[168], die Impulsantwort ist jetzt halt ein $f \in \ell(\mathbb{Z}^2)$ und die Filterung ergibt sich als

$$Fc = f * c = \sum_{\alpha \in \mathbb{Z}^2} f(\cdot - \alpha) \, c(\alpha), \tag{6.1}$$

[168] Für irgendwas müssen unsere Vorarbeiten ja gut sein.

was für einen FIR–Filter mit $f \in \ell_{00}(\mathbb{Z}^2)$ auch wieder "nur" eine endliche Summe ist. Besonders einfach wird die Filterung, wenn f ein *Tensorprodukt* ist, das heißt, wenn

$$f = f_1 \otimes f_2, \qquad \text{d.h.} \qquad f(\alpha) = f_1(\alpha_1)\ f_2(\alpha_2)\,, \tag{6.2}$$

ist, denn dann ist

$$\begin{aligned} f * c &= \sum_{\alpha\in\mathbb{Z}^2} f(\cdot-\alpha)\, c(\alpha) = \sum_{\alpha_1,\alpha_2\in\mathbb{Z}} f_1(\cdot-\alpha_1)\, f_2(\cdot-\alpha_2)\ c(\alpha_1,\alpha_2) \\ &= \sum_{\alpha_1\in\mathbb{Z}} f_1(\cdot-\alpha_1) \sum_{\alpha_2\in\mathbb{Z}} f_2(\cdot-\alpha_2)\ c(\alpha_1,\alpha_2) = \sum_{\alpha_1\in\mathbb{Z}} f_1(\cdot-\alpha_1)\ (f_2 * c(\alpha_1,\cdot))\,. \end{aligned} \tag{6.3}$$

Das liefert uns bereits ein Schema, wie wir so einen Filter anwenden: Für jedes feste α_1 filtern wir die "Zeile" $c(\alpha_1,\cdot)$ mit f_2 und stecken das Ergebnis dann in den Filter f_1. Schematisch sieht das dann wie folgt aus:

$$\begin{array}{cccccc} \ddots & \vdots & \vdots & \ddots & & \vdots \\ \dots & c(0,0) & c(0,1) & \dots & \to\ f_2 * c(0,\cdot) = & c'(0) \\ \dots & c(1,0) & c(1,1) & \dots & \to\ f_2 * c(1,\cdot) = & c'(1) \\ \ddots & \vdots & \vdots & \ddots & & \vdots \\ & & & & & \downarrow \\ & & & & & f_1 * c' \\ & & & & & \downarrow \\ & & & & & f * c \end{array}$$

Hat nun ein Filter F bzw. seine Impulsantwort f Tensorproduktstruktur, dann ist

$$f^*(z) \ := \ \sum_{\alpha\in\mathbb{Z}^2} f(\alpha) z^{-\alpha} = \sum_{\alpha\in\mathbb{Z}^2} f_1(\alpha_1)\ z^{-\alpha_1} f_2(\alpha_2)\ z^{-\alpha_2} = f_1^*(z_1)\ f_2^*(z_2)$$

und natürlich auch

$$\widehat{f}(\xi) = \widehat{f_1}(\xi_1)\ \widehat{f_2}(\xi_2)\,.$$

Anders gesagt: Die Faltung mit Tensorproduktfiltern ist besonders einfach! Das überträgt sich auch auf die schnelle Faltung: Man berechnet $\widehat{c}(\xi)$ und bildet dann für festes ξ_1 die punktweisen Produkte $c'(\xi) = \widehat{f_2}(\xi)\, c(\xi_1,\xi)$ und dann $(f * c)^\wedge(\xi) = \widehat{f_1}(\xi)\, c'(\xi)$.

Übung 6.3 Leiten Sie die FFT für Signale in zwei Variablen her, programmieren Sie den Algorithmus und geben Sie dessen Komplexität an. ♢

6.2 Einfache Filter für Bilder

Es gibt in der (medizinischen) Bildverarbeitung, siehe z.B. [25] eine ganze Menge von "Standardfiltern", die wir uns kurz ansehen wollen. Dabei ist es so, daß diese Filter oftmals *kontinuierlich*, also für Funktionen[169] $\phi : \mathbb{R}^2 \to \mathbb{R}$ konzipiert sind und für diskrete Daten einfach diskretisiert werden.

[169] Wir drücken uns hier vor der Festlegung von Eigenschaften wie Stetigkeit, Differenzierbarkeit oder Integrierbarkeit.

Der erste Filter ist der *Mittelwertfilter*

$$f = \frac{1}{|\Omega|}\chi_\Omega, \qquad \Omega \subset \mathbb{R}^2,$$

wobei Ω eine *kompakte* Teilmenge sein sollte. Die Filterung

$$\phi \mapsto f * \phi(x) = \frac{1}{|\Omega|}\int_{\mathbb{R}^2} \chi_\Omega(t)\phi\left(x-t\right)\,dt = \frac{1}{|\Omega|}\int_{\Omega} \phi\left(x-t\right)\,dt = \frac{1}{|\Omega|}\int_{x+\Omega} \phi(t)\,dt$$

ordnet also an jeder Stelle x der Funktion den Mittelwert über die Menge $x+\Omega$ zu. Die Diskretisierung dieses Filters ist einfach: Man nimmt einfach $f = S_h\chi_\Omega$ mit passender Abtastung h und normalisiert den Filter, indem man durch die Anzahl der in Ω enthaltenen Abtastpunkt teilt.

Ein wesentlicher Vorteil von Mittelungsfiltern ist, daß sie Rauschen unterdrücken. Rauschen ist ein unangenehmer Bestandteil gemessener Daten, der sich normalerweise nur stochastisch beschreiben lässt. Das Standardmodell für verrauschte Daten ist

$$\phi(x) = \psi(x) + \epsilon(x),$$

wobei ψ die eigentlichen Daten sind und ϵ das Rauschen. Eine allgemein übliche[170] Annahme ist, daß das Rauschen *mittelwertfrei* ist, das heißt, daß

$$E(\epsilon) = \int_{\mathbb{R}^2} \epsilon(x)\,dx = 0.$$

Bei der Filterung mit einem Mittelwerfilter erhält man dann also

$$F\phi(x) = \frac{1}{|\Omega|}\int_{x+\Omega} \phi(t)\,dt + \frac{1}{|\Omega|}\int_{x+\Omega} \epsilon(t)\,dt$$

und das mittelwertfreie Rauschen sollte zu einer geringeren Störung führen. Hier sieht man auch schon das Problem mit solchen Filtern: Der Träger sollte klein und "symmetrisch" genug sein, daß $F\psi \sim \psi$, was wegen

$$\phi(x) = \lim_{h\to 0^+} \frac{1}{h|\Omega|}\int_{x+h\Omega} \phi(t)\,dt, \qquad \phi \in L_1\left(\mathbb{R}^2\right)$$

zu schaffen ist, aber auch groß genug, damit das Rauschen auch wirklich ausgemittelt wird.

Will man es etwas vornehmer haben, dann kann man die charakteristische Funktion beispielsweise durch den *Gaußkern*

$$f(x) = \frac{1}{2\pi\sigma^2}\, e^{-\frac{\|x\|_2^2}{2\sigma^2}}, \qquad \sigma > 0$$

mit Standardabweichung σ ersetzen – so hat man sogar noch einen Parameter zur Verfügung. Für die Diskretisierung tastet man nun wieder f, genauer gesagt $f\chi_{[-N,N]^2}$ ab, wobei wir die

[170] Und in vielen Fällen genauso plausible wie unrealistische.

charakteristische Funktion benötigen, um einen FIR–Filter zu erhalten, denn die Exponentialfunktion klingt zwar schnell ab, hat aber trotzdem unendlichen Träger. Das diskrete Gegenstück dazu sind die Binomialfilter

$$f(j,k) = 2^{-m-n}\binom{m}{j}\binom{n}{k}, \qquad \begin{matrix} j = 0,\ldots,m, \\ k = 0,\ldots,n, \end{matrix}$$

den man natürlich für gerade m, n auch zentrieren kann. Zum Beispiel hat der zentrierte $2,2$–Binomialfilter die Form

$$\frac{1}{16}\begin{pmatrix} 1 & 2 & 1 \\ 2 & \boxed{4} & 2 \\ 1 & 2 & 1 \end{pmatrix},$$

wobei der eingerahmte Wert die Stelle $(0,0)$ bezeichnet.

Gradientenfilter werden bei Bildern verwendet, um Konturen hervorzuheben, denn da, wo der Unterschied zwischen zwei benachbarten Punkten groß ist, ist auch die Steigung groß und bei einem richtigen Sprung sogar unendlich. Der *Gradient* einer Funktion ist als

$$\nabla\phi = \begin{bmatrix} \dfrac{\partial}{\partial x}\phi(x,y) \\ \\ \dfrac{\partial}{\partial y}\phi(x,y) \end{bmatrix}$$

definiert und läßt sich durch

$$\nabla c = \begin{bmatrix} \nabla_1 c \\ \nabla_2 c \end{bmatrix} = \begin{bmatrix} c(\cdot+1,\cdot) - c(\cdot,\cdot) \\ c(\cdot,\cdot+1) - c(\cdot,\cdot) \end{bmatrix}$$

diskretisieren, und die beiden partiellen Differenzen sind nun wieder durch Faltungen mit den Impulsantworten

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & \boxed{-1} & 1 \\ 0 & 0 & 0 \end{pmatrix} \qquad \text{und} \qquad \begin{pmatrix} 0 & 1 & 0 \\ 0 & \boxed{-1} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

realisierbar.

Pure Gradientenverfahren sind allerdings sehr empfindlich gegen Rauschen, schließlich verbirgt sich ja hinter dem Gradienten ein Differenzenquotient mit Schrittweite h, wobei h die Abtastgenauigkeit ist, was Rauschen um einen Faktor h^{-1} verstärkt. Deswegen kombiniert man Gradienten gerne mit Mittelungsverfahren, z.B. dem Mittelungsoperator

$$f := \frac{1}{9}\begin{pmatrix} 1 & 1 & 1 \\ 1 & \boxed{1} & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Die Faltung $f * \nabla_j$ berechnet man am einfachsten über die z–Transformation als

$$\begin{aligned} (f * \nabla_1)^*(z) &= f^*(z)\,\nabla_1^*(z)\frac{1}{9}\sum_{\|\alpha\|_\infty \le 1} z^{-\alpha}\,(z_1 - 1) \\ &= \frac{1}{9}\sum_{\|\alpha\|_\infty \le 1} z^{-\alpha+\epsilon_1} - \sum_{\|\alpha\|_\infty \le 1} z^{-\alpha} = \frac{1}{9}\left(z_1^2 - z_1^{-1}\right)\left(z_2 + 1 + z_2^{-1}\right), \end{aligned}$$

und der gemittelte Gradientenfilter hat die Gestalt

$$\frac{1}{9}\begin{pmatrix} -1 & 0 & 0 & 1 \\ -1 & \boxed{0} & 0 & 1 \\ -1 & 0 & 0 & 1 \end{pmatrix} \qquad \text{und} \qquad \frac{1}{9}\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & \boxed{0} & 0 \\ -1 & -1 & -1 \end{pmatrix}$$

Und wenn wir schon bei Ableitungen sind, dann können wir auch den Laplaceoperator

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

verwenden; die zweiten partiellen Ableitungen kann durch symmetrische zweite Differenzen $c(\cdot + 1) - 2c + c(\cdot - 1)$ annähern und man erhält so den Filter

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & \boxed{-4} & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

als einfachste Diskretisierung des Laplaceoperators. Eine weitere Variante des Filters, der auch die Diagonalrichtungen berücksichtigt, ist

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & \boxed{-8} & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Wie in [25] ausgeführt ist, ist der Laplaceoperator besonders empfindlich gegen Rauscheffekte, weswegen man ihn meistens mit Glättungsfiltern, also Mittelung oder Gauß kombiniert.

Bleibt noch ein Klassiker, nämlich der *Medianfilter*

$$Mc(j) = \text{Median}\left\{c(k) \ : \ k \in j + \Omega\right\}, \qquad \Omega \subset \mathbb{Z}^2,$$

der alle Werte $c\left(j + \Omega\right)$ der Größr nach sortiert und dann den Wert in mittlerer Position wählt – eben einen Median berechnet. Dieser Filter hat allerdings einige Eigenheiten: Er ist nicht linear und fordert einen recht hohen Rechenaufwand, auch wenn Sortieren zu den "einfacheren" Operationen gehört. Der Medianfilter hat den Vorteil, daß er völlig unbeeindruckt von Ausreißern arbeitet und im wesentlichen Kanten erhält.

6.3 Tensorproduktwavelets für Bilder

Der einfachste und auch meistbenutzte Ansatz bei eine MRA für zweidimensionale Objekte wie Bilder ist sicherlich der der *Tensorproduktwavelets*.

Definition 6.2 *Seien ϕ_1, ϕ_2 Skalierungsfunktionen*[171]*, dann ist die* Tensorprodukt–Skalierungsfunktion *definiert als*

$$\Phi(x, y) = \phi(x)\,\phi(y), \qquad (x, y) \in \mathbb{R}^2.$$

[171] Es ist in keinster Weise verboten, in unterschiedliche Richtungen unterschiedliche Wavelets zu verwenden.

