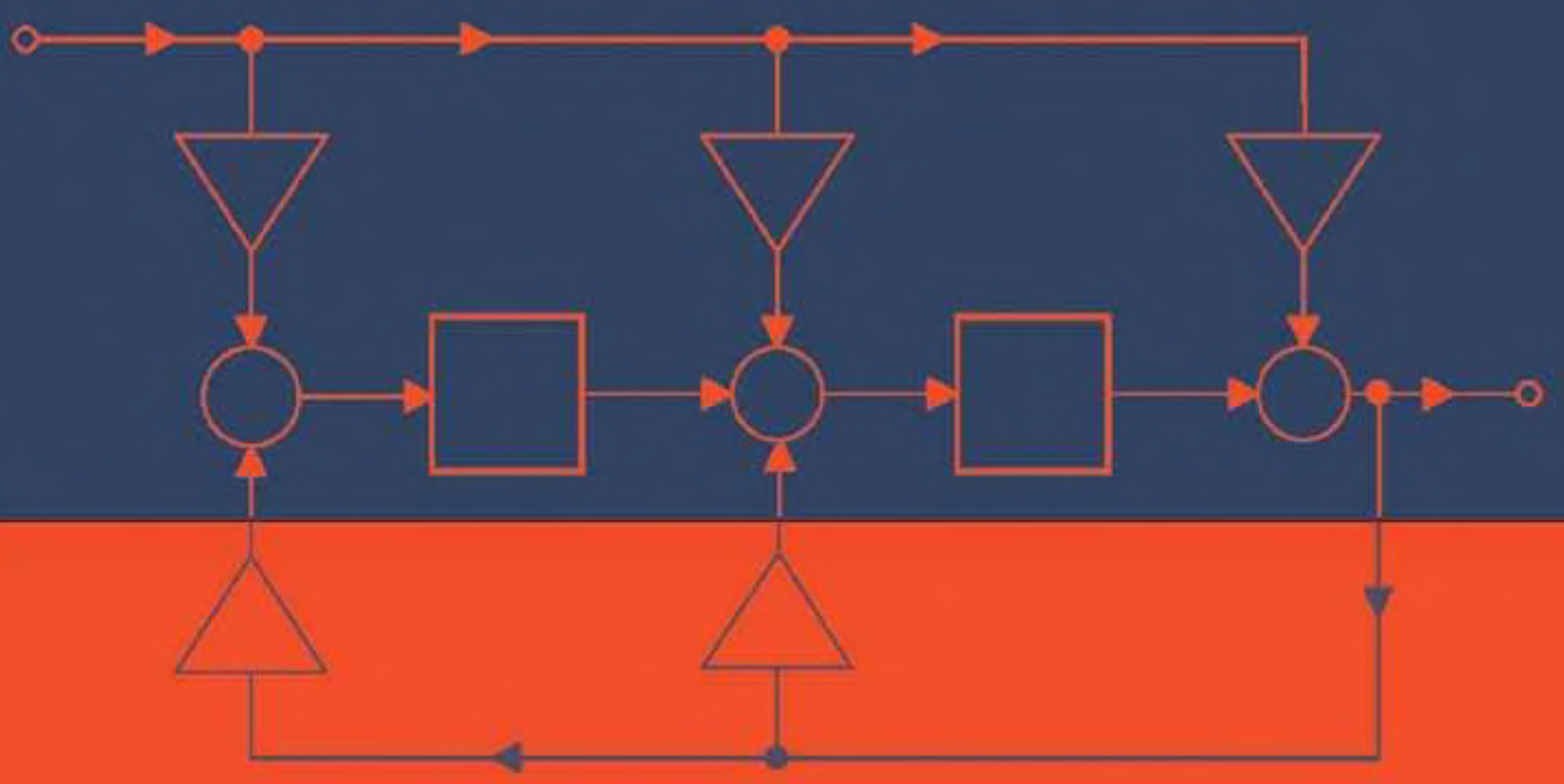
Wendemuth

Grundlagen der digitalen Signalverarbeitung

Ein mathematischer Zugang



Springer

Andreas Wendemuth

Grundlagen der digitalen Signalverarbeitung

Ein mathematischer Zugang

Unter Mitarbeit von
Edin Andelic, Sebastian Barth, Marcel Katz, Sven Krüger,
Mathias Mamsch, Martin Schafföner

Mit 58 Abbildungen

Springer

Prof. Dr. rer. nat. Andreas Wendemuth
Otto-von-Guericke-Universität Magdeburg
Institut für Elektronik, Signalverarbeitung & Kommunikationstechnik
Professur für Kognitive Systeme
Postfach 41 20
39016 Magedeburg
wendemu@iesk.et.uni-magdeburg.de
http://iesk.et.uni-magdeburg.de/ko/

ISBN 3-540-21885-8 Springer Berlin Heidelberg New York

Bibliografische Information der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

Springer. Ein Unternehmen von Springer Science+Business Media
springer.de

Satz: Gelieferte Daten des Autors
Einbandgestaltung: design & production, Heidelberg
Gedruckt auf säurefreiem Papier 7/3020/M - 5 4 3 2 1 0

Für Allegra und Franca

For every complex problem there is a simple solution that is wrong.
George Bernard Shaw

Everything should be explained as simple as possible, but no simpler.
Albert Einstein

Vorwort

Der Entwurf, die Beschreibung oder zumindest die Nutzung von signalverarbeitenden Systemen – seien es biologische, elektronische, mechanische oder andere Systeme – stellt eine Grundlage nahezu jeder technischen Anwendung dar. Kein anderes Gebiet wurde dabei in den letzten Jahrzehnten so intensiv bearbeitet und weiterentwickelt, wie die Signalverarbeitung mit Hilfe von digitalen elektronischen Systemen. Das populärste Ergebnis dieser Entwicklung sind mikroprozessorgesteuerte Systeme, in Form von Computern, Industrie-Automatisierungsanlagen, Kommunikationseinrichtungen oder Consumer-Geräten, um nur einige Kategorien zu nennen. Da Mikroprozessoren, Digitale Signalprozessoren (DSP) und sogar von Betriebssystemen unterstützte Rechner immer billiger werden, werden immer weitere Verfahren digitalisiert werden.

All diese Systeme nutzen ein und den selben erfolgreichen Ansatz: die Digitaltechnik. Dieses Verfahren ist bekanntermaßen vielen zuvor verwendeten technischen Lösungen deutlich überlegen und erlaubt eine enorme Steigerung von Präzision, Robustheit, Komplexität und Effizienz in unzähligen Anwendungen. Die Möglichkeit, das Design kompletter signalverarbeitender Systeme auf die softwareseitige Implementierung der benötigten Algorithmen zu beschränken, vereinfacht den Entwurf komplexer Systeme derart, dass sich in einer ständig wachsenden Anzahl von Fällen eine „Digitalisierung“ von Problemstellungen in der üblicherweise nicht digitalen Umwelt lohnt.

Bezeichnenderweise spielt es im Umgang mit digitalen Signalverarbeitungssystemen selten eine Rolle, ob es sich um ein digitales Signal handelt oder nicht. Der wichtigste Unterschied zwischen analogen Signalen und den Signalen, die mit üblichen digitalen Systemen – beispielsweise Mikroprozessoren – verarbeitet werden, ist deren *zeitliche* Quantisierung (Abtastung, Taktung). Digitale Signale oder digitale Systeme sind jedoch nicht notwendigerweise zeitdiskret, denn die Eigenschaft „digital “ bezieht sich auf die *Amplituden*quantisierung. Letztere kann jedoch heutzutage aufgrund gesteigerter Wortbreiten oftmals getrost vernachlässigt werden und daher werden im Folgenden fast ausschließlich Effekte behandelt, die aus der *zeitlichen* Quanti-

sierung eines analogen Signals resultieren. Somit würde ein besserer Titel für dieses Buch eher „Grundlagen der Verarbeitung zeitdiskreter Signale“ lauten. Es ist jedoch verbreitet, dies mit der Bezeichnung „digital“ zusammenzufassen und man sollte sich darüber im Klaren sein, dass die alleinige Behandlung *zeitdiskreter* Signale ein großes Gebiet der Digitaltechnik ausschließt, nämlich die kombinatorische Logik, die heutzutage zumeist in Form von programmierbaren Logikbausteinen (PLDs) einen großen Stellenwert unter den digitalen Systeme ausmacht. Im Folgenden wollen wir den Unterschied zwischen zeit– und amplitiudendiskreten Systemen genau auseinander halten.

Dieses Buch vermittelt die Grundlagen, die benötigt werden, um zeitdiskrete Signalverarbeitungssysteme zu berechnen und zu entwerfen. Ein großer Teil der digitalen Signalverarbeitungssysteme ist, wie eben geschildert, zeitdiskret. Das Buch beschreibt in Kap. 1 zunächst Eigenschaften zeitdiskreter Signale. Es folgt mit den Kapiteln 2 und 3 eine Gegenüberstellung von analogen und zeitdiskreten Systemen, wobei gewisse Parallelen aufgezeigt werden, die dem Leser, der mit der analogen Signalverarbeitung vertraut ist, den Einstieg in die digitale (zeitdiskrete) Signalverarbeitung erleichtern. In Kap. 4 folgt die Umwandlung von analogen in digitale Signale (und umgekehrt) mit den dabei zu beachtenden Forderungen an die Abtastung des ursprünglichen Signals und dessen Rekonstruktion. Dabei werden Aspekte der Informations– und Codierungstheorie gestreift. Wichtige Methoden zur Verarbeitung zeitdiskreter Signale, vor allem die einseitige Z-Transformation, werden ausführlich in Kap. 5 behandelt. Zusammen mit der diskreten Fouriertransformation in Kap. 6 ermöglicht dies die Analyse und Synthese der Eigenschaften von linearen, zeitinvarianten Systemen.

Wurden bisher die Eingangssignale als deterministisch betrachtet, gelangen wir schließlich in Kap. 7 zu den zunehmend an Bedeutung gewinnenden stochastischen Signalen und ihrer Beschreibung. Nach der Behandlung von Korrelationsverfahren schliessen sich in Kap. 8 Modellsysteme zur Verarbeitung solcher Signale an. Als prototypisches Anwendungsgebiet werden digitale Systeme der automatischen Sprachsignalverarbeitung vorgestellt.

Alle beschriebenen Methoden werden ausführlich vorgestellt. Mathematische Zusammenhänge werden ausführlich erläutert, was das Buch ohne Zuhilfenahme von Sekundärliteratur lesbar macht. Damit legt dieses Buch stärkeren Wert auf die mathematische Verständlichkeit und Konsistenz der Digitalen Signalverarbeitung, als auf deren Anwendungen. In jedem Kapitel werden gelöste Beispielaufgaben vorgestellt, die den Stoff unmittelbar vertiefen. Am Ende der Kapitel sind Übungsaufgaben zum Selbststudium angegeben.

Das Buch richtet sich an Studierende im Hauptstudium, die Digitale Signalverarbeitung als eigenständiges Fach oder als Teilgebiet der Mustererkennung, des Filterentwurfs oder anderer Fächer studieren. Es richtet sich ebenso an gestandene Ingenieure, Informatiker und Naturwissenschaftler, die sich die Grundlagen der Digitalen Signalverarbeitung selbständig aneignen möchten.

Magdeburg, im Sommer 2004 Andreas Wendemuth

Inhaltsverzeichnis

1

Zeitdiskrete Signale und Signalparameter

Eine physikalische Größe kann als *Träger* einer *Information* dienen, wenn man für ihren *Zustand* oder dessen zeitlichen bzw. räumlichen Verlauf eine Bedeutung vereinbart. Beispielsweise kann der Luftdruck in einem begrenzten Bereich eines Raumes Träger von Sprachinformation sein, wenn eine nahe gelegene Schallquelle entsprechende Luftdruckschwankungen auf der Basis einer Sprachinformation hervorruft. Eine physikalische Größe, für deren Zustand man eine Bedeutung vereinbart hat, bezeichnet man als *„Signal“* , also als ein Merkmal, das dafür geeignet ist, die eingeprägte Information zeitlich oder räumlich zu übertragen.

Lokale physikalische Zustände sind im Allgemeinen nie von ihrer Umgebung isoliert. Die Kopplung zu benachbarten physikalischen Größen ergibt ein zusammenhängendes *System* von Zuständen, durch die sich eine lokale *Änderung eines Zustandes* unter Berücksichtigung des Relativitäts- und des Kausalitätsprinzips ausbreitet.

Ausgehend von dieser Überlegung soll das folgende Kapitel die grundlegenden Eigenschaften zeitdiskreter Signale im Vergleich zu kontinuierlichen analogen Signalen in zunächst sehr einfacher Form verdeutlichen. Danach wagen wir uns an eine mathematische Notation solcher Signale und entwickeln damit die Grundlagen, auf die die weiteren Kapitel dieses Buches in etwas anspruchsvollerer Form aufbauen. Leser, denen zeitdiskrete Signale und deren Eigenschaften bereits bekannt sind, sollten sich die in diesem Buch verwendete Nomenklatur im Kapitel 1.3 ansehen und können dann mit einem der nächsten Kapitel fortfahren.

1.1 Klassifikation von Signalen

Je nach Betrachtungsweise lassen sich Signale in unterschiedliche Klassen aufteilen. Mögliche Unterscheidungskriterien können sein:

- Zeitsignal $\leftrightarrow$ Ortssignal
- analog $\leftrightarrow$ (amplituden)diskret
- kontinuierlich $\leftrightarrow$ diskret (diskontinuierlich)
- periodisch $\leftrightarrow$ nicht periodisch
- (streng) monoton $\leftrightarrow$ nicht (streng) monoton
- stetig $\leftrightarrow$ unstetig
- bandbegrenzt $\leftrightarrow$ nicht bandbegrenzt $\leftrightarrow$ harmonisch
- kausal $\leftrightarrow$ akausal
- energiebegrenzt $\leftrightarrow$ nicht energiebegrenzt
- amplitudenbegrenzt $\leftrightarrow$ nicht amplitudenbegrenzt
- deterministisch $\leftrightarrow$ stochastisch

Die für unsere Betrachtungen relevanten Signaleigenschaften werden im Folgenden näher erläutert.

1.2 Signaleigenschaften

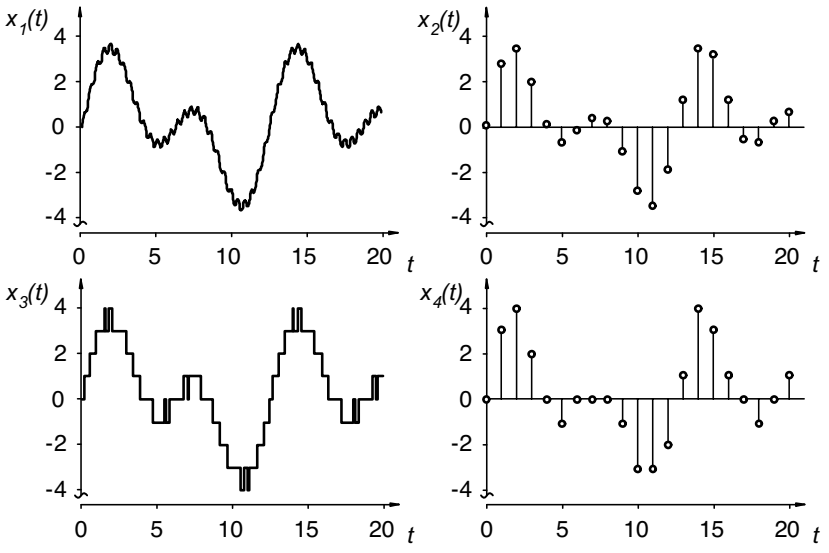


Abbildung 1.1. Kontinuierliche und quantisierte Signale: x_1: kontinuierliches analoges Originalsignal x_2: zeitliche Quantisierung, x_3: Amplitudenquantisierung x_4: zeitliche und Amplitudenquantisierung

In Abb. 1.1 sind die Signale x_2 und x_3 dargestellt, die durch zeitliche Quantisierung sowie durch Amplitudenquantisierung eines Zeitsignals x_1 entstanden sind. Das Signal x_4 ergibt sich durch Zeit- und Amplitudenquantisierung des Signales x_1. Dieses ist

- ein Zeitsignal: x_1 ist eine Funktion des eindimensionalen Signalparameters Zeit.
- kontinuierlich: Es existiert zu jedem Zeitpunkt ein Signalwert $x_1(t)$.
- analog: Jeder Signalwert – evtl. zwischen zwei Grenzen – kann im Signal auftreten.
- skalar: Jeder Signalwert ist ein Skalar, d.h. lässt sich durch eine reelle bzw. eine komplexe Zahl ausdrücken.

Signale sind nicht notwendigerweise *Zeitsignale*, also Signale, die durch Beobachtung eines Umweltparameters im Verlaufe der Zeit entstehen. Beispielsweise kann ein Kamerabild interpretiert werden als ein zweidimensionales Signal, in dem beispielsweise die Farbe eine Funktion über dem Ort (x, y) des Bildes darstellt. Im Folgenden werden wir allerdings von Zeitsignalen ausgehen, sofern dies nicht explizit anders erwähnt wird. Sämtliche Sachverhalte lassen sich jedoch verallgemeinern auf Signale, die keine Funktion der Zeit sind. Die Zeit bzw. der Ort werden als *Signalparameter* des Signals bezeichnet.

1.2.1 Zeitdiskrete Signale

Die *zeitliche Quantisierung* analoger Signale wird häufig als *„Abtastung“* bezeichnet und führt zu (zeit-)diskreten Signalen, von denen dieses Buch hauptsächlich handelt. Auf den Vorgang der Abtastung werden wir noch ausführlich eingehen und beschränken uns daher zunächst auf eine intuitive Darstellung zeitdiskreter Signale. Diskrete Signale unterscheiden sich von den kontinuierlichen Signalen durch ihren quantisierten Signalparameter. Die Beobachtung eines diskreten Zeitsignals liefert also eine abzählbare Folge von Signalwerten — jeweils nach einem bestimmten Zeitschritt einen neuen Signalwert, während ein kontinuierliches Signal in einem bestimmten Zeitintervall aus unbegrenzt vielen Signalwerten besteht.

1.2.2 Amplitudenquantisierte Signale

Amplitudenquantisierte Signale nehmen im Gegensatz zu den analogen Signalen zu jedem Zeitpunkt nur genau einen Zustand aus einer vordefinierten abzählbaren Menge von Zuständen ein. Werden beispielsweise zwei Zustände (Quantisierungsstufen) mit den Bezeichnungen H=„high“ und L=„low“ unterschieden, spricht man von einem *Binärsignal*, das im zeitdiskreten Fall so aussehen könnte: $\ldots \underline{L}LHLLHHHLHHH \ldots$. Dabei werde zum Zeitpunkt $t = 0$ der unterstrichene Zustand $\underline{L}$ eingenommen und in gewissen Zeitschritten folgt eine neue Informationseinheit — nämlich ein neues L oder ein neues H.

Nun stehen jedoch in elektronischen Schaltungen keine binären Signalträger zur Verfügung, sondern lediglich Spannungen, Ströme, Ladungen, Temperaturen, Lichtstärken etc. Da Signale immer an einen Signalträger gebunden

sind, ist ein Verfahren erforderlich, ein amplitudenquantisiertes Signal mit einem analogen Signalträger zu übertragen. Die Möglichkeiten dafür sind vielfältig und sind Gegenstand der Nachrichtentechnik. In Abb. 1.2 sind einige gebräuchliche Varianten skizziert, mit denen ein zeitdiskretes binäres Signal mit einem analogen Signalträger (hier eine Spannung) codiert werden kann.

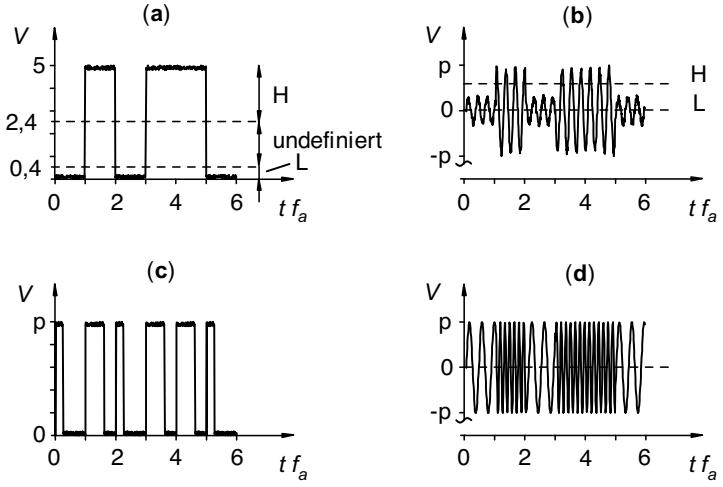


Abbildung 1.2. Varianten der Codierung eines zeitdiskreten Binärsignals „LHLHHL“ mit einen analogen Signalträger

1.2.3 Digitale Signale

Ein *digitales* Signal ist ein speziell codiertes amplitudendiskretes Signal. Ist ein Signal nicht nur auf zwei, sondern allgemein auf n Quantisierungsstufen quantisiert worden, kann wiederum jeder Stufe ein Code, beispielsweise eine Zahl $0...n-1$ zugeordnet werden. Das quantisierte Signal ist somit beschrieben durch eine fortlaufende Folge von je einem Codewort pro Zeitschritt. Von einem digitalen Signalverarbeitungssystem kann ein amplitudendiskretes Signal verarbeitet werden, wenn die Codierung der Quantisierungsstufen digital, d.h. mit einer Kombination binärer Teilsignale erfolgt. Die Amplitudenquantisierung und die verschiedenen Methoden der Digitalcodierung sind Gegenstand ausführlicherer Betrachtungen und wir werden im Kapitel 4 darauf zurückkommen, nachdem wir uns zunächst mit den mathematischen Grundlagen zur Beschreibung von zeitdiskreten Signalen und den zeitdiskreten Systemen im Kapitel 3 auseinandergesetzt haben. In Abb. 1.3 ist die Umwandlung eines analogen Signals x in ein digitales Signal x_d veranschaulicht. Abb. 1.4 zeigt die später beschriebenen Komponenten, die dies technisch realisieren.

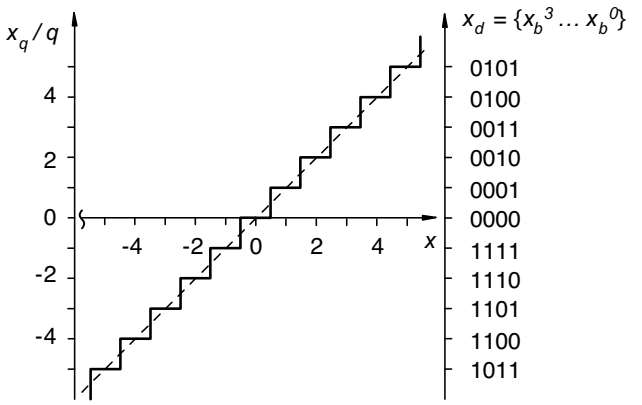


Abbildung 1.3. Digitalumsetzung: Das analoge Signal x geht durch die Quantisierung in ein amplitudendiskretes Signal x_d über. Die einzelnen Quantisierungsstufen (hier elf) werden mittels einer Kombination der vier Binärsignale x_b^0 bis x_b^3 digital codiert. Dies beschreibt die Abbildung von x auf das Digitalsignal x_d.

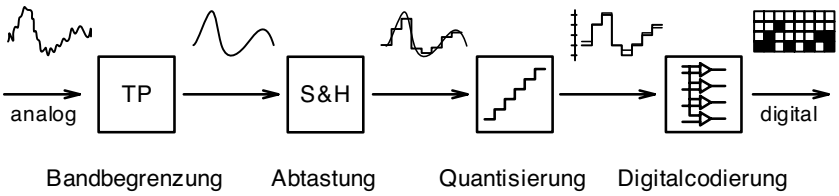


Abbildung 1.4. Umwandlung eines analogen Signals in ein digitales Signal mit Bandbegrenzung durch Tiefpaß (TP) und Abtastung durch Sample & Hold (S&H)

1.3 Diskrete Signale und deren Notation

Diskrete Signale entstehen im Allgemeinen durch Abtastung eines analogen kontinuierlichen Signals. Man erhält eine Sequenz von analogen oder quantisierten Signalwerten x_n, die sich im Intervall $[a, b]$ als Folge notieren lässt:

$$x[n] = \{x_a, x_{a+1}, \ldots, \underline{x_0}, \ldots, x_{b-1}, x_b\} = \{x_n\}_{n=a}^{b} \tag{1.1}$$

Diese Folge charakterisiert eine diskrete Funktion

$$X : \mathbb{Z} \Rightarrow \mathbb{R} \text{ oder } \mathbb{C}, \tag{1.2}$$

die einen ganzzahligen Index n auf die einzelnen Signalwerte x_n abbildet:

$$n \Rightarrow x(n) = x_n \tag{1.3}$$

Die Abtastung kann äquidistant, zufällig oder in beliebiger anderer Weise erfolgen. Wird das zeitkontinuierliche Signal $x^a(t)$ äquidistant abgetastet – wovon im Folgenden ausgegangen werden soll – lässt sich eine Abtastfrequenz f_a definieren:

$$f_a = \frac{1}{T_a}$$
$$x_n = x(n) = x^a(nT_a) \tag{1.4}$$

Wir legen fest, dass mit

$$x[n] \tag{1.5}$$

die *gesamte Folge* x der Variablen n bezeichnet wird, während wir hingegen mit einer der beiden Darstellungen

$$x(n) = x_n \tag{1.6}$$

ein *einzelnes Element* – nämlich das n-te Element – der Folge $x[k]$ bezeichnen. Um eine Folge durch Aneinanderreihen der Folgen-Elemente zu notieren, verwenden wir geschweifte Klammern:

$$x[n] = \{0, \ldots, 0, \underline{1}, 0, \ldots, 0\} \tag{1.7}$$

In dieser Darstellung kennzeichnen wir das Element $x[0]$ durch Unterstreichen.

Beispiel 1.1 – Notation zeitdiskreter Folgen.

Notieren Sie das Signal x_4 aus Abb. 1.1 als Folge und geben Sie die Abtastfrequenz an. Die Zahlen der Abszisse in der Abbildung sollen die Einheit ms besitzen. Es werde vereinbart, dass das zu $n = 0$ gehörige Element x_0 unterstrichen dargestellt wird.