Die zugehörigen Wavelets[172] *sind dann*

$$\Psi_1(x,y) = \phi(x)\,\psi(y), \qquad \Psi_1(x,y) = \psi(x)\,\phi(y), \qquad \Psi_3(x,y) = \psi(x)\,\psi(y).$$

Damit dieses Φ und seine Wavelets eine MRA genieren, brauchen wir in erster Linie Verfeinerbarkeit und die Zerlegung. Also beweisen wir es mal. Dazu nehmen wir an, daß wir die Zwei–Skalen–Gleichungen

$$\phi_j = a_j * \phi\,(2\cdot) = \sum_{k\in\mathbb{Z}} a_j(k)\,\phi\,(2\cdot - k)\,, \qquad j = 1,2,$$

und die Zerlegungsformel

$$\phi_j\,(2\cdot) = c'_j * \phi_j + d'_j * \psi_j, \qquad j = 1,2,$$

zur Verfügung haben, siehe Definition 4.23.

Satz 6.3 *Die Funktion Φ ist verfeinerbar:*

$$\Phi = \sum_{\alpha\in\mathbb{Z}^2} (a_1 \otimes a_2)\,(\alpha)\Phi\,(2\cdot - \alpha)\,, \tag{6.4}$$

und es gilt

$$\Phi\,(2\cdot) = c * \Phi + \sum_{j=1}^{3} d_j * \Psi_j, \tag{6.5}$$

wobei

$$c = c'_1 \otimes c'_2, \qquad d_1 = c'_1 \otimes d'_2, \qquad d_2 = d'_1 \otimes c'_2, \qquad d_3 = d'_1 \otimes d'_2. \tag{6.6}$$

Beweis: Die Identität (6.4) folgt sehr einfach aus dem univariaten Fall,

$$\begin{aligned}\widehat{\Phi}\,(\xi) &= \widehat{\phi}_1\,(\xi_1)\;\widehat{\phi}_2\,(\xi_2) = \widehat{a}_1\,(\xi_1/2)\,\widehat{\phi}_1\,(\xi_1/2)\;\widehat{a}_2\,(\xi_2/2)\,\widehat{\phi}_2\,(\xi_2/2)\\ &= \left[\widehat{a}_1\,(\xi_1/2)\;\widehat{a}_2\,(\xi_2/2)\right]\left[\widehat{\phi}_1\,(\xi_1/2)\;\widehat{\phi}_2\,(\xi_2/2)\right] = (a_1\otimes a_2)^{\wedge}\left(\frac{\xi}{2}\right)\;\widehat{\Phi}\left(\frac{\xi}{2}\right).\end{aligned}$$

Für (6.5) setzen wir die univariaten Zerlegungsformeln

$$\phi_j\,(2\cdot) = c'_j * \phi_j + d'_j * \psi_j, \qquad j = 1,2,$$

in die Definition $\Phi = \phi_1 \otimes \phi_2$ ein und erhalten, daß

$$\begin{aligned}\Phi\,(2x,2y) &= \phi_1(x)\,\phi_2(y) = (c'_1 * \phi_1(x) + d'_1 * \psi_1(x))\;(c'_2 * \phi_2(y) + d'_2 * \psi_2(y))\\ &= (c'_1 \otimes c'_2)\,\phi_1(x)\phi_2(y) + (c'_1 \otimes d'_2)\,\phi_1(x)\psi_2(y) + (d'_1 \otimes c'_2)\,\psi_1(x)\phi_2(y)\\ &\qquad + (d'_1 \otimes d'_2)\,\psi_1(x)\psi_2(y)\end{aligned}$$

□

Wir haben es jetzt also mit *drei* Wavelets zu tun: Zwei der Wavelets ergeben sich als Tensorprodukt einer Skalierungsfunktion mit einem Wavelet, und eines als Tensorprodukt der beiden Wavelets.

[172] Ja, hier steht der Plural. Und zwar mit Absicht!

Beispiel 6.4 *Sehen wir uns das einmal im allereinfachsten Fall*[173] *an, daß* $\phi_1 = \phi_2 = \chi_{[0,1]}$ *und damit* $\psi_1 = \psi_2 = \chi_{\left[0,\frac{1}{2}\right]} - \chi_{\left[\frac{1}{2},1\right]}$. *Die resultierenden Funktionen sind dann in Abb. 6.1 zu sehen.*

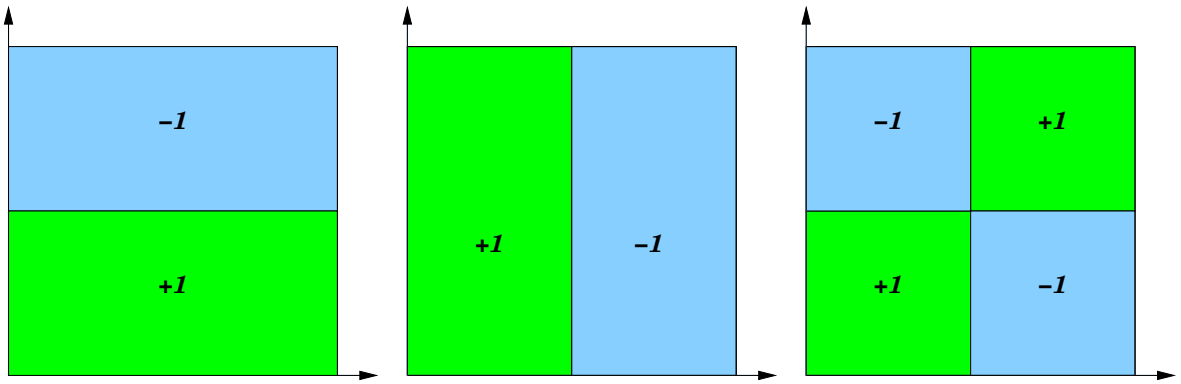


Abbildung 6.1: Die drei Tensorprodukte zum Haar–Wavelet: Skalierungsfunktion in x und Wavelet in y *(links)*, Wavelet in x und Skalierungsfunktion in y *(mitte)* und Wavelet in beiden Variablen *(rechts)*.

Die drei Wavelets haben nun unterschiedliche Fähigkeiten: Ψ_1 erkennt Kanten parallel zur x–Achse, Ψ_2 solche parallel zur y–Achse, während sich Ψ_3 um die diagonalen Kanten kümmert. Die Zerlegung eines Bildes läßt sich dann schematisch wie in Abb 6.2 darstellen, die "Durchführung"

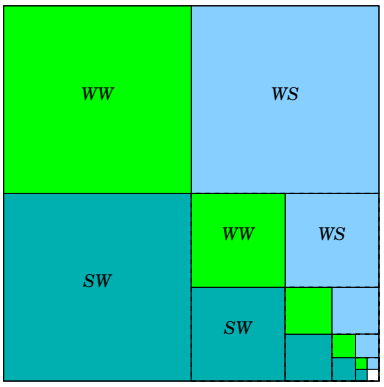


Abbildung 6.2: Die Waveletzerlegung für Bilder: Man zerlegt in drei Anteile mit Waveletbeteiligung und einen Anteil der "vergröberten" Skalierungsfunktion und zerlegt, wie bei eindimensionalen Signalen, letzteren dann weiter.

dieser Zerlegung landet wieder bei den Tensorprodukt–Filtern, die sich ja recht einfach realisieren lassen, indem man einfach die univariaten Filter in eine "Kaskade" schaltet.

[173] Das ist das gute alte haar–Wavelet.

Dies führt zwar einerseits zu recht ordentlichen Zerlegungen von Bildern un ermöglicht ein Übertragen des gesamten Filterbanktheorie aus den verhergegangen Kapiteln, hat aber den Nachteil, daß wirklich nur senkrechte, waagerechte und (positiv) diagonale Konturen wirklich erkannt werden. Leichte Drehungen hingegen können die Anzahl der großen Waveletkoeffizienten dramatisch steigen lassen. Deswegen gibt es auch diverse Ansätze wie "Ridgelets", "Curvelets" oder "Shearlets", die versuchen, einen geometrischeren Zugang beispielsweise zur Kantenerkennung zu ermöglichen, sich dann aber auch nicht mehr als einfache Filterbänke darstellen lassen.

6.4 Viele Basen, Wörterbücher und der Nutzen der Gier

Wie wir gerade gesehen haben[174], sind "einfache" Wavelets bei der Bildverarbeitung in ihren Richtungen limitiert. Deswegen könnte man naiv den Ansatz verfolgen, nicht nur eine Waveletbasis zu verwenden, sondern ein Signal mit verschiedenen, beispielsweise gedrehten Waveletbasen der Form

$$\Phi_A = \Phi\left(A\cdot\right), \qquad A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \quad \theta \in [0, 2\pi)$$

zu approximieren, oder, was dasselbe ist, die Filter entsprechend auf gedrehte Versionen des Bildes anzuwenden. Das führt natürlich nun zu einer dramatischen Überrepräsentation, die man besser etwas in den Griff bekommen sollte. Aber zuerst mal formal.

Definition 6.5 *Ein* Hilbertraum $\mathscr{H}$ *ist ein Banachraum*[175]*, dessen Norm durch ein Skalarprodukt definiert ist:*

$$\|f\| = \sqrt{\langle f, f\rangle}.$$

Ein Wörterbuch[176] *in $\mathscr{H}$ ist eine Menge $\mathscr{D} = \{f_\gamma \,:\, \gamma \in \Gamma\} \subset \mathscr{H}$, wobei Γ eine beliebige endliche oder unendliche*[177] *Menge ist. Wie nehmen außerdem an, daß $\mathscr{D}$ mindestens $N \leq \#\Gamma$ linear unabhängige Elemente enthält.*

Nur zur Warnung: Wie nehmen **nicht** an, daß $\mathscr{D}$ ein linearer Raum wäre, die Summe zweier Elemente des Wörterbuchs muss also nicht notwendigerweise wieder zum Wörterbuch gehören.

Beispiel 6.6 *1. Die endlichen Waveletzerlegungen können auch als Wörterbuch aufgefasst werden:*

$$\begin{aligned} \mathscr{D} \;=\; & \{\phi\left(\cdot - k\right) \,:\, k = -M, \ldots, M\} \\ & \cup \left\{\psi\left(2^j \cdot - k\right) \,:\, k = -2^j M, \ldots, 2^j M,\, j = 0, \ldots, n-1\right\}. \end{aligned}$$

Dieses Wörterbuch ist sogar eine Basis eines linearen Raums!

[174] Eigentlich haben wir es **nicht** gesehen, sondern müssen es einfach glauben.

[175] Also ein vollständiger linearer Raum.

[176] Englisch *"Dictionary"*.

[177] Diese sind in praktischen Anwendungen allerdings etwas unhandlich – das Durchsuchen unendlicher Mengen ist normalerweise mit einem etwas unerfreulichen Aufwand verbunden.

2. *Fügt man nun noch Rotationen, beispielsweise um Winkel $2\pi\ell/p$, $\ell = 0, \dots, p-1$, $p \in \mathbb{N}_0$, hinzu, dann erhält man ein Wörterbuch mit starker Überrepräsentation, das auch keine Basis mehr ist.*

Definition 6.7 *Sei $\mathscr{D} \subset \mathscr{H}$ ein Wörterbuch. Für $g \in \mathscr{H}$ und $m \in \mathbb{N}$ ist die* beste m–Term–Approximation *von $\mathscr{D}$ an g definiert als Lösung des Minimierungsproblems*

$$\min \|f - g\|, \qquad f = \sum_{\gamma \in \Gamma'} c_\gamma f_\gamma, \quad \Gamma' \subseteq \Gamma,\ \#\Gamma' \leq m.$$

Anders gesagt: Wir suchen die beste Näherung an g, die höchstens m von Null verschiedene Koeffizienten bezüglich des Wörterbuchs hat.

Die beste m–Term–Approximation ist in der Tat das beste, was wir mit unserem Wörterbuch in Sachen g anfangen können, wenn beispielsweise der Speicherplatz beschränkt ist und erinnert schon stark an einen Thresholding–Ansatz, bei dem wir nur die m größten Koeffizienten behalten würden. Allerdings wird die Theorie komplizierter, denn diese Approximationen sind weder linear noch konvex! Seien beispielsweise

$$f_1 = \sum_{\gamma \in \Gamma_1} c_{1,\gamma}\, f_\gamma \qquad \text{und} \qquad f_2 = \sum_{\gamma \in \Gamma_2} c_{2,\gamma}\, f_\gamma,$$

die besten Approximationen an $g_1, g_2 \in \mathscr{H}$, dann hat

$$f_1 + f_2 = \sum_{\gamma \in \Gamma_1 \cup \Gamma_2} c'_\gamma\, f_\gamma = \sum_{\gamma \in \Gamma_1 \setminus \Gamma_2} c_{1,\gamma}\, f_\gamma + \sum_{\gamma \in \Gamma_2 \setminus \Gamma_1} c_{2,\gamma}\, f_\gamma + \sum_{\gamma \in \Gamma_1 \cap \Gamma_2} (c_{1,\gamma} + c_{1,\gamma})\, f_\gamma$$

normalerweise deutlich mehr als m von Null verschiedene Koeffizienten und ist damit keine m–Term–Approximation mehr.

Ein weiteres Problem ist natürlich die *Bestimmung* dieser Approximation, denn normalerweise ist natürlich $N \gg m$, wobei N immer noch die "Dimension" von $\mathscr{D}$, also die Anzahl der linear unabhängigen Elemente in $\mathscr{D}$ bezeichnet. Der Algorithmus, der diese Bestimmung *näherungsweise* erledigt, wird als *matching pursuit* bezeichnet und gehört zur großen Familie der *"gierigen Algorithmen"*, auf Englisch als *"greedy algorithms"* bezeichnet. Die Idee besteht darin, in jedem Schritt den Fehler so weit wie möglich zu reduzieren[178] – natürlich, so wie es sich für einen Hilbertraum gehört, unter Verwendung von Orthogonalität.

Für ein beliebiges $\gamma \in \Gamma$ schreiben wir $g = c_\gamma f_\gamma + r$, $r \perp f_\gamma$, d.h. $\langle r, f_\gamma \rangle = 0$, dann ist

$$\langle f_\gamma, g \rangle = \langle f_\gamma, c_\gamma f_\gamma + r \rangle = c_\gamma \langle f_\gamma, f_\gamma \rangle + \underbrace{\langle f_\gamma, r \rangle}_{=0} = c_\gamma \langle f_\gamma, f_\gamma \rangle,$$

also

$$c_\gamma = \frac{\langle f_\gamma, g \rangle}{\langle f_\gamma, f_\gamma \rangle}. \tag{6.7}$$

[178] Das ist die Gier!

Der Fehler Rg, den wir bei dieser Projektion machen, ist dann

$$Rg = g - \frac{\langle f_\gamma, g\rangle}{\langle f_\gamma, f_\gamma\rangle} f_\gamma \qquad \Rightarrow \qquad Rg \perp f_\gamma,$$

und hat die Norm

$$\begin{aligned} \|Rg\|^2 &= \left\langle g - \frac{\langle f_\gamma, g\rangle}{\langle f_\gamma, f_\gamma\rangle} f_\gamma, g - \frac{\langle f_\gamma, g\rangle}{\langle f_\gamma, f_\gamma\rangle} f_\gamma \right\rangle \\ &= \|g\|^2 - 2\frac{\langle f_\gamma, g\rangle}{\langle f_\gamma, f_\gamma\rangle}\langle f_\gamma, g\rangle + \frac{\langle f_\gamma, g\rangle^2}{\langle f_\gamma, f_\gamma\rangle^2}\langle f_\gamma, f_\gamma\rangle = \|g\|^2 - \frac{\langle f_\gamma, g\rangle^2}{\langle f_\gamma, f_\gamma\rangle}, \end{aligned}$$

was unter der Voraussetzung[179] $\|f_\gamma\| = 1$

$$\|Rg\|^2 = \|g\|^2 - \langle f_\gamma, g\rangle^2 \tag{6.8}$$

liefert, was wiederum genau dann minimiert wird, wenn $\langle f_\gamma, g\rangle$ maximiert wird.

Definition 6.8 *Ein Wörterbuch $\mathscr{D}$ heißt* normiert*, wenn $\|f_\gamma\| = 1$ für alle $\gamma \in \Gamma$ gilt.*

Und jetzt können wir auch schon unseren gierigen Algorithmus für ein normiertes Wörterbuch $\mathscr{D} \subset \mathscr{H}$ und $g \in \mathscr{H}$ formulieren. Dazu legen wir noch eine Optimalitätskonstante $\alpha \in (0, 1]$ fest, setzen

$$R^0 g = g, \qquad \Gamma_0 = \emptyset$$

und bestimmen im k–ten Schritt, $k = 1, \ldots, m$, ein $\gamma^k \in \Gamma$ so, daß

$$\left|\left\langle R^{k-1}g, f_{\gamma^k}\right\rangle\right| \geq \alpha \max_{\gamma\in\Gamma} \left|\left\langle R^{k-1}g, f_\gamma\right\rangle\right|. \tag{6.9}$$

Dann erweitern wir nur noch zu $\Gamma_k = \Gamma_{k-1} \cup \left\{\gamma^k\right\}$ und bilden das neue Residuum nach Projektion

$$R^k g := R^{k-1}g - \left\langle R^{k-1}g, f_{\gamma^k}\right\rangle f_{\gamma^k}. \tag{6.10}$$

Über den Fehler können wir sofort eine Aussage machen! Unter Verwendung von (6.8) und (6.10) ist dann

$$\begin{aligned} \left\|R^k g\right\|^2 &= \left\|R^{k-1}g\right\|^2 - \left\langle R^{k-1}, f_{\gamma^k}\right\rangle^2 = \left\|R^{k-2}g\right\|^2 - \left\langle R^{k-2}, f_{\gamma^{k-1}}\right\rangle^2 - \left\langle R^{k-1}, f_{\gamma^k}\right\rangle^2 \\ &= \|g\|^2 - \sum_{j=1}^{k} \left\langle R^{j-1}, f_{\gamma^j}\right\rangle^2. \end{aligned}$$

Bemerkung 6.9 *Es ist bei diesem Algorithmus nicht gefordert, daß man in jedem Schritt ein "neues" f_γ verwendet, sondern es kann durchaus vorkommen, daß $\gamma^j = \gamma^k$ für $1 \leq j < k \leq m$ ist. Um also "wirklich" die beste m–Term–Approximation zu erhalten, wird man die Iteration so lange fortsetzen, bis $\#\Gamma_k = m + 1$ ist, und sich dann mit der Näherung aus Γ_k zufriedengeben.*

[179] Und das ist lediglich eine kleine Normierung …

Man kann nun sogar zeigen, daß dieser Algorithmus recht gut funktioniert so lange nur das Wörterbuch passend gewählt ist und eine Approximation der Funktion g ermöglicht.

Satz 6.10 *Ist $g \in \operatorname{span} \mathscr{D}'$ für eine endliche Teilmenge $\mathscr{D}' \subset \mathscr{D}$, dann gibt es eine Konstante $\lambda > 0$, so daß für alle $k \in \mathbb{N}_0$ die Abschätzung*

$$\left\| R^k g \right\| \leq 2^{-\lambda k} \, \|g\| \tag{6.11}$$

erfüllt ist.

Beweis: Wir zeigen zuerst, daß es eine Konstante $\beta > 0$ gibt, so daß für jede Funktion $g \in \operatorname{span} \mathscr{D}'$

$$\sup_{\gamma \in \Gamma} |\langle g, f_\gamma \rangle| > \beta \, \|g\| \tag{6.12}$$

gilt. Wäre das nicht der Fall, so gäbe es eine Folge $g_n \in \operatorname{span} \mathscr{D}'$, $n \in \mathbb{N}$, von Funktionen mit der Eigenschaft, daß

$$\|g_n\| = 1 \qquad \text{und} \qquad \lim_{n \to \infty} \sup_{\gamma \in \Gamma} |\langle g_n, f_\gamma \rangle| = 0. \tag{6.13}$$

Da $\operatorname{span} \mathscr{D}'$ endlichdimensional ist, ist die zugehörige Einheitskugel kompakt [34, 65][180] und eine Teilfolge der g_n konvergiert gegen eine Funktion g mit der Eigenschaft, daß[181]

$$\|g\| = 1 \qquad \text{und} \qquad \langle g, f_\gamma \rangle = \lim_{n \to \infty} \langle g_n, f_\gamma \rangle = 0, \quad \gamma \in \Gamma,$$

was insbesondere für eine Basis von $\operatorname{span} \mathscr{D}'$ gilt. Mit anderen Worten: $g \in \operatorname{span} \mathscr{D}' \cap (\operatorname{span} \mathscr{D}')^\perp$, weswegen $g = 0$ sein muesste, was einen Widerspruch zu $\|g\| = 1$ darstellt.

Der Rest ist wieder (6.8), genauer die Identität

$$\left\| R^k g \right\|^2 = \left\| R^{k-1} g \right\|^2 - \left\langle R^{k-1}, f_{\gamma^k} \right\rangle^2.$$

Wegen (6.9) und (6.12) ist

$$\begin{aligned} \left\| R^k g \right\|^2 &= \left\| R^{k-1} g \right\|^2 - \left\langle R^{k-1} f_{\gamma^k} \right\rangle^2 \\ &\leq \left\| R^{k-1} g \right\|^2 - \alpha^2 \max_{\gamma \in \Gamma} \left| \left\langle R^{k-1} g, f_\gamma \right\rangle \right|^2 \leq \left\| R^{k-1} g \right\|^2 - \alpha^2 \beta^2 \, \left\| R^{k-1} g \right\|^2 \\ &= \left(1 - \alpha^2 \beta^2 \right) \, \left\| R^{k-1} g \right\|^2, \end{aligned}$$

also

$$\left\| R^k g \right\| \leq \sqrt{1 - \alpha^2 \beta^2} \, \left\| R^{k-1} g \right\| \leq \cdots \leq \sqrt{1 - \alpha^2 \beta^2}^k \, \|g\|,$$

was uns $\lambda = -\frac{1}{2} \log_2 \left(1 - \alpha^2 \beta^2 \right)$ liefert, was wohldefiniert ist, da wir nichts verlieren, wenn wir $\beta < 1$ wählen[182] □

[180] Zur Erinnerung: Die Einheitskugel eines Banachraums ist genau dann kompakt, wenn dieser endlichdimensional ist!

[181] Hier geht noch Stetigkeit linearer Funktionale und Normen mit ein …

[182] Was wir ohnehin tun müssen!

Übung 6.4 Zeigen Sie, daß man β in (6.12) immer < 1 wählen **muss**. $\diamondsuit$

In dem Algorithmus müssen ja in jedem Schritt eigentlich **alle** inneren Produkte $\langle R^{k-1}, f_\gamma \rangle$, $\gamma \in \Gamma$, ausgerechnet werden, was einen ziemlichen Aufwand darstellt. Was man allerdings wirklich braucht, das sind

$$\boldsymbol{g} = [\langle g, f_\gamma \rangle \ : \ \gamma \in \Gamma] \in \mathbb{R}^\Gamma \qquad \text{und} \qquad \boldsymbol{F} = [\langle f_\gamma, f_{\gamma'} \rangle \ : \ \gamma, \gamma' \in \Gamma] \in \mathbb{R}^{\Gamma \times \Gamma}.$$

Nach (6.10) ist dann

$$\boldsymbol{r}^k := \big[\langle R^k g, f_\gamma \rangle \ : \ \gamma \in \Gamma\big] = \boldsymbol{r}^{k-1} - \boldsymbol{r}^{k-1}_{\gamma^k} \boldsymbol{F}\, \boldsymbol{e}_{\gamma^k}, \qquad \boldsymbol{r}^0 = \boldsymbol{g},$$

was "lediglich" eine Addition eines Vektors der Länge $\#\Gamma$ bedeutet. Besonders effizient wird das, wenn die Matrix $\boldsymbol{F}$ dünn besetzt ist, wenn also unter den Elementen des Wörterbuchs viel Orthogonalität besteht, denn dann benötigt man in jedem Iterationsschritt nur so viele Rechenoperationen wir die Spalte $\boldsymbol{F}\, \boldsymbol{e}_{\gamma^k}$ von Null verschiedene Einträge hat.

Was seltsam ist, bleibt selten lange unerklärt. Das Unerklärliche ist gewöhnlich nicht mehr seltsam, und ist es vielleicht nie gewesen.

G. Chr. Lichtenberg

7 Rauschen, Zufall und inverse Probleme

Leider ist es in der Realität mit der Messung von Signalen und Bildern nicht so einfach! Nur sehr selten und mit sehr großem Aufwand kann man genau messen und auch wirklich das messen, was man messen will. Oftmals sucht man ein Signal f, kann aber nur

$$g = Tf + \epsilon \tag{7.1}$$

messen, wobei T ein (hoffentlich linearer) Operator ist[183] und ϵ *zufälliges* Rauschen. Die Zufälligkeit sorgt dafür, daß ein gewisses Maß an Stochastik nicht ganz vermeidbar ist.

7.1 Zufallsprozesse, Hauptkomponenten und Approximation

Als erstes beschäftigen wir uns mit *stochastischen diskreten Signalen*, also Signalen, die nicht mehr deterministisch, sondern zufällig sind. Diese modelliert man als einen *Zufallsprozess*, eine Art Bildungsgesetz für Zufallsvariablen, also Dichtefunktionen für eine Wahrscheinlichkeit, siehe z.B. [42]. Etwas formaler[184]: Ein *Zufallsprozess* ist eine Funktion $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, also $f(s,t)$, die zu jedem festen s eine Dichtefunktion $f(s,\cdot)$ liefert, die man als *Realisierung* des Zufallsprozesses bezeichnet. Der Erwartungswert der Realisierung ist dann

$$\mathbb{E}_s(f) = \int_{\mathbb{R}} f(s,t)\,dt,$$

die Varianz

$$\mathbb{V}_s(f) = \int_{\mathbb{R}} \left(f(s,t) - \mathbb{E}_s(f)\right)^2\,dt$$

und die Standardabweichung $\sigma_s(f) = \mathbb{V}_s(f)^{1/2}$. Unser Zufallssignal f ist dann eine Folge $f(j) = f\left(s_j,\cdot\right)$ von Zufallsvariablen, das heißt, jede *Messung* ist eine Realisierung des Zufallsprozessen. Die *Kovarianz* dieser, der Einfachheit halber als *mittelwertfrei* — $\mathbb{E}\left(f_j\right) = 0$ — angenommenen, Messungen ist die Matrix

$$\boldsymbol{K}(f) = \left[\mathbb{E}\left(f(j)\,f(k)\right)\ :\ j,k\right].$$

[183] Was den Sonderfall $T = I$ direkter Messungen ja nicht ausschließt.

[184] Aber immer noch nicht im Detail – es geht uns hier um die Idee, nicht um die volle Theorie.