Lösung:

$$x_4[n] = \{\underline{0}, 3, 4, 2, 0, -1, 0, 0, 0, -1, -3, -3, -2, 1, 4, 3, 1, 0, -1, 0, 1\} \tag{1.8}$$

Der zeitliche Bezug, wie er in der Abbildung dargestellt ist, geht verloren. An seine Stelle tritt ein ganzzahlig durchnummerierter Index: $x(0) = 0, x(1) = 3$, usw.

$$f_a = \frac{1}{1ms} = 1kHz \tag{1.9}$$

□

1.4 Spezielle Folgen

Besondere Bedeutung besitzen die folgenden fünf Folgen:

- **Impulsfolge** (Deltafolge, Dirac):

$$\delta[n] = \{\ldots, 0, \underline{1}, 0, \ldots\} = \left\{ \begin{matrix} 1 & n = 0 \\ 0 & n \neq 0 \end{matrix} \right\}_{n=-\infty}^{\infty} \tag{1.10}$$

- **Sprungfolge (Heavisidefolge):**

$$\sigma[n] = \{\ldots, 0, \underline{1}, 1, \ldots\} = \left\{ \begin{matrix} 1 & n \geq 0 \\ 0 & n < 0 \end{matrix} \right\}_{n=-\infty}^{\infty} \tag{1.11}$$

- **Rampenfolge:**

$$\rho[n] = \{\ldots 0, \underline{0}, 1, 2, 3, \ldots\} = \left\{ \begin{matrix} n & n \geq 0 \\ 0 & n < 0 \end{matrix} \right\}_{n=-\infty}^{\infty} \tag{1.12}$$

- **reelle kausale Exponentialfolge:**

$$x_a[n] = \begin{cases} a^n & n \geq 0 \\ 0 & n < 0 \end{cases} \quad |a| < 1 \tag{1.13}$$

- **komplexe Exponentialfolge**:

$$x_e[n] = e^{jn\omega T} = \cos(n\omega T) + j \sin(n\omega T) \tag{1.14}$$

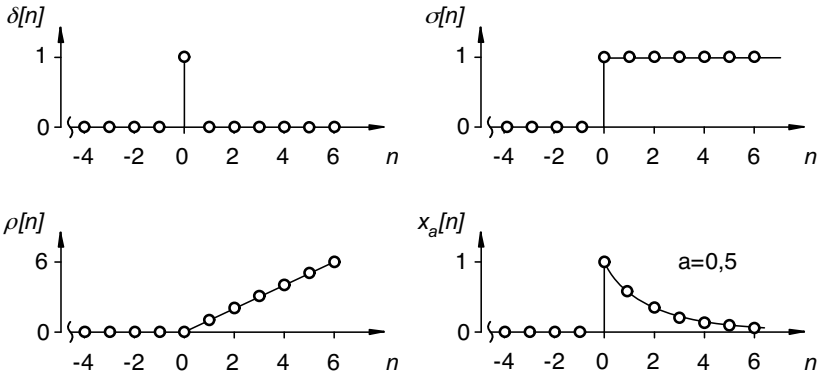


Abbildung 1.5. Darstellung der Impulsfolge, Sprungfolge, Rampenfolge und der rellen, kausalen Exponentialfolge

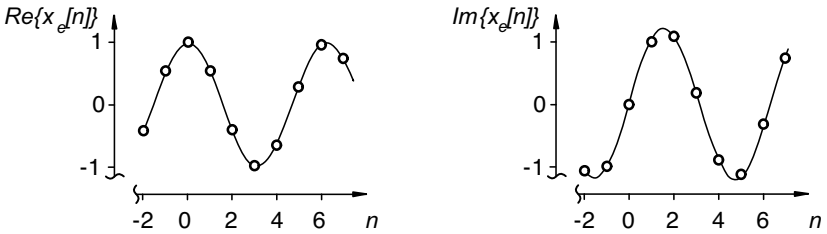


Abbildung 1.6. Realteil und Imaginärteil der komplexen Exponentialfolge mit $\omega T = \frac{2\pi}{10}$

Abb. 1.5 zeigt diese Folgen grafisch. In Abb. 1.6 sind der Real- und der Imaginärteil der komplexen Exponentialfolge dargestellt.

Die komplexe Exponentialfolge wird verwendet, um das Verhalten von Systemen im Frequenzbereich genauer zu untersuchen. Jedes diskrete Signal kann als Summe von gewichteten und zeitverschobenen Deltafolgen, Sprungfolgen oder Rampenfolgen geschrieben werden. Dies ist bei den Deltafolgen sofort einsichtig, da eine Folge $x[n]$ wie folgt geschrieben werden kann:

$$x[n] = \sum_{i=-\infty}^{\infty} x_i \delta[n-i] \tag{1.15}$$

Die Deltafolge verschwindet für alle $i \neq n$. Berechnen wir z.B. das k-te Element der Folge in Gl. (1.15), so ergibt der Wert der Summe $x_k \delta[0] = x_k$. Nach dieser Gleichung lassen sich auch die Rampen- und die Sprungfolge aus Diracimpulsen zusammensetzen. Es ist ersichtlich, dass die Dirac-, Rampen- und Sprungfolge wie folgt ineinander übergehen:

$$\sigma[n] = \sum_{i=0}^{\infty} \delta[n-i] \overset{\delta[k]=\delta[-k]}{=} \sum_{i=0}^{\infty} \delta[i-n] \tag{1.16}$$

$$\rho[n] = \sum_{i=1}^{\infty} \sigma[n-i] = \sum_{i=1}^{\infty} i\delta[i-n] \tag{1.17}$$

In Abb. 1.7 ist eine anschauliche Darstellung vom rechten Teil der Gl. (1.16) dargestellt.

Die Gleichungen (1.16) und (1.17) ergeben sich direkt aus Gl. (1.15). Weiterhin kann man die Diracfolge auch wie folgt aus einer Sprungfolge zusammensetzen:

$$\delta[n] = \sigma[n] - \sigma[n-1] = \sigma[n]\sigma[-n] \tag{1.18}$$

oder einer Rampenfolge:

$$\sigma[n] = \rho[n+1] - \rho[n] \tag{1.19}$$

$$\delta[n] = \sigma[n] - \sigma[n-1] = \rho[n+1] - 2\rho[n] + \rho[n-1] \tag{1.20}$$

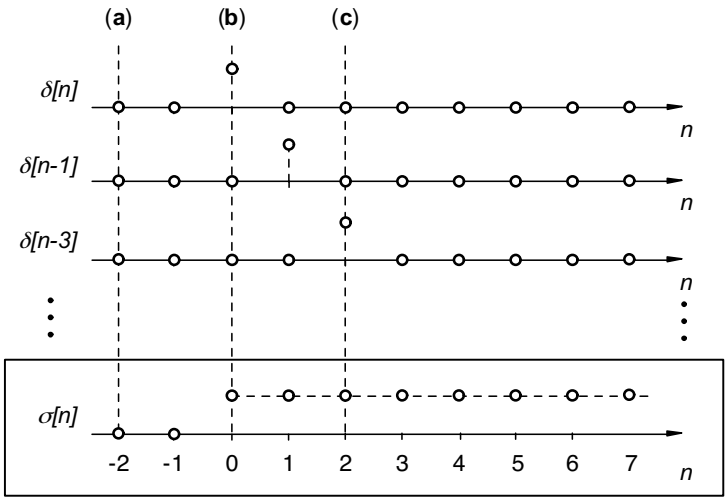


Abbildung 1.7. Darstellung des rechten Teils von Gl. (1.16). Die vertikalen, gestrichelten Linien deuten die Summation an:

Position a): Für alle $n < 0$ ist $\sigma(n) = 0$:
Der Index $(i - n)$ der rechten Seite durchläuft für $i = 0 \ldots \infty$ die Werte $(-n) > 0 \ldots \infty$. Damit sind die aufsummierten Impulsfolgen für alle Werte von i ebenfalls 0.

Positionen b,c): Für alle $n \geq 0$ ist $\sigma[n] = 1$:
Der Index $(i - n)$ der rechten Seite durchläuft für $i = 0 \ldots \infty$ die Werte $(-n) \leq 0 \ldots \infty$, also $\{-n, \ldots, 0, \ldots\}$. Für genau einen Wert von i im Bereich $0 \ldots \infty$, nämlich $i = n$ (für die Position b): $n = 0$, für die Position c): $n > 0$), wird eine der aufsummierten Impulsfolgen an der Stelle 0 betrachtet und liefert damit den Wert 1. Alle anderen Impulsfolgen der rechten Seite liefern den Wert 0. Damit nimmt die Summe der Impulsfolgen auf der rechten Seite für jedes beliebige, aber feste $n \geq 0$ den Wert 1 an.

Man erkennt eine gewisse Verwandtschaft zwischen der Integration kontinuierlicher Signale und der Summation bei diskreten Signalen. Der einfachste Weg zur Überprüfung der Gültigkeit der Gl. (1.19) ist die Berechnung eines beliebigen Elements k. Laut Gl. (1.19) und Gl. (1.12) gilt zum Beispiel:

$$\sigma_k = \begin{cases} 0 & k < -1 \\ 0 - 0 = 0 & k = -1 \\ k + 1 - k = 1 & k >= 0 \end{cases} \tag{1.21}$$

Gl. (1.20) folgt direkt durch Einsetzen aus Gl. (1.19). Das bedeutet, dass jede beliebige Folge als gewichtete Summe von Diracfolgen, Sprungfolgen oder Rampenfolgen geschrieben werden kann. Die Inverse Diskrete Fourier-

Transformation (Kap. 6) stellt die Zerlegung einer beliebigen Folge als Summe von gewichteten, komplexen Exponentialfunktionen dar.

Beispiel 1.2 – Notation eines Signals als Summe von Rampen-Folgen.

Notieren Sie die Folge

$$x[n] = \begin{cases} n-1 & 0 < n < 4 \\ 0 & \text{sonst} \end{cases} \tag{1.22}$$

als gewichtete Summe von Rampen-Folgen!

Lösung:

$$x[n] = \rho[n-1] - 3\rho[n-3] + 2\rho[n-4] \tag{1.23}$$

Anschaulich zeigt dies das folgende Bild.

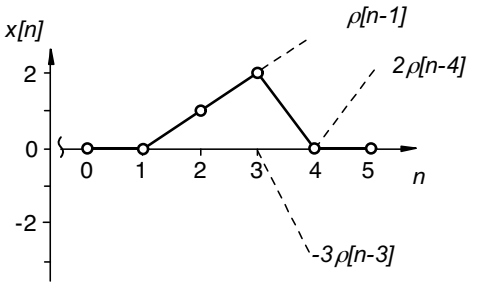


Abbildung 1.8. Zusammensetzung aus Rampenfolgen

□

An dieser Stelle sei noch einmal angemerkt, dass mit „$\delta[n]$“ die komplette Folge bezeichnet wird und nicht etwa nur das n-te Element. Dies gilt entsprechend auch für andere Folgen. Alle Gleichungen, die in diesem Abschnitt hergeleitet wurden, beziehen sich also jeweils auf die *gesamte Folge*.

1.5 Signalmaße zeitdiskreter Signale

Zur Untersuchung von Signalen lassen sich geeignete Parameter definieren, die aus dem Signal unter Verwendung eines eindeutigen Algorithmus errechnet werden. Einige dieser als „Signalmaße“ bezeichneten Parameter sollen bereits

an dieser Stelle definiert werden – weitere folgen an geeigneten Stellen im Buch.
Für zeitdiskrete Signale $x[n]$ seien folgende Signalmaße definiert:

- **Diskrete Summe:**

$$s_D = \sum_{n=-\infty}^{+\infty} x_n \tag{1.24}$$

- **Absolute Summe:**

$$s_a = \sum_{n=-\infty}^{+\infty} |x_n| \tag{1.25}$$

  Wenn für s_a ein endlicher Grenzwert existiert, nennt man $x[n]$ „**absolut summierbar**".
- **Signalenergie:**
  Sei $x(t)$ ein **zeitkontinuierliches** Zeitsignal, so bezeichnet man mit dem Integral

$$E = \int_{-\infty}^{+\infty} |x(t)|^2 dt \tag{1.26}$$

  die **Energie** von $x(t)$. Ist dieses Integral endlich, nennt man $x(t)$ ein **Energiesignal**.
  Analog dazu erfolgt auch die Berechnung der Energie eines **zeitdiskreten** Signals (von Grünigen, 2001). Die Integration wird durch eine Summation ersetzt:

$$E = \sum_{n=-\infty}^{\infty} |x_n|^2 \tag{1.27}$$

- **Mittelwert:**

$$\bar{x} = \lim_{\substack{a\to-\infty \\ b\to\infty}} \frac{1}{|a-b|+1} \sum_{n=a}^{b} x_n \tag{1.28}$$

  Sinnvoll ist diese Berechnung insbesondere für periodische Signale (Periode N). Liegt ein solches Signal vor, vereinfacht sich die Mittelwertbildung zu

$$\bar{x} = \frac{1}{N} \sum_{n=k}^{k+N-1} x_n \qquad , k \in \mathbb{Z} \tag{1.29}$$

- **Signal-Leistung:**
  Mit dem Grenzwert über das Integral

$$\bar{P} = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} |x(t)|^2 dt \tag{1.30}$$

wird die **mittlere Leistung** eines **kontinuierlichen** Zeitsignals berechnet. Ist $x(t)$ kein Energiesignal aber P endlich, zählt $x(t)$ zu den **Leistungssignalen**.
Auch hier reduziert sich die Integration auf eine Summation im **zeitdiskreten** Bereich:

$$\bar{P} = \lim_{\substack{a\to-\infty \\ b\to\infty}} \frac{1}{|a-b|+1} \sum_{n=a}^{b} |x_n|^2 \tag{1.31}$$

Liegt zur Berechnung der Leistung ein periodisches zeitdiskretes Signal vor (Periode N), wird dessen Leistung berechnet mit

$$\bar{P} = \frac{1}{N} \sum_{n=0}^{N-1} |x_n|^2. \tag{1.32}$$

Beispiel 1.3 – Berechnung unendlicher Summen.

Berechnen Sie folgende unendliche Summe der zweiseitigen Exponentialfolge:

$$S = \sum_{n=-\infty}^{\infty} a^{-|n|} \tag{1.33}$$

Lösung:
Aufgrund der Symmetrie des Signals vereinfacht sich S zu:

$$S = 1 + 2\sum_{n=1}^{\infty} a^{-n} \tag{1.34}$$

Es ist somit ein Grenzwert $N \to \infty$ für die Summe $\Theta = \sum_{n=1}^{N} a^{-n}$ zu berechen. Dazu wird durch eine Indexverschiebung $n = m + 1$ substituiert und anschliessend aufgelöst:

$$\Theta = \sum_{m=0}^{N-1} a^{-(m+1)} = \frac{1}{a}\left[1 + \sum_{m=1}^{N} a^{-m} - a^{-N}\right] = \frac{1}{a}\left[1 + \Theta - a^{-N}\right] \tag{1.35}$$

Durch Umstellen erhält man (für $a \neq 1$)

$$\Theta = \frac{1 - a^{-N}}{a - 1}. \tag{1.36}$$

Der Grenzwert $N \to \infty$ für diesen Ausdruck wird verwendet, um die Grenzwertaufgabe zur Berechnung der unendlichen Summe S zu formulieren:

$$S = 1 + 2 \lim_{N \to \infty} \left(\frac{1 - a^{-N}}{a - 1} \right) \tag{1.37}$$

Für $|a| < 1$ divergiert der Zähler gegen $-\infty$ bei negativem Nenner.Für $a = 1$ divergiert die Summe Θ offensichtlich. Somit erhält man folgende Lösung:

$$S = \begin{cases} 1 + \frac{2}{a-1} & |a| > 1 \\ \infty & \{|a| < 1 \text{ oder } a = 1\} \end{cases}. \tag{1.38}$$

Zum Verhalten für S für $\{|a| = 1 \text{ und } a \neq 1\}$ siehe Übung 1.1. □

Beispiel 1.4 – Berechnung der Energie und der mittleren Leistung zeitdiskreter Folgen.

Berechnen Sie die Energie und Leistung der zweiseitigen Exponentialfolge:

$$x[n] = a^{-|n|} \tag{1.39}$$

Lösung:

$$E = \sum_{n=-\infty}^{\infty} \left(a^{-|n|} \right)^2 = \sum_{n=-\infty}^{\infty} (a^2)^{-|n|} \tag{1.40}$$

Diese unendliche Summe wurde bereits in Beispiel 1.3 berechnet, so dass hier eine direkte Lösung angegeben werden kann:

$$E = \begin{cases} 1 + \frac{2}{a^2-1} & |a| > 1 \\ \infty & |a| < 1 \text{ oder } a = 1 \end{cases}. \tag{1.41}$$

Ebenfalls unter Zuhilfenahme der in Bsp. 1.3 verwendeten Methoden kann die Leistung berechnet werden:

$$P = \lim_{N \to \infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \left(a^{-|n|} \right)^2 = \lim_{N \to \infty} \frac{1}{2N+1} \left[1 + 2 \sum_{n=1}^{N} (a^2)^{-n} \right]$$

$$P = \begin{cases} 0 & |a| > 1 \\ 1 & a = 1 \\ \infty & |a| < 1 \end{cases}. \tag{1.42}$$

Der Fall $\{|a| = 1 \text{ und } a \neq 1\}$ für Energie und Leistung wird in Übung 1.2 behandelt. □

1.6 Eigenschaften diskreter Signale

Die Definition von Signal-Eigenschaften dient der theoretischen Verallgemeinerung von Algorithmen, die auf Signale angewendet werden können. An dieser Stelle soll auf die Kausalität und die Periodizität eingegangen werden.

- **Kausalität:** Eine zeitdiskrete Folge $x[n]$ heißt „**kausal**“, wenn sie nur für nichtnegative ganzzahlige Argumente definiert ist:

$$x_{\text{kausal}}[n] : \mathbb{N}_0 \Rightarrow \mathbb{R} \vee \mathbb{C} \tag{1.43}$$

  Ist sie dagegen auch für negative ganzzahlige Argumente definiert, heißt die Folge „**antikausal**“:

$$x_{\text{antikausal}}[n] : \mathbb{Z} \Rightarrow \mathbb{R} \vee \mathbb{C} \tag{1.44}$$

  Im Unterschied dazu nennt man Folgen, die im Bereich der ganzen Zahlen definiert sind „**rechtsseitig**“ bzw. „**linksseitig**“, wenn sämtliche Werte unterhalb bzw. oberhalb einer Grenze k null sind:

$$x_{\text{rechtsseitig}}[n] : \{\mathbb{Z}; \forall x[n < k \in \mathbb{Z}] = 0\} \Rightarrow \mathbb{R} \vee \mathbb{C} \tag{1.45}$$

$$x_{\text{linksseitig}}[n] : \{\mathbb{Z}; \forall x[n > k \in \mathbb{Z}] = 0\} \Rightarrow \mathbb{R} \vee \mathbb{C} \tag{1.46}$$

  Kausalität bezeichnet den zeitlich vorwärts gerichteten Zusammenhang von Ursache und Wirkung. So zeigt es sich an vielen Punkten der Signalverarbeitung, dass ein berechenbares theoretisches System praktisch nicht realisierbar ist, wenn es nichtkausalen Charakter besitzt. Als Beispiel sei ein Anwendungsfall aus der Regelungstechnik genannt: Bekanntermaßen zeigen Regelsysteme im Falle einer unerwarteten Störung eine vorübergehende Regelabweichung, die a priori kompensiert werden könnte, wenn man in der Lage wäre, bereits vor dem Einwirken der Störung die Stellglieder geeignet anzusteuern. Ist die Herkunft der Störung jedoch nicht früh genug abzusehen, gibt es aus Gründen der Kausalitätsbedingung zwar theoretische aber keine praktischen Regelsysteme, die in der Lage wären, eine Regelabweichung zu vermeiden.

- **Periodizität:**
  Eine Folge $x[n]$, für die gilt

$$x[n] = x[n \pm Nk] \quad k = 0, 1, 2, \ldots \tag{1.47}$$

  heißt „**periodisch**“ mit der Periodenlänge N.
  Durch Periodizität vereinfachen sich sowohl die Beschreibung eines Signals als auch arithmetische Operationen mit ihm. Zumindest bleibt der theoretische Informationsgehalt im Gegensatz zu zeitlich nicht begrenzten und

nicht periodischen Signalen beschränkt, wodurch es überhaupt erst möglich wird, endliche Algorithmen für bestimmte Problemstellungen zu finden. Es werden Algorithmen vorgestellt, die explizit für periodische Signale verwendbar sind oder aber zu periodischen Signalen führen. Die weit verbreitete Fouriertransformation beispielsweise bildet jedes Zeitsignal in eine Summe (bzw. ein Integral) aus (periodischen) harmonischen Signalen ab. Ein Algorithmus, der die Fouriertransformation bzw. ihr zeitdiskretes Pendant - die DFT - numerisch ausführen soll, terminiert somit nur für ein periodisches Signal.

Übungen

Übung 1.1 – Unendliche Summe.

Diskutieren Sie folgenden Spezialfall einer unendlichen Summe der zweiseitigen Exponentialfolge:

$$S = \lim_{N\to\infty} \sum_{n=-N}^{N} a^{-|n|} \quad \{|a| = 1 \text{ und } a \neq 1\}$$

In Präzisierung von Gl. (1.33) wird hier genau angegeben, dass der obere und untere Grenzwert gemeinsam zu bilden sind.

Nach Gl. (1.37) lautet das Ergebnis

$$S = 1 + 2 \lim_{N\to\infty} \left(\frac{1 - a^{-N}}{a - 1} \right) \quad .$$

Da $a \neq 1$, existiert der Grenzwert. Da $|a| = 1$, ist er aber nicht eindeutig, denn a^{-N} konvergiert nicht und wird mit steigendem N viele verschiedene Werte annehmen. Untersuchen Sie a^{-N}. Sind die Werte von a^{-N} immer andere, oder unterliegen sie einer bestimmten Gesetzmässigkeit oder Regelmässigkeit? Wenn ja, gibt es einen *mittleren Grenzwert* $\bar{S}$? Welchen Wert hat er?

Einige Lösungshinweise (aber versuchen Sie es erst einmal ohne diese!):

Mit $j = +\sqrt{-1}$ schreiben wir: $a = \exp(j\ q\ 2\pi)$. Damit ist die Bedingung $|a| = 1$ sichergestellt. Ferner sei $q \neq k$, $k \in \mathbb{Z}$, womit die Bedingung $a \neq 1$ sichergestellt ist.

Nun benötigen wir folgende Fallunterscheidung:

Fall 1: q ist rational. Dann lässt sich q als Quotient $q = U/V$ schreiben, mit ganzen, teilerfremden Zahlen U, V und $U \neq k\ V$. Dann ist

$$a^{-N} = \exp(-j\ 2\pi\ N\ U/V) = \exp(-j\ 2\pi\ (N + V)\ U/V) = a^{-(N+V)}$$

Argumentieren Sie, dass sich dadurch für a^{-N} mit steigendem N nur V mögliche Werte ergeben und dass a^{-N} diese Werte zyklisch durchläuft. Ein *mittlerer Grenzwert* kann also durch Mittelung über beliebige V aufeinanderfolgende Werte von a^{-N} ermittelt werden. Möge diese Folge mit n^* beginnen, so ergibt diese Mittelung:

$$\frac{1}{V} \sum_{n=n^*}^{n^*+V-1} \exp(-j\ 2\pi\ n\ U/V)$$

Begründen Sie, warum diese Summe (für beliebiges n^*) Null ergibt. Insgesamt lautet das Ergebnis für den mittleren Wert $\bar{S}$ also

$$\bar{S} = 1 + 2\frac{1}{a-1} = \frac{a+1}{a-1} \quad .$$

Als Beispiel betrachten wir den einfachen Fall $a = -1$. Man erhält $\bar{S} = 0$. Verifizieren Sie dieses Ergebnis durch direktes Bestimmen von

$$S(N) = \sum_{n=-N}^{N} a^{-|n|} \quad \text{für } a = -1$$

für wachsende N. Die Werte, die Sie auf diese Weise erhalten, sollten um $\bar{S} = 0$ als mittleren Wert schwanken.

Führen Sie eine äquivalente Verifikation durch für $a = j$.

Fall 2: q ist nicht rational. Dann ist obiges Argument nicht möglich. Reicht es aus, die irrationale Zahl q beliebig genau durch einen rationalen Quotienten U/V mit (ggf. grossen) Zahlen U, V anzunähern? Gibt es einen mittleren Grenzwert? Kann man dann argumentieren, dass ebenfalls $\bar{S} = \frac{a+1}{a-1}$ das Ergebnis ist? Ist das Ergebnis für $S(N)$ zyklisch?

Können Sie Ihr Resultat anders als über die approximative Rationalitätseigenschaft von q begründen? Wie lässt sich der Mittelwert von a^{-N} über ein beliebiges (grosses) Intervall V für allgemeine q näherungsweise berechnen?

Übung 1.2 – Energie und Leistung.

Berechnen Sie die mittlere Energie und die mittlere Leistung der zweiseitigen Exponentialfolge für den Fall

$$\{|a| = 1 \text{ und } a \neq 1\}$$

Beachten Sie dazu, wenn nötig, die Lösungshinweise aus Übung 1.1.

Übung 1.3 – Zusammensetzen von Signalen.

a) Wie lautet die Funktionsvorschrift einer Rampen-Folge?
b) Stellen Sie die Sprungfolge durch eine Rampenfolge dar.
c) Notieren Sie folgende Funktion als gewichtete Summe von Rampen-Folgen:

$$x(n) = \begin{cases} n-1 & \text{für } 0 < n < 4 \\ 0 & \text{sonst} \end{cases}$$

d) Leiten Sie die resultierende Funktion ab und skizieren Sie das Ergebnis.
e) Stellen Sie die Sprungfolge durch eine Impulsfolge dar, und die Rampenfolge durch eine Sprungfolge.

Übung 1.4 – Signalmaße.

Gegeben sei die Funktion

$$x(n) = \begin{cases} 0 & \text{für } n < 0 \\ (-0.5)^n & \text{sonst} \end{cases} .$$

Berechnen Sie

a) die absolute Summe $S_A = \sum_{n=-\infty}^{\infty} |x_n|$

b) die diskrete Summe $S_D = \sum_{n=-\infty}^{\infty} x_n$

c) und die Signalenergie $E = \sum_{n=-\infty}^{\infty} |x_n|^2$

2

Analoge zeitkontinuierliche Systeme

Ein **System** ist eine Menge von meist vielen Elementen, die in irgendeiner Form zusammenarbeiten um ein ganz bestimmtes Verhalten aufrechtzuerhalten. Eine Definition des Begriffs System im Sinne der Systemtheorie (Fliege, 1991) ist „ein abgegrenzter Teil eines größeren Ganzen“ .

Aus diesen beiden Definitionen können wir ein System als Teil eines Wirkungsgefüges, wie in Abb. 2.1 gezeigt, auffassen, in dem verschiedene Teilsysteme in Wechselwirkung miteinander stehen. Die Wechselwirkung zweier Teilsysteme ist durch eine Verbindung dargestellt. Ändert sich z.B. der Zustand des Teilsystems C in Abb. 2.1 so wird dies den Zustand der Teilsysteme B,D,E und F beeinflussen. Ändert sich der Zustand von System B, so beeinflusst dies wieder den Zustand der Teilsysteme A,F und C. Weiterhin ist anzumerken, dass sich jedes System wiederum in kleinere Teilsysteme aufteilen lässt oder mehrere Teilsysteme zu einem größeren Teilsystem zusammengefasst werden können. Daher ist die Definition eines Systems in jedem Fall willkürlich.

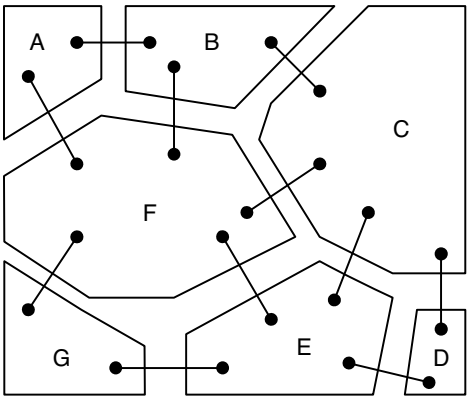


Abbildung 2.1. Ein System als Wirkungsgefüge von Teilsystemen

Aus den obigen Anmerkungen kommen wir zum Schluss, dass die *Eingangsgrößen* und *Ausgangsgrößen* eines Teilsystems davon abhängen, wo wir in ein Wirkungsgefüge eingreifen. Beobachten wir nun den Verlauf einer physikalischen Größe in einem System so können wir diesen als *Signal* auffassen. Wir sprechen von einem *Eingangssignal* eines Systems wenn der *Zustand* des Systems von diesem Signal abhängt. Um festzustellen in welchem Zustand sich ein System befindet müssen wir bestimmte *Zustandsgrößen* im System beobachten bzw. messen. Diese werden auch willkürlich von uns festgelegt. Bei realen Systemen wählt man möglichst solche Zustandsgrößen, die leicht zu beobachten *und* aussagekräftig sind.