Unser erstes Ziel besteht nun darin, das Zufallssignal so mit endlicher Information anzunähern, daß der zu erwartende Fehler möglichst klein wird. Dazu seien g_j, $j \in \mathbb{N}$, eine Orthonormalbasis von $\ell(\mathbb{Z})$ bezüglich des inneren Produkts[185]

$$\langle f, g \rangle = \sum_{j \in \mathbb{Z}} f(j)\, g(j),$$

dann ist die Projektion auf den endlichdimensionalen Teilraum, der von $g_1, \ldots, g_m$ aufgespannt wird, wieder

$$f_m := \sum_{j=1}^{m} \langle f, g_j \rangle\, g_j = \sum_{j=1}^{m} \sum_{k \in \mathbb{Z}} f(k)\, g_j(k)\, g_j.$$

Die Frage ist nun:

Wie wählt man $g_1, \ldots, g_m$ so, daß der zu erwartende Fehler minimal wird?

Mit anderen Worten: Wir müssen den Fehler

$$\varepsilon_m := \mathbb{E}\left(\|f - f_m\|^2\right) = \mathbb{E}\left(\left\| \sum_{j=m+1}^{\infty} \langle f, g_j \rangle\, g_j \right\|^2\right) = \sum_{j=m+1}^{\infty} \mathbb{E}\left(\langle f, g_j \rangle^2\right)$$

minimieren, wobei wir ja angenommen haben, daß die g_j eine Orthormalbasis bilden und sich somit f als

$$f = \sum_{j=1}^{\infty} \langle f, g_j \rangle\, g_j$$

schreiben lässt. Für einen beliebigen deterministischen Vektor x ist nun aber

$$\begin{aligned} \mathbb{E}\left(\langle f, x \rangle^2\right) &= \mathbb{E}\left(\left[\sum_{j \in \mathbb{Z}} f(j)\, x(j)\right]^2\right) = \mathbb{E}\left(\sum_{j,k \in \mathbb{Z}} f(j)\, f(k)\, x(j)\, x(k)\right) \\ &= \mathbb{E}\left(x^T \left[f(j)\, f(k) \;:\; j,k \in \mathbb{Z}\right] x\right) = x^T \left[\mathbb{E}\left(f(j)\, f(k)\right) \;:\; j,k \in \mathbb{Z}\right] x \\ &= x^T \boldsymbol{K}(f)\, x = \langle \boldsymbol{K}(f)\, x, x \rangle, \end{aligned}$$

und daher ist

$$\varepsilon_m = \sum_{j=m+1}^{\infty} \langle \boldsymbol{K}(f)\, g_j, g_j \rangle \tag{7.2}$$

Übung 7.1 Zeigen Sie: Die Kovarianzmatrix $\boldsymbol{K}(f)$ ist positiv semidefinit. ♢

Da $\boldsymbol{K}(f)$ eine symmetrische und positiv semidefinite Matrix ist, existiert eine Orthonormalbasis von Eigenvektoren falls f endlichen Träger hat. Jede solche Basis heißt *Karhunen–Loève–Basis*. Für jedes Element einer derartigen Basis ist dann

$$\langle \boldsymbol{K}(f)\, g_j, g_j \rangle = \langle \lambda_j\, g_j, g_j \rangle = \lambda_j \underbrace{\|g_j\|^2}_{=1} = \lambda_j,$$

[185] Konvergenzfragen lassen wir hier außen vor, beispielsweise, indem wir annehmen, daß f nur ein *endlicher* Vektor ist.

hängt also "modulo Normierung" nur vom zugehörigen Eigenwert ab.

Satz 7.1 *Hat f den Träger $[1, N]$, dann minimiert eine Basis $G = \{g_j \,:\, j = 1, \ldots, N\}$, den Fehler ε_m genau dann, wenn G eine Karhunen-Loève-Basis mit*

$$\langle \boldsymbol{K}(f) g_j, g_j \rangle \geq \langle \boldsymbol{K}(f) g_{j+1}, g_{j+1} \rangle$$

ist.

Definition 7.2 *Die Eigenvektoren zu den größten Eigenwerten der Kovarianzmatrix $\boldsymbol{K}(f)$ bezeichnet man auch als die* Hauptkomponenten *von f.*

Beweis: Für *jede* Orthonormalbasis h_j, $j = 1, \ldots, m$ ist

$$\text{trace}\, \boldsymbol{K}(f) = \sum_{j=1}^{N} \lambda_j = \sum_{j=1}^{N} \left(\begin{bmatrix} h_1^T \\ \vdots \\ h_N^T \end{bmatrix} \boldsymbol{K}(f) \, [h_1 \ \ldots h_N] \right)_{jj} = \sum_{j=1}^{N} \langle \boldsymbol{K}(f)\, h_j, h_j \rangle,$$

so daß eine Basis ε_m genau dann minimiert, wenn sie

$$\sum_{j=1}^{N} \lambda_j - \varepsilon_m = \sum_{j=1}^{N} \langle \boldsymbol{K}(f)\, h_j, h_j \rangle - \underbrace{\sum_{j=m+1}^{N} \langle \boldsymbol{K}(f)\, h_j, h_j \rangle}_{=\varepsilon_m} = \sum_{j=1}^{m} \langle \boldsymbol{K}(f)\, h_j, h_j \rangle,$$

maximiert.

Sei nun G eine Karhunen-Loève-Basis entsprechend den Voraussetzungen des Satzes und H eine beliebige andere Orthogonalbasis, die wir bezüglich G als

$$h_j = \sum_{k=1}^{N} \langle h_j, g_k \rangle \, g_k$$

schreiben können, weswegen

$$\begin{aligned} \langle \boldsymbol{K}(f)\, h_j, h_j \rangle &= \sum_{k,\ell=1}^{N} \langle h_j, g_k \rangle \langle h_j, g_\ell \rangle \langle \boldsymbol{K}(f)\, g_k, g_\ell \rangle = \sum_{k,\ell=1}^{N} \langle h_j, g_k \rangle \langle h_j, g_\ell \rangle \lambda_k \langle g_k, g_\ell \rangle \\ &= \sum_{k=1}^{N} \langle h_j, g_k \rangle^2 \lambda_k \end{aligned}$$

und damit auch

$$\sum_{j=1}^{m} \langle \boldsymbol{K}(f)\, h_j, h_j \rangle = \sum_{j=1}^{m} \sum_{k=1}^{N} \langle h_j, g_k \rangle^2 \lambda_k = \sum_{k=1}^{N} \underbrace{\left(\sum_{j=1}^{m} \langle h_j, g_k \rangle^2 \right)}_{=:q_k \leq 1} \lambda_k$$

ist, mit $0 \leq q_k \leq 1$ und

$$\sum_{k=1}^{N} q_k \;=\; \sum_{k=1}^{N}\sum_{j=1}^{m} \langle h_j, g_k\rangle^2 = \sum_{j=1}^{m} \underbrace{\sum_{k=1}^{N} \langle h_j, g_k\rangle^2}_{=\|h_j\|^2=1} = m.$$

Also ist

$$\begin{aligned}
&\sum_{j=1}^{m} \langle \boldsymbol{K}(f)\, h_j, h_j\rangle - \sum_{j=1}^{m} \langle \boldsymbol{K}(f)\, g_j, g_j\rangle = \sum_{j=1}^{m} \langle \boldsymbol{K}(f)\, h_j, h_j\rangle - \sum_{j=1}^{m} \lambda_j \\
&\quad= \sum_{k=1}^{N} q_k\, \lambda_k - \sum_{j=1}^{m} \lambda_j = \sum_{k=1}^{N} q_k\, \lambda_k - \sum_{j=1}^{m} \lambda_j + \lambda_m \underbrace{\left(m - \sum_{k=1}^{N} q_k\right)}_{=0} \\
&\quad= \sum_{j=1}^{m} (q_j - 1)\, \lambda_j + \sum_{j=m+1}^{N} \lambda_j\, q_j + \lambda_m \sum_{j=1}^{m} (1 - q_j) - \lambda_m \sum_{j=m+1}^{N} q_j \\
&\quad= \sum_{j=1}^{m} \underbrace{(\lambda_j - \lambda_m)}_{\geq 0} \underbrace{(q_j - 1)}_{\leq 0} + \sum_{j=m+1}^{N} q_j \underbrace{(\lambda_j - \lambda_m)}_{\leq 0} \leq 0
\end{aligned}$$

mit Gleichheit genau dann[186], wenn $N = m$ und $q_j = 1$, $j = 1, \ldots, N$. Anders gesagt: Die Karhunen–Loève–Basis ist immer besser als andere Basen und das beste Ergebnis erhält man natürlich genau dann, wenn man die Basis so anordnet, daß $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ ist. □

7.2 Regularisierung oder wie man unterbestimmte Probleme löst

So, wir wissen also jetzt, wie wir einen reinen Zufallsprozess am besten approximieren, aber es gibt auch noch eine wichtige deterministische Komponente in (7.1), nämlich den Operator T, der alles sein muss, aber nicht invertierbar! Na gut, wenn T eine Matrix ist[187], dann können wir natürlich *jede* Matrix "invertierbar" machen, indem wir die Singulärwertzerlegung

$$T = U\Sigma V^T, \qquad \sigma_{11} \geq \sigma_{22} \geq \cdots \geq 0$$

mit orthogonalen Matrizen U und V und einer Diagonalmatrix[188] Σ verwenden, die auch für nichtquadratische Matrizen definiert ist, siehe z.B. [21, 31, 44]. Die *Pseudoinverse* oder *Moore–Penrose–Inverse* zu T ist dann als

$$T^+ = V\Sigma^+ U^T, \qquad \left(\Sigma^+\right)_{jk} = \begin{cases} 0, & \sigma_{jk} = 0, \\ \sigma_{jk}^{-1}, & \sigma_{jk} > 0, \end{cases}$$

[186] Zumindest in dem Fall, daß $\boldsymbol{K}(f)$ *strikt* positiv definit ist – aber wenn gewisse der Eigenwerte Null wären, dann könnte man das zugehörige Element der Karhunen–Loève–Basis auch gleich unter den Tisch fallen lassen, also ist diese Annahme keine wirkliche Restriktion!

[187] Und bei endlichen diskreten Signalen sind alle linearen Operatoren Matrizen oder zumindest als solche darstellbar.

[188] Genau gesagt ist Σ eine Matrix, bei der alle Nicht–Diagonalelemente verschwinden, schließlich muss Σ nicht quadratisch sein!

definiert, und man könnte als Lösung $f = T^+ g$ wählen, denn wenn es eine Lösung des linearen Systems $g = Tf$, dann muß auch $f = T^+ g$ eine sein. Normalerweise ist das Gleichungssystem auch *unterbestimmt*, das heißt, das zugrundeliegende Modell ist komplexer als die gemessenen Größen.

Beispiel 7.3 *Wir betrachten ein univariates Signal und als Operator T die Mittelung von 4 benachbarten Werten gefolgt von Downsampling um Faktor 2, also*

$$\begin{aligned} T &= \begin{bmatrix} \ddots & & & & & & \\ & 1 & 0 & 0 & 0 & 0 & \\ & 0 & 0 & 1 & 0 & 0 & \\ & 0 & 0 & 0 & 0 & 1 & \\ & & & & & & \ddots \end{bmatrix} \frac{1}{4} \begin{bmatrix} \ddots & & & & & & & \\ & 1 & 1 & 1 & 1 & & & \\ & & 1 & 1 & 1 & 1 & & \\ & & & 1 & 1 & 1 & 1 & \\ & & & & & & & \ddots \end{bmatrix} \\ &= \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 & & & \\ & & 1 & 1 & 1 & 1 & & \\ & & & & & & \ddots \end{bmatrix} \in \mathbb{R}^{n \times 2n+2}. \end{aligned}$$

Beschränken wir uns auf $n = 3$, dann ist beispielsweise

$$T\,[1\,0\,0\,0\,0\,0\,0\,1]^T = \begin{bmatrix} 1/4 \\ 0 \\ 1/4 \end{bmatrix}, \qquad T\,[1\,2\,3\,4\,5\,6\,7\,8]^T = \begin{bmatrix} 2.5 \\ 4.5 \\ 6.5 \end{bmatrix}$$

und die Rekonstruktionen über die Pseudoinverse ergeben

$$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \qquad \textit{und} \qquad [2.5\ 2.5\ 2.5\ 2.5\ 6.5\ 6.5\ 6.5\ 6.5],$$

was nicht sonderlich berauschend ist.

Das grundlegende Problem bei der ganzen Geschichte besteht darin, daß das *inverse Problem* $Tf \mapsto f$ im Normalfall *schlecht gestellte*, auf Englisch *ill posed*, ist und daß die Pseudoinverse keine wirkliche Lösung des Problems darstellen kann – ganz abgesehen davon, daß man es oft auch mit "richtigen Operatoren" und nicht "nur" mit Matrizen zu tun hat.

Der beste Ansatz besteht nun darin, entweder unter f unter Einhaltung der Nebenbedingung $Tf = g$ ein möglichst problembezogenes Energiefunktional E minimieren zu lassen, also

$$\min_f E(f), \qquad Tf = g, \tag{7.3}$$

zu bestimmen, oder aber Nebenbedingung und Energiefunktional in ein parametrisiertes Minimierungsproblem zu packen:

$$\min_f \|Tf - g\| + \lambda E(f), \qquad \lambda \in \mathbb{R}_+, \tag{7.4}$$

wobei man beim zweiten Ansatz sogar auf die "exakte" Lösung $Tf = g$ verzichtet, die in Anwesenheit von Rauschen ohnehin nicht wirklich sinnvoll ist, wenn man dafür die Funktion f "einfacher" oder "glatter" bekommen kann.