Jedes physikalische System, das in beliebiger Form Signale verändert, verarbeitet oder weiterleitet, bezeichnen wir als „System“ oder „*Filter*“ im Sinne der Signalverarbeitung. Gegenstand der Betrachtungen ist erstens, in welcher Weise der Zustand eines Systems von verschiedenen Eingangssignalen (ggf. inklusive Störgrößen) abhängt und zweitens, wie sich diese Zustandsänderung auf die beobachteten Ausgangsgrößen auswirkt.

In diesem Kapitel soll zunächst (im Sinne einer Wiederholung) auf *zeitkontinuierliche* Systeme eingegangen werden, da für diese Klasse von Systemen die wichtigen Begriffe „Impulsantwort“ , „Faltung“ und „Übertragungsfunktion“ größtenteils schon bekannt sein sollten.

2.1 Systembeschreibung

Das Verhalten eines Systems lässt sich auf verschiedene Arten beschreiben. Eine Form ist die *Zustandsbeschreibung*. Besonders bei physikalischen Systemen können die Wechselwirkungen innerhalb eines Systems oder das Verhalten des Systems sehr genau durch die entsprechenden physikalischen Gesetze vorhergesagt werden. Für eine Vorhersage ist jedoch meistens der aktuellen Zustand des Systems nötig, der gemessen werden muss.

Beispiel 2.1 – System fliegende Metallkugel.

Ist das Gewicht einer fliegenden Metallkugel und deren Geschwindigkeit in Richtung und Betrag bekannt und wissen wir, wie die Schwerkraft auf diese Kugel einwirkt, so wird das Verhalten dieser Kugel sehr gut durch die Newtonsche Mechanik beschrieben. Über die Gleichung $\mathbf{F} = \dot{\mathbf{p}}$, die den Zusammenhang zwischen der zeitlichen Änderung des Impulses der Kugel $\dot{\mathbf{p}}$ und der Summe der an der Kugel angreifenden Kräfte $\mathbf{F}$ beschreibt, ist das System „Fliegende Kugel“ ausreichend gut beschrieben. Die Kugel befindet sich in jedem Zeitpunkt in einem bestimmten Zustand, ihre Zustandsgrößen sind z.B. ihre Geschwindigkeit $\mathbf{v}$ und ihre Masse m.
Es soll nochmal betont sein, dass die *Eingangs- und Ausgangsgrößen* des Systems völlig willkürlich gewählt werden können. Man könnte die

Geschwindigkeit der Kugel beeinflussen (Eingangsgröße) und als Ausgangsgröße an der auf die Kugel wirkende Kraft interessiert sein. Man könnte jedoch auch die auf die Kugel wirkenden Kräfte beeinflussen (z.B. Magnetfeld) und als Ausgangsgröße an der Flugbahn der Kugel interessiert sein. In beiden Fällen beschreibt die obige Gleichung das System korrekt. Man wählt nur andere Zustandsgrößen, andere Eingangsgrößen und verschiedene Ausgangsgrößen. □

Ein System wird in einer Zustandsbeschreibung daher allgemein beschrieben durch eine Systemgleichung

$$S\{\mathbf{x}, \mathbf{e}\} = 0 \tag{2.1}$$

mit den Zustandsgrößen $\mathbf{x}$ und den Eingangsgrößen $\mathbf{e}$ und durch eine Beschreibung der interessierenden Ausgangsgrößen $\mathbf{y}$

$$f\{\mathbf{y}, \mathbf{x}, \mathbf{e}\} = 0. \tag{2.2}$$

Die weniger allgemeine Form der Zustandsbeschreibung ist die explizite Form, in der die Systemgleichungen durch Differentialgleichungen beschrieben werden können, (hier sei die unabhängige Variable die Zeit t) die dann in die Form

$$\dot{\mathbf{x}}(t) = S\{\mathbf{x}(t), \mathbf{e}(t)\} \tag{2.3}$$
$$\mathbf{y}(t) = f(\mathbf{x}(t), \mathbf{e}(t)) \tag{2.4}$$

gebracht werden können. Die Gl. (2.3) beschreibt hierbei den Zusammenhang zwischen der Änderung der Zustandsgrößen und dem aktuellen Zustand des Systems und den Eingangsgrößen.

Durch Elimination der Zustandsgrößen $\mathbf{x}$ aus dem Gleichungssystem erhält man den Zusammenhang zwischen Eingangsgrößen und Ausgangsgrößen, vorrausgesetzt man kennt den Anfangszustand des Systems. Geht man davon aus, dass der Anfangszustand des Systems $= 0$ ist, kann nun die Übertragungsfunktion des Systems definiert werden.

2.2 Arten von Systemen

Wie bei den Signalen in Kap. 1 Abschn. 1.1 kann man auch Systeme nach ihren Eigenschaften verschieden klassifizieren. Mögliche Unterscheidungskriterien können hier sein:

- offenes System $\leftrightarrow$ (ab)geschlossenes oder isoliertes System
- dynamisches System $\leftrightarrow$ statisches System
- kontinuierliches System $\leftrightarrow$ diskretes System
- determiniertes System $\leftrightarrow$ stochastisches System
- stabiles System $\leftrightarrow$ instabiles System

- lineares System $\leftrightarrow$ nichtlineares System
- kausales System $\leftrightarrow$ akausales System
- System mit Gedächtnis $\leftrightarrow$ System ohne Gedächtnis
- zeitvariantes System $\leftrightarrow$ zeitinvariantes System

Die für dieses Kapitel relevanten Eigenschaften von Systemen werden im folgenden näher erläutert.

2.3 Linearität

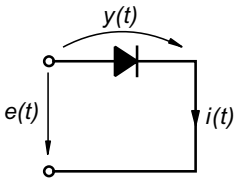
Für jedes lineare System gilt die *Superpositionseigenschaft*: Wird an ein lineares System S eine Linearkombination beliebiger Einzelsignale x_i angelegt, so ergibt sich das Ausgangssignal y aus derselben Linearkombination der Systemantworten y_i auf die Einzelsignale:

$$S(\alpha x_1 + \beta x_2) \overset{lin.}{=} \alpha S(x_1) + \beta S(x_2). \tag{2.5}$$

Wir wollen nun einige Beispiele für elektronische Schaltungen auf Linearität untersuchen. Dies soll helfen, zunächst an einfachen, bekannten (analogen) Schaltungen zu sehen, wie diese auch im Systembegriff untersucht werden können. Vor allem können wir die hier eingeführten Begriffe und Konzepte immer wieder an uns bekannten Resultaten aus der Elektronik überprüfen und uns damit Schaltungstechnisch veranschaulichen.

Wir betrachten zunächst 2 elektronische Beispiele, um Kenntnisse aus der Elektronik direkt in „Sytemschreibweise“ angeben zu können.

Beispiel 2.2 – Linearität einer Diode.



Untersuchen Sie die Linearität der abgebildeten Diodenschaltung mit:
$i(t) = S\{e(t)\}$

Lösung:

Um die Linearität dieses Systems zu überprüfen, stellen wir zunächst die Systemgleichung auf. Sie lautet

$$i(t) = I_S \left(e^{\frac{y(t)}{k}} - 1 \right) \tag{2.6}$$

In dieser Gleichung ist „k “ eine temperaturabhängige Konstante.
Sei $e_1(t) = E$ eine Gleichspannung, so gilt

$$i_1 = I_S(e^{\frac{E}{k}} - 1). \tag{2.7}$$

Für eine sinusförmige Wechselspannung $e_2(t) = U\sin(\omega t)$ ergibt sich

$$i_2 = I_S(e^{\frac{U\sin(\omega t)}{k}} - 1). \tag{2.8}$$

Wählen wir nun die Summe der beiden Spannungen $e_1(t) + e_2(t)$ als Eingangssignal, so erhält man

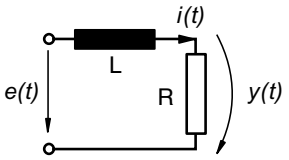
$$i = I_S(e^{\frac{E+U\sin(wt)}{k}} - 1) \neq i_1 + i_2 \tag{2.9}$$

oder in Systemschreibweise unter Verwendung von (2.5)

$$S\{e_1(t) + e_2(t)\} \neq S\{e_1(t)\} + S\{e_2(t)\} \tag{2.10}$$

Die Diode ist also kein lineares System. Selbstverständlich ist ein Gegenbeispiel hinreichend um die Nichtlinearität einer Diode zu zeigen. Um die Linearität eines Systems jedoch allgemein zu beweisen, muss man mit beliebigen Eingangssignalen arbeiten, wie im nächsten Beispiel gezeigt. □

Beispiel 2.3 – Linearität eines LR-Gliedes.



Untersuchen Sie die Linearität des abgebildeten LR-Glieds mit:
$i(t) = S\{e(t)\}$

Lösung:

Wir wählen als Zustandsgröße den Strom i. Die Differentialgleichung (DGL) für diese Schaltung lautet:

$$L\frac{d\,i(t)}{dt} + Ri(t) = e(t) \tag{2.11}$$

mit

$$y(t) = Ri(t). \tag{2.12}$$

Da es sich hierbei um eine *lineare Differentialgleichung* handelt könnte man ja bereits vermuten, dass sie auch ein *lineares System* beschreibt! Eine kurze Rechnung kann dies überprüfen: Gegeben seien 2 Eingangssignale $e_1(t)$ und $e_2(t)$ mit den dazugehörigen Ausgangssignalen $i_1(t)$ und $i_2(t)$. Dann gilt für jedes dieser Signale:

$$L\frac{di_1(t)}{dt} + Ri_1(t) = e_1(t) \tag{2.13}$$

$$L\frac{di_2(t)}{dt} + Ri_2(t) = e_2(t) \tag{2.14}$$

Eine Addition führt auf

$$L\frac{di_1(t)}{dt} + Ri_1(t) + L\frac{di_2(t)}{dt} + Ri_2(t) = e_1(t) + e_2(t). \tag{2.15}$$

Daraus folgt dann:

$$L\frac{di^*(t)}{dt} + Ri^*(t) = e^*(t) \tag{2.16}$$

mit

$$i^*(t) = i_1(t) + i_2(t) \quad \text{und} \quad e^*(t) = e_1(t) + e_2(t) \tag{2.17}$$

Wir haben damit gezeigt: $S\{e1 + e2\} = S\{e1\} + S\{e2\}$
Das System ist linear. Wir können das an Beispielen mit der Lösung der DGL überprüfen.
Als homogene Lösung dieser DGL erhält man:

$$i_h(t) = ce^{-\frac{t}{\tau}} \tag{2.18}$$

mit einer noch unbestimmten Konstanten c und der Zeitkonstante $\tau = L/R$. Die Gesamtlösung der DGL ergibt sich als Summe der homogenen und der partikulären Lösung, wobei die letztere sowie die Konstante c sich nur bei Kenntnis der Eingangsspannung angeben lassen (Auch wenn die homogene Lösung eigentlich nicht von $e(t)$ abhängt, so tun das in diesem Fall die Anfangsbedingungen, da wir hier ein *sprungfähiges* System haben).

a. Wählen wir als Eingangsspannung wieder wie bei der Diode eine Gleichspannung der Größe E, die aber erst zum Zeitpunkt $t = 0$ angelegt werden möge, so gilt mit der Sprungfunktion $u(t)$
$e_1(t) = Eu(t)$
Die partikuläre Lösung lautet damit für $t < 0$:

$$i_{1,p}(t) = 0 \tag{2.19}$$

und für $t \geq 0$:

$$i_{1,p}(t) = \frac{E}{R} \tag{2.20}$$

Der Strom an der Spule kann nicht springen, denn er ist proportional zum Integral über die Spannung an der Spule. Daher ist die homogene Lösung für $t < 0$:

$$i_{1,h}(t) = 0 \tag{2.21}$$

Wir haben die homgene und partikuläre Lösung for $t < 0$ und für $t \geq 0$ bestimmt. Nach Zusammenfassen unter Verwendung der

Sprungfunktion $u(t)$ erhalten wir nach Bestimmung der Konstanten c:

$$i_{1,h}(t) = \frac{-E}{R} e^{-\frac{t}{\tau}} u(t) \tag{2.22}$$

$$i_{1,p}(t) = \frac{E}{R} u(t). \tag{2.23}$$

Damit erhält man folgende Ausgangsspannung

$$y_1(t) = R i_1(t) = E(1 - e^{-\frac{t}{\tau}}) u(t). \tag{2.24}$$

b. Wählen wir eine sinusförmige Wechselspannung in komplexer Darstellung

$$e_2(t) = \hat{U} e^{j\omega t}, \tag{2.25}$$

so gilt mit dem Ansatz

$$i_{2,p}(t) = A e^{j\omega t} \tag{2.26}$$

nach Einsetzen in die DGL:

$$AL j\omega + AR = \hat{U} \tag{2.27}$$

Damit ergeben sich für den Maschenstrom und die Ausgangsspannung die folgenden Formeln:

$$i_{2,p} = \frac{\hat{U}}{R(1 + j\omega\tau)} e^{j\omega t} \tag{2.28}$$

$$y_{2,p} = \frac{\hat{U}}{1 + j\omega\tau} e^{j\omega t}. \tag{2.29}$$

Da in diesem Fall die Eingangsspannung seit unendlich langer Zeit anliegt, ist der Wert der Konstante $c = 0$.

c. Wir überzeugen uns nun, dass bei der Anregung mit der Summe der eben betrachteten Einzelsignale

$$e_3(t) = e_1(t) + e_2(t) = E u(t) + \hat{U} e^{j\omega t} \tag{2.30}$$

am Ausgang die Spannung

$$y_3(t) = E(1 - e^{-\frac{t}{\tau}}) u(t) + \frac{\hat{U}}{1 + j\omega\tau} e^{j\omega t} \quad , \quad i_3(t) = \frac{u_3(t)}{R} \tag{2.31}$$

anliegt, dies können wir durch einfaches Einsetzen von i_3 und e_3 in die DGL überprüfen. Da wir ja aber wissen, dass es jeweils für die

einzelnen Summanden von y_3 gilt, brauchen wir diese Rechnung nun nicht noch einmal länglich ausführen!
Bemerkung: Da eine Zeitverschiebung des Eingangssignals ebenfalls eine Zeitverschiebung des Ausgangssignals ergibt ist ein LR-Glied auch zeitinvariant und damit ein linear time-invariant (LTI) System!

□

2.4 Die Impulsfunktion (Delta-Funktion) als Distribution

Bekanntermaßen bildet die Spannung an einer Induktivität die zeitliche Ableitung des durch sie fließenden Stromes. Eine Unstetigkeit des Stromes, die beispielsweise durch einen Einschaltvorgang ausgelöst werden kann, führt zu der Problematik, dass eine Differentation im mathematischen Sinne zu nicht endlichen Werten führt. Aus diesem Grunde soll eine verallgemeinerte Größe eingeführt werden, die nur im Distributionensinn existiert - die oftmals auch als „*Dirac-Impuls*“ oder „**Impulsfunktion**“ bezeichnete **Delta-Funktion**.

Wir definieren die Delta-Funktion $\delta(t)$ (vgl. Kap. 1: zeitdiskrete Delta-Folge) im Distributionensinne als eine Funktion, die folgenden Eigenschaften genügt:

1.
$$\delta(t) = \begin{cases} 0 & t \neq 0 \\ \infty & t = 0 \end{cases} \tag{2.32}$$

2.
$$\int_a^b \delta(t)dt = 1 \qquad \text{mit } 0 \in [a,b] \tag{2.33}$$

Man könnte also auch sagen, die Delta–Funktion ist eine Funktion mit der Fläche 1 , welche bei $x = 0$ konzentriert ist. Aus diesen beiden Eigenschaften folgt sofort eine dritte, die als **Ausblend-** oder **Abtasteigenschaft** bezeichnet wird:

3.
$$\begin{aligned} \int_{-\infty}^{\infty} f(t)\delta(t-t_0)dt &= \int_{-\infty}^{\infty} f(t)\delta(t_0-t)dt \\ &= \int_{-\infty}^{\infty} f(t+t_0)\delta(t)dt \\ &= \lim_{c\to 0}\int_{-c}^{+c} f(t+t_0)\delta(t)dt = f(t_0) \end{aligned} \tag{2.34}$$

Um die Ausblendeigenschaft einzusehen, kann man sich vorstellen, dass das Produkt der beiden Funktionen $f(t+t_0)\delta(t)$ überall ausser an der Stelle 0 verschwindet und der Funktionswert von $f(0+t_0)$ wie eine Konstante vor

das Integral gezogen werden (auch wenn das streng mathematisch natürlich nicht ganz korrekt ist).

Die Impulsfunktion kann durch Folgen angenähert werden (sog. **Fundamentalfolgen**). Z.B. kann die Impulsfunktion als Grenzwert (für $n \to \infty$) einer Folge von Rechtecken mit der Fläche 1 dargestellt werden, deren Breite $1/n$ und Höhe n ist. Für wachsende n konvergiert diese Funktion gegen die Impulsfunktion. Wir werden dies im folgenden noch benutzen.

Diese Delta-Funktion stellt die Idealisierung eines praktisch nicht realisierbaren Impulses unbegrenzter Höhe, unendlich kurzer Dauer und begrenzter Energie dar. Auch wenn somit für eine praktische Untersuchung realer Systeme Delta-Funktionen nicht verwendbar sind, ist eine derartige Distribution für theoretische Überlegungen nützlich, wie in folgendem Beispiel gezeigt werden soll.

Beispiel 2.4 – Beschreibung der Leistung eines LR-Glieds mittels Delta-Funktion.

Wir betrachten die Leistung und die Energie des homogenen Stroms durch den Widerstand R eines LR-Gliedes. Wir werden sehen, dass die Leistung sich bei entsprechender Konfiguration von L und R einer Delta-Funktion annähert und dass die Energie dabei konstant bleibt, wie dies auch bei der Delta-Funktion der Fall ist. Die Leistung ist damit eine physikalische Größe, die sich wie die Delta-Funktion verhält! Dies gilt streng natürlich nur solange, wie die hier angenommenen Gleichungen der Bauelemente Gültigkeit haben.
Die homogene Lösung $i_{1,h}(t)$ war für einen Einschaltvorgang an einem LR-Glied im Beispiel 2.3 berechnet worden:

$$i_{1,h}(t) = \frac{-E}{R} e^{-\frac{t}{\tau}} u(t) \quad ; \quad \tau = L/R. \tag{2.35}$$

Die Leistung P am Widerstand ist dann

$$P = R i_{1,h}^2(t) = \frac{E^2}{R} e^{-\frac{2t}{\tau}} u(t) \tag{2.36}$$

und die Energie am Widerstand berechnet sich zu

$$H = \int_{-\infty}^{\infty} P dt = \frac{E^2}{R} \int_0^{\infty} e^{\frac{-2 \cdot t}{\tau}} dt = E^2 \frac{\tau}{2R} \tag{2.37}$$

Nun soll die Leistung als eine Delta-Distribution aufgefasst werden. Dazu halten wir die Energie H und den Zeitfaktor τ konstant fest, und wählen L und R als Funktionen von H und τ:

$$R(H, \tau) = E^2 \frac{\tau}{2H} \tag{2.38}$$

sowie

$$L(H,\tau) = \tau R = E^2 \frac{\tau^2}{2H} \tag{2.39}$$

Dann ist

$$P = \frac{2 \cdot H}{\tau} e^{-\frac{2t}{\tau}} \tag{2.40}$$

mit

$$H = \text{const.} \tag{2.41}$$

Das heißt, wir können uns bei konstanter Energie H einen Zeitfaktor τ vorgeben, und mit der angegeben Wahl von R und L resultiert daraus das berechnete Zeitverhalten der Leistung. In Abb. 2.2(a) ist der Verlauf der Leistung für drei verschiedene Werte von τ gezeigt.

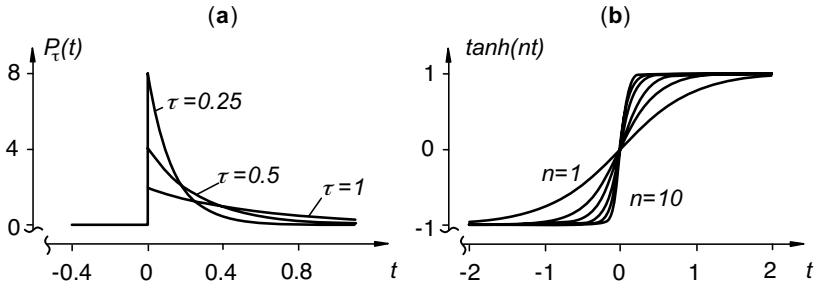


Abbildung 2.2. Distributionen: (**a**) Approximation der Delta-Funktion gemäß Beispiel 2.4, (**b**) Approximation der Signumfunktion durch $\tanh(nt)$

Betrachten wir den relativen Anteil an der Gesamtenergie über dem Widerstand R im Zeitintervall $[0, n\,\tau]$ so erhalten wir:

$$\frac{\int_0^{n\,\tau} P dt}{H} = 1 - e^{-2n} \tag{2.42}$$

Für $n = 3$ bedeutet dies, dass sich 98% der Energie des Ausgangssignals im Zeitintervall $[0, 3\tau]$ konzentriert. Machen wir τ immer kleiner, so nähert sich die Leistung am Widerstand immer weiter der Delta-Funktion, wie sie oben definiert wurde, an. Wir haben uns also überzeugt, dass es physikalische Größen gibt, die gegen das Verhalten von Delta-Funktionen konvergieren.
Dieses Verhalten hat seine physikalische Grenze, wenn die Linearitätseigenschaften der Bauelemente nicht mehr gewährleistet sind. Da nach Gl. (2.38) $R \propto \tau$ und nach Gl. (2.39) $L \propto \tau^2$ sind, tritt dieses Problem in der Tat für kleine τ auf: die Einschaltströme durch die

Bauelemente werden dann so gross, dass ihre Linearitätseigenschaften verletzt werden bzw. die Bauelemente zerstört werden. □

2.5 Die Impulsfunktion (Delta-Funktion) als Ableitung der Sprungfunktion

Die Sprungfunktion $u(t)$ kann mit Hilfe der Vorzeichenfunktion $\mathrm{sgn}(t)$ folgendermaßen dargestellt werden:

$$u(t) = \frac{1}{2}(1 + \mathrm{sgn}(t)) \tag{2.43}$$

mit

$$\mathrm{sgn}(t) = \frac{t}{|t|} \quad (\text{für } t \neq 0) \text{ und } \mathrm{sgn}(0) = 0. \tag{2.44}$$

Die Vorzeichenfunktion $\mathrm{sgn}(t)$ wiederum kann man als Grenzwert einer Folge von $\tanh(nt)$-Funktion für $n \to \infty$ darstellen:

$$\mathrm{sgn}(t) = \lim_{n\to\infty} \tanh(n \cdot t), \tag{2.45}$$

s. Abb. 2.2(b) zur Verdeutlichung. Da diese Folge gleichmäßig gegen $\mathrm{sgn}(t)$ konvergiert, können wir nun schreiben

$$\begin{aligned} \frac{du(t)}{dt} &= \frac{d}{dt}\left[\frac{1}{2}(1 + \mathrm{sgn}(t))\right] = \frac{1}{2}\frac{d}{dt}\left[\lim_{n\to\infty} \tanh(nt)\right] = \frac{1}{2}\lim_{n\to\infty}\frac{d}{dt}\left[\tanh(nt)\right] \\ &= \frac{1}{2}\lim_{n\to\infty}\frac{n}{\cosh^2(nt)} \\ &= \delta(t). \end{aligned} \tag{2.46}$$

Das bedeutet, dass wir im Sinne der Distributionstheorie als Ableitung der Sprungfunktion die Delta-Funktion annehmen können (obwohl die Sprungfunktion an sich an der Stelle 0 nicht differenzierbar ist!)

2.6 Die Impulsantwort

Die Impulsantwort eines Systems ist die Antwort dieses Systems auf eine Anregung mit der Impulsfunktion:

$$h(t) = S\{\delta(t)\}. \tag{2.47}$$

Hier ist zu beachten, dass die o.a. Darstellung der Impulsfunktion als Funktionenfolge bei Systemen mit unbeschränkt wachsendem Ausgang nicht verwendet werden kann. Wir betrachten nämlich formal

$$h(t) = S\{\delta(t)\} = S\left\{\lim_{n\to\infty} f_n(t)\right\} = \lim_{n\to\infty} S\{f_n(t)\}\,. \tag{2.48}$$

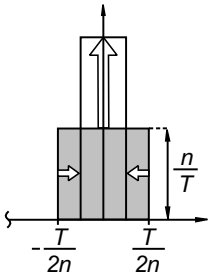
Die Vertauschung des Grenzwertes mit der Systemfunktion ist nur für beschränktes S erlaubt (notwendige Bedingung). Es gibt noch weitere, hinreichende Bedingungen für diese Vertauschung. Bei beschränkten LTI-Systemen ist die Vertauschung erlaubt.

Beispiel 2.5 – Impulsantwort eines LR-Gliedes.

Berechnen Sie die Impulsantwort eines LR-Gliedes erster Ordnung!

Lösung:

Da das direkte Bestimmen der Impulsantwort schwierig ist, bestimmen wir sie über eine Funktionenfolge (Fundamentalfolge im Sinne der Funktionentheorie). Dazu setzen wir mit der Sprungfunktion $u(t)$ eine Folge (in n) von Rechteckfunktionen mit Fläche 1 an:



$$\delta(t) = \lim_{n\to\infty} \frac{n}{T}\left[u\left(t+\frac{T}{2n}\right) - u\left(t-\frac{T}{2}n\right)\right] \tag{2.49}$$

Wir betrachten die Sprungantwort y eines LR-Systems erster Ordnung. Das System ist beschrieben durch

$$L\frac{di}{dt} + Ri = u(t). \tag{2.50}$$

Es gilt

$$y = (1 - e^{-\frac{t}{\tau}})u(t). \tag{2.51}$$

Mit $\tau = L/R$ gilt für den Systemausgang

$$\begin{aligned} h_n(t) &= \frac{n}{T}\left[1 - e^{-\frac{t+T/2n}{\tau}}\right]u(t) - \frac{n}{T}\left[1 - e^{-\frac{t-T/2n}{\tau}}\right]u(t) \\ &= \frac{n}{T}u(t)e^{-t}\left[e^{T/2n} - e^{-T/2n}\right]. \end{aligned} \tag{2.52}$$

und mit $\sinh(x) = \frac{1}{2}(e^x - e^{-x})$ und Erweitern mit τ:

$$h_n(t) = \frac{1}{\tau} e^{-\frac{t}{\tau}} u(t) \frac{2n\tau}{T} \sinh\left(\frac{T}{2n\tau}\right) \tag{2.53}$$

Wegen $\sinh(x) = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!} \underset{|x|\to 0}{\to} x$ gilt

$$\frac{2n\tau}{T} \sinh\left(\frac{T}{2n\tau}\right) \underset{n\to\infty}{\to} 1, \tag{2.54}$$

und folglich lautet die Impulsantwort

$$h(t) = \frac{1}{\tau} e^{-\frac{t}{\tau}} u(t). \tag{2.55}$$

□

2.7 Impulsantwort als Ableitung der Sprungantwort bei linearen Systemen

Wir haben bereits festgestellt, dass die Impulsfunktion (Delta-Funktion) als Ableitung der Sprungfunktion interpretiert werden kann. Die Systemantwort eines *linearen* Systems auf die Impulsfunktion ist identisch mit der zeitlichen Ableitung der Systemantwort auf einen Einheitssprung. Im folgenden zeigen wir diesen Sachverhalt allgemein.