Das einfachste Energiefunktional ist sicherlich $E(f) = \|f\|_2^2$. In diesem Fall wird (7.3) für $f \in \ell_{00}(\mathbb{Z})$ zu

$$\min \sum_{j=0}^{N} f(j)^2, \qquad Tf = g,$$

was sich mittels Lagrange–Multiplikatoren, siehe z.B. [28, 40, 47, 57], in

$$0 = \nabla E(f) - \nabla^T (g - Tf)\, \mu = 2f + T^T \mu, \qquad Tf = g,$$

also als das lineare Gleichungssystem

$$\begin{bmatrix} 2I & T^T \\ T & 0 \end{bmatrix} \begin{bmatrix} f \\ \mu \end{bmatrix} = \begin{bmatrix} 0 \\ g \end{bmatrix}$$

ergibt. Für unsere Daten aus Beispiel erhalten wir auch genau dasselbe Ergebnis wie mit der Pseudoinversen, nicht weiter verwunderlich, wenn man sich die Rolle der Pseudoinversen bei der Least–Squares–Aproximation in Erinnerung ruft, siehe [2, 21]. Der Ansatz aus (7.4) hingegen liefert uns als zu minimierende Funktion

$$\|Tf - g\|_2^2 + \lambda \|f\|_2^2 = (Tf - g)^T (Tf - g) + \lambda f^T f = f^T \left(T^T T + \lambda I\right) f - 2g^T Tf + g^T g,$$

was nach f abgeleitet und gleich Null gesetzt das Gleichungssystem

$$\left(T^T T + \lambda I\right) f = T^T g$$

liefert. Für $\lambda \to 0$ erhalten wir dann dieselbe Lösung wir für das Optimierungsproblem (7.3), für große λ eher eigenwillige Werte.

So weit also nichts neues. Versuchen wir uns also einmal an anderen Energiefunktionalen, beispielsweise an

$$\|\nabla f\|_2^2 = \sum_{j=0}^{N-1} (f(j+1) - f(j))^2 = \|Df\|^2 = f^T D^T D f, \qquad D = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}$$

Da nun $\nabla E(f) = 2D^T Df$ ist[189], erhalten wir jetzt die Gleichungssysteme

$$\begin{bmatrix} 2D^T D & T^T \\ T & 0 \end{bmatrix} \begin{bmatrix} f \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ g \end{bmatrix} \qquad \text{und} \qquad \left(T^T T + \lambda D^T D\right) f = T^T g.$$

Während hier der erste Vektor immer noch schlecht rekonstruiert wird, das Ergebnis ist der `Octave`–Vektor

[189] Im Gegensatz zu $\nabla E(f) = 2I$ vorher.

```
 0.583333
 0.416667
 0.083333
-0.083333
-0.083333
 0.083333
 0.416667
 0.583333
```

erhalten wir für unseren linearen Vektor `(1:8)` schon das deutlich bessere Ergebnis

```
1.5909
1.9545
2.6818
3.7727
5.2273
6.3182
7.0455
7.4091
```

bei der Rekonstruktion. Verwenden wir hingegen den "Laplace–Operator" $\Delta = f(\cdot + 2) - 2f(\cdot + 1) + f(\cdot)$, dann ist

$$\|\Delta f\|_2^2 = \|D_2 f\|_2^2 = f^T D_2^T D_2 f, \qquad D = \begin{bmatrix} 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \end{bmatrix},$$

und die Ergebnisse sind

```
 0.653846
 0.346154
 0.076923
-0.076923
-0.076923
 0.076923
 0.346154
 0.653846
```

und **exakte** Rekonstruktion von `(1:8)`.

Der Vorteil quadratischer Funktionale ist offensichtlich: Man kann sie sehr leicht und systematisch in lineare Gleichungssysteme umformen, wobei die Anteile $D^T D$ in diesen Gleichungssystemen auch noch eine recht nette Struktur haben: Sie sind bandiert. Normalerweise verwendet man allerdings eher die (7.4), da dieser Ansatz wesentlich "rauschtoleranter" ist, denn wir haben ja eben nicht Tf gemessen, sondern $Tf + \epsilon \ldots$ Außerdem ist es hier auch völlig egal, ob die Nebenbedingung $Tf = g$ überhaupt erfüllbar ist, sie wird einfach so gut erfüllt wie möglich.

Die Auswahl des jeweiligen Energiefunktionals[190] ist problemabhngig und wird in der Bildverarbeitung auch lebhaft diskutiert. Das einfachste Funktional, das in der Bildverarbeitung häufig verwendet wird, ist in der kontinuierlichen Formulierung von (7.4)

$$E(f) = \int_\Omega \|\nabla f\|_2^2 + \lambda \int_\Omega (g - Tf)^2 \,,$$

meist mit $T = I$ verwendet.

Ein anderer Ansatz verwendet die sogenannte *TV–Norm*, die die *totale Variation* misst[191],

$$\|f\|_{TV} := \int \|\nabla f\|_1 \,, \tag{7.5}$$

und wesentlich sensibler für Ecken ist. Der Funktionen mit endlicher TV–Norm bilden einen Banachraum, oft als $BV(\Omega)$ geschrieben, nämlich den Raum der Funktionen *von beschränkter Variation*. In der Tat kann man zeigen, siehe [38, S. 37, Theorem 2.7], daß es eine enge Beziehung zwischen dieser Norm und den Konturkurven eines Bildes gibt. In unserem diskreten Beispiel ist nun

$$\|f\|_{TV} = \sum_{j=0}^{N-1} |f(j+1) - f(j)| = \|Df\|_1 \,.$$

Die Lösung von

$$\min_f \|f\|_{TV} = \|Df\|_1 \,, \qquad Tf = g$$

erhält man als Lösung eines linearen Optimierungsproblems

$$\min \sum_{j=0}^{N-1} |f(j+1) - f(j)| \,, \qquad Tf = g.$$

Um zu sehen, daß das wirklich ein "normales" lineares Programm ist, setzen wir $y_i = f(j+1) - f(j)$ und die zu minimierende Summe

$$\sum_{j=0}^{N-1} |y_j| = \sum_{j=0}^{N-1} y_j + u_j, \qquad u_j = \begin{cases} 0, & y_j \geq 0, \\ -2\,y_j, & y_j < 0, \end{cases}$$

was wir in die Ungleichungsnebenbedingungen

$$0 \leq u_j, \qquad -2y_j \leq u_j, \qquad j = 0, \ldots, N-1,$$

[190] Man kann diese natürlich auch wieder kombinieren, also beispielsweise die Norm des Vektors mit denen seiner ersten und zweiten Ableitung "ausbalancieren".

[191] Als Vektornorm wird, je nach Literaturstelle, entweder $\|\cdot\|_1$ oder $\|\cdot\|_2$ verwendet; für einen Variationsansatz ist das nicht so relevant, denn da wird das Optimierungsproblem in eine Differentialgleichung umgeformt, für den Optimierungsansatz allerdings schon, weil man dann eben nach der Diskretisierung ein lineares Programm erhält oder nicht.

umschreiben lässt – die Minimierung sorgt dann schon dafür, daß u_j automatisch den richtigen Wert annimmt. Die Zielfunktion ist dann

$$\sum_{j=0}^{N-1} (f(j+1) - f(j)) + u_j = f(N) - f(0) + \sum_{j=0}^{N-1} u_j$$

und das[192] lineare Problem hat die Form

$$\min_{f,u} f(N) - f(0) + \sum_{j=0}^{N-1} u_j, \qquad \begin{bmatrix} T & 0 \\ -T & 0 \\ 0 & I \\ 2D & I \end{bmatrix} \begin{bmatrix} f \\ u \end{bmatrix} \geq \begin{bmatrix} g \\ -g \\ 0 \\ 0 \end{bmatrix} \tag{7.6}$$

mit

$$D = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & -1 \end{bmatrix},$$

was sich nun wirklich mit dem Siplexalgorithmus in Angriff nehmen lässt. Zumindest im diskreten Fall und ohne Auftreten von Rauschen kann man sich also den Variationsansatz für die Minimierung der totalen Variation erst einmal sparen. Da ℓ_1–minimale Vektoren eine viel stärkere Tendenz zu kleinen Trägern haben, könnten wir hier eine bessere Rekonstruktion unseres ersten Beispielvektors erwarten.

7.3 Es rauscht mal wieder

Jetzt aber zurück zu unseren verrauschten Daten aus (7.1). Es ist klar, daß man ohne Annahmen an das Rauschen keine vernünftigen Aussagen machen kann. Und wenn man schon annimmt, dann kann man auch gleich **die** Standardannahme machen, nämlich daß das Rauschen normalverteilt mit Mittelwert Null und Standardabweichung σ ist, daß also

$$\int \epsilon = 0, \qquad \int \epsilon^2 = \sigma^2$$

gilt, denn dann sind die Nebenbedingungen unserer inversen Probleme nur unwesentlich anders und ergeben sich zu

$$0 = \int \epsilon = \int g - Tf \qquad \text{und} \qquad \sigma^2 = \int \epsilon^2 = \int (g - Tf)^2 .$$

Und wenn wir σ nicht kennen? Dann könnten wir es einfach als einen weiteren Parameter wählen und so schätzen, daß das Ergebnis möglichst gut wird – wir "erklären" die Daten unter der Annahme des "freundlichsten" Rauschens:

$$\min_f E(f), \qquad \int g - Tf = 0, \quad \int (g - Tf)^2 = \sigma^2. \tag{7.7}$$

[192] Zuerst einmal unbeschränkte, siehe z.B. [47].

Allerdings hat das einen kleinen Nachteil, denn neben dem Rauschen könnten auch andere "hochfrequente" Anteile eines Bildes, nämlich Texturen, nur sehr schwer vom Rauschen zu unterscheiden sein – ein verrauschtes Fernsehbild eines Zebras ist ein schönes Beispiel. Nun ist die Definitionsgleichung des Fehlers

$$\epsilon = g - Tf$$

ja eigentlich sehr einfach, und wann immer wir eine Schätzung für f bestimmt haben, ist das geschätzte[193] Rauschen ebenfalls festgelegt. Daher verwendet man Verfahren, die *iterativ* aus dem geschätzten Rauschen eine neue Approximation bestimmen, daraus dann wieder ein neues Rauschen und so weiter.

Der Algorithmus selbst setzt einfach $\epsilon_0 = 0$ und bestimmt für $k = 1, 2, \ldots$ die Approximation f_k als Lösung des Minimierungsproblems

$$\min_f \int_\Omega \|\nabla f\|_1 + \lambda \int_\Omega (g - Tf + \epsilon_{k-1})^2 \tag{7.8}$$

und setzen dann $\epsilon_k = g - Tf_k$. In [41] wird gezeigt, daß für $T = I$, also den Fall "reinen" Entrauschens, die Folge der f_k gegen das verrauschte Signal $g + \epsilon$ konvergiert, aber in der BV–Norm sich zuerst einmal dem *unverrauschten* Originalsignal f annähert, zumindest so lange, bis $\|f_k - f\| \leq \sigma^2$ ist. Deswegen führt man Algorithmus so lange durch, bis das Rauschen der f_k, also das ϵ_k, größer wird oder bis man bis auf σ^2 an der gesuchten Lösung heran ist.

7.4 Ein bisschen Variationsrechnung

Zuletzt hatten wir es bei der Minimierung immer wieder mit Funktionalen der Form

$$E(f) = \int_\Omega \|\mathscr{A} f\| + \lambda \left(g - Tf\right)^2 \, d\omega,$$

wobei $\mathscr{A}$ ein Differentialoperator ist, zu tun. Wie löst man eigentlich derartige Minimierungsprobleme? Lagrange–Multiplikatoren sind hier nicht hilfreich, weil es a priori erst einmal nichts gibt, wonach wir $E(f)$ partiell ableiten könnten, und wenn wir f als Linearkombination einer Basis ansetzen[194], dann muß diese Basis erst einmal gewährleisten, daß die Minimallösung in diesem endlichdimensionalen Raum auch wirklich nahe genug an die wirkliche Optimallösung herankommt.

Die Minimierung von Funktionalen ist genau das Thema der *Variationsrechnung* – im Übrigen ein sehr klassisches Gebiet der Analysis. Sehen wir uns einmal kurz die wesentlichen Ideen an, und zwar in der Darstellung von [19]. Ein *Variationsproblem* befasst sich zuerst einmal mit der Minimierung eines Funktionals der Form

$$J[f] = \int_a^b F\left(x, f, f'\right) dx \tag{7.9}$$

[193] Klingt besser als "geraten".

[194] **Dann** könnten wir nach den Koeffizienten in dieser Darstellung ableiten und Lagrange–Multiplikatoren verwenden.

bezüglich der Funktion f, wobei $F(x, y, y') \quad : \quad \mathbb{R}^3 \to \mathbb{R}$ eine Funktion in den formalen Variablen x, y, y' ist.