Die Sprungfunktion sei $u(t)$, die Systemantwort darauf werde als Sprungantwort $a(t)$ bezeichnet:

$$a(t) = S\{u(t)\} \tag{2.56}$$

Da wir lineare Systeme betrachten, und da die Differentiation linear ist, gilt:

$$\frac{da(t)}{dt} = \frac{dS\{u(t)\}}{dt} = S\left\{\frac{du(t)}{dt}\right\} = S\{\delta(t)\} = h(t) \tag{2.57}$$

Diese Beziehung ist nützlich, da sich die Sprungantwort in analogen Systemen sehr viel einfacher bestimmen lässt als die Impulsantwort, denn wie bereits erwähnt wurde, ist die Delta-Funktion technisch nur als Näherung realisierbar.

Beispiel 2.6 – Impulsantwort für ein LR-Glied durch Ableiten der Sprungantwort.

Berechnen Sie die Impulsantwort eines LR-Gliedes erster Ordnung! Leiten Sie dazu die Sprungantwort dieses Systems ab!

Lösung:
Wir betrachten (wie im vorigen Beispiel) wiederum die DGL des LR-Gliedes erster Ordnung bei Anregung mit der Sprungfunktion:

$$L\frac{di}{dt} + Ri = u(t). \tag{2.58}$$

Diesmal wählen wir nicht den Übergang mit einer Funktionenfolge, sondern erhalten direkt mit $\tau = L/R$ für $t > 0$ für die Sprungantwort

$$a(t) = \left(1 - e^{-\frac{t}{\tau}}\right) u(t). \tag{2.59}$$

Daraus erhalten wir unter Mißachtung, daß $a(t)$ bei $t = 0$ nicht differenzierbar ist für die Impulsantwort

$$h(t) = \frac{da}{dt} = \frac{1}{\tau} e^{-\frac{t}{\tau}} u(t) + \underbrace{\delta(t)\left(1 - e^{-\frac{t}{\tau}}\right)}_{\text{undefiniert für t = 0}} \equiv \frac{1}{\tau} e^{-\frac{t}{\tau}} u(t). \tag{2.60}$$

Die komplizierte Herleitung über eine Funktionenfolge haben wir also durch eine einfache Ableitung der leicht zu bestimmenden Sprungantwort ersetzen können. $a(0)$ müsste exakterweise dabei jedoch separat bestimmt werden, es ergäbe sich jedoch $a(0) = 0$. □

2.8 Analoge Faltung

Wir haben die Antworten eines Systems für verschiedene Anregungen bestimmt. Es stellt sich die Frage, ob für jede neue Anregung auch eine neue Systemantwort durch Aufstellen und Lösen der Differentialgleichungen (partikuläre Lösung) berechnet werden muss, um die Antwort des Systems zu finden. Die Antwort auf diese Frage lautet „nein". Es ist ausreichend, die Impuls- oder die Sprungantwort des Systems zu kennen.

Theorem 2.1 (Analoger Faltungssatz).
Die Antwort eines LTI-Systems auf jede Anregung $x(t)$ ergibt sich als Faltung mit der Impulsantwort $h(t)$:

$$S\{x(t)\} = x(t) * h(t) = \int_{-\infty}^{\infty} x(T)h(t-T)dT \tag{2.61}$$

Wir leiten diesen Zusammenhang jetzt her. Es ist $h(t) = S\{\delta(t)\}$ und wegen der Zeitinvarianz des Systems auch, für festes T:

$$h(t-T) = S\{\delta(t-T)\}. \tag{2.62}$$

Sei nun $x(T)$ der (zunächst feste) Wert einer anregenden Funktion zum Zeitpunkt T. Dann gilt wegen der Linearität des Systems [Multiplikation mit einem konstanten Faktor $x(T)$]:

$$x(T)h(t-T) = S\{x(T)\delta(t-T)\}. \tag{2.63}$$

Da auch die Integration eine lineare Operation ist, gilt ferner

$$\int_{-\infty}^{\infty} x(T)h(t-T)dT = S\left\{\int_{-\infty}^{\infty} x(T)\delta(t-T)dT\right\}. \tag{2.64}$$

Das Integral auf der rechten Seite ist nach (2.34) gleich $x(t)$, so dass wir schließlich schreiben können:

$$\int_{-\infty}^{\infty} x(T)h(t-T)dT = S\{x(t)\}. \tag{2.65}$$

Dieses Integral wird als **Faltungsintegral** bezeichnet. Damit ist Theorem 2.1 bewiesen. □

Es ist manchmal hilfreich, T durch $(t-T)$ zu substituieren. Das Ergebnis lautet:

$$\int_{-\infty}^{\infty} x(t-T)h(T)dT = S\{x(t)\}. \tag{2.66}$$

Man sagt für Gln. (2.65) und (2.66): „$x(t)$ wird mit $h(t)$ gefaltet". Die Faltung ist ebenfalls eine lineare Operation. Die Faltung ist wegen o.a. Substitution kommutativ, es gilt also:

$$S\{x(t)\} = h(t) * x(t) = x(t) * h(t). \tag{2.67}$$

Bisher haben wir noch nicht die **Kausalität** des Systems betrachtet. Für kausale Systeme können nur solche Werte des Eingangssignals den Ausgang beeinflussen, die bis zum Zeitpunkt t auf den Systemeingang gelangen. Es gilt also

$$h(t) = 0 \quad \forall t < 0 \tag{2.68}$$

und damit für die beiden Darstellungen der Faltung:

$$\int_{-\infty}^{t} x(T)h(t-T)dT = S\{x(t)\}, \tag{2.69}$$

$$\int_{0}^{\infty} x(t-T)h(T)dT = S\{x(t)\}. \tag{2.70}$$

Beispiel 2.7 – Berechnung der Impulsantwort für ein LR-Glied mittels Faltung.

Berechnen Sie die Impulsantwort eines LR-Gliedes erster Ordnung! Verwenden Sie dazu die Faltung.

Lösung:

In Beispiel 2.5 hatten wir als Antwort des LR-Gliedes auf einen Spannungssprung $x(t) = u(t)$ direkt hergeleitet (wir lassen in den folgenden

Rechnungen das $u(t)$ weg, da wir das LR Glied als ein kausales System annehmen. Für ein kausales System ist die Sprungantwort für $t < 0$ immer $a(t < 0) = 0$.
Die Sprungantwort des Systems schreiben wir daher als:

$$a(t) = 1 - e^{-\frac{t}{\tau}} \tag{2.71}$$

Wir hatten gesehen, dass für ein lineares System die Stammfunktion der Impulsantwort die Sprungantwort darstellt. Das gleiche Ergebnis wird nun durch die Faltung des Einheitssprunges mit der Impulsantwort hergeleitet. Formal erhalten wir für die Sprunganregung eine Übereinstimmung von Faltungsintegral und Stammfunktion. (Das ist im allgemeinen nicht dasselbe!) Hierbei wird der Unterschied zwischen kausalen und akausalen Systemen betont.
Die Impulsantwort $h(t)$ lautet nach Beispiel 2.6:

$$h(t) = \frac{1}{\tau} e^{-\frac{t}{\tau}} \tag{2.72}$$

Beginnen wir für den Fall nichtkausaler Systeme. Wir benutzen das Faltungsintegral, schließen jedoch das $u(t)$ aus der Impulsfunktion $h(t)$ aus. Wir erhalten

$$\begin{aligned} S\{x(t)\} &= \int_{-\infty}^{\infty} x(t-T)h(T)dT \\ &= \int_{-\infty}^{\infty} u(t-T)\frac{1}{\tau}e^{-\frac{T}{\tau}}dT \\ &= -e^{-\frac{T}{\tau}}|_{-\infty}^{t} \to \infty \end{aligned} \tag{2.73}$$

Bevor wir dieses Ergebnis kommentieren, leiten wir die Systemantwort für kausale Systeme her. Jetzt benutzen wir die kausale Variante:

$$\begin{aligned} S\{x(t)\} &= \int_{0}^{\infty} x(t-T)h(T)dT \\ &= \int_{-\infty}^{\infty} u(t-T)\frac{1}{\tau}e^{-\frac{T}{\tau}}u(T)dT \\ &= \int_{0}^{\infty} u(t-T)\frac{1}{\tau}e^{-\frac{T}{\tau}}dT \\ &= -e^{-\frac{T}{\tau}}|_{0}^{t} = 1 - e^{-\frac{T}{\tau}} \end{aligned} \tag{2.74}$$

Wir müssen also exakt rechnen. Für kausale Systeme können wir in der Rechnung das $u(t)$ weglassen, müssen uns jedoch bei der Faltung daran erinnern, dass wir ein kausales System meinen. Für kausale Systeme müssen wir Gl. (2.66) entsprechend anpassen, wenn wir das

$u(t)$ weglassen. Die falsche Lösung ergibt sich in 2.73, weil wir auch über die gesamte „Vergangenheit“ $T = -\infty..0$ integrieren, wodurch negative Zeiten in der negativen Exponentialfunktion berücksichtigt wurden. Diese bringen die Exponentialfunktion zur Divergenz. □

Beispiel 2.8 – Impulsantwort eines LR-Gliedes mit Exponentialanregung.

Berechnen Sie die Systemantwort eines LR-Gliedes erster Ordnung! Verwenden Sie dazu die Impulsantwort, die Faltung und eine Anregung mit der Exponentialfunktion $x(t) = e^{-\frac{t}{g}}u(t)$.

Lösung:

Die kausale Impulsantwort $h(t)$ lautet nach Beispiel 2.6:

$$h(t) = \frac{1}{\tau}e^{-\frac{t}{\tau}} \tag{2.75}$$

Im Falle eines kausalen Systems kann das Faltungsintegral entsprechend vereinfacht werden. Es gilt:

$$\begin{aligned} S\{x(t)\} &= \int_0^\infty x(t-T)h(T)dT \\ &= \int_0^\infty e^{-\frac{t-T}{g}}u(t-T)\frac{1}{\tau}e^{-\frac{T}{\tau}}dT \\ &= \frac{1}{\tau}e^{-\frac{t}{g}}\int_0^t e^{T[1/g-1/\tau]}dT \\ &= \frac{1}{\tau}\frac{1}{(1/g-1/\tau)}e^{-\frac{t}{g}}\left[e^{T(1/g-1/\tau)}\right]_0^t \\ &= \frac{1}{\tau/g-1}\left[e^{-t/\tau}-e^{-t/g}\right] \end{aligned} \tag{2.76}$$

Für $g \to \infty$ geht das wieder über in die Sprungantwort $a(t)$. Wir konnten also ohne erneutes Untersuchen des Systems die Systemantwort für eine neue Erregung berechnen. □

Übungen

Übung 2.1 – Systemreaktion aus Impulsantwort.

Sei $h(t) = u(t)te^{-t}$ die Impulsantwort eines Systems ($u(t)$ ist die Sprungfunktion). Berechen Sie

a) die Sprungantwort des Systems,
b) mit dem Faltungsintegral $\int_{-\infty}^{\infty} x(T)h(t-T)dT$ die Systemreaktion auf das Eingangssignal

$$x(t) = \begin{cases} e^{t} & t < 0 \\ e^{-t} & t > 0 \end{cases}.$$

Übung 2.2 – Systemreaktion bei RC-Schaltung.

Die Impulsantwort einer RC-Schaltung ist (mit Sprungfunktion $u(t)$)

$$h(t) = u(t)\frac{1}{RC}e^{-t/(RC)}.$$

Berechnen Sie mit dem Faltungsintegral $\int_{-\infty}^{\infty} x(T)h(t-T)dT$ die Systemreaktion auf das Signal

$$x(t) = u(t)\frac{\alpha}{T}t, \quad \text{mit } \alpha > 0.$$

3

Zeitdiskrete LTI-Systeme

Es zeigt sich, dass es in den meisten technischen Belangen ausreicht, lineare zeitinvariante Systeme (LTI-Systeme, linear time invariant) einzusetzen. Dabei spart man sich die komplexe Analytik nichtlinearer Systeme und kann einen überschaubaren Satz an mathematischen Methoden der Signalverarbeitung und der Regelungs- und Steuerungstechnik nutzen, der speziell für LTI-Systeme entwickelt wurde. Aus diesem Grunde werden im Folgenden fast ausschließlich LTI-Systeme behandelt. An gegebener Stelle wird jedoch darauf hingewiesen, was zu beachten ist, sollte ein System nicht linear oder nicht zeitinvariant sein.

3.1 Mathematische Grundlagen zeitdiskreter Systeme

Die Beschreibung von zeitdiskreten Systemen erfordert einige Grundbegriffe und mathematische Methoden, von denen die wichtigsten in diesem Unterkapitel zusammengefasst werden sollen.

3.1.1 Funktionenräume und Normen

Die Signalverarbeitung ist an zwei Begriffe gebunden: erstens an die Signale und zweitens die signalverarbeitenden Systeme. Eine abstrakte aber axiomatisch aufgebaute mathematische Definition dieser beiden Begriffe könnte folgendermaßen aussehen:

Ein Signal ist ein Vektor $\mathbf{x}$, der einem linearen Vektorraum X entstammt: $\mathbf{x} \in X$. Eine Signalklasse stelle einen verallgemeinerten Prototypen für sämtliche zunächst nicht genauer spezifizierten Signale dar. In der umgekehrten Beziehung ist ein Signal eine Instanz einer Signalklasse. Demzufolge entstammt auch diese Signalklasse einem Vektorraum, der als Signalraum bezeichnet werde.

Weiterhin werde eine Klasse von Abbildungen definiert: $S : \mathbf{y} \Leftarrow S(\mathbf{x})$ Ein signalverarbeitendes System ist aus theoretischer Sicht vollständig durch

seine Abbildungseigenschaft charakterisiert: Es bildet eine Instanz x der Signalklasse (das Eingangssignal) auf eine Instanz y einer (eventuell anderen) Signalklasse (nämlich das Ausgangssignal) ab. Somit beschreibt diese Klasse S vollständig die signalverarbeitenden Systeme.

Als Norm $||\bullet||$ auf einen Vektor wird eine Operation auf diesen Vektor bezeichnet, die folgenden Axiomen genügt:

- $\|\mathbf{x}\| \geq 0$; $\|\mathbf{x}\| = 0$ nur dann, wenn $x = 0$
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (Dreiecksungleichung)
- $\|\alpha\mathbf{x}\| = |\alpha|\,\|\mathbf{x}\|$ (Linearität der Norm)

Wichtig im Bereich der Signalverarbeitung sind die L_p-Normen. Sie sind folgendermaßen definiert:

$$\|x\|_p := \left[\int_a^b |x(t)|^p \mathrm{d}t\right]^{\frac{1}{p}} \tag{3.1}$$

Einen Spezialfall der L_p-Normen stellt die Euklidische Norm (L_2-Norm) dar:

$$\|x\|_2 = \sqrt{\int_a^b |x(t)|^2 \mathrm{d}t} \tag{3.2}$$

Beispiel 3.1 – Funktionenräume.

Ist die Funktion $f(x) = \frac{1}{\sqrt{x}}$ Element der folgenden Funktionenräume?

a) $L_1(0,1)$

$$\int_0^1 |f(x)|^1 \mathrm{d}x = \int_0^1 \frac{1}{\sqrt{x}} \mathrm{d}x = 2[\sqrt{x}]_0^1 = 2$$
$$\Rightarrow \frac{1}{\sqrt{x}} \in L_1(0,1)$$

b) $L_2(0,1)$

$$\int_0^1 |f(x)|^2 \mathrm{d}x = \int_0^1 \frac{1}{x} \mathrm{d}x = [\ln(x)]_0^1 = \infty$$
$$\Rightarrow \frac{1}{\sqrt{x}} \notin L_2(0,1)$$

□

3.1.2 Diskrete Faltung

In ähnlicher Weise, wie im Kap. 2.8 auf Seite 32 die analoge Faltung erläutert wurde, kann auch für zeitdiskrete Signale und Systeme eine Faltung definiert werden. Auch an dieser Stelle soll diese Operation im Zeitbereich veranschaulicht werden.

Betrachtet man ein relaxiertes lineares zeitinvariantes System, so gilt auch im zeitdiskreten Fall, dass die Antwort eines Systems auf einen skalierten Einheitsimpuls $\delta[n]$ (aus Gründen der Linearität des Systems) die um den gleichen Faktor skalierte Impulsantwort $h[n]$ ist:

$$x[n] = x_0\delta[n] \quad \Rightarrow \quad y[n] = x_0 h[n] \tag{3.3}$$

Für einen um k Zeitschritte verschobenen Einheitsimpuls gilt folglich aufgrund der Zeitinvarianz

$$x[n] = x_1\delta[n-k] \quad \Rightarrow \quad y[n] = x_1 h[n-k] \tag{3.4}$$

Jedes beliebige Eingangssignal lässt sich aus einer Summe von gewichteten und zeitverschobenen Einheitsimpulsen zusammensetzen:

$$x[n] = \sum_{k=-\infty}^{\infty} x_k\delta[n-k] \tag{3.5}$$

Mit Superposition folgt aus obigen Überlegungen die Antwort des relaxierten zeitdiskreten linearen Systems:

Theorem 3.1 (Diskreter Faltungssatz).
Die Antwort eines diskreten LTI-Systems auf jede Anregung $x[n]$ ergibt sich als Faltung mit der Impulsantwort $h[n]$:

$$y[n] = \sum_{k=-\infty}^{\infty} x_k h[n-k] = x[n] * h[n] \tag{3.6}$$

Diese Summe wird in Analogie zum Faltungsintegral (2.65) als „Faltungssumme" bezeichnet, die die diskreten Folgen $x[n]$ und $h[n]$ miteinander „faltet".

Auch für die diskrete Faltung gilt:

$$x[n] * h[n] = h[n] * x[n] \tag{3.7}$$
$$(\alpha x[n]) * h[n] = \alpha(x[n] * h[n]) \tag{3.8}$$
$$(x_1[n] + x_2[n]) * h[n] = x_1[n] * h[n] + x_2[n] * h[n] \tag{3.9}$$
$$x[n] * \delta[n] = x[n] \tag{3.10}$$

Beispiel 3.2 – Diskrete Faltung.

Ein System habe die Impulsantwort $h[n] = \{\underline{2}, -1, 3\}$. Berechnen Sie die Systemantwort auf das Signal $x[n] = \{\underline{1}, 3, 2, 1\}$.

Lösung:
Die Systemantwort ergibt sich aus der diskreten Faltung von $h[n]$ und $x[n]$: $y[n] = h[n] * x[n]$. Betrachtet man die Faltungssumme, so könnte man die auszuführenden Rechenschritte beispielsweise in folgender übersichtlicher Form darstellen:

$h[n]$	$x[n]$	1	3	2	1				
2		2	6	4	2				
-1	+		-1	-3	-2	-1			
3	+			3	9	6	3		
	0	2	5	4	9	5	3	0	

$y[n] = \{\ldots, 0, \underline{2}, 5, 4, 9, 5, 3, 0, \ldots\}$ □

3.1.3 Periodische Faltung

Eine Faltungssumme gemäß (3.1) konvergiert im allgemeinen nicht für periodische Signale. Man reduziert deswegen die Summengrenzen und definiert bei Periodenlänge N eine periodische Faltung der beiden periodischen Signale $x_p[n]$ und $h_p[n]$ als:

$$y_p[n] = x_p[n] \tilde{*} h_p[n] = \frac{1}{N} \sum_{k=0}^{N-1} x_p[k] h_p[n-k] \tag{3.11}$$

Die periodische Faltung ist kommutativ. Das Resultat $y_p[n]$ ist selbst auch wieder periodisch mit N.

Zu den selben durchzuführenden Rechenschritten und zum selben Ergebnis führt eine andere Notationsform der Faltung zweier periodischer Signale $x_p[n]$ und $h_p[n]$. Soll eine Faltung vorgenommen werden, kann dies auch eine formale Vektormultiplikation mit einer sog. Zirkulantenmatrix vorgenommen werden. Dazu wandelt man eines der beiden periodischen Signale (Periodenlänge N) formal in einen Vektor um:

$$x_p[n] = \{\ldots, x_0, x_1, \ldots, x_{N-1}, x_0, x_1, \ldots\} \rightarrow \mathbf{x} = \begin{pmatrix} x_0 \\ \vdots \\ x_{N-1} \end{pmatrix} \tag{3.12}$$

Das andere periodische Signal verwenden wir zur Bildung der Zirkulantenmatrix:

$$h_p[n] \rightarrow \mathbf{C} = \begin{pmatrix} h_0 & h_{N-1} & h_{N-2} & \cdots & h_1 \\ h_1 & h_0 & h_N - 1 & \cdots & h_2 \\ \vdots & \vdots & \vdots & & \vdots \\ h_{N-1} & h_{N-2} & h_{N-3} & \cdots & h_0 \end{pmatrix} \tag{3.13}$$

Das Ergebnis der Faltung $y = x_p[n] \tilde{*} h_p[n]$ erhalten wir aus der Multiplikation

$$\mathbf{y} = \mathbf{C} \cdot \mathbf{h} \tag{3.14}$$

und einer anschließenden Normierung mit N:

$$\frac{1}{N}\mathbf{y} = \frac{1}{N}\begin{pmatrix} y_0 \\ \vdots \\ y_{N-1} \end{pmatrix} \quad \rightarrow \quad \left\{\ldots, \underline{\frac{y_0}{N}}, \ldots, \frac{y_{N-1}}{N}, \ldots\right\} = y_p[n] \tag{3.15}$$

Beispiel 3.3 – Diskrete Faltung periodischer Signale mit der Zirkulantenmatrix.

Falten Sie die beiden 3-periodischen Signale $h[n] = \{\underline{1}, 0, 2\}$ und $x[n] = \{\underline{1}, 2, 3\}$!

Lösung:

$$x[n] = \{\ldots 3, \underline{1}, 2, 3, 1, \ldots\} \rightarrow \mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

$$h[n] \rightarrow \mathbf{C} = \begin{pmatrix} 1\ 2\ 0 \\ 0\ 1\ 2 \\ 2\ 0\ 1 \end{pmatrix}$$

$$\mathbf{y} = \begin{pmatrix} 1\ 2\ 0 \\ 0\ 1\ 2 \\ 2\ 0\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 5 \\ 8 \\ 5 \end{pmatrix}$$

Wir normieren mit 3 und erhalten:

$$y_p[n] = \left\{\ldots, \underline{\frac{5}{3}}, \frac{8}{3}, \frac{5}{3}, \ldots\right\}$$

□

3.1.4 Z-Transformation

In vielen Fällen vereinfacht sich die Beschreibung und die Berechnung von Signal- und Systemverhalten, wenn man eine geeignete Transformation in den Frequenzbereich vornimmt. Beispielsweise wird die im vorigen Abschnitt beschriebene Faltung im Frequenzbereich auf eine Multiplikation abgebildet. Während die Laplace- bzw. Fouriertransformation im Bereich der zeitkontinuierlichen Signale und Systeme ein geeignete Abbildung darstellt, ist die Z-Transformation das entsprechende Pendant für die Klasse der zeitdiskreten Signale und Systeme.

Tatsächlich ließe sich auch die Fourier-Reihenentwicklung oder – sollte diese nicht konvergent sein – die Laplace-Transformation für zeitdiskrete Signale verwenden. Es zeigt sich jedoch, dass dabei periodische Mehrdeutigkeiten auftreten, die im Umgang mit zeitdiskreten Signalen unpraktikabel sind. Stattdessen verwendet man eine nichtlineare Transformation, die die s-Ebene der Laplace-Transformation in die z-Ebene transformiert. Dadurch gelangt die linke s-Halbebene in das Innere des Einheitskreises in der z-Ebene und

die periodischen Mehrdeutigkeiten verschwinden, weil sie durch die nunmehr kreisförmige Struktur mehrfach auf sich selbst abgebildet werden. Es gäbe mehrere Möglichkeiten, eine solche Transformation vorzunehmen, jedoch hat sich die Z-Transformation durchgesetzt.

In den folgenden Abschnitten sollen die Grundlagen erläutert werden, die notwendig sind, um ein zeitdiskretes Signal sowie die Übertragungsfunktion eines Systems von der Zeitbereichs-Darstellung in eine Darstellung im z-Bereich hin und wieder zurück zu transformieren.

Ist $x[n]$ ein zeitdiskretes Signal, so bezeichnet man mit

$$X(z) = \sum_{n=-\infty}^{\infty} x_n z^{-n} \tag{3.16}$$

die Z-Transformierte von $x[n]$ und mit

$$X_e(z) = \sum_{n=0}^{\infty} x_n z^{-n} \tag{3.17}$$

die einseitig Z-Transformierte von $x[n]$. Die einseitige Z-Transformierte berücksichtigt somit nur den kausalen Teil des Signals $x[n]$. Es werde vereinbart, dass die Z-Transformierte eines Signals mit dem entsprechenden Großbuchstaben bezeichnet wird.

Das transformierte Signal verwendet anstelle der Zeit n eine andere Basis im Signalraum - die Z-Transformierte ist keine Funktion mehr von n, sondern nunmehr eine Funktion der komplexen Zahl z. Trotzdem handelt es sich noch um das selbe Signal. Ein Signal in Zeitdarstellung und die entsprechende Z-Transformierte lassen sich ohne Informationsverlust ein-eindeutig ineinander umwandeln.

Die Z-Transformation ist unter anderem durch folgende Eigenschaften gekennzeichnet:

Linearität

$$Z\{ax[n] + by[n]\} = aX(z) + bY(z) \tag{3.18}$$

Verschiebung

$$Z\{x[n-i]\} = z^{-i}X(z) \tag{3.19}$$

Faltung

$$Z\{x[n] * y[n]\} = X(z)Y(z) \tag{3.20}$$

Modulation

$$Z\left\{e^{anT}x[n]\right\} = X(e^{-aT}z) \tag{3.21}$$

Besondere Beachtung verlangt die Konvergenz der Summe (3.16). Transformiert man eine beliebige Folge $x[n]$ in den z-Bereich, wird man feststellen, dass $X(z)$ im allgemeinen Fall nur für bestimmte z konvergent ist. Fasst

man z als komplexe Zahl auf, kann der Konvergenzbereich (**ROC** - region of convergence) anschaulich in einem zweidimensionalen Diagramm, der sog. z-Ebene, dargestellt werden. Nur mit Angabe des Konvergenzbereiches ist die Z-Transformation eindeutig, siehe Beispiel 3.5 auf der nächsten Seite. Welche Bedeutung die Konvergenz bei der Verwendung der Z-Transformation für die Digitale Signalverarbeitung hat, wird weiter unten erläutert.

Zur Frage der Konvergenz stellen wir zunächst fest, dass Geichung (3.16) die Laurent-Reihe (Bronstein u. a., 1999) von $F(z)$ ist. Diese konvergiert, wenn der reguläre Teil $F_{\text{reg}}(z)$ und der Hauptteil $F_{\text{haupt}}(z)$ konvergieren.

$$F_{\text{reg}}(z) = \sum_{n=0}^{\infty} f(-n)z^n \tag{3.22}$$

$$F_{\text{haupt}}(z) = \sum_{n=1}^{\infty} f(n)z^{-n} \tag{3.23}$$

Stellt man das Konvergenzgebiet der Laurent-Reihe in der z-Ebene dar, so wird es durch maximal zwei Kreise mit den Radien r_{reg} bzw. r_{haupt} um den Ursprung der z-Ebene begrenzt; das Konvergenzgebiet liegt innerhalb des Kreises mit dem Radius

$$r_{\text{reg}} = \frac{1}{\lim_{n\to\infty} \sup |f(-n)|^{1/n}} \tag{3.24}$$

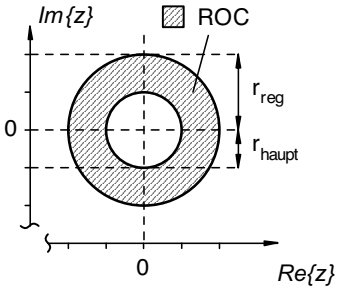


Abbildung 3.1. Konvergenzgebiet der Laurent-Reihe

und außerhalb des Kreises mit dem Radius

$$r_{\text{haupt}} = \lim_{n\to\infty} \sup |f(n)|^{1/n} \tag{3.25}$$

Es gilt also:

$$r_{\text{haupt}} < |z| < r_{\text{reg}} \tag{3.26}$$

Beispiel 3.4 – Konvergenz der Z-Transformation.

Bestimmen Sie das Konvergenzgebiet der reellen kausalen Exponentialfolge $f(n) = a^n$ für $n > 0$, $f(n) = 0$ sonst

Lösung:
$F(z)$ besteht wegen $f(n) = 0$ für $n < 0$ nur aus dem Hauptteil.
Es ist $r_{\text{haupt}} = a$, also insgesamt ROC: $|z| > a$. Den Wert der z-Transformierten bestimmen wir weiter unten. □

Auch wenn sich Z-Transformation und deren Konvergenzbereich für einfache Signale leicht berechnen lassen, bietet es sich an, Korrespondenztabellen zu verwenden. Hierbei lässt sich die Tatsache nutzen, dass sich beliebige Signale in eine Summe von Folgen zerlegen lassen, für die die Z-Transformation bereits bekannt ist. Grundlage dafür ist wiederum die Linearitätseigenschaft der Z-Transformation.

Die Z-Transformation ist eine *lineare* Transformation. Somit wird eine Linearkombination zweier Signale auf die Linearkombination ihrer Z-Transformierten abgebildet:

$$x[n] = x_1[n] + x_2[n] \Leftrightarrow X(z) = X_1(z) + X_2(z) \tag{3.27}$$

Somit lässt sich ein komplexes Signal in eine Summe von einfacheren Signalen (beispielsweise in eine Summe gewichteter Dirac-Folgen) zerlegen. Die Z-Transformierte des komplexen Signals ergibt sich dann aus der Summe der Z-Transformierten der einzelnen Signale.

Beispiel 3.5 – Beispiele zur Z-Transformation.

a) Reelle kausale Exponentialfolge: $x(n) = \sigma(n)a^n$

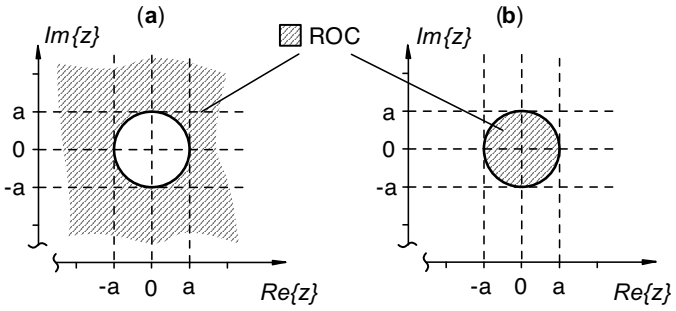
$$X(z) = \sum_{n=0}^{\infty} a^n z^{-n} = \sum_{n=0}^{\infty} (az^{-1})^n = \frac{z}{z-a} \quad ; \quad a < |z| \leq \infty$$

b) Reelle nichtkausale Exponentialfolge: $x(n) = \begin{cases} -a^n, & n \leq -1 \\ 0, & n > -1 \end{cases}$

$$X(z) = \sum_{n=-\infty}^{-1} -a^n z^{-n} = -\sum_{n=1}^{\infty} a^{-n} z^n = 1 - \sum_{n=0}^{\infty} (a^{-1}z)^n \tag{3.28}$$

$$= 1 - \frac{a}{a-z} = \frac{z}{z-a} \quad ; \quad 0 \leq |z| < a \tag{3.29}$$

Die Z-Transformierten sind im Fall a) und b) identisch, besitzen jedoch ein unterschiedliches Konvergenzgebiet (ROC), wie die folgende Abbildung zeigt. Nur mit Angabe des Konvergenzgebiets ist die Z-Transformation eindeutig.



□

3.1.5 Inverse Z-Transformation

Um die Z-Transformation für eine Berechnung von Problemen im Zeitbereich nutzen zu können, ist eine arithmetische Methode der Rücktransformation notwendig. Dafür gibt es mehrere Möglichkeiten:

1. Direkte Integration
2. Anwendung des Residuensatzes
3. Zerlegung in Teile, deren Inverse bekannt ist (Korrespondenztabelle)
4. Bei algebraischen Funktionen: Partialbruchzerlegung
5. Komplexe Transformation $z = 1/w$ und Durchführung einer der o.a. Methoden im w-Raum.

Diese Verfahren sollen im Folgenden erläutert werden, wobei Beispiele zeigen sollen, welches jeweils die zweckmäßige Methode ist.

Inverse Z-Transformation mittels Integration

Auf alle Fälle anwendbar ist folgende Vorgehensweise: Sei $X(z)$ ein Signal im z-Bereich, so ist $x[n]$ durch eine Integration über die geschlossene Kurve Γ um den Nullpunkt in der komplexen Ebene zu berechnen mit Hilfe des folgenden

Theorem 3.2 (Inversionssatz der Z-Transformation).

$$x[n] = \frac{1}{2\pi j} \oint_{\Gamma} X(z) z^{n-1} \mathrm{d}z \tag{3.30}$$

Dabei entstehen je nach Wahl des Konvergenzbereiches unterschiedliche Rücktransformationen.

Die Inversionsformel wollen wir jetzt beweisen. Wir gehen von der Definition der Z-Transformierten aus und multiplizieren beide Seiten mit z^{m-1}:

$$z^{m-1} F(z) = \sum_{n=-\infty}^{\infty} f(n) z^{m-n-1} \tag{3.31}$$

Nun integrieren wir entlang einer geschlossenen Kontur innerhalb des Konvergenzgebietes ROC. Da die ROC Ringe sind, umfassen diese geschlossenen Konturen den Nullpunkt.

$$\oint z^{m-1}F(z)\mathrm{d}z = \oint \sum_{n=-\infty}^{\infty} f(n)z^{m-n-1}\mathrm{d}z \tag{3.32}$$

Da die Reihe gleichmässig konvergiert, vertauschen wir

$$\oint z^{m-1}F(z)\mathrm{d}z = \sum_{n=-\infty}^{\infty} f(n) \oint z^{m-n-1}\mathrm{d}z \tag{3.33}$$

Wir wählen als Integrationskontur einen Kreis

$$z = re^{i\phi};\ \mathrm{d}z = jre^{j\phi}\mathrm{d}\phi = zj\mathrm{d}\phi \tag{3.34}$$

und erhalten damit

$$\begin{aligned}\oint z^{m-1}F(z)\mathrm{d}z &= \sum_{n=-\infty}^{\infty} f(n)jr^{m-n}\int_0^{2\pi} e^{j(m-n)\phi}\mathrm{d}\phi \\ &= \sum_{n=-\infty}^{\infty} f(n)jr^{m-n}2\pi\delta(m-n) \\ &= 2\pi j f(m)\end{aligned} \tag{3.35}$$

Daraus folgt die gesuchte Umkehrformel nach Ersetzung von m mit n, wobei die Kontur den Nullpunkt umfasst und innerhalb einer (Teil-)ROC liegt. □

Das Integral der inversen Z-Transformation ist im allgemeinen wie im letzten Abschnitt angegeben zu berechnen: Man wähle einen Integrationsweg (Kreis), parametrisiere das Integral nach Φ und führe es aus. Dabei muss besondere Sorgfalt verwendet werden auf den Radius des gewählten Kreises, aus ihm ergeben sich im allgemeinen unterschiedliche Konvergenzbereiche.

Der Residuensatz – eine anschauliche Herleitung

Wir stellen hier zunächst den Residuensatz vor, der aus der Funktionentheorie bekannt ist. Er lautet für eine beliebige Funktion $\Phi(z)$:

Theorem 3.3 (Residuensatz).

$$\oint \phi(z)\mathrm{d}z = (2\pi i) * \textit{ Summe der Residuen von } \phi(z) \textit{ in den von der Kontur umschlossenen singulären Stellen} \tag{3.36}$$

Wir wollen an dieser Stelle eine „anschauliche“ Begründung für den Residuensatz und die Werte der Residuen geben (insoweit das möglich ist) und dabei einen Satz für die Werte der Residuen herleiten. Für eine genaue Herleitung wird auf einschlägige Literatur der Funktionentheorie verwiesen, z.B. (Remmert, 2001).

Der Wert der Residuen hängt ab von der Ordnung r des Pols an der singulären Stelle z_0. Bildlich gesprochen lässt das Residuum sich bestimmen, indem man die Funktion $\Phi(z)$ mit $(z-z_0)^r$ multipliziert, um die Singularität „zum Verschwinden zu bringen“.

Beispiel 3.6 – Illustration des Residuums.

(Die Analogie ist nicht wörtlich zu nehmen) Das Feld einer Punktladung Q am Orte z_0 ist im z-Raum proportional zu $\frac{Q}{(z-z_0)^2}$. Das Feld hat also einen Pol der Ordnung 2 bei z_0. Das Residuum erhalten wir durch Multiplikation mit $(z-z_0)^2$, und es beträgt gerade Q, also den Wert der Ladung. Dies macht anschaulich die „Polstärke“ (das Residuum) an diesem Punkt klar. □

Etwas formaler betrachten wir eine Funktion $G(z)$ wie folgt:
$\Phi(z)$ habe einen Pol der Ordnung r bei z_0, und $G(z)$ sei die „ausmultiplizierte“ Funktion

$$G(z) = \Phi(z)(z-z_0)^r \tag{3.37}$$

Damit hat $G(z)$ nun keinen Pol mehr, und wir können $G(z)$ um die Stelle $z = z_0$ in eine normale Taylor-Reihe entwickeln:

$$G(z) = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{d^n G}{dz^n}\bigg|_{(z=z_0)} (z-z_0)^n \tag{3.38}$$

Dann ist für $z \neq z_0$

$$\phi(z) = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{d^n G}{dz^n}\bigg|_{(z=z_0)} (z-z_0)^{(n-r)} \tag{3.39}$$

Nun wollen wir das Ringintegral über $\Phi(z)$ lösen, wobei die Kontur die Stelle z_0 enthalten möge. Wir wählen als Kontur einen Kreis um z_0 und schreiben

$$\phi(z)\mathrm{d}z = \phi(z)\mathrm{d}(z-z_0) = \phi(z)|z-z_0| j e^{(j\phi)}\mathrm{d}\phi \tag{3.40}$$

Das Einsetzen der Reihenentwicklung für $\Phi(z)$ und Berücksichtigung des Kreises liefert

$$\oint \phi(z)\mathrm{d}z = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{d^n G}{dz^n}\bigg|_{(z=z_0)} |z-z_0|^{(n-r+1)} j \int_0^{2\pi} e^{(j(n-r+1)\phi)}\mathrm{d}\phi \tag{3.41}$$

Die Integration liefert nur einen Term (2π) für $n = r-1$, also

$$\oint \phi(z)\mathrm{d}z = (2\pi j)\frac{1}{(r-1)!}\frac{d^{(r-1)}G(z)}{\mathrm{d}z^{(r-1)}}\Big|_{(z=z_0)} \tag{3.42}$$

Dies gilt für *einen* Pol der Ordnung r. Bestehen weitere Pole, so gilt entsprechend eine Summe. Wir können das so verstehen, dass Konturen um die einzelnen Singularitäten gebildet und durch „schmale Kanäle" miteinander verbunden werden. Die Konturintegration entlang des „Hin- und Rückwegs" an beiden Seiten dieser schmalen Kanäle hebt sich gerade auf.

Wir vergleichen jetzt mit dem oben schon angegebenen Residuensatz (3.36) und stellen die formale Analogie fest. Zusätzlich erhalten wir

Theorem 3.4 (Werte der Residuen).
Die Residuen von $\phi(z)$ an der singulären Stelle z_0 mit Ordnung r sind gegeben durch

$$\frac{1}{(r-1)!}\frac{d^{(r-1)}G(z)}{dz^{(r-1)}}|_{(z=z_0)} \tag{3.43}$$

wobei

$$G(z) = \phi(z)(z-z_0)^r \tag{3.44}$$

Beachten Sie nochmals, dass es sich hier nicht um eine strikte mathematische Herleitung handelt, sondern um eine anschauliche Erläuterung der Formel für die Werte der Residuen. Die aus anschaulicher Herleitung resultierenden Sätze sind indes mathematisch exakt.

Inverse Z-Transformation mittels Residuensatz

Nun soll der Residuensatz für die inverse Z-Transformation eingesetzt werden. Wir setzen dazu den Residuensatz (3.36) in die Formel (3.30) für die inverse Z-Transformation ein und erhalten

Theorem 3.5 (Inverse Z-Transformation mit Residuensatz).

$$\begin{aligned} f(n) &= \frac{1}{(2\pi j)}\oint z^{n-1}F(z)\mathrm{d}z \\ &= \textit{Summe der Residuen von } \phi(z) = z^{n-1}F(z) \textit{ in den} \\ &\quad \textit{von der Kontur im ROC umschlossenen} \\ &\quad \textit{singulären Stellen } z_0 \textit{ mit der Ordung } r[z_0] \\ &= \sum_{z_0}\frac{1}{(r-1)!}\frac{d^{(r-1)}G(z,z_0)}{dz^{(r-1)}}|_{(z=z_0)} \\ &\quad \textit{mit } G(z,z_0) = (z-z_0)^{r[z_0]}z^{n-1}F(z) \end{aligned} \tag{3.45}$$

Beispiel 3.7 – Inverse Z-Transformation über den Residuensatz.

Transformieren Sie $X(z) = 1$ in den Zeitbereich.

Lösung:
(über die Inversionsformel mit Residuensatz):

$$x[n] = \text{Residuum von } \left\{\phi(z) = z^{(n-1)}X(z)\right\}$$

Die Funktion $\phi(z)$ hat für $n > 0$ keine Pole, so dass unmittelbar gilt

$$x[n] = 0 \quad \text{für } n > 0$$

Die Funktion $\phi(z)$ hat für $n \leq 0$ einen Pol der Ordnung $r = 1 - n$ bei $z_0 = 0$. Nach Theorem 3.5 ist dann $G(z, z_0) = 1$, und es gilt für die Integration geschlossener Kurven um diese Polstelle

$$x[n] = \frac{1}{(r-1)!}\frac{d^{(r-1)}G(z,z_0)}{dz^{(r-1)}}|_{(z=z_0)} = \frac{1}{(-n)!}\frac{d^{(-n)}\,1}{dz^{(-n)}}|_{(z=0)}$$

Für $n = 0$ ist nichts abzuleiten, und wir erhalten $x[0] = 1$. Für $n < 0$ liefern alle Ableitungen den Wert 0. Zusammengefasst erhalten wir

$$x[n] = \begin{cases} 0 & n > 0 \\ 1 & n = 0 \\ 0 & n < 0 \end{cases}$$

Also

$$x[n] = \{\ldots, 0, \underline{1}, 0, \ldots\} = \delta[n]$$

□

Beispiel 3.8 – Vergleich zwischen direkter Integration und Residuensatz.

Bilden Sie die inverse Z-Transformierte für die in obigem Beispiel schon bestimmte Funktion $F(z) = \frac{z}{(z-a)}$.

Lösung:
a) Direkte Integration liefert:

$$f(n) = \frac{1}{(2\pi j)}\oint z^{n-1}F(z)\mathrm{d}z = \frac{1}{(2\pi j)}\oint \frac{z^n}{z-a}\mathrm{d}z$$

Wähle als Kontur einen Kreis mit Radius r um 0, also

$$z = re^{j\phi}$$
$$\mathrm{d}z = jre^{j\phi}\mathrm{d}\phi = jz\mathrm{d}\phi$$

Dann ist

$$f(n) = -\frac{1}{2\pi a}\int_0^{2\pi} \frac{r^{n+1}}{(1-\frac{r}{ae^{j\phi}})} e^{j(n+1)\phi} \mathrm{d}\phi$$

a1) Die Kontur umfasst $z = a$ nicht. Dann ist $|z| < |a|$ und wir können für $|a| > r = |z|$ schreiben (geometrische Reihe)

$$\frac{1}{(1-\frac{r}{ae^{j\phi}})} = \sum_{i=0}^{\infty} \left(\frac{r}{ae^{j\phi}}\right)^i$$

und damit

$$f(n) = -\frac{1}{2\pi a}\sum_{i=0}^{\infty} r^{n+1} \left(\frac{r}{a}\right)^i \int_0^{2\pi} e^{j(n+1+i)\phi} \mathrm{d}\phi$$

Die Integrale ergeben für $i, n \geq 0$ sämtlich den Wert 0, und somit für die kausale Folge $f(n) = 0(n > 0)$. Für $n < 0$ gibt das Integral den Wert 2π für $i = -n - 1$, und damit $f(n) = -a^n(n < 0)$. Dies ist die vollständig akausale Folge aus Bsp. 3.5 auf Seite 44. Wieder haben wir gesehen, dass die Wahl des ROC entscheidend für die Berechnung der inversen Transformation ist und zu unterschiedlichen Folgen bei gleicher Z-Transformierten führt.

a2) Die Kontur umfasst $z = a$. Dann ist $|z| > |a|$ und wir müssen, um die geometrische Reihe anwenden zu können, den Bruch zunächst umschreiben:

$$\frac{1}{(1-\frac{z}{a})} = \frac{-\left(\frac{a}{z}\right)}{(1-\frac{a}{z})}$$

bzw.

$$\frac{1}{(1-\frac{r}{ae^{j\phi}})} = \frac{-\frac{a}{re^{-j\phi}}}{(1-\frac{a}{re^{-j\phi}})}$$

Nun können wir für $|a| < r = |z|$ schreiben (geometrische Reihe)

$$\frac{1}{(1-\frac{r}{ae^{j\phi}})} = -\frac{a}{re^{-j\phi}}\sum_{i=0}^{\infty}(\frac{a}{re^{-j\phi}})^i$$

und damit

$$f(n) = \frac{1}{(2\pi)}\sum_{i=0}^{\infty} r^n \left(\frac{a}{r}\right)^i \int_0^{2\pi} e^{j(n-i)\phi} \mathrm{d}\phi$$

Die Integrale ergeben nur für $n = i$ den Wert 2π, sonst 0, und damit innerhalb der Summationsgrenzen $f(n) = a^n$,$(n \geq 0)$. Dies ist die kausale Folge in ihrem ROC.

Fazit der inversen Z-Transformation mittels Integration: Es dürfte jedem Leser bei diesem sehr einfachen Beispiel bereits deutlich geworden sein, dass solche Berechnungen über das direkte Integral sehr zeitaufwändig sind und jedesmal einen neuen Ansatz erfordern. Wir möchten

das nicht gerne öfter durchführen. Allerdings haben wir mit der Transformation

$$\frac{1}{(1-\frac{z}{a})} = \frac{-(\frac{a}{z})}{(1-\frac{a}{z})}$$

bereits ein sehr nützliches Hilfsmittel kennen gelernt, um die komplexe Transformation $z = 1/w$ und Durchführung einer der Methoden im w-Raum anzuwenden, siehe unten im Abschnitt 3.1.5. Wir haben dies hier implizit getan, indem wir die Variable der geometrischen Reihe z/a mit ihrem Kehrwert a/z ersetzt haben, womit wir gleichzeitig das Konvergenzgebiet $|z| > a$ in $|z| < a$ umgekehrt haben.

b) Berechnung mit Hilfe des Residuensatzes: Der Integrand des Ringintegrals lautet

$$\frac{z^n}{(z-a)}$$

b1) Wählen wir eine Kontur, die z_0 nicht umschließt, also z.B. einen Kreis um den Nullpunkt mit Radius $< |a|$, so haben wir $f(n) = 0$ für $n \geq 0$. Für $n < 0$ haben wir einen Pol der Ordnung $(-n)$ bei $z = 0$. Es ist

$$G(z, z_0) = \frac{1}{z-a}$$

und es muß $(-n-1)$ mal abgeleitet werden. Wir erhalten

$$\begin{aligned} f(n) &= \frac{1}{(-n-1)!} \frac{d^{(-n-1)} G(z, z_0)}{dz^{(-n-1)}}|_{(z=z_0=0)} \\ &= \frac{1}{(-n-1)!} \frac{(-1)^{-n-1}(-n-1)!}{(z-a)^{-n}}|_{(z=z_0=0)} \\ &= (-1)^{-n-1}(z-a)^n|_{(z=z_0=0)} = -a^n \end{aligned}$$

also gerade die vollständig akausale Folge.
b2) Wir wählen eine Kontur, die $z_0 = a$ umschließt, und haben für $n \geq 0$ nur einen einfachen Pol bei $z = a$, also

$$G(z, z_0) = (z-a)^1 z^{(n-1)} \frac{z}{z-a} = z^n$$

Es muss 0 mal abgeleitet werden, womit sofort folgt

$$f(n) = G(z = z_0) = a^n$$

Für $n < 0$ haben wir einen zusätzlichen Pol bei $z = 0$, dessen residualer Anteil gerade zu $f(n) = -a^n$ berechnet wurde. Damit ist für $n < 0$ die Summe der Residuen $f(n) = a^n - a^n = 0$. Zusammen für alle n ergibt sich also für eine Kontur, die $z_0 = a$ umschließt, die vollständig kausale Folge.

Diese Rechnungen waren schon wesentlich systematischer und schneller. Die Frage mag sich jedoch stellen, ob wir durch Rechnung mit dem Residuensatz die akausale Folge einfacher „sehen“ können. Hierzu werden wir weiter unten den Residuensatz im $1/z$ -Bereich herleiten. □

Inverse Z-Transformation mittels Korrespondenztabelle

Eine Alternative zur Lösung des Inversionsintegrals (direkt oder über den Residuensatz) ist eine Zerlegung von $X(z)$ in additive Terme, deren Rücktransformationen bekannt sind. Dies funktioniert insbesondere dann, wenn $X(z)$ ein Signal darstellt und bereits als Summe vorliegt.

Beispiel 3.9 – Inverse Z-Transformation eines Signals.

Beschreibt $X(z)$ ein Signal im z-Bereich, kann es in einer charakteristischen Form geschrieben werden:

$$X(z) = 2z^{-1} + 5z^{-2} + 3z^{-4}$$

Nutzt man die Kenntnis der Z-Transformierten der Dirac-Folge

$$x[n] = \delta[n] \Leftrightarrow X(z) = 1$$

und der shift-Eigenschaft der Z-Transformation

$$x[n - N] \Leftrightarrow X(z)z^{-N}$$

so lässt sich $x[n]$ unter Verwendung des Überlagerungssatzes bestimmen:

$$x[n] = 2\delta[n-1] + 5\delta[n-2] + 3\delta[n-4] = \{\ldots, \underline{0}, 2, 5, 0, 3, 0, \ldots\}$$

□

Inversion im $1/z$-Bereich

Im Beispiel 3.4 auf Seite 43 hatten wir die geometrische Reihe im z- und im $1/z$-Bereich betrachtet. Dies wollen wir hier formalisieren.

$F(z)$ konvergiert, wie im Abschnitt 3.1.4 auf Seite 41 gezeigt, für

$$r_{\text{haupt}} < |z| < r_{\text{reg}} \tag{3.46}$$

mit

$$F_{\text{reg}}(z) = \sum_{n=0}^{\infty} f(-n)z^n \tag{3.47}$$

und

$$F_{\text{haupt}}(z) = \sum_{n=1}^{\infty} f(n) z^{-n} \tag{3.48}$$

Wir betrachten nun

$$F(1/z) = \sum_{n=-\infty}^{\infty} f(n) z^{n} \tag{3.49}$$

In dieser Darstellung vertauschen sich also gerade Hauptteil und regulären Teil der Laurent-Reihe, und damit gilt

$$\frac{1}{r_{\text{reg}}} < |z| < \frac{1}{r_{\text{haupt}}} \tag{3.50}$$

Wir leiten nun die Inversionsformel für $F(1/z)$ ab:

Theorem 3.6 (Inversionsformel im 1/z -Bereich).

$$f(n) = \frac{1}{(2\pi j)} \oint z^{-n-1} F(1/z) \mathrm{d}z \tag{3.51}$$

wobei die Kontur in dem angegebenen ROC Gl. (3.50) *für* $F(1/z)$ *verläuft.*

Zum Beweis setzen wir $w = 1/z$ und erhalten

$$F(z^{-1}) = F(w) = \sum_{n=-\infty}^{\infty} f(n) w^{-n} \tag{3.52}$$

Die Inversionsformel für $F(w)$ können wir wie oben angeben:

$$f(n) = \frac{1}{(2\pi j)} \oint w^{n-1} F(w) \mathrm{d}w \tag{3.53}$$

Wir ersetzen wieder $w = 1/z$, $dw = -z^{-2} dz$, und durch Umdrehen des Richtungssinns der Ringintegration bei dieser Transformation erhalten wir einen weiteren Faktor -1. Insgesamt erhalten wir die gesuchte Inversionsformel im 1/z -Bereich. □

Beispiel 3.10 – Inverse Z-Transformation im $1/z$-Bereich.

Wir betrachten wieder die Inversion von $F(z) = \frac{z}{(z-a)}$
Oben hatten wir die Inversion für $|z| > |a|$ mit dem Residuensatz aus der Inversionsformel für $F(z)$ hergeleitet. Analog können wir nun für $|w| = |1/z| > |1/a|$, also im komplementären Bereich $|z| < |a|$, die Inversion mit der Formel für $F(1/z)$ bestimmen.

Lösung:
Wir erhalten

$$F(z^{-1}) = \frac{\frac{1}{z}}{\frac{1}{z} - a} = \frac{1}{(1 - az)}$$

und somit

$$f(n) = \frac{1}{(2\pi j)} \oint \frac{z^{-n-1}}{(1-az)} \mathrm{d}z$$

Der Integrand hat für $n < 0$ einen einfachen Pol bei $z = 1/a$. Anwendung des Residuensatzes liefert:

$$f(n) = \mathrm{Res}\left\{ (z - \frac{1}{a}) \frac{z^{-n-1}}{(1-az)} \right\} = -(\frac{1}{a}) z^{-n-1}|_{z=\frac{1}{a}} = -a^n$$

Für $n \geq 0$ wählen wir zweckmäßigerweise nicht die Transformation im $1/z$-Bereich, sondern die im obigen Beispiel verwendete Tansformation im z-Bereich, die sofort $f(n) = 0$ liefert.
Beachten Sie nochmals, dass in diesem obigen Beispiel für $n < 0$ $(-n-1)$-mal abgeleitet werden musste, während hier (im $1/z$-Bereich) wegen des einfachen Poles gar nicht abgeleitet werden muss. Dieses Ergebnis gilt im ROC für $1/z$, also für $|z| < |a|$. Wir haben damit also den akausalen Teil der Folge hergeleitet. □

Inverse Z-Transformation durch Partialbruchzerlegung

Beschreibt $X(z)$ die Übertragungsfunktion eines LTI-Systems, so handelt es sich in den meisten Fällen um eine gebrochen-rationale Funktion. Mittels Partialbruchzerlegung (PBZ) kann solch ein Bruch in Summanden aufgeteilt werden, die einzeln in den Zeitbereich transformiert werden können.