Beispiel 7.4 (Kürzester Weg) *Gesucht ist eine Funktion f mit $f(a) = f_a$, $f(b) = f_b$ und kürzester Bogenlänge*

$$\int_a^b \sqrt{1 + f'(x)^2}\, dx, \qquad \Rightarrow \qquad F(x, y, y') = \sqrt{1 + (y')^2}.$$

Das wesentliche Konzept ist das *Differential* eines Funktionals: Für ein Funktional J fixieren wir f und betrachten für Funktionen h die Differenz

$$\Delta J[h] = \Delta J[h, f] = J[f + h] - J[f].$$

Wir nennen J differenzierbar an f, wenn es ein lineares Funktional[195] $\delta J[h]$ gibt, so daß

$$\Delta J[h] = \delta J[h] + o(\|h\|) = \delta J[h] + \varepsilon_h \|h\|, \qquad \lim_{\|h\| \to 0} \varepsilon_h = 0.$$

Wenn so ein Funktional $\delta J[h]$ exsitiert, dann ist es eindeutig, denn gäbe es zwei Differentiale $\delta_1 J[h]$ und $\delta_2 J[h]$, dann wäre

$$\delta_1 J[h] - \delta_2 J[h] = \Delta J[h] - \varepsilon_{1,h} \|h\| - \Delta J[h] + \varepsilon_{2,h} \|h\| = (\varepsilon_{2,h} - \varepsilon_{1,h}) \|h\|,$$

also

$$\lim_{\|h\| \to 0} \frac{\delta_1 J[h] - \delta_2 J[h]}{\|h\|} = \lim_{\|h\| \to 0} \frac{\varepsilon_{2,h} - \varepsilon_{1,h}}{\|h\|} = 0,$$

aber ein lineares Funktional[196] mit dieser Eigenschaft muss das Nullfunktional sein.

Bemerkung 7.5 *Das Funktional $J[f]$ ist normalerweise* nicht *differenzierbar, sondern "nur" konvex, d.h.*

$$J[\alpha f + (1 - \alpha) g] \leq \alpha\, J[f] + (1 - \alpha)\, J[g], \qquad \alpha \in [0, 1].$$

In diesem Fall betrachtet man nicht das Differential, sondern das sogenannte Subdifferential*, siehe z.B. [43], was die Dinge zwar schon etwas schwerer aber nicht unmöglich macht.*

Proposition 7.6 *Wenn das Funktional $J[f]$ an f^* ein Minimum hat, dann ist $\delta J[h, f^*] = 0$ für alle h.*

Beweis: Ist f^* eine Minimalstelle, dann ist für hinreichend kleines $\|h\|$

$$0 \leq J[f^* + h] - J[f^*] = \delta J[h] + \varepsilon_h \|h\|,$$

weswegen $\delta J[h] \geq 0$ sein muß, da der andere Term gegen Null geht. Nun ist aber

$$0 \leq \delta J[h] = -\delta J[-h] \leq 0$$

[195] Diese Funktional entspricht der Richtungsableitung in Richtung h.

[196] Die Differenz zwischen zwei linearen Funktionalen ist wieder ein lineares Funktional

nur mit $\delta J[h] = 0$ zu erreichen. □

Die Bedingung an das Differential ist ja ganz schön, aber es ist in keinster Weise klar, wie man sowas praktisch ausnutzen könnte. Daher verwenden wir einmal eine Taylor–Entwicklung von F und erhalten, daß

$$\begin{aligned}\Delta J[h] &= J[f+h] - J[f] = \int_a^b F(x, f+h, f'+h') - F(x, f, f')\, dx \\ &= \int_a^b h\,\frac{\partial F}{\partial y}(x, f, f') + h'\,\frac{\partial F}{\partial y'}(x, f, f')\, dx + \cdots,\end{aligned}$$

weswegen

$$\delta J[h] = \int_a^b h\,\frac{\partial F}{\partial y}(x, f, f') + h'\,\frac{\partial F}{\partial y'}(x, f, f')\, dx \tag{7.10}$$

ist.

Satz 7.7 *Ist eine Funktion f Minimallösung, dann ist*

$$\frac{\partial F}{\partial y}(x, f, f') - \frac{\partial}{\partial x}\frac{\partial F}{\partial y'}(x, f, f') = 0. \tag{7.11}$$

Die Differentialgleichung (7.11) heißt *Euler–Lagrange–Gleichung* des Variationsproblems. Das ist eine Differentialgleichung in f, die man mit den verschiedensten Verfahren numerisch angehen kann.

Der Beweis von Satz 7.7 basiert auf dem folgenden Lemma.

Lemma 7.8 *Seien $f_0, \ldots, f_n \in C[a,b]$ so, daß*

$$\int_a^b \sum_{j=0}^{n} f_j(x)\, h^{(j)}(x) = 0 \tag{7.12}$$

für alle $h \in C^n[a,b]$ mit $h^{(j)}(a) = h^{(j)}(b) = 0$, $j = 0, \ldots, n$, dann gilt

$$\sum_{j=0}^{n} (-1)^j\, f_j^{(j)}(x) = 0, \qquad x \in [a,b]. \tag{7.13}$$

Beweis: Partielle Integration von

$$\int_a^b f(x)\, h^{(j)}(x)\, dx = (-1)^j \int_a^b f^{(j)}(x)\, h(x)\, dx$$

zeigt, daß genau dann $\int_a^b f(x)\, h^{(j)}(x)\, dx = 0$ für alle h, die am Rand verschwinden, gilt, wenn

$$0 = f^{(j)} \qquad \Leftrightarrow \qquad f \in \Pi_{j-1}$$

ist. Sei F_j die $(n-j)$–te Stammfunktion von f_j, also

$$F_j = \int_a^x \cdots \int_a^{t_2} f_j(t)\, dt_1\, dt_2 \cdots dt_{n-j},$$

dann ist $f_j = F_j^{(n-j)}$ und partielle Integration liefert wieder einmal

$$\begin{aligned} 0 &= \int_a^b \sum_{j=0}^n f_j(x)\, h^{(j)}(x)\, dx = \sum_{j=0}^n \int_a^b F_j^{(n-j)}(x)\, h^{(j)}(x)\, dx \\ &= (-1)^n \sum_{j=0}^n \int_a^b (-1)^j F_j(x)\, h^{(n)}(x)\, dx = (-1)^n \int_a^b \sum_{j=0}^n (-1)^j F_j(x)\, h^{(n)}(x)\, dx, \end{aligned}$$

also

$$\sum_{j=0}^n (-1)^{n-j} F_j(x) \in \Pi_{n-1} \qquad \Rightarrow \qquad f_n = -\sum_{j=0}^{n-1} (-1)^j F_j + p, \quad p \in \Pi_{n-1}.$$

Damit ist f_n mindestens eine C^1–Funktion, denn alle Funktionen F_j, $j = 0, \ldots, n-1$, auf der rechten Seite sind Stammfunktionen stetiger Funktionen. Nimmt man nun die erste Ableitung, dann ist

$$f_n' - f_{n-1} = -\sum_{j=0}^{n-2} (-1)^j F_j' + p'$$

wieder eine C^1–Funktion, kann also noch einmal differenziert werden und damit ist

$$(f_n' - f_{n-1})' = f_n'' - f_{n-1}'$$

eine wohldefinierte stetige Funktion. Setzen wir dieses Argument induktiv fort, so erhalten wir schließlich, daß

$$\sum_{j=0}^n (-1)^j\, f_j^{(j)}$$

eine stetige Funktion ist und jetzt liefert und partielle Integration, daß

$$0 = \int_a^b \left(\sum_{j=0}^n (-1)^j\, f_j^{(j)} \right)(x)\, h(x)\, dx = 0, \qquad h(a) = h(b) = 0,$$

ist, und da das für alle stetigen h gelten muß, muß die stetige Funktion in (7.13) in der Tat auf ganz $[a, b]$ verschwinden. □

Übung 7.2 Zeigen Sie: Ist f stetig und erfüllt

$$\int_a^b f(x) h(x)\, dx = 0, \qquad h \in C[a, b], \quad h(a) = h(b) = 0,$$

dann ist $f = 0$. ♢

Beweis von Satz 7.7: Wir wenden einfach Lemma 7.8 auf die Identität

$$0 = \delta J[h] = \int_a^b h \, \frac{\partial F}{\partial y}\left(x, f, f'\right) + h' \, \frac{\partial F}{\partial y'}\left(x, f, f'\right) \, dx$$

an. □

Beispiel 7.9 (Kürzester Weg, Teil II) *In diesem Fall ist* $F\left(x, y, y'\right) = \sqrt{1 + y'^2}$ *und damit*

$$\frac{\partial F}{\partial y} = 0, \qquad \frac{\partial F}{\partial y'} = \frac{1}{2} \frac{2y'}{\sqrt{1 + y'^2}} = \frac{y'}{\sqrt{1 + y'^2}}.$$

Die Euler–Lagrange–Gleichung lautet also

$$\frac{\partial}{\partial x} \frac{\partial F}{\partial y'} = 0 \Rightarrow \frac{\partial F}{\partial y'} = C,$$

und somit ist

$$C^2 = \frac{y'^2}{1 + y'^2} \qquad \Rightarrow \qquad y' = \pm \sqrt{\frac{C^2}{1 - C^2}},$$

also ist y' *eine Konstante und der Weg eine Gerade – die Konstante* C *bestimmt sich aus den Randbedingungen!*

Nun, um Satz 7.7 beweisen zu können, hätten wir nicht diese allgemeine Form von Lemma 7.8 benötigt. Allerdings hatten wir es bei unseren Beispielen ja auch schon mit Funktionalen zu tun, die höhere Ableitungen als nur erste Ableitungen benötigt haben, und da reicht halt Satz 7.7 nicht mehr aus. Nicht so schlimm: Ist $F\left(x, y_0, \ldots, y_n\right)$ eine stetige Funktion in $n + 1$ Variablen und betrachten wir das Funktional

$$J[f] = \int_a^b F\left(x, f, \ldots, f^{(n)}\right) \, dx,$$

dann erhalten wir mit genau derselben Argumentation wie oben, daß

$$\delta J[h] = \int_a^b \sum_{j=0}^{n} h^{(j)}(x) \, \frac{\partial F}{\partial y_j}\left(x, f, \ldots, f^{(n)}\right) \, dx$$

und die Euler–Lagrange–Gleichungen nehmen die Form

$$0 = \sum_{j=0}^{n} (-1)^j \frac{\partial^j}{\partial x^j} \frac{\partial F}{\partial y_j}\left(x, f, \ldots, f^{(n)}\right) \tag{7.14}$$

an. Bezüglich f ist das eine nichtlineare Differentialgleichung der Ordnung $2n$ – nicht gerade das, was man im Vorübergehen erledigt, aber eben auch nicht unlösbar.

Beispiel 7.10 (Entrauschen, Teil I) *Nehmen wir jetzt aber einmal ein Variationsproblem aus der Signalverarbeitung, und zwar das Entrauschen mit dem Funktional*

$$J[f] = \int_a^b \left(f'(x)\right)^2 + \lambda \left(g(x) - f(x)\right)^2 \, dx,$$

dann ist $F(x, y, y') = y'^2 + \lambda \left(g(x) - y\right)^2$ *und somit*

$$\frac{\partial F}{\partial y} = -2\lambda \left(g(x) - y\right), \qquad \frac{\partial F}{\partial y'} = 2y', \qquad \frac{\partial}{\partial x}\frac{\partial F}{\partial y'} = 2y'',$$

was zu der Differentialgleichung

$$y'' = \lambda \left(y - g(x)\right) \tag{7.15}$$

führt, deren Fouriertransformierte[197] *nach Satz 1.8*

$$-\xi^2 \widehat{y}(\xi) = \lambda \left(\widehat{y} - \widehat{g}(\xi)\right) \qquad \Leftrightarrow \qquad \left(\lambda + \xi^2\right) \widehat{y}(\xi) = \lambda \widehat{g}(\xi)$$

ist und uns die Lösung[198]

$$y = \left(\frac{\lambda \widehat{g}}{\lambda + (\cdot)^2}\right)^\vee = \left(\frac{\widehat{g}}{1 + (\cdot)^2/\lambda}\right)^\vee$$

liefert. Das ist für $\lambda > 0$ *auch alles wohldefiniert.*

Das kann man sich mal an einem Beispiel ansehen, nämlich einem verrauschten Signal mit zwei "Stufen" wie in Abb. 7.1, das wir mit Hilfe dieser Methode entrauschen wollen. Beispiele für das Entrauschen sieht man in Abb. 7.1 und Abb. 7.2, die Funktionen werden in der Tat mit $\lambda \to 0^+$ *immer glatter*[199]*, zeigen aber auch eine Tendenz, immer mehr zu verschmieren. Man muß also wissen, wann man mit der Variationsminimierung aufzuhören hat.*

Bemerkung 7.11 *Es gilt*

$$\lim_{\lambda \to 0} \widehat{y}(\xi) = \lim_{\lambda \to 0} \frac{\widehat{g}(\xi)}{1 + \xi^2/\lambda} = \widehat{g}(x)\, \delta(\xi),$$

und damit ist y *eine konstante Funktion mit Integral* $\widehat{g}(0)$*.*

Beispiel 7.12 *Mit dem "Laplace"–Operator* f''*, also dem Funktional*

$$J[f] = \int_a^b \left(f'(x)\right)^2 + \lambda \left(g(x) - f(x)\right)^2 \, dx,$$

[197] Die Differentialgleichung (7.15) sieht ziemlich nach der sogenannten *Wärmeleitungsgleichung* aus (in der sich aber Orts- und Zeitableitugen finden) und so ist es vielleicht jetzt etwas weniger Überraschend, daß Fouier mit "seiner" Transformierten in erster Linie die Lösung der Wärmeleitungsgleichung im Auge hatte.