Wir betrachten der Einfachheit halber nur gebrochen rationale Funktionen $F(z)$ mit K einfachen Polen bei $z_k \neq 0$. Nun betrachten wir die Funktion $F(z)/z$ und führen für diese Funktion eine Partialbruchzerlegung durch:

$$\frac{F(z)}{z} = \sum_{k=1}^{K} \frac{A_k}{(z - z_k)} \tag{3.54}$$

Damit ist

$$F(z) = \sum_{k=1}^{K} \frac{A_k z}{(z - z_k)} \tag{3.55}$$

und mit der Inversion aus den schon bekannten Beispielen gilt im kausalen Bereich das

Theorem 3.7 (Z-Transformation mit Partialbruchzerlegung).

$$f(n) = u(n) \sum_{k=1}^{K} A_k z_k^n, \quad (|z| > |z_k|) \; \forall k = 1 \ldots K \tag{3.56}$$

Die A_k ergeben sich dabei zu

$$A_k = (z - z_k) \frac{F(z)}{z} \bigg|_{z=z_k} \tag{3.57}$$

Beispiel 3.11 – Partialbruchzerlegung.

Wir betrachten die Funktion

$$F(z) = \frac{1}{(z-a)(z-b)}$$

Partialbruchzerlegung liefert

$$\frac{F(z)}{z} = \frac{A_1}{z} + \frac{A_2}{(z-a)} + \frac{A_3}{(z-b)}$$

mit

$$A_1 = \frac{1}{(ab)},\ A_2 = \frac{1}{(a(a-b))},\ A_3 = \frac{1}{(b(b-a))}$$

Mit dem Satz über die Partialbruchzerlegung erhalten wir sofort

$$f(n) = u(n-1)\frac{[a^{n-1} - b^{n-1}]}{(a-b)}$$

Dieses Resultat gilt nur, wenn alle Pole im ROC liegen, also für $|z| > \max\{|a|, |b|\}$. □

3.1.6 Parseval'sche Gleichung

Wir leiten zunächst folgenden Ausdruck für die Z-Transformation des Quadrates einer Folge her.

Theorem 3.8 (Z-Transformation des Quadrates einer Folge).

$$Z\{f(n)f(n)\} = \sum_{n=-\infty}^{\infty} f(n)f(n)z^{-n} = \frac{1}{(2\pi j)} \oint F(v)\frac{1}{v}F\left(\frac{z}{v}\right) \mathrm{d}v \qquad (3.58)$$

Zum Beweis gehen wir von der Definition der Z-Transformation aus:

$$Z\{f(n)f(n)\} = \sum_{n=-\infty}^{\infty} f(n)f(n)z^{-n} \qquad (3.59)$$

Eine Folge $f(n)$ wird darin mit Hilfe des Inversions-Theorems 3.6 auf Seite 53, Gl. (3.51), ausgedrückt:

$$Z\{f(n)f(n)\} = \frac{1}{(2\pi j)} \sum_{n=-\infty}^{\infty} f(n)z^{-n} \oint F(v)v^{n-1}\mathrm{d}v \qquad (3.60)$$

Wir vertauschen Summation und Integration und erhalten

$$Z\{f(n)f(n)\} = \frac{1}{(2\pi j)} \oint F(v)\frac{1}{v} \sum_{n=-\infty}^{\infty} f(n)\left(\frac{z}{v}\right)^{-n} \mathrm{d}v \qquad (3.61)$$

Nach Definition der Z-Transformation ist das gerade

$$Z\{f(n)f(n)\} = \frac{1}{(2\pi j)} \oint F(v)\frac{1}{v}F\left(\frac{z}{v}\right) \mathrm{d}v \tag{3.62}$$

Setzen wir Gl. (3.59) in Gl. (3.62) ein, so ergibt sich gerade die in Theorem 3.8 verlangte Z-Transformierte des Quadrates einer Summe. □

Aus Theorem 3.8 können wir leicht als Spezialfall die Parseval'sche Gleichung im z-Bereich ermitteln:

Theorem 3.9 (Parseval'sche Gleichung im z-Bereich).

$$\sum_{n=-\infty}^{\infty} f(n)f(n) = \frac{1}{(2\pi j)} \oint F(z)\frac{1}{z}F\left(\frac{1}{z}\right) \mathrm{d}z \tag{3.63}$$

Zum Beweis vertauschen wir in Theorem 3.8 die Variablen z und v:

$$\sum_{n=-\infty}^{\infty} f(n)f(n)v^{-n} = \frac{1}{(2\pi j)} \oint F(z)\frac{1}{z}F\left(\frac{v}{z}\right) \mathrm{d}z \tag{3.64}$$

Für $v = 1$ ergibt sich sofort Theorem 3.9. □

Hieraus können wir die „übliche" Parseval'sche Gleichung herleiten:

Theorem 3.10 (Parseval'sche Gleichung im Frequenzbereich).

$$\sum_{n=-\infty}^{\infty} |f(n)|^2 = \frac{1}{\Omega} \int_{-\frac{\Omega}{2}}^{\frac{\Omega}{2}} \left|F(e^{j\,\omega\,T})\right|^2 \mathrm{d}\omega \quad \text{mit } \Omega = \frac{2\pi}{\mathrm{T}} \tag{3.65}$$

Dazu ersetzen wir eine der Folgen $f(n)$ durch die konjugiert komplexe Folge $f^*(n)$, und wir betrachten Theorem 3.9 an der Stelle $z = e^{j\,\omega\,T}$. Damit ist $dz/z = jT\mathrm{d}\omega$. Nach Einsetzen in Theorem 3.9 folgt das gewünschte Resultat. □

Die Parseval'sche Gleichung besagt also, dass die „Energie der Folge"

$$\sum_{n=-\infty}^{\infty} |f(n)|^2 \tag{3.66}$$

entweder durch Summation über die Einzelenergien berechnet werden kann oder durch Integration über die spektralen Energien.

Beispiel 3.12 – Parseval'sche Gleichung.

Wir betrachten wieder die Folge $f(n) = a^n u(n)$ mit $|a| < 1$ und berechnen ihre Energie.

Mit direkter Berechnung gilt (geometrische Reihe)

$$\sum_{n=-\infty}^{\infty} f^2(n) = \sum_{n=0}^{\infty} (a^2)^n = \frac{1}{1-a^2}$$

Mit $F(z) = z/(z-a)$ gilt mit Theorem 3.9, Parsevalsche Gleichung,

$$\sum_{n=-\infty}^{\infty} f^2(n) = \frac{1}{(2\pi j)} \oint F(z) \frac{1}{z} F(\frac{1}{z}) \mathrm{d}z = \frac{1}{(2\pi j)} \oint \frac{1}{((z-a)(1-az))} \mathrm{d}z$$

und über den Residuensatz, da $|a| < 1$, geht nur der Pol bei $z = a$ ein:

$$\sum_{n=-\infty}^{\infty} f^2(n) = \frac{1}{1-az} \Big|_{z=a} = \frac{1}{1-a^2}$$

□

3.1.7 Nichtlineare Systeme

Eine analytische Untersuchung *nichtlinearer* zeitdiskreter Systeme ist insbesondere im Falle einer zusätzlichen Zeitvarianz mathematisch nur mit hohem Aufwand möglich. Nichtlineare rekursive Systeme zeigen in vielen Fällen chaotisches Verhalten. Eine analytische Lösung der Differenzengleichung ist damit nicht mehr möglich.

In der Praxis versucht man aus diesen Gründen nichtlineare Systeme weitestgehend zu vermeiden. Deshalb sei an dieser Stelle lediglich exemplarisch aufgeführt, auf welche Probleme man bei der Untersuchung nichtlinearer zeitdiskreter Systeme stößt.

Als Beispiel sei ein zeitdiskretes System mit folgender Differenzengleichung gegeben:

$$x_{n+1} = a x_n (1 - x_n) \tag{3.67}$$

Es handelt sich hierbei um eine nichtlineares, zeitinvariantes, rekursives System mit einem freien Parameter a. Obwohl eine vergleichbare Differentialgleichung im zeitkontinuierlichen Bereich explizit lösbar ist, führt das vergleichbare zeitdiskrete System zu einer extrem komplizierten Lösungsmenge. Einen tieferen Einblick erhält der geneigte Leser in dem klassischen Aufsatz (Feigenbaum, 1980).

Wie man in Abb. 3.2 und 3.3 sieht, konvergiert das System nur für $a \leq 3$ gegen einen festen Wert. Wählt man den freien Parameter größer, so oszilliert x_n zwischen $2, 4, 8, \ldots, 2^{f(a)}$ Werten. Für Werte oberhalb ca. $3,8$ wird die Lösungsmenge derart komplex, dass das Verhalten des Systems als „deterministisch chaotisch“ bezeichnet wird.

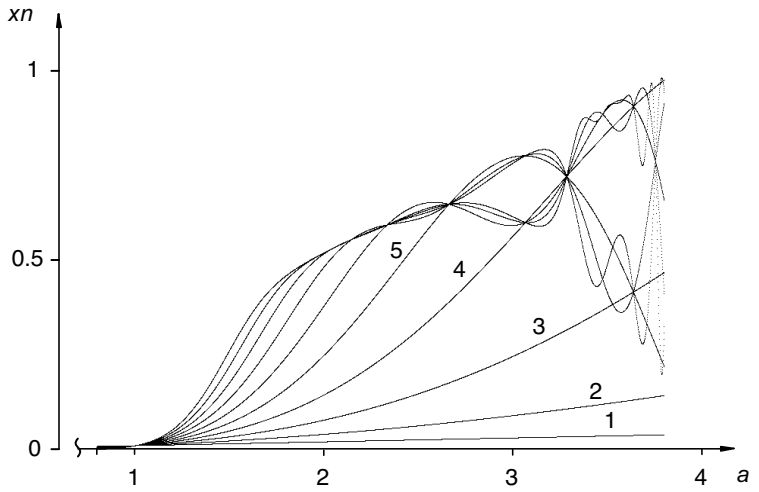


Abbildung 3.2. Bifurkationsdiagramm für die ersten zehn Iterationen (Startwert: 0.01)

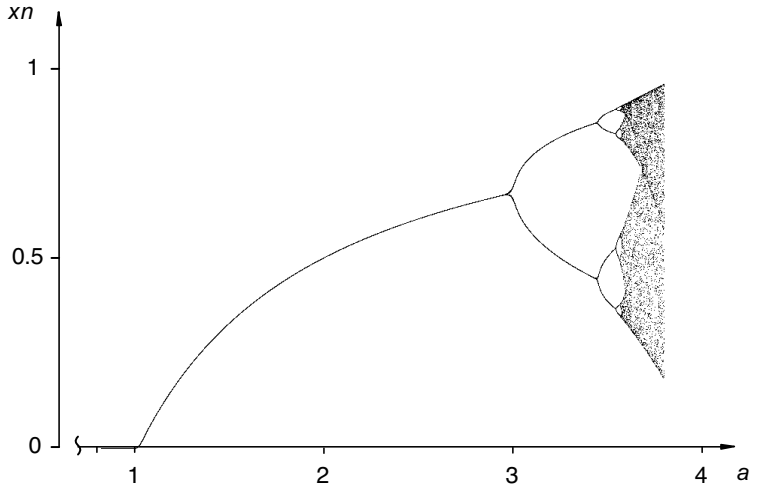


Abbildung 3.3. Bifurkationsdiagramm der 50. bis 100. Iteration. Nur für Werte $a < 3$ konvergiert die Rekursion gegen einen festen Wert

3.2 Beschreibungsformen zeitdiskreter LTI-Systeme

Wie noch gezeigt werden soll, haben unterschiedliche Beschreibungsformen eines LTI-Systems je nach Problemstellung bzw. je nach Anwendungsfall Vor- und Nachteile. Sie alle beschreiben ein System mehr oder weniger vollständig und sind demzufolge eindeutig ineinander umwandelbar – andernfalls ist gesondert zu beachten, welche Informationen über das System durch die Umwandlung verloren gehen. Auf folgende Darstellungsformen und ihre Verwendung soll in den nächsten Kapiteln genauer eingegangen werden:

- Operatordarstellung
- funktionale Darstellung
- Impulsantwort und Sprungantwort
- Differenzengleichung
- Übertragungsfunktion (im z-Bereich)
- Pol- und Nullstellendarstellung
- Blockschaltbild

3.2.1 Operatordarstellung und funktionale Darstellung

Zentraler Punkt der theoretischen Beschreibung eines zeitdiskreten Systems ist die Unterscheidung zwischen Eingangssignalen (hier x), Ausgangssignalen (y) und dem Systemzustand (z). Alle drei Größen können vektoriell auftreten. Die Anzahl der skalaren Elemente von x entspricht somit der Anzahl der betrachteten Eingänge in das System, und die Anzahl der Elemente von y entspricht der Anzahl der Ausgänge. Die Anzahl der Elemente des Zustandsvektors z entspricht der Zahl der unabhängigen Speicher im System, und bestimmt die **Ordnung** des Systems.

Der zweite wesentliche Punkt der Systembeschreibung ist der *System-Operator* (hier Θ), der den eigentlichen mathematischen Zusammenhang zwischen Eingangssignal x, dem Systemzustand z zu einem Zeitpunkt t mit dem Ausgangssignal y herstellt:

$$\mathbf{y}[n] = \Theta(\mathbf{x}[n], \mathbf{z}(\mathbf{x}[n], \mathbf{z}[n-1], t), t) \tag{3.68}$$

In dieser kompakten und eher theoretischen Schreibweise wird ausgedrückt, dass das Element n der Ausgangsfolge y berechnet wird aus dem zugehörigen Element des Eingangssignals und dem aktuellen Systemzustand. Dieser Systemzustand wird wiederum bestimmt aus dem Eingangssignal und dem vorangegangen Systemzustand. Im Falle eines zeitvarianten Systems, ist der System-Operator Θ und ggf. auch der Systemzustand eine Funktion von t.

Betrachtet man ein System von außen, so stellt der Zustand z einen nicht direkt beobachtbaren Vektor dar, dessen Kenntnis oftmals nicht von Interesse ist, da lediglich ein Zusammenhang zwischen Eingangs- und Ausgangssignal gesucht wird. Aus dieser Sichtweise ist es wünschenswert, ein LTI-System als

eine Funktion (hier ebenfalls mit Θ bezeichnet) betrachten zu können, die eine beliebige Eingangsfolge $x[n]$ auf eine Ausgangsfolge $y[n]$ abbildet:

$$\mathbf{y}[n] = \mathbf{\Theta}(\mathbf{x}[n]) = \begin{pmatrix} \Theta_1(\mathbf{x}[n]) \\ \Theta_2(\mathbf{x}[n]) \\ \vdots \\ \Theta_a(\mathbf{x}[n]) \end{pmatrix} \tag{3.69}$$

Es darf jedoch nicht vernachlässigt werden, dass unterschiedliche Anfangszustände z des Systems bei gleichen Eingangssignalen im Allgemeinen zu unterschiedlichen Ausgangssignalen führen. Um einen definierten (und technisch praktikablen) Anfangszustand festzulegen, können beispielsweise sämtliche Elemente des Zustandsvektors auf Null gesetzt werden. Ein LTI-System mit $z = 0$ („zero state“) nennt man „*relaxiertes LTI-System*“.

Ein LTI-System ist durch folgende Eigenschaften gekennzeichnet

- Linearität: Für jedes lineare System gilt die Superpositionseigenschaft und umgekehrt:
$$\Theta(x_1[n] + x_2[n]) \overset{LTI}{=} \Theta(x_1[n]) + \Theta(x_2[n]) \tag{3.70}$$
- Homogenität: Identisch zur Linearitätseigenschaft und aus (3.6) unmittelbar abzuleiten ist der für LTI-Systeme geltende Zusammenhang
$$\Theta(\alpha x[n]) = \alpha\Theta(x[n]) \tag{3.71}$$
- Zeitinvarianz: In diesem Fall gilt, dass zwei von der Form her identische Signale bei gleichen Anfangsbedingungen (gleichem Systemzustand) ebenfalls zwei in der Form identische Ausgangssignale erzeugen (Verschiebungsinvarianz):
$$\Theta(x[n]) = y[n] \;\Leftrightarrow\; \Theta(x[n - n_0]) = y[n - n_0] \tag{3.72}$$

Diese drei Eigenschaften – hier für zeitdiskrete Signale und Systeme angegeben – gelten in vergleichbarer Weise auch für zeitkontinuierliche Signale und Systeme. Es wird sich zeigen, dass insbesondere die Linearitätseigenschaft eine wesentliche Vereinfachungsmöglichkeit bei der Analyse technischer Systeme zeigt.

3.2.2 Problemstellungen der Systemanalyse

Bevor auf weitere Beschreibungsformen von LTI-Systemen eingegangen wird, soll zunächst erläutert werden, wofür eine mathematische Beschreibung eines allgemeinen Systems notwendig ist. Aus dieser Überlegung heraus soll es möglich werden, die unterschiedlichen Beschreibungsformen hinsichtlich ihrer Brauchbarkeit zur Problemlösung einzuschätzen.

Betrachtet man die Systemgleichung eines LTI-Systems in der oben erläuterten Operatordarstellung

$$\mathbf{y}[n] = \boldsymbol{\Theta}(\mathbf{x}[n]) \tag{3.73}$$

so lassen sich drei verschiedene analytische Aufgabenstellungen formulieren:

1. $x[n]$ und Θ sind gegeben, berechne $y[n]$
2. $y[n]$ und Θ sind gegeben, berechne $x[n]$
3. $x[n]$ und $y[n]$ sind bekannt, ermittle Θ

Während sich im ersten Problemfall lediglich numerische Probleme ergeben können, erfordert der zweite Fall ein Umstellen der Gleichung (3.8) nach dem implizit verknüpften Eingangssignal $x[n]$, um einen inversen Operator Θ^{-1} zu erhalten. Weitaus schwieriger kann sich eine Lösung der dritten Aufgabenstellung gestalten: Es ist ein System zu finden, um einen Satz von Eingangssignalen in bestimmte Aussgangssignale umzuwandeln. Im Falle der Sprachverarbeitung kann es sich beispielsweise um ein System handeln, das ein analoges Sprachsignal in geschriebenen Text umwandelt. Hier wird deutlich, dass für die Lösung dieser Problemstellung keine allgemein gültige algorithmische Vorgehensweise angegeben werden kann.

3.2.3 Impulsantwort

Jede diskrete Folge lässt sich zu einer Summe von gewichteten und verschobenen δ-Folgen zusammensetzen:

$$x[n] = \{x_n\}_{n=a}^{b} = \sum_{i=a}^{b} x(i)\delta[n-i] \tag{3.74}$$

mit

$$\delta[n] = \{\ldots, 0, \underline{1}, 0, \ldots\} = \left\{ \begin{matrix} 1; n = 0 \\ 0; n \neq 0 \end{matrix} \right\}_{n=-\infty}^{\infty} \tag{3.75}$$

Beispiel 3.13 – Notation eines Signals als Summe von gewichteten Delta-Folgen.

Gegeben sei eine Folge wie folgt:

$$x[n] = \{x_n\}_{n=-1}^{3} = \{2, \underline{1}, 3, 5, 7\}$$

Diese lässt sich darstellen als Summe von gewichteten Delta-Folgen gemäß

$$x[n] = 2\delta(n+1) + 1\delta(n) + 3\delta(n-1) + 5\delta(n-2) + 7\delta(n-3)$$

Dies entspricht der Darstellung

$$x[n] = \sum_{i=-1}^{3} x(i)\delta[n-i]$$

□

Als Impulsantwort $h[n]$ (auch: „Gewichtsfunktion“) wird wie im zeitkontinuierlichen Fall die Antwort eines relaxierten LTI-Systems (Zustandsvektor Null) auf den Einheits-Impuls die diskrete Delta-Folge $\delta[n]$ bezeichnet.

$$h[n] = \Theta_{\text{relaxiert}}(\delta[n]) \tag{3.76}$$

Ist $h[n]$ für ein LTI-System bekannt, so kann unter Verwendung der drei im Abschnitt 3.2.1 auf Seite 59 erläuterten LTI-System-Eigenschaften Linearität, Homogenität und Zeitinvarianz die Systemantwort $y[n]$ eines relaxierten LTI-Systems auf eine beliebige Eingangsfolge $x[n]$ bestimmt werden:

$$x[n] = \sum_{i=-\infty}^{\infty} x(i)\delta[n-i] \;\Leftrightarrow\; y[n] = \sum_{i=-\infty}^{\infty} x(i)h[n-i] \tag{3.77}$$

Die Angabe der Impulsantwort eines LTI-Systems reicht demzufolge aus, das LTI-System vollständig zu beschreiben. Insbesondere bei der Untersuchung realer LTI-Systeme bietet sich hiermit eine praktische Möglichkeit, das Verhalten unbekannter Systeme zu ermitteln[1]. Wie in Abschnitt 1.4 gezeigt, lassen sich beliebige Eingangssignale ebenso als Summe gewichteter Einheits-Sprungfunktionen $u[n]$ oder Einheits-Rampen-Folgen $\rho[n]$ darstellen. Somit ist die Angabe der *Übergangsfunktion* (Systemantwort auf $u[n]$) oder die Angabe der Systemantwort auf $\rho[n]$ ebenso zur Systembeschreibung möglich.

3.2.4 Die Differenzengleichung

Auch wenn die funktionale Darstellung (3.2) aus systemtheoretischer Sicht grundlegend ist, ist die Operatorschreibweise in vielen Fällen nicht praktikabel. Auch eine Systemberechnung anhand einer gegebenen Impulsantwort ist für viele arithmetische Problemstellungen ungeeignet. Eine weitere allgemeine Beschreibung eines zeitdiskreten LTI-Systems kann durch die Angabe einer *Differenzengleichung* erfolgen, die den funktionalen Zusammenhang zwischen dem Eingangs- und dem Ausgangssignal in folgender Weise beschreibt:

$$y_n + a_1 y_{n-1} + \ldots + a_N y_{n-N} = b_0 x_m + b_1 x_{m-1} + \ldots + b_M x_{m-M} \tag{3.78}$$

Anmerkungen zur Differenzengleichung:

- Sind x und y (wie bisher angenommen) ein *Vektor* (mehrere Ein-/Ausgänge), dann besitzen auch die Koeffizienten a_n und b_m vektoriellen Charakter. Da dies jedoch keinen prinzipiellen Einfluss auf die im Folgenden vorgestellten mathematischen Methoden zur Berechnung eines LTI-Systems hat, soll hier von skalaren reellwertigen Koeffizienten ausgegangen werden.

[1] Zur praktischen Untersuchung von zeitkontinuierlichen Systemen wird die Impulsantwort selten verwendet, da hierfür ein Eingangssignal unendlicher Amplitude und verschwindender Dauer benötigt würde. Zusätzlich zeigt sich, dass nur wenige analoge Systeme überhaupt „sprungfähige“ Systeme darstellen - somit bleibt diese Anregung im Allgemeinen wirkungslos. Im zeitdiskreten Falle hingegen ist ein Einheitsimpuls sehr einfach zu erzeugen und bietet sich für eine Systemanalyse an.

- Sind die Koeffizienten a_i und b_i in der Differenzengleichung zeitinvariant, nennt man auch das System „zeitinvariant".
- Ist einer der a-Koeffizienten nicht Null - wird also das Ausgangssignal auf den Eingang des Systems rückgekoppelt - erhält man ein *rekursives LTI-System*, auf dessen spezielle Eigenschaften weiter unten eingegangen werden soll.
- In der Differenzengleichung wird der Systemzustand nicht erfasst. Um ein konkretes LTI-System mit einer Differenzengleichung komplett zu beschreiben, ist also auch hier die Angabe von Anfangsbedingungen notwendig.
- Um aus einem gegebenen Eingangssignal das Ausgangssignal zu berechnen (oder umgekehrt), ist die Differenzengleichung zu lösen. Ähnlich dem Lösen von Differentialgleichungen sind dafür verschiedene Methoden möglich, auf die im Kap. 5 auf Seite 133 eingegangen wird. Einen Vergleich von Differential- und Differenzengleichungen bringt Übung 3.9.

3.2.5 Übertragungsfunktionen

Wir leiten jetzt 3 Darstellungen der Übertragungsfunktion im z-Bereich her:

1. Aus (3.11) ist bekannt, daß $y[n]$ sich als Faltung der Eingangsfolge $x[n]$ mit der Impulsantwort-Folge $h[n]$ darstellen läßt. Benutzen wir die Äquivalenz von Faltung im Zeit- zur Multiplikation im z-Bereich, so erhalten wir nach Z-Transformation $Y(z) = H(z)X(z)$ bzw. $H(z) = Y(z)/X(z)$, das ist dasselbe $H(z)$ für alle Anregungen $X(z)$ (1. Darstellung der Übertragungsfunktion).
2. Speziell können wir $x[n] = \delta[n]$ wählen, dann ist $y[n]$ die Impulsantwort. Nach Z-Transformation erhalten wir $X(z) = 1$ und damit $Y(z) = H(z)$. $H(z)$ ist also auch die Z-Transformierte der Impulsantwort (2. Darstellung der Übertragungsfunktion).
3. Eine weitere, mathematisch angenehme Form von $H(z)$ erhalten wir, wenn wir formal eine Eingangsfolge $x[n] = z^n$ wählen. Dann ist die Systemantwort durch Faltung gegeben zu

$$y[n] = \sum_{k=-\infty}^{\infty} h[k] z^{n-k} = z^n H(z) \tag{3.79}$$

   also $H(z) = y[n]/z^n$, wobei die Abhängigkeit von n auf der rechten Seite in jedem Fall verschwindet. Speziell gilt $H(z) = y[0]$ bei (formaler) Anregung des Sytems mit $x[n] = z^n$ (3. Darstellung der Übertragungsfunktion).

3.2.6 Blockschaltbilder

Eine graphische Form der Darstellung eines LTI-Systems bieten Blockschaltbilder. Wir verwenden für die Blockschaltbild-Darstellung folgende (und andere) Symbole:

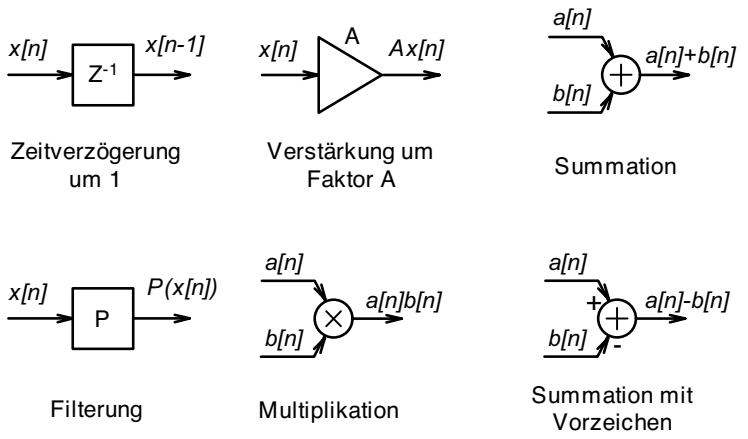


Abbildung 3.4. Blockschaltbilder für sechs verschiedene Operationen mit den Signalen a[n] und b[n]

Die allgemeine Differenzengleichung des LTI-Systems

$$y_n + a_1 y_{n-1} + a_2 y_{n-2} + \ldots + a_N y_{n-N} = b_0 x_m + b_1 x_{m-1} + b_2 x_{m-2} + \ldots + b_M x_{m-M} \quad (3.80)$$

führt zu folgendem (Abb. 3.5) *kanonischen* Blockschaltbild, wobei ohne Beschränkung der Allgemeinheit $N \geq M$ angenommen wurde (ansonsten vertausche N und M). Die Bezeichnung kanonisch rührt davon her, dass in dieser Anordnung die minimale Anzahl, nämlich N Verzögerer benötigt werden, wohingegen bei direktem Umsetzen der Gl. (3.80) in ein Blockschaltbild (siehe Abb. 8.3(c)) $(N + M)$ Verzögerer benötigt werden. Wir können nun zeigen:

Theorem 3.11 (Äquivalenz allgemeiner LTI-Differenzengleichung und kanonischer Schaltung).
Die allgemeine LTI-Differenzengleichung (3.80) *und die kanonische Schaltung gemäss Abb. 3.5 sind äquivalent.*

Die Äquivalenz von Gl. (3.80) und Abb. 3.5 können wir wie folgt zeigen: Wir definieren als $s[n]$ die Grösse im oberen Zentrum der Abb. 3.5, vor dem b_0-Multiplizierer. Für sie können wir im rechten und im linken Zweig der Abb. 3.5 folgende zwei Differenzengleichungen direkt ablesen, wobei wir ohne Beschränkung der Allgemeinheit annehmen, von den N dargestellten Koeffizienten b_i seien die letzten $(N - M)$ viele vom Werte $b_i = 0$.

$$\begin{aligned} y_n &= b_0 s_n + b_1 s_{n-1} + b_2 s_{n-2} + \ldots + b_M s_{n-M} \\ s_n &= x_n - a_1 s_{n-1} - a_2 s_{n-2} - \ldots - a_N s_{n-N} \end{aligned} \quad (3.81)$$

Nach Z-Transformation ergibt sich:

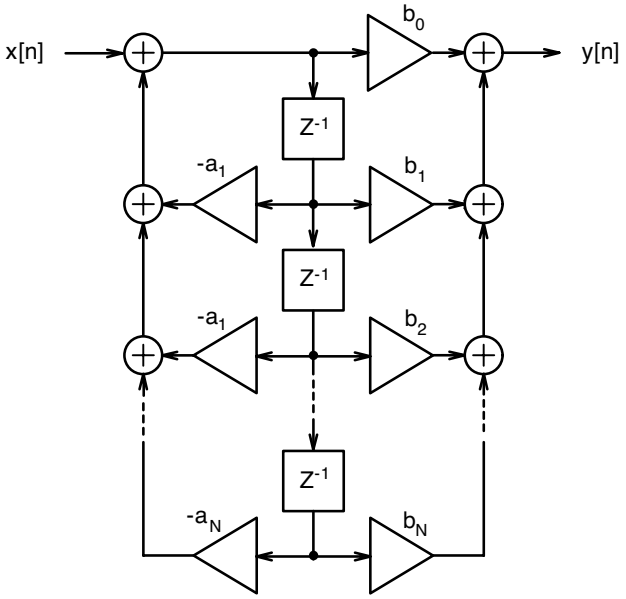


Abbildung 3.5. Kanonische Blockschaltbild-Darstellung eines digitalen Filters

$$\begin{aligned} Y(z) &= S(z)[b_0 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_M z^{-M}] \\ X(z) &= S(z)[1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_N z^{-N}] \end{aligned} \tag{3.82}$$

Nach Einsetzen der beiden Gleichungen ineinander wird $S(z)$ eliminiert, und es ergibt sich unmittelbar die folgende Gleichung:

$$\begin{aligned} Y(z)[1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_N z^{-N}] \\ = X(z)[b_0 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_M z^{-M}] \end{aligned} \tag{3.83}$$

Dieselbe Gleichung ergibt sich durch Transformation beider Seiten der allgemeinen LTI-Differenzengleichung (3.80). Dabei machen wir von der Linearitätseigenschaft (3.18) und der Verschiebungseigenschaft (3.19) der Z-Transformation Gebrauch. Damit ist Theorem 3.11 bewiesen.

Mit der Definition einer Hilfsgrösse $s[n]$ haben wir zum ersten Mal eine innere *Zustandsgrösse* (engl. state variable, daher die Bezeichnung $s[n]$) eingeführt. Diese ist von aussen nicht beobachtbar. Wir werden solche Zustandsgrössen im Kapitel 5 noch ausführlich benutzen. □

Wir fragen nun nach der Übertragungsfunktion der Schaltung in Abb. 3.5. Mit der 1. Darstellung der Übertragungsfunktion (siehe Abschnitt 3.2.5 auf Seite 63) folgt aus Gl. (3.83) sofort

$$H(z) = \frac{Y(z)}{X(z)} = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_M z^{-M}}{1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_N z^{-N}} \tag{3.84}$$

Wir können also bei Differenzengleichungen die Darstellung in der Z-Transformation direkt ablesen!

3.2.7 Pol-Nullstellen-Darstellung, Analyse und Synthese von Systemen

Wir bringen jetzt die Polynomen-Darstellung der Übertragungsfunktion (3.84) in eine Pol-Nullstellen-Form. Durch Umschreiben erhalten wir

$$H(z) = z^{N-M} \frac{b_0 z^M + b_1 z^{M-1} + b_2 z^{M-2} + \ldots + b_M}{z^N + a_1 z^{N-1} + a_2 z^{N-2} + \ldots + a_N} \tag{3.85}$$

Nach dem Fundamentalsatz der Algebra (Argand, 1813, 1815)[2] läßt sich nun jedes Polynom k-ten Grades in ein Produkt aus den k Nullstellenfaktoren des Polynoms und ggf. einen globalen Faktor umschreiben. Es gilt also z.B. für den Zähler

$$b_0 z^M + b_1 z^{M-1} + b_2 z^{M-2} + \ldots + b_M = b_0 \prod_{i=1}^{M} (z - z_{0i}) \tag{3.86}$$

Wir führen dies auch für den Nenner durch, dessen Nullstellen wir mit z_∞ bezeichnen. Die Übertragungsfunktion lautet dann

$$H(z) = b_0 z^{N-M} \frac{\prod_{i=1}^{M} (z - z_{0i})}{\prod_{i=1}^{N} (z - z_{\infty i})} \tag{3.87}$$

Die Nullstellen des Nenners sind die Pole der Übertragungsfunktion. Diese Darstellung heisst daher Pol-Nullstellen-Darstellung der Übertragungsfunktion.

Es ist klar, dass die Umwandlung der Polynom-Darstellung in die Pol-Nullstellen-Darstellung der Übertragungsfunktion rechnerisch aufwendig ist. Der Fundamentalsatz der Algebra garantiert zwar die Pol-Nullstellen-Darstellung, die tatsächliche Berechnung der Nullstellen der Zähler- und Nennerpolynome ist jedoch für Grade > 5 nur numerisch möglich und erfordert erheblichen Aufwand. Anders verhält es sich mit dem umgekehrten Weg, der Umwandlung der Pol-Nullstellen-Darstellung in die Polynomen-Darstellung der Übertragungsfunktion. Hier müssen die Produkte lediglich ausmultipliziert werden.

[2] Der Fundamentalsatz der Algebra wurde bereits von d'Alembert 1746 fast vollständig eingeführt. Normalerweise wird der erste vollständige Beweis Gauß zugeschrieben, der ihn in seiner Doktorarbeit (!) bei J. Pfaff in Helmstedt 1799 erbracht hat - dafür wurde ihm die mündliche Prüfung erlassen. Gauß behauptet in seiner Arbeit übrigens nicht, er habe den ersten richtigen Beweis erbracht, er nennt seinen Beweis lediglich „neu". Nach heutigen Standards ist Gauß' Beweis nicht vollständig, sondern der erste lückenlose Beweis (1814) stammt von Jean Robert Argand. Zur Geschichte des Fundamentalsatzes siehe (Petrova, 1973).

Diese beide Umwandlungswege entspechen 2 verschiedenen Aufgabenstellungen der Systembeschreibung. Ist die Polynom-Darstellung bekannt, z.B. aus einer vorgegebenen Schaltung, so dient die Pol-Nullstellen-Darstellung der Analyse dieser Schaltung. Wie wir sehen werden, sind solche Eigenschaften wie Stabilität oder Phasenverhalten der Schaltung von der Lage der Pole und Nullstellen abhängig, diese müssen also berechnet werden. Dies ist also eine rechnerisch aufwendige Prozedur.

In der Synthese eines Systems soll hingegen ein bestimmtes Systemverhalten erzeugt werden. Dazu werden die Pole und Nullstellen der Übertragungsfunktion festgelegt. Die anschliessende Umwandlung in die Polynom-Darstellung und damit in eine Schaltung ist rechnerisch einfach. Allerdings wird man nicht in jedem Fall diejenige Schaltung wählen, die direkt aus der algebraischen Struktur der Übertragungsfunktion folgt. Diesen Aspekten widmen wir uns noch bei der Betrachtung der *numerischen* Stabilität von digitalen Schaltungen.

Analyse wie Synthese machen die extreme Nützlichkeit der Z-Transformation deutlich. Die Systemeigenschaften wie Stabilität oder Phasenverhalten können aus der Systembeschreibung im Zeitbereich, also z.B. aus der Impulsantwort $h[n]$, praktisch nicht abgelesen werden. Ebensowenig ist eine Umsetzung der durch das System erzeugten Faltung $y[n] = h[n] * x[n]$ in eine Schaltung direkt ablesbar. Diese Aufgaben vereinfachen sich im z-Bereich beträchtlich bzw. sind dort erst möglich.

3.3 Eigenschaften zeitdiskreter LTI-Systeme

3.3.1 Stabilität

Stabilität im Zeitbereich

Ein System wird „BIBO-stabil“ (bounded input - bounded output) genannt, wenn jedes amplitudenbegrenzte Eingangssignal auch ein amplitudenbegrenztes Ausgangssignal liefert. Für die Stabilität eines LTI-Systems gelten folgende Gesetzmäßigkeiten:

- Jedes nicht rekursive System ist BIBO-stabil
- Rekursive LTI-Systeme sind BIBO-stabil, wenn die Nullstellen λ_i des charakteristischen Polynoms der zugehörigen Differenzengleichung der Bedingung $|\lambda_i| < 1$ genügen.
- Ein äquivalentes Kriterium für die Stabilität leitet sich aus der Impulsantwort eines Systems her:

Theorem 3.12 (Stabilitätskriterium im Zeitbereich). *Ein LTI-System ist dann und nur dann BIBO-stabil, wenn gilt:*

$$\sum_{k=-\infty}^{\infty} |h_k| < \infty \tag{3.88}$$

Wir zeigen, dass aus begrenztem input begrenzter output (BIBO) des Systems dann und nur dann folgt, wenn

$$\sum_{n=-\infty}^{\infty} |h[n]| = M < \infty \tag{3.89}$$

Zunächst beweisen wir die Notwendigkeit. Dazu reicht es, ein Gegenbeispiel anzugeben. Wir nehmen also für folgende beschränkte Eingangsfolge

$$x[n] = \text{sign}(h[-n]) \tag{3.90}$$

an, die Ausgangsfolge wäre auch dann begrenzt, wenn die BIBO-Bedingung **nicht** erfüllt ist. Der Wert der Ausgangsfolge des Systems, speziell bei $n = 0$, errechnet sich dann zu

$$y[0] = \sum_{k=-\infty}^{\infty} x[k]h[-k] = \sum_{k=-\infty}^{\infty} |h[-k]| = \sum_{k=-\infty}^{\infty} |h[k]| \tag{3.91}$$

und da nach Voraussetzung die BIBO-Bedingung nicht erfüllt ist, gilt

$$y[0] \Rightarrow \infty \tag{3.92}$$

Damit ist die Notwendigkeit der BIBO-Bedingung bewiesen.

Wir zeigen nun direkt, dass die BIBO-Bedingung hinreichend ist. Gegeben sei also eine beschränkte Eingangsfolge mit

$$|x[n]| < Q \tag{3.93}$$

Dann gilt

$$\begin{aligned} |y[n]| &= |\sum_{k=-\infty}^{\infty} x[k]h[n-k]| \\ &\leq \sum_{k=-\infty}^{\infty} |x[k]h[n-k]| \\ &\leq Q \sum_{k=-\infty}^{\infty} |h[n-k]| \\ &< Q\,M < \infty \end{aligned} \tag{3.94}$$

wobei der letzte Übergang aus der BIBO-Bedingung folgt. Damit ist die BIBO-Bedingung auch hinreichend. □

Stabilität im z-Bereich

Wir zeigen nun eine zur BIBO-Stabilität (im Zeitbereich) äquivalente Formulierung für die Systemfunktion (im z-Bereich).

Theorem 3.13 (Stabilitätskriterium im Z-Bereich).
Ein LTI-System ist dann und nur dann BIBO-stabil, wenn die Pole seiner Übertragungsfunktion im Einheitskreis der z-Ebene liegen.

Wir zeigen diesen Zusammenhang für eine Systemfunktion mit einfachen Polen z_k. Nach dem im Abschnitt 3.1.5 hergeleiteten Theorem über die Partialbruchzerlegung (Theorem 3.7 auf Seite 54) kann diese dargestellt werden als

$$h[n] = C\delta[n] + u[n] \sum_{k=1}^{K} A_k z_k^n \quad \text{für } z > z_k \tag{3.95}$$

Die A_k ergeben sich dabei aus der Systemfunktion $H(z)$ zu

$$A_k = (z - z_k) \frac{H(z)}{z}|_{[z=z_k]} \tag{3.96}$$

In dieser Darstellung lautet die BIBO-Bedingung nun

$$\sum_{n=-\infty}^{\infty} |h[n]| = \sum_{n=-\infty}^{\infty} |C\delta[n] + u[n] \sum_{k=1}^{K} A_k z_k^n| \leq |C| + \sum_{k=1}^{K} |A_k| \sum_{n=0}^{\infty} |z_k^n| \tag{3.97}$$

Damit die am Ende stehende geometrische Reihe konvergiert, muss $|z_k| < 1$ für alle k gelten, mithin müssen alle Pole der Systemfunktion im Einheitskreis liegen. Die Lage der Pole ist somit hinreichend.

Die Notwendigkeit der Lage der Pole zeigt man an einem Gegenbeispiel: Die Übertragungsfunktion habe nur einen einzigen einfachen Pol bei $z_1 = 1$. Dann ist das BIBO-Kriterium

$$\sum_{n=-\infty}^{\infty} |h[n]| = \sum_{n=-\infty}^{\infty} |C\delta[n] + u[n] \sum_{k=1}^{K} A_k z_k^n| = |C + A_1| + \sum_{n=1}^{\infty} |A_1| \to \infty \tag{3.98}$$

verletzt, und somit ist das System nicht BIBO-stabil.

Für die Stabilität ist die Lage der Nullstellen der Systemfunktion ohne Bedeutung. □

Die Lage der Pole im Einheitskreis hat eine Korrespondenz zu der bei der Laplace-Transformation geforderten Stabilitätseigenschaft, nach der alle Pole s_k in der linken Halbebene liegen müssen. Die komplexe Frequenzgrösse $s = \sigma + jw$ im Laplaceraum korrepondiert zum z-Raum via $z = e^{\sigma + jw}$. Damit ergibt die linke Halbebene $\sigma < 0$ im Laplaceraum nach Exponentierung das Innere des Einheitskreises im z-Raum, und ein vertikales Linienstück $\{w = w_0 \ldots w_0 + 2\pi\}$ mit $\sigma = \text{const.}$ im Laplaceraum wird in einen Kreisumlauf mit Radius e^{σ} im z-Raum abgebildet. Wir sehen, dass weitere vertikale Linienstücke $\{w = w_0 + n2\pi \ldots w_0 + (n+1)2\pi\}$ mit demselben $\sigma = \text{const.}$ im Laplaceraum für ganzzahlige n in denselben Kreisumlauf mit Radius e^{σ} im z-Raum abgebildet werden, die Abbildung $s \to z$ ist also nur innerhalb eines Streifens im Laplaceraum mit der imaginären Ausdehnung 2π umkehrbar. Solche Streifen der linke Halbenene im Laplaceraum werden jeweils auf das Innere des Einheitskreises im z-Raum abgebildet.

3.3.2 Kausalität

Kausalität in Laurent-Reihen-Darstellung

Wir wissen bereits, dass ein System kausal ist, wenn für seine Impulsantwort gilt

$$h[n] = 0 \quad \forall n < 0 \tag{3.99}$$

Daraus entnehmen wir eine Darstellung für die Kausalität im z-Bereich.

Theorem 3.14 (Kausalität im Zeitbereich).
Ein System ist dann und nur dann kausal, wenn die Entwicklung der Übertragungsfunktion in eine Laurentreihe nur aus dem Hauptteil besteht.

Um dies zu zeigen, stellen wir die Übertragungsfunktion dar als Z-Transformierte der Impulsantwort und erhalten

$$H(z) = \sum_{n=0}^{\infty} h[n] z^{-n} \tag{3.100}$$

Es ergibt sich also nur der Hauptteil. □

Beispiel 3.14 – Kausalität der Übertragungsfunktion.

Wir betrachten die Übertragungsfunktion $H(z) = z/(z-a)$. Für $|z| < |a|$ lautet die Entwicklung dieser Funktion in eine Laurentreihe

$$H(z) = -\left(\frac{z}{a}\right) \frac{1}{1-\frac{z}{a}} = -\frac{z}{a} \sum_{n=0}^{\infty} \left(\frac{z}{a}\right)^n = -\sum_{n=1}^{\infty} a^{-n} z^n$$

Die Laurentreihe enthält also nur den Regulärteil, das System ist vollständig akausal. Für $|z| > |a|$ gilt:

$$H(z) = \frac{1}{1-\frac{a}{z}} = \sum_{n=0}^{\infty} \left(\frac{a}{z}\right)^n = \sum_{n=0}^{\infty} a^n z^{-n}$$

Die Laurentreihe enthält also nur den Hauptteil, die Koeffizienten sind $h[n] = a^n$, das System ist kausal. □

Besteht die Laurentreihe der Übertragungsfunktion aus dem Hauptteil sowie einer endlichen Anzahl von Koeffizienten des Regulärteils mit maximaler Potenz z^m, so ist jede Funktion $G(z) = z^{-r} H(z)$ mit $r \geq m$ kausal. Im Zeitbereich entspricht dies einer Verschiebung der Impulsantwort, also einer zeitlichen Verzögerung um r Schritte.

Beispiel 3.15 – Kausalität der verschobenen Übertragungsfunktion.

Wir betrachten die Übertragungsfunktion $H(z) = z^m z/z - a$. Für $|z| < |a|$ lautet die Entwicklung dieser Funktion in eine Laurentreihe

$$H(z) = -\frac{z^{m+1}}{a}\frac{1}{1-\frac{z}{a}} = -\frac{z^{m+1}}{a}\sum_{n=0}^{\infty}\left(\frac{z}{a}\right)^n = -a^m \sum_{n=m+1}^{\infty} a^{-n}z^n$$

Die Laurentreihe enthält also unendlich viele Terme des Regulärteils. Damit kann keine modifizierte kausale Übertragungsfunktion angegeben werden. Für $|z| > |a|$ gilt

$$H(z) = z^m \frac{1}{1-\frac{a}{z}} = z^m \sum_{n=0}^{\infty}\left(\frac{a}{z}\right)^n = a^m \sum_{n=-m}^{\infty} a^n z^{-n}$$

Die Laurentreihe enthält also nur den Hauptteil, sowie endlich (m) viele Koeffizienten des Regulärteils. Offensichtlich sind alle modifizierten Funktionen $G(z) = z^{-r}H(z)$ mit $r \geq m$ kausal, denn

$$G(z) = z^{m-r}\frac{1}{1-\frac{a}{z}} = z^{m-r}\sum_{n=0}^{\infty}\left(\frac{a}{z}\right)^n = a^{m-r}\sum_{n=r-m}^{\infty} a^n z^{-n}$$

d.h. die modifizierte Übertragungsfunktion $G(z)$ hat nur einen Hauptteil. □

Kausalität in Pol-Nullstellen-Darstellung

Vor dem Hintergrund der Kausalität in Laurentreihen-Darstellung können wir nun die Übertragungsfunktion in Pol-Nullstellen-Form untersuchen. Sie lautet:

$$H(z) = C\frac{\prod_{q=1}^{Q}(z-z_{0q})}{\prod_{n=1}^{N}(z-z_{\infty n})} \tag{3.101}$$

Multipliziert man die Terme im Zähler und Nenner aus und untersucht man das Ergebnis dahingehend, ob nur der Hauptteil resultiert, so gilt folgendes: Der Term mit der höchsten Potenz im Zähler, z^Q, muß mindestens von einem Term derselben Potenz im Nenner dividiert werden. Damit muß offensichtlich für Kausalität gelten $N \geq Q$. Kausale Systeme haben also mindestens soviele Polstellen wie Nullstellen.

Der kausale Bereich in der z-Ebene ergibt sich dann durch Entwicklung aller Teil-Übertragungsfunktionen H_a für die Polstellen a in geometrische Reihen. Es gilt für je ein H_a und $|z| > |a|$:

$$H_a(z) = \frac{1}{z-a} = \frac{1}{z}\frac{1}{1-\frac{a}{z}} = \frac{1}{z}\sum_{n=0}^{\infty}\left(\frac{a}{z}\right)^n = \frac{1}{a}\sum_{n=1}^{\infty} a^n z^{-n} \tag{3.102}$$

und damit für $|z| > \max_n |z_{\infty_n}|$:

$$H(z) = C\frac{\prod_{q=1}^{Q}(z-z_{0q})}{\prod_{n=1}^{N}(z-z_{\infty n})} = C\prod_{q=1}^{Q}(z-z_{0q})\prod_{n=1}^{N}\left(\frac{1}{z_{\infty n}}\sum_{k=1}^{\infty} z_{\infty n}^k z^{-k}\right) \tag{3.103}$$

Wir multiplizieren alle Terme in dem Q-Produkt mit $1/z$ und alle Terme in dem N-Produkt mit z, was im ganzen einen weiteren Faktor z^{Q-N} ergibt:

$$H(z) = Cz^{Q-N} \prod_{q=1}^{Q} \left(1 - \frac{z_{0q}}{z}\right) \prod_{n=1}^{N} \left(\sum_{k=0}^{\infty} z_{\infty n}^{k} z^{-k}\right) \tag{3.104}$$

Wie man sieht, enthalten für $N \geq Q$ die 4 Faktoren des Produkts keine Terme mit positiven Potenzen von z. Wir betrachten die Ordnung der Terme:

$$\begin{gathered} C = Ord(z^0),\ z^{Q-N} = Ord(z^{Q-N}) \leq Ord(z^0), \\ \prod_{q=1}^{Q} \left(1 - \frac{z_{0q}}{z}\right) = Ord(z^0), \\ \prod_{n=1}^{N} \left(\sum_{k=0}^{\infty} z_{\infty n}^{k} z^{-k}\right) = Ord(z^0) \end{gathered} \tag{3.105}$$

Die Übertragungsfunktion enthält also ausschließlich den Hauptteil und ist damit kausal. Wir fassen noch einmal zusammen:

Theorem 3.15 (Kausalität in Pol-Nullstellen-Darstellung).
Eine Übertragungsfunktion in Pol-Nullstellendarstellung ist kausal für $|z| > \max_n |z_{\infty_n}|$, *und wenn sie mindestens soviele Polstellen wie Nullstellen enthält. Sie kann dann geschrieben werden als*

$$H(z) = C \frac{\prod_{q=1}^{Q}(z - z_{0q})}{\prod_{n=1}^{N}(z - z_{\infty n})} = Cz^{Q-N} \prod_{q=1}^{Q}(1 - \frac{z_{0q}}{z}) \prod_{n=1}^{N} \left(\sum_{k=0}^{\infty} z_{\infty n}^{k} z^{-k}\right) \tag{3.106}$$

woraus sich nach Ausmultiplizieren und Identifikation der Koeffizienten mit $h[n]$ *die Darstellung im Zeitbereich ergibt.*

Einen Fall, in dem eine verletzte Kausalität schon durch einen Pol mit $|z| < |z_{\infty_n}|$ verursacht ist, finden Sie in Übung 3.13. Insbesondere gilt dann auch, dass in diesem z-Bereich jede verschobene Übertragungsfunktion (siehe obiges Beispiel 3.15) ebenfalls nicht kausal sein kann.

3.3.3 Allpässe

Allpass-Eigenschaft in Pol-Nullstellen-Form

Ein Allpass ist ein System, bei dem der Betrag der Systemfunktion für alle Frequenzen konstant ist, also

$$|H[z = e^{jwt}]| = \text{const}(w) \tag{3.107}$$

Wir wollen nun untersuchen, welche Auswirkungen diese verlangte Eigenschaft auf Pole und Nullstellen hat. Betrachten wir die typischerweise auftretende

rationale Systemfunktion, so lässt sich diese darstellen als Quotient eines Zählerpolynoms $Q[z]$ und Nennerpolynoms $N[z]$ mittels

$$H(z) = \frac{Q(z)}{N(z)} \tag{3.108}$$

Betrachten wir diese Darstellung in faktorisierter Form an den Stellen $z = e^{jwt}$, so ergibt sich

$$H(z) = C\frac{\prod_{q=1}^{Q}(z - z_{0q})}{\prod_{n=1}^{N}(z - z_{\infty n})} \quad \text{für } z = e^{jwt} \tag{3.109}$$

Soll die Allpass-Bedingung gelten, so muß

$$C|Q(e^{jwt})| = |N(e^{jwt})| \quad \forall w \tag{3.110}$$

gelten. Dies ist insbesondere dann der Fall, wenn Zähler- und Nennerpolynom konjugiert komplex zueinander sind, also

$$CQ(e^{jwt}) = N^*(e^{jwt}) \tag{3.111}$$

Es kann gezeigt werden, dass dieser Zusammenhang sowie die (uninteressante, weil $H = C$) Identität $CQ = N$ in der Tat die einzig möglichen Darstellungen zur Erfüllung der Allpass-Bedingung sind. Wir zeigen dies im Anschluss an diesen Abschnitt. Damit sind alle folgenden Sätze sowohl notwendig wie auch hinreichend („genau dann, wenn"). Bei konjugierter Komplexheit gilt:

$$\begin{aligned} H(z) &= C\frac{\prod_{q=1}^{Q}(e^{-jwt} - z_{0q}^*)}{\prod_{q=1}^{Q}(e^{jwt} - z_{0q})} \\ &= C\prod_{q=1}^{Q}(-e^{-jwt}z_{0q}^*)\frac{\prod_{q=1}^{Q}(e^{jwt} - \frac{1}{z_{0q}^*})}{\prod_{q=1}^{Q}(e^{jwt} - z_{0q})} \end{aligned} \tag{3.112}$$

und damit gilt für den Betrag:

$$|H(z)| = |C|\prod_{q=1}^{Q}|z_{0q}||\frac{\prod_{q=1}^{Q}(e^{jwt} - \frac{1}{z_{0q}^*})}{\prod_{q=1}^{Q}(e^{jwt} - z_{0q})}| \tag{3.113}$$

Für die Frequenzbetrachtung ist nur der Quotient entscheidend. Durch Vergleich mit der oben angegebenen allgemeinen Darstellung für die Systemfunktion folgt:

Theorem 3.16 (Allpass-Eigenschaft für Übertragungsfunktionen in Pol-Nullstellen-Formulierung).