[198] So ganz richtig ist das noch nicht, denn schließlich nimmt die Fouriertransformierte ja im Moment *irgendein* Verhalten von f und g außerhalb von $[a, b]$ an, das man noch irgendwie unterbringen muss.

[199] Denn je kleiner λ ist, desto mehr gilt die Devise "Glattheit vor Genauigkeit".

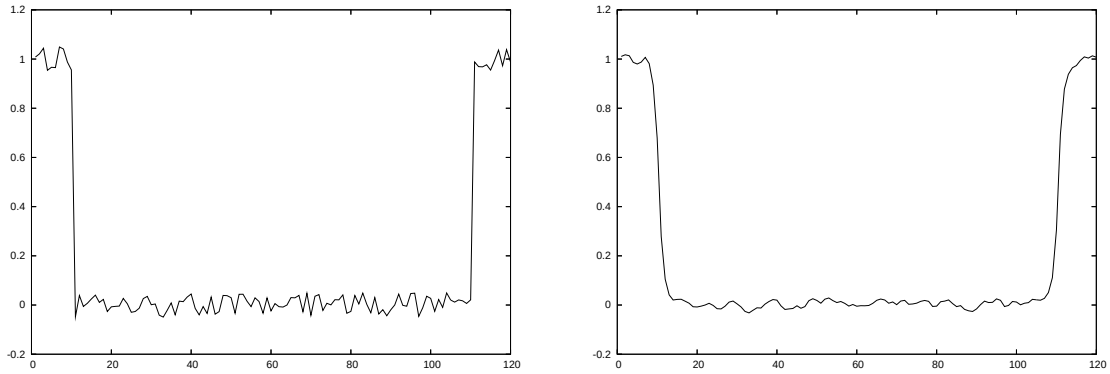


Abbildung 7.1: Ein zufällig verrauschtes Signal *(links)* und die Variationslösung für $\lambda = 1$ *(rechts)*.

ist die Sache systematisch nicht groß anders, nur hat die Euler–Lagrange–Gleichung jetzt die Form

$$0 = \frac{\partial F}{\partial y} + \frac{\partial^2}{\partial x^2}\frac{\partial F}{\partial y''} \qquad \Leftrightarrow \qquad y^{(4)} = \lambda(g - y) \qquad \Leftrightarrow \qquad \left(\lambda + \xi^4\right)\widehat{y}(\xi) = \lambda\widehat{g}(\xi),$$

und somit ist

$$y = \left(\frac{\lambda\widehat{g}}{\lambda + (\cdot)^4}\right)^{\vee}.$$

Wie das aussieht, kann man schließlich in Abb. 7.3 erkennen.

Bemerkung 7.13 *Die Fouriertransformationen*

$$\widehat{y}(\xi) = \frac{g}{1 + \xi^2/\lambda} \qquad \textit{bzw.} \qquad \widehat{y}(\xi) = \frac{g}{1 + \xi^4/\lambda} \tag{7.16}$$

lassen sich auch noch anders interpretieren, nämlich als

$$y = g * f, \qquad f = \left(\frac{1}{1 + (\cdot)^2/\lambda}\right)^{\vee} \quad \textit{bzw.} \quad f = \left(\frac{1}{1 + (\cdot)^4/\lambda}\right)^{\vee}, \tag{7.17}$$

also einfach nur als Filterung der Daten. Da man Faltung mit Filtern sowieso am effizientesten über die FFT durchführt[200]*, ist diese Beobachtung allerdings nicht wirklich praxisrelevant. Wie dem auch sei: Diese Form der Glättung lässt sich mit einem Aufwand von $O\left(N \log N\right)$ bei N Abtastpunkten von g durchführen und das ist schon ziemlich "billig".*

[200] Und dann auch keine Probleme mit der Tatsache hat, daß es sich um einen IIR–Filter handelt.

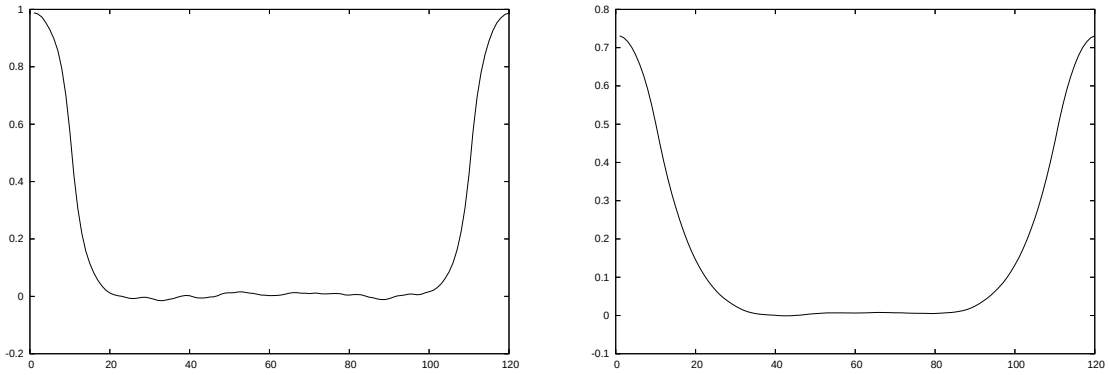


Abbildung 7.2: Lösungen des Variationsproblems für $\lambda = 0.1$ *(links)* und $\lambda = 0.01$ *(rechts)*. Die Funktionen werden immer glatter, verschmieren aber auch mehr und mehr die Gestalt der Ausgangsfunktion und konvergieren gegen die konstante Funktion mit demselben Integral wie g.

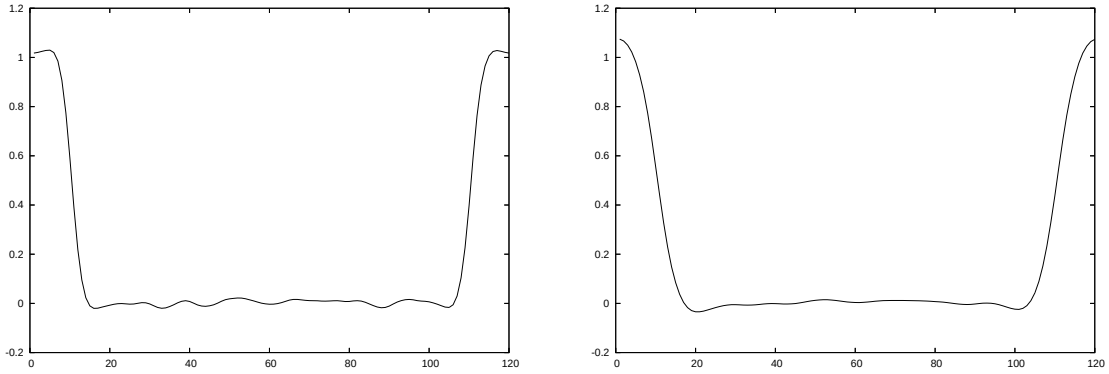


Abbildung 7.3: Lösungen des Variationsproblems mit zweiter Ableitung für $\lambda = 0.1$ *(links)* und $\lambda = 0.01$ *(rechts)*. Was vorher über immer massivere Glättung gesagt wurde, gilt natürlich auch hier.

Uns ist in alten mæren
wunders viel geseit
von Helden lobebæren
von grôzer arebeit

Das Nibelungenlied

7 Literatur

[1] N. I. Akhieser, *Lectures on integral transforms*, Translations of Mathematical Monographs, vol. 70, AMS, 1988.

[2] A. Björck, *Numerical methods for least squares problems*, SIAM, 1996.

[3] W. L. Briggs and V. E. Henson, *The dft: An owner's manual for the discrete fourier transform*, SIAM, 1995.

[4] J. W. Cooley, *The re–discovery of the Fast Fourier Transform*, Mikrochimica Acta **3** (1987), 33–45.

[5] _____, *How the FFT gained acceptance*, A History of Scientific Computing (S. G. Nash, ed.), ACM–Press and Addison–Wesley, 1990, pp. 133–140.

[6] J. W. Cooley and J. W. Tukey, *An algorithm for machine calculation of complex Fourier series*, Math. Comp. **19** (1965), 297–301.

[7] M. Cotronei, M. L. Lo Casio, and T. Sauer, *Dual non-negative rational symbols with arbitrary approximation order*, Appl. Numer. Anal. **51** (2004), 497–510.

[8] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Commun. on Pure and Appl. Math. **41** (1988), 909–996.

[9] _____, *Ten lectures on wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61, SIAM, 1992.

[10] J. M. DeVilliers, C. A. Micchelli, and T. Sauer, *Building refinable functions from their values at integers*, Calcolo **37** (2000), no. 3, 139–158.

[11] R. A. DeVore and G. G. Lorentz, *Constructive approximation*, Grundlehren der mathematischen Wissenschaften, vol. 303, Springer, 1993.

[12] E. Doblhofer, *Zeichen und wunder. geschichte und entzifferung verschollener schriften und sprachen*, Paul Neff Verlag, Wien. Lizenzausgabe Weltbild Verlag, 1990.

[13] FFTW, *FFTW – the Fastest Fourier Transform in the West*, `http://www.fftw.org`, 2003.

[14] J. Foley, A. van Dam, S. Feiner, and J. Hughes, *Computer graphics*, 2nd ed., Addison Wesley, 1990.

[15] O. Föllinger, *Laplace-, Fourier- und z–Transformation*, Hüthig, 2000.

[16] O. Forster, *Analysis 3. Integralrechung im $\mathbb{R}^n$ mit Anwendungen*, 3. ed., Vieweg, 1984.

[17] C. Gasquet and P. Witomski, *Fourier analysis and applications. Filtering, numerical computation, wavelets*, Texts in Applied Mathematics, vol. 30, Springer, 1998.

[18] J. von zur Gathen and J. Gerhard, *Modern computer algebra*, Cambridge University Press, 1999.

[19] I. M. Gelfand and S. V. Fomin, *Calculus of variations*, Prentice–Hall, 1963, Dover reprint, 2000.

[20] J. W. von Goethe, *Götz von berlichingen mit der eisernen hand. ein schauspiel*, Selbstverlag Goethe und Merck, 1773.

[21] G. Golub and C. F. van Loan, *Matrix computations*, 3rd ed., The Johns Hopkins University Press, 1996.

[22] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete mathematics*, 2nd ed., Addison–Wesley, 1998.

[23] D. Ch. von Grüningen, *Digitale Signalverarbeitung*, VDE Verlag, AT Verlag, 1993.

[24] R. W. Hamming, *Digital filters*, Prentice–Hall, 1989, Republished by Dover Publications, 1998.

[25] H. Handels, *Medizinische bildverarbeitung*, B. G. Teubner, 2000.

[26] G. H. Hardy and W. W. Rogosinsky, *Fourier series*, 3. ed., Cambridge University Press, 1956, Republished by Dover Publications, 1999.

[27] C. Herley and M. Vetterli, *Wavelets and recursive filter banks*, IEEE Transactions of Signal Processing **41** (1993), 2536–2556.

[28] H. Heuser, *Lehrbuch der Analysis. Teil 2*, 2. ed., B. G. Teubner, 1983.

[29] N. J. Higham, *Accuracy and stability of numerical algorithms*, SIAM, 1996.

[30] E. Hille, *Analytic function theory*, 2nd ed., Chelsea Publishing Company, 1982.

[31] R. A. Horn and C. R. Johnson, *Matrix analysis*, Cambridge University Press, 1985.

[32] K. D. Kammeyer and K. Kroschel, *Digitale Signalverarbeitung*, Teubner Studienbücher Elektrotechnik, B. G. Teubner, Stuttgart, 1998.

[33] Y. Katznelson, *An introduction to harmonic analysis*, 2. ed., Dover Books on advanced Mathematics, Dover Publications, 1976.

[34] E. Kreyszig, *Introductionary functional analysis with applications*, John Wiley & Sons, 1978.

[35] G. G. Lorentz, *Approximation of functions*, Chelsea Publishing Company, 1966.

[36] A. K. Louis, P. Maaß, and A. Rieder, *Wavelets*, 2. ed., B. G. Teubner, 1998.

[37] MacTutor, *The MacTutor History of Mathematics archive*, `http://www-groups.dcs.st-and.ac.uk/~history`, 2003, University of St. Andrews.

[38] S. Mallat, *A wavelet tour of signal processing*, 2. ed., Academic Press, 1999.

[39] H. M. Mhaskar and D. V. Pai, *Fundamentals of Approximation Theory*, Narosa Publishing House, 2000.

[40] J. Nocedal and S. J Wright, *Numerical optimization*, Springer Series in Operations Research, Springer, 1999.

[41] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, *An iterative regularization method for total variation based image restoration*, Tech. report, UCLA, 2004.

[42] P. Z. Peebles, *Probability, random variables and random signal principles*, McGraw–Hill, 1980.

[43] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[44] T. Sauer, *Numerische Mathematik II*, Vorlesungsskript, Friedrich–Alexander–Universität Erlangen–Nürnberg, Justus–Liebig–Universität Gießen, 2000, `http://www.math.uni-giessen.de/tomas.sauer`.

[45] ———, *Computeralgebra*, Vorlesungsskript, Justus–Liebig–Universität Gießen, 2001, `http://www.math.uni-giessen.de/tomas.sauer`.

[46] ———, *Approximationstheorie*, Vorlesungsskript, Justus–Liebig–Universität Gießen, 2002, `http://www.math.uni-giessen.de/tomas.sauer`.

[47] ———, *Optimierung*, Vorlesungsskript, Justus–Liebig–Universität Gießen, 2002, `http://www.math.uni-giessen.de/tomas.sauer`.

[48] _____, *Splinekurven und –flächen in Theorie und Anwendung*, Vorlesungsskript, Friedrich–Alexander–Universität Erlangen–Nürnberg, Justus–Liebig–Universität Gießen, 2007, `http://www.math.uni-giessen.de/tomas.sauer`.

[49] I. J. Schoenberg, *Contributions to the problem of approximation of equidistant data by analytic functions. part B. – on the second problem of osculatory interpolation. a second class of analytic approximation formulae*, Quart. Appl. Math. **4** (1949), 112–141.

[50] _____, *Cardinal spline interpolation*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 12, SIAM, 1973.

[51] A. Schönhage, *Multiplikation großer Zahlen*, Computing **1** (1966), 182–196.

[52] _____, *Schnelle Multiplikation von Polynomen über Körpern der Charakteristik* 2, Acta Informatica **7** (1977), 395–398.

[53] A. Schönhage and V. Strassen, *Schnelle Multiplikation großer Zahlen*, Computing **7** (1971), 281–292.

[54] H. W. Schüßler, *Digitale Signalverarbeitung*, 3. ed., Springer, 1992.

[55] C. E. Shannon, *Communications in the presence of noise*, Proc. of the IRE **37** (1949), 10–21.

[56] W. Skrandies and A. Jedynak, *Associative learning in human brains – conditioning of sensory–evoked brain activity*, Behavioural Brain Research **107** (2000), 1–8.

[57] P. Spellucci, *Numerische Verfahren der nichtlinearen Optimierung*, Internationale Schriftenreihe zu Numerischen Mathematik, Birkhäuser, 1993.

[58] A. Steger, *Diskrete Strukturen 1. Kombinatorik – Graphentheorie – algebra*, Springer, 2001.

[59] G. Strang and G. Fix, *A Fourier analysis of the finite element variational method*, Constructive aspects of functional analysis, C.I.M.E, Il Ciclo 1971, 1973, pp. 793–840.

[60] W. Sweldens, *The lifting scheme: a custom–design construction of biorthogonal wavelets*, ACHA **3** (1996), 186–200.

[61] G. P. Tolstov, *Fourier series*, Prentice–Hall, 1962, Republished by Dover Publications, 1972.

[62] C. van Loan, *Computational frameworks for the Fast Fourier Transform*, SIAM, 1992.

[63] M. Vetterli and J. Kovačević, *Wavelets and subband coding*, Prentice Hall, 1995.

[64] J. Whittaker, *Interpolatory function theory*, Cambridge Tracts in Math. and Math. Physics, vol. 33, 1935.

[65] K. Yosida, *Functional Analysis*, Grundlehren der mathematischen Wissenschaften, Springer–Verlag, 1965.

7 Index

$L_1(\mathbb{R})$, 4
$L_2(\mathbb{R})$, 4
$L_2(\mathbb{R})$, 12
$L_\infty(\mathbb{R})$, 4
$L_{00}(\mathbb{R})$, 4
$\mathbb{T}_N$, 55
δ–Puls, 5
$\ell_1(\mathbb{Z})$, 4
$\ell_2(\mathbb{Z})$, 4
$\ell_\infty(\mathbb{Z})$, 4
$\ell_{00}(\mathbb{Z})$, 4

Abtastrate, *siehe* Rate, Abtast 55
Abtastsatz, 16, 26
Addierer, 23, 34
Algorithmus
    gieriger, 129, 131
    greedy, 129
Aliasing, 18
Approximation
    m–Term, 129
    beste, 129
Approximationsgüte, 109
    translationsinvarianter Räume, 109
Approximationsordnung, 109
Auflosung
    Frequenz-, 55
Ausschlag, 107
Autokorrelation, 85, 86

B–Spline
    kardinaler, 10, 103
    zentrierter, 11
Bézout–Identitat, *42*
Banachraum, 5, 128
Bandbreite, 15, 19
    effektive, 19
Basis
    Karhunen–Loève–, 134
Bedingungen
    Strang–Fix, 101, *101*, 103, 105, 109
BELL, A. G., 68
Bernoulli–Zahlen, 41
Bestapproximation, 27
Bit Reversal, 62
Blaschke–Produkt, 73
BONAPARTE, N., 6
Butterfly–Element, 61

CAUCHY, A., 32
CHAMPOLLION, J. F., 6
Chinesischer Restsatz, 72

Dauer
    effektive, 19
dB, 68–70
Delayed Feedback, 34
Dezibel, *siehe* dB 68
DFT, 48, *48*, 52
    inverse, 50
    Matrix, 50
    naive Realisierung, 57
Differential, 143
DIRAC, P., 5
Dirac–Puls, 5
Distribution, 5
    Dirac-, 66
    temperierte, 6

Eckenerkennung, 98, 119
Einheit, 31
Einheitswurzel, 50
    primitive, 56
Energie

endliche, 4, 12
Energiefunktional, 137
Entrauschen, 98, 119
erweiterter euklidischer Algorithmus, 43

Faltung, *7*, 22, 32
zyklische, 52, 59
Fejérsche Mittel, 28
Fenster, 64, 66
Bartlett-, 67
Blackmann, 70
Dreiecks-, 67
Gauß, 69
Hamming-, 67
Hann-, 67
Hanning-, *siehe* Fenster, Hann 67
Kaiser-, 70
Rechteck-, 66
FFT, 48, 57, 60
Blockschaltbild, 62
Komplexität, 57
Radix-p, 59
Radix–2, 58
Realisierung, 61
Filter, 21, *21*
-bank, *siehe* Filterbank 75
-länge, 115
Allpass-, *71*, 72, 73, 84–86
antikausaler, 22
Bandpass-, 53, 74
Binomial-, 124
Butterworth-, 87
digitaler, 21
diskreter, 21
ergieerhaltender, 21
FIR-, 22, 27, 30, 38, 39, 85
Gradienten-, 124
IIR-, 22, 35
Integrations-, 37
Kaskadenbild, 24
kausaler, 21, 33
linearer, 21
LTI-, 21, 33
Median-, 125
Mittelwert-, 123
rationaler, 33, 38, 71, 72, 80, 84, 86
Dämpfung, 39, *39*, 41
Realisierung, 34
reeller, 83
rekursiver, *siehe* Filter, rationaler 40
schnelle Berechnung, 60
Spiegel-, 82, 84
stabiler, 40, *40*
steilflankiger, 29
Summations-, 34, 35
Tensorprodukt-, 122
Tiefpass-, 27
zeitinvarianter, 21
Filterbank, 75, 76, *79*
Analyse-, 75, 89, 96
Artefakte, 115, 116
biorthogonale, 81
orthogonale, 81, 86
rationale, 80, 86
Strang–Fix–Bedingungen, 105
Synthese-, 78, 89
FIX, G., 101
Folge
beschränkte, 4
mit endlichem Träger, 4
periodische, 49
periodisiserte, 49
quadratsummierbare, 4
summierbare, 4
Formel
Taylor-, 110
FOURIER, J. B., 6, 12
Fourierkoeffizienten, 13
Fourierreihe, 13, 14, 27
Partialsumme, 27
Fouriertransformation, 6
diskrete, *siehe* DFT 48
inverse, 8
kontinuierliche, 65
schnelle, *siehe* FFT 48
Fouriertransformierte, 6, *6*, 7, 12, 32, 77

auf $L_2(\mathbb{R})$, 12
Normalisierung, 6, 16
Fouriertrtransformation
mehrdimensionale, 121
Frequenz
Abtast-, 16
Nyquist-, 16
Funktion
analytische, 20
bandbeschränkte, 15, 16, 19, 20
beschränkte, 4
Dach-, 113
gleichmäßig stetige, 11
kardinale, 54
mit endlichem Träger, 4, 19
quadratsummierbare, 4
rationale, 71
sigmoidale, 25
Skalierungs-, 96
stabile, 105, 107
summierbare, 4
Transfer-, *siehe* Transferfunktion 22
verfeinerbare, 91, 92, 105, 126

Ganzzahlmultiplikation, 60
GAUSS, C.–F., 11
GIBBS, W., 29
Gibbs–Phänomen, 29
Gleichung
Euler–Lagrange, 144
Gradient, 124
Gruppe
duale, 7

HANN, J. V., 67
Hauptkomponenten, 135
HEISENBERG, W., 19
Hilbertraum, 128

Impulsantwort, 21, 38
endliche, 22
Integrierer, 35
Interpolation, 43
Inverse
Moore–Penrose-, 136

Jackson–Satze, 109

Kaskaden–Schema, 94
Kern
Féjer-, 14
Fejér-, 9
Gauß, 123
Klassifizierung, 119
Kompression, 98, 119
Korrelation, 85
Kovarianz, 133
KRONECKER, L., 5
Kronecker–Delta, 5

LANDAU, CH., 101
LAPLACE, P. S., 14
Latenzzeit, 30
Leakage Pheneomenon, 64
LEBESGUE, H., 11
Lebesgue–Punkt, 5
Leck–Effekt, 64, 66
Leibniz–Regel, 105
Lemma
Riemann–Lebesgue, 11
Liftig scheme, 43

Master Theorem, 58
Matrix
Dreiecks-, 106
Polyphase, 78
Matrixmultiplikation, 60
Maß
Haar-, 7
Median, 118
Modulationsmatrix, *75*, 76, 78, 79
Invertierbarkeit, 80
Momente, 112
verschwindende, 112
MRA, 95
Multiplikatoren
Lagrange-, 138
Multiplizierer, 23, 34

Multiresolution Analysis, *siehe* MRA 95

Norm
    TV-, 140

Operator
    Abtast-, *5*, 16
    Downsampling-, *74*, 75, 76
    Laplace-, 125
    Schoenberg, 54, 55
    stationärer, 21
    Transfer-, 94
    Translations-, 21
    Upsampling-, *74*, 76
Oversampling, 16

PALEY, R., 20
Parseval, 12
PARSEVAL, M.–A., 12
Partialbruchzerlegung, 38, 42, *44*
Perfect Reconstruction, 79
Perfekte Rekonstruktion, 79, 80, 83, 95
Periodisierung, 18
Pixel, 120
Plancherel, 12
POISSON, S., 14
Polymom, 31
Polynom
    faktorisiertes, 72
    Generierung, 98
    Laurent-, 31, 42
    Nullstellen, 72
    Reproduktion, 101
    schnelle Multiplikation, 60
    Taylor-, 98, 110
    trigonometrisches, 27
Primfaktorzerlegung, 59
Problem
    schlecht gestelltes, 137
    Variations-, 142
Pseudoinverse, 136
Puls, 5
Pyramidenschema, 96

Quadraturformel, 36
Quantisierung, 5
Quasiinterpolant, 54, 107

RADEMACHER, H., 101
Rate
    Abtast-, 55
    Sampling, *siehe* Rate, Abtast 55
Raum
    Banach-, 128
    Hilbert-, 128
    translationsinvarianter, 99, 109
Rauschen
    mittelwertfreies, 123
Realisierung, 133
Refinement Equation, 91
Reihe
    Fourier-, *siehe* Fourierreihe 13
    Laurent-, 31
    Potenz-, 31
    trigonometrische, 13
Reihenfolge
    alphabetische, 101
Restglied
    Integral-, 110
RIEMANN, B., 11

Samplingrate, *siehe* Rate, Abtast- 55
SCHOENBERG, I. J., 101
SHANNON, C., 15
Signal, *3*
    diskretes, 3, 26
    EEG, 114
    kontinuierliches, 3
    stochastisches, 133
    zeitabhängiges, 3
Signalraum, 4
Simpson–Regel, 36
sinc, 15, *15*, 16, 66, 68, 97
Sinus Cardinalis, *siehe* sinc 15
Skalierung, 7
Spiegelung, 72
Spline
    kardinaler, 54
Stabilität, 107

STRANG, G., 101
Subband Coding, 74, 75
Subdivision–Schema, 89, 94
Summenformel
    Poisson-, 14, 102
Supremum
    wesentliches, 4
Symbol, 32
System
    instabiles, 40

Tap, 30
Tensorprodukt, 122
Thresholding, 116
    hard, 116
    soft, 116, 117
Toeplitz–Matrix, 60
Torus
    diskreter, 55
Transferfunktion, 22, 24, 26, 71
    relle, 23
Transformation
    z-, 31, 32, 85
      inverse, 32
    Fourier, *siehe* Fouriertransformierte 6
    Wavelet-, 71
Translation, 7
Translationsinvarianz, 95, 99
Trapezregel, 36

Unimodular, 81
Unschärferelation, 19, *19*, 66

Variation
    beschrankte, 140
    totale, 140
Verfeinerungsgleichung, 91
Verzögerer, 23, 34

Wavelet, 96
    -koeffizienten, 96, 109
      Abklingrate, 109
    -zerlegung, 96, *96*, 109
    Littlewood–Paley-, 97
    Packages, 89
    Tensorprodukt-, 125
WIENER, N., 20
Worterbuch, 128
    normiertes, 130

Zeit
    Latenz-, 30
Zeitverzögerung, 79
Zerlegung
    Singularwert-, 136
Zufallsprozess, 133
Zweiskalenbeziehung, 91