Ein System mit rationaler Systemfunktion hat genau dann Allpass-Eigenschaften, wenn Zähler- und Nennerpolynom gleichen Grad Q haben und bis

auf eine Konstante konjugiert komplex zueinander sind, womit Pole und Nullstellen paarweise am Einheitskreis gespiegelt sein müssen:

$$\frac{1}{z_{0q}^*} = z_{\infty q} \quad \forall q = 1 \dots Q \tag{3.114}$$

Für stabile Systeme liegen alle Pole innerhalb des Einheitskreises. Sind diese Systeme auch allphasig, bedeutet dies, dass alle Nullstellen außerhalb des Einheitskreises liegen.

Allpass-Eigenschaft in Polynom-Form

Eine andere (algebraische) Formulierung ergibt sich wie folgt. Wir betrachten eine Systemfunktion in Polynom-Form. Diese lässt sich wie folgt darstellen:

$$H(z) = C\frac{\sum_{q=1}^{Q} a_q e^{jwTq}}{\sum_{n=1}^{N} b_n e^{jwTn}} \tag{3.115}$$

Jetzt betrachten wir den Allpass. Nach dem o.a. Satz über die Allpass-Eigenschaft haben Zähler- und Nennerpolynom den gleichen Grad und sind konjugiert komplex zueinander, somit kann diese Systemfunktion auch ausmultipliziert werden und lautet dann

$$H(z) = C\frac{\sum_{q=1}^{Q} b_q^* e^{-jwTq}}{\sum_{q=1}^{Q} b_q e^{jwTq}} \tag{3.116}$$

Dies läßt sich umformen zu

$$H(z) = Ce^{-jwTQ}\frac{\sum_{q=1}^{Q} b_{Q-q}^* e^{jwTq}}{\sum_{q=1}^{Q} b_q e^{jwTq}} \tag{3.117}$$

und damit gilt für den Betrag:

$$|H(z)| = |C|\frac{|\sum_{q=1}^{Q} b_{Q-q}^* e^{jwTq}|}{|\sum_{q=1}^{Q} b_q e^{jwTq}|} \tag{3.118}$$

Daraus kann nach Vergleich mit der oben angegebenen allgemeinen Formulierung für die Systemfunktion sofort eine Koeffizientenbedingung für die Allphasigkeit bei Darstellung der Systemfunktion in ausmultiplizierter Form angegeben werden. Diese lautet:

Theorem 3.17 (Allpasseigenschaft für Übertragungsfunktionen in Polynom-Form).

Ein System mit rationaler Systemfunktion hat genau dann Allpass-Eigenschaften, wenn Zähler- und Nennerpolynom gleichen Grad Q haben und wenn für die Koeffizienten gilt:

$$a_q = b_{Q-q}^* \quad \forall q = 1 \dots Q \tag{3.119}$$

Beachten Sie, dass man an dieser Formulierung nicht sehen kann, ob der Allpass auch stabil ist.

Mögliche Darstellungen zur Erfüllung der Allpass-Bedingung

Oben wurde gezeigt, dass für die Gültigkeit der Allpass-Bedingung gelten muss:

$$C|Q(e^{jwt})| = |N(e^{jwt})| \quad \forall w$$

Dies ist insbesondere dann der Fall, wenn Zähler- und Nennerpolynom konjugiert komplex zueinander sind, also

$$CQ(e^{jwt}) = N^{*(e^{jwt})}$$

Wir wollen nun zeigen, dass dieser Zusammenhang sowie die (uninteressante, weil $H = C$) Identität $CQ = N$ in der Tat die einzig möglichen Darstellungen zur Erfüllung der Allpass-Bedingung sind. Damit gilt der Satz über Allpässe in Eineindeutigkeit, also „genau dann, wenn".

Zunächst zeigen wir die Eineindeutigkeit algebraisch, dann geben wir dafür eine anschauliche Erklärung. Wir betrachten ein Pol-Nullstellen-Paar in $H(z)$, $\frac{(e^{jwt}-z_{0q})}{(e^{jwt}-z_{\infty q})}$. Für den Betrag muss gelten

$$|\frac{e^{j\Omega} - z_{0q}}{e^{j\Omega} - z_{\infty q}}| = \text{const}(\Omega) \text{ mit } \Omega = wt \tag{3.120}$$

Wir multiplizieren das Betragsquadrat aus und erhalten

$$\begin{aligned}\left|\frac{e^{j\Omega} - z_{0q}}{e^{j\Omega} - z_{\infty q}}\right|^2 &= \frac{1 + |z_{0q}|^2 - 2\cos\Omega \text{Re}z_{0q} - 2\sin\Omega \text{Im}z_{0q}}{1 + |z_{\infty q}|^2 - 2\cos\Omega \text{Re}z_{\infty q} - 2\sin\Omega \text{Im}z_{\infty q}} \\ &= k = \text{const}(\Omega)\end{aligned} \tag{3.121}$$

Dies schreiben wir um zu

$$\begin{aligned}&-(1 + |z_{0q}|^2) + k(1 + |z_{\infty q}|^2) \\ &\quad = 2\cos\Omega(k\text{Re}z_{\infty q} - \text{Re}z_{0q}) + 2\sin\Omega(k\text{Im}z_{\infty q} - \text{Im}z_{0q})\end{aligned} \tag{3.122}$$

Damit das für alle Ω gilt, müssen die beiden Klammern auf der rechten Seite verschwinden. Dies ist die einzige Möglichkeit, Gleichungen vom Typ $a + b\cos\Omega = c\sin\Omega$ für alle Ω zu erfüllen. Man sieht das, indem man $\cos\Omega = x$ und damit $\sin^2\Omega = 1 - x^2$ setzt, was nach Quadrieren $a^2 + 2abx + b^2x^2 = c^2(1 - x)^2$ bzw. $a^2 - c^2 + 2abx + (b^2 + c^2)x^2 = 0$ liefert. Diese algebraische Gleichung ist für alle x nur koeffizientenweise erfüllbar, was nur für $a = b = c = 0$ möglich ist. Wir erhalten also

$$k = \frac{\text{Re}z_{0q}}{\text{Re}z_{\infty q}} = \frac{\text{Im}z_{0q}}{\text{Im}z_{\infty q}} \tag{3.123}$$

also

$$\frac{\text{Re}z_{0q}}{\text{Im}z_{0q}} = \frac{\text{Re}z_{\infty q}}{\text{Im}z_{\infty q}} \tag{3.124}$$

also

$$\Phi_{0q} = \Phi_{\infty q} \tag{3.125}$$

Pole und Nullstellen haben also dieselbe Phase. Mit dem Ergebnis für k folgt weiter

$$(1 + |z_{0q}|^2) = \frac{\mathrm{Re} z_{0q}}{\mathrm{Re} z_{\infty q}}(1 + |z_{\infty q}|^2) \tag{3.126}$$

bzw.

$$\frac{\mathrm{Re} z_{0q}}{1 + |z_{0q}|^2} = \frac{\mathrm{Re} z_{\infty q}}{1 + |z_{\infty q}|^2} \tag{3.127}$$

Da Pole und Nullstellen dieselbe Phase haben, folgt daraus auch

$$\frac{|z_{0q}|}{1 + |z_{0q}|^2} = \frac{|z_{\infty q}|}{1 + |z_{\infty q}|^2} \tag{3.128}$$

Dies ist eine quadratische Gleichung in den Nullstellen, wenn der Pol festliegt (und umgekehrt). Sei also z.B. z_{0q} gegeben, so ist die linke Seite der Gleichung fest und werde gleich $-1/C = -1/C(z_{0q})$ gesetzt. Dann folgt aus der rechten Seite

$$1 + C|z_{\infty q}| + |z_{\infty q}|^2 = 0 \tag{3.129}$$

Diese quadratische Gleichung für den Betrag der Polstelle hat nur 2 Lösungen (für eine festgehaltene Nullstelle). Da wir für jede Nullstelle z_{0q} diese Lösungen für die Pole bereits mit

$$z_{\infty q} = \frac{1}{z_{0q}^*} \tag{3.130}$$

und

$$z_{\infty q} = z_{0q} \tag{3.131}$$

angegeben hatten, gibt es keine weiteren Lösungen. Dass diese Pole die Gl. (3.129) auch wirklich lösen, wird in Übung 3.16 gezeigt. Der Satz über Allpässe ist somit eineindeutig. □

Wir können jetzt eine anschauliche geometrische Begründung geben, welche Auswirkungen die Bedingungen des Theorems 3.16 für Pol-Nullstellenpaare in der z-Ebene hat. Dazu setzen wir die algebraische Lösung in ein Pol-Nullstellen-Paar ein und formen um:

$$\frac{(e^{jwt} - z_{0y})}{(e^{jwt} - \frac{1}{z_{0y}^*})} = -z_{0y} e^{-jwt} \frac{(1 - \frac{e^{jwt}}{z_{0y}})}{(1 - \frac{e^{-jwt}}{z_{0y}^*})} \tag{3.132}$$

Für den Betrag gilt dann:

$$|\frac{(e^{jwt} - z_{0y})}{(e^{jwt} - \frac{1}{z_{0y}^*})}| = |z_{0y}||\frac{(1 - \frac{e^{jwt}}{z_{0y}})}{(1 - \frac{e^{jwt}}{z_{0y}})^*}| \tag{3.133}$$

Der erste Term auf der rechten Seite ist konstant bzgl. w, der zweite stellt gerade den Betrag eines Bruches dar, dessen Nenner das konjugiert komplexe des Zählers ist. Zähler und Nenner haben also gleichen Betrag und laufen gegenphasig, damit sind für alle w die Beträge gleich. Man überlegt sich, dass diese beiden Bedingungen: gleicher Betrag und gegenläufige Phase bei gleicher Winkelgeschwindigkeit w und gleicher Anfangsphase, die einzige Möglichkeit darstellen, für alle w den Betrag des Bruches konstant zu halten. Z.B. zeichne man sich die Situation in der z-Ebene auf. Der Beweis folgt dann durch Gegenbeispiel: Ist eine dieser Bedingungen verletzt, so lässt sich immer ein bestimmtes w finden, bei dem sich der Betrag des Bruches ändert.

3.3.4 Minimalphasigkeit

Wir wenden uns nun der Frage der Minimalphasigkeit zu. Dazu suchen wir Systeme, deren Phasendrehung (bei gleichem $H(z)$) minimal ist. Die Frage lautet, welche Eigenschaften die Systemfunktion solcher Systeme haben muss.

Wir betrachten eine stabile Systemfunktion $H(z)$ mit Nullstellen sowohl innerhalb (z_{0x}) wie auch ausserhalb (z_{0y}) des Einheitskreises. Diese stellen wir wie folgt dar:

$$H(z) = C\frac{\prod_{x=1}^{X}(z - z_{0x})\prod_{y=1}^{Y}(z - z_{0y})}{\prod_{n=1}^{N}(z - z_{\infty n})} \tag{3.134}$$

Erweitern liefert:

$$\begin{aligned} &H(z) \\ &= Cz^{X+Y-N}\frac{\prod_{x=1}^{X}(1 - \frac{z_{0x}}{z})\prod_{y=1}^{Y}(1 - 1/(z_{0y}^{*}z))}{\prod_{n=1}^{N}(1 - \frac{z_{\infty n}}{z})}\ \frac{\prod_{y=1}^{Y}(z - z_{0y})}{\prod_{y=1}^{Y}(z - \frac{1}{z_{0y}^{*}})} \end{aligned} \tag{3.135}$$

Der erste Bruch enthält nun nur Nullstellen innerhalb des Einheitskreises. Der zweite Bruch ist ein stabiler Allpass. Die Phasendrehung eines Pol-Nullstellenpaares des Allpasses kann abgelesen werden aus:

$$\begin{aligned} \frac{(e^{jwt} - z_{0y})}{(e^{jwt} - \frac{1}{z_{0y}^{*}})} &= -z_{0y}e^{-jwt}\frac{(1 - \frac{e^{jwt}}{z_{0y}})}{(1 - \frac{e^{-jwt}}{z_{0y}^{*}})} \\ &= -z_{0y}e^{-jwt}\frac{(1 - \frac{e^{jwt}}{z_{0y}})}{(1 - \frac{e^{jwt}}{z_{0y}})^{*}} \end{aligned} \tag{3.136}$$

Der letzte Bruch in (3.135) liefert wegen $|z_{0y}| > 1$ nur eine begrenzte Phasendrehung von maximalem Betrag

$$2\Phi(1 - \frac{e^{jwt}}{z_{0y}}) = 2\arctan\frac{|z_{0y}|}{|z_{0y}|^2 - 1} < \pi\ [\text{ für } |z_{0y}| = 1] \tag{3.137}$$

Die Phase des Allpasses dreht also im wesentlichen mit wt und wächst damit von 0 bis 2π, ist also nicht beschränkt und damit nicht minimalphasig.

Der erste Bruch in (3.135) hat nur Nullstellen und Pole innerhalb des Einheitskreises. Seine Phasendrehung ist begrenzt, da die Drehungen durch Pole und Nullstellen sich aufheben können, und da in jedem Fall ebenso wie in dem eben behandelten letzten Bruch der Gleichung die Phasendrehung begrenzt ist, und zwar pro Nullstelle und Pol durch

$$\Phi(1 - e^{-jwt} z_{0x}) = \arctan \frac{|z_{0x}|}{1 - |z_{0x}|^2} < \frac{\pi}{2} \text{ [für } |z_{0y}| = 1] \tag{3.138}$$

mithin für die ganzen ersten Bruch von $H(z)$ durch $M < (X + Y + N)\pi/2$. M sollte mit Hilfe der Pole und Nullstellen so gewählt werden, dass tatsächlich eine Beschränkung der Phasendrehung erreicht wird. Essentiell ist, dass diese Beschränkung *möglich* ist, während sie für einen Allpass-Term *nicht möglich* ist.

Zusammengefasst ergibt sich, dass minimalphasige Systeme keine Allpassfaktoren enthalten dürfen, denn diese führen zu unbegrenzter Phasendrehung. Nach der o.a. Herleitung ergibt sich:

Theorem 3.18 (Minimalphasigkeit).
Ein stabiles minimalphasiges System hat nur Pole und Nullstellen innerhalb des Einheitskreises.

In Umkehrung dieses Satzes werden nichtminimalphasige Systeme auch als *allpasshaltige Systeme* bezeichnet.

Minimalphasigkeit bei Systemen mit reellen Multiplizierern

Wir betrachten jetzt diskrete Systeme, wie wir sie oben in Blockschaltbildern schon gesehen haben. Treten in diesen Systemen nur reelle Multiplizierer auf, so hat die Übertragungsfunktion in Polynomen-Darstellung nur reelle Koeffizienten. Daraus folgt, dass die Übertragungsfunktion in Pol-Nullstellen-Darstellung nur relle Nullstellen und Pole, oder konjugiert komplexe Pol-Paare bzw. Nullstellen-Paare haben kann. Dies sieht man leicht aus der Polynomen-Darstellung:

$$H(z) = C \frac{\sum_{q=1}^{Q} a_q z^{-q}}{\sum_{n=1}^{N} b_n z^{-n}} \tag{3.139}$$

Betrachten wir speziell reellwertige $z = x$, so ist $H(z = x)$ bei reellwertigen Koeffizienten a_q, b_n eine reelle Funktion. Damit das auch für die Pol-Nullstellen-Darstellung gilt, müssen wir haben

$$H(z = x) = C \frac{\prod_{q=1}^{Q} (x - z_{0q})}{\prod_{n=1}^{N} (x - z_{\infty n})} \Rightarrow \text{reell} \tag{3.140}$$

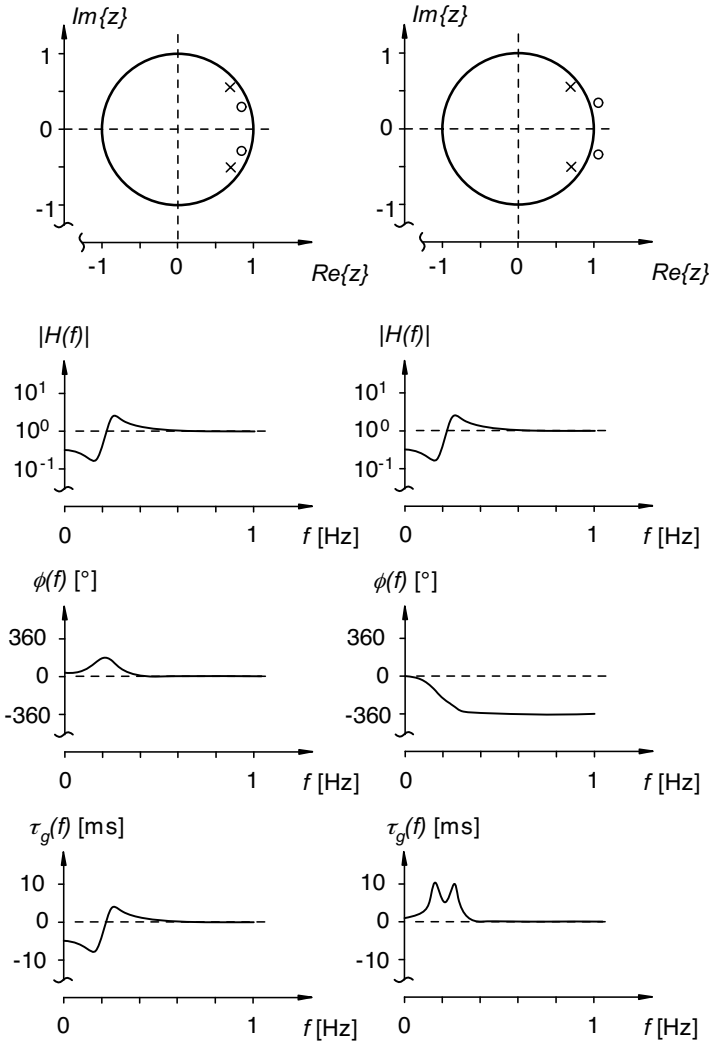


Abbildung 3.6. Links: Minimalphasiges System. Rechts: Allpasshaltiges System nach Spiegelung der Nullstellen. Bild nach (von Grünigen, 2001).

Dies geht nur, wenn Pole und Nullstellen reell sind, oder wenn sie in konjugiert komplexen Paaren auftreten, da dann für ein solches Paar (hier am Beispiel einer Nullstelle)

$$\begin{aligned}(x - z_{0q})(x - z_{0q}^*) &= (x - \mathrm{Re}z_{0q} - j\mathrm{Im}z_{0q})(x - \mathrm{Re}z_{0q} + j\mathrm{Im}z_{0q}) \\ &= (x - \mathrm{Re}z_{0q})^2 + (\mathrm{Im}z_{0q})^2 = \text{rell}\end{aligned} \tag{3.141}$$

Wir betrachten jetzt ein Beispiel für ein solches System, wobei wir den **Frequenzgang** untersuchen. Dies ist die Übertragungsfunktion an der Stelle $z = e^{jwt}$.

Beispiel 3.16 – Nicht-Minimalphasigkeit durch Spiegelung der Nullstellen.

Das System in Abb. 3.6 mit Polen und Nullstellen im Einheitskreis ist minimalphasig. Nach Spiegelung, d.h. Betragsinvertierung und Konjugation, der Nullstellen am Einheitskreis entsteht daraus ein allpasshaltiges System mit -bis auf einen konstanten Faktor- gleichem Betrag der Übertragungsfunktion, aber nichtminimalphasigem Phasengang, der in der Tat von 0 bis 2π dreht.
In der Übung 3.15 wird dieses Verhalten explizit berechnet. □

Für Systeme mit reellen Multiplizierern haben wir also gesehen, dass Minimalphasigkeit erreicht werden kann, wenn eventuell ausserhalb des Einheitskreises vorhandene Nullstellenpaare am Einheitskreis gespiegelt werden. Der Betrag des Frequenzganges bleibt dabei bis auf einen Faktor $|z_0|^2$ für jedes gespiegelte Nullstellenpaar unverändert.

Übungen

Übung 3.1 – Systemeigenschaften.

Sind folgende Systeme LTI-Systeme?

- $y(t) = x(t) + b$
- $y(t) = x(t)\ m(t)$

Übung 3.2 – Diskrete Faltung.

- Berechnen und skizzieren Sie die lineare Faltung für

$$x[n] = \{\ldots, 0, \underline{2}, 7, -5, 3, 4, 0, \ldots\}$$

und

$$h[n] = \{\ldots, 0, \underline{2}, -5, 4, 1, 0, \ldots\}$$

- Nehmen Sie an, die Signale für $n = 0, 1, 2, 3$ würden periodisch fortgeführt. Berechnen Sie die periodische Faltung.

Übung 3.3 – Z-Transformation und ROC.

Berechnen Sie die z-Transformation $Z\{x[n]\} = X(z) = \sum_{n=-\infty}^{\infty} x_n z^{-n}$ der Signale

- $x[n] = \begin{cases} a^n & n \geq 0 \\ 0 & \text{sonst} \end{cases}$
- $x[n] = \begin{cases} -a^n & n \leq -1 \\ 0 & \text{sonst} \end{cases}$

Vergleichen Sie die Ergebnisse!

Übung 3.4 – Z-Transformation und Faltungssatz.

Zeigen Sie, daß für die z-Transformation
$Z\{x[n]\} = X(z) = \sum_{n=-\infty}^{\infty} x_n z^{-n}$

- der Faltungssatz $Z\{x(n) \star h(n)\} = X(z) \cdot H(z)$,
- und für die Ableitung der z-Transformierten $-z\frac{dF(z)}{dz} = Z\{n \cdot f[n]\}$ gilt.

Übung 3.5 – Kausalität, Z-Transformation, Pol-Nullstellen.

Gegeben sei das Signal $x[n] = \begin{cases} 1/4 & 0 < n < 5 \\ 0 & \text{sonst} \end{cases}$. Ist das Signal kausal?
Bestimmen Sie die z-Transformierte $X(z)$ des Signals. Wie lauten die Pol- und Nullstellen von $X(z)$?

Übung 3.6 – Partialbruchzerlegung, inverse Z-Transformation.

Führen Sie die Partialbruchzerlegung von $F(z) = \frac{z}{(z+0.5)(z-1)}$ durch! Bestimmen Sie damit und mit Hilfe der Relation $X(z) = \frac{1}{(z-z_0)}, \quad |z| > |z_0| \quad \Leftrightarrow \quad x[n] = u[n-1]z_0^{n-1}$ die inverse z-Transformierte von $F(z)$.

Übung 3.7 – Inverse Z-Transformation, Residuensatz.

Berechnen Sie die inverse z-Transformierte der Funktion
$F(z) = \frac{1}{z^3(2z-1)}$.

a) Benutzen Sie den Residuensatz!
Führen Sie außerdem für den Fall $n < 4$ die Inversion im $1/z$-Bereich (mit Residuensatz) durch!

b) Benutzen Sie Eigenschaften der z-Transformation und die Relation $X(z) = \frac{1}{(z-z_\infty)}, \quad |z| > |z_\infty| \quad \Leftrightarrow \quad x[n] = u[n-1]z_\infty^{n-1}$
($u[n]$ ist die Sprungfolge).

Übung 3.8 – Partialbruchzerlegung bei mehrfacher Polstelle.

Berechnen Sie die inverse z-Transformierte der Funktion $F(z) = \frac{z}{(z+0.5)^2(z-1)}$. Benutzen Sie Partialbruchzerlegung und die Beziehung

$$X(z) = \frac{1}{(z-z_0)^k} \quad \Leftrightarrow \quad x[n] = u(n-k)\binom{n-1}{k-1} z_0^{n-k}, \qquad |z| > |z_0|.$$

Übung 3.9 – Übergang Differentialgleichung zu Differenzengleichung.

Im Bereich der zeitkontinuierlichen Systeme sind die Differentialgleichungen (DGL) eine verbreitete Form, Systemverhalten mathematisch zu erfassen. In welcher Hinsicht stellen Differenzengleichungen ein vergleichbares Äquivalent zu den DGLs dar? Betrachten Sie den Übergang vom Differentialquotienten zum Differenzenquotienten. Diskutieren Sie Parallelen und Unterschiede.

Übung 3.10 – Pol-Nullstellen, Differenzengleichung.

Das PN-Schema der Übertragungsfunktion zeigt eine Nullstelle bei $z = 0$ und Polstellen bei $z = 3/4 \pm j/2$.
Berechnen Sie $H(z)$!
Geben Sie die Differenzengleichung für das System und eine Realisierungsschaltung an!

Übung 3.11 – Stabilität.

Ist das System mit der Diff-Gl. $y[n] = ay[n-1] + 2\delta[n]$ stabil? Überlegen Sie sich die Lösung sowohl im Zeit- als auch im Z-Bereich.

Übung 3.12 – Kausalität.

Gegeben ist die Übertragungsfunktion

$$H(z) = \frac{1}{(z-1)(z-3)}.$$

Untersuchen Sie durch Entwicklung in eine Laurentreihe für die beiden Bereiche $1 < |z| < 3$ und für $|z| > 3$ ob das System kausal ist!
Untersuchen Sie die Stabilität durch Betrachtung der PN-Darstellung des Systems.

Übung 3.13 – Verletzung der Kausalität in Pol-Nullstellen-Darstellung.

Zeigen Sie, dass der Satz über die Kausalität in Pol-Nullstellen-Darstellung (Theorem 3.15 auf Seite 72) schon dann verletzt ist, wenn nur für eine (!) Polstelle p gilt

$$|z| < |z_{\infty_p}|$$

Dazu entwickeln Sie die Teilübertragungsfunktion H_a für $a = z_{\infty_p}$ in ihre entsprechende geometrische Reihe und zeigen, dass dann auf keinen Fall (für beliebige Wahl von N und Q) die Gesamtübertragungsfunktion nur aus dem Hauptteil bestehen kann.

Insbesondere gilt dann auch, dass jede verschobene Übertragungsfunktion (siehe Beispiel 3.15 auf Seite 70) ebenfalls nicht kausal sein kann. Die Verschiebung der Übertragungsfunktion um m Zeitschritte entsteht ja durch Multiplizieren von m weiteren Nullstellentermen $(z - z_{0q})$ mit $z_{0q} = 0$, ist also in Ihrer obigen Überlegung bereits enthalten.

Übung 3.14 – Allpass.

Die Übertragungsfunktion $H(z)$ eines Systems habe genau zwei Polstellen bei $z_{\infty 1} = 1/2$ und bei $z_{\infty 2} = 1+j$. Wählen Sie die Nullstellen so, daß das System einen Allpass darstellt. Berechnen Sie dann dieses $H(z)$, und überprüfen Sie die Allpass-Relationen der Koeffizienten der Zähler- und Nennerpolynome.

Übung 3.15 – Rechnung für Nichtminimalphasigkeit durch Spiegelung der Nullstellen.

Zeigen Sie durch Rechnung für das (konjugiert komplexe!) Nullstellenpaar (z_0, z_0*), dass in dem Beispiel 3.16 auf Seite 80 tatsächlich der Betrag des Frequenzganges bis auf einen konstanten Faktor $|z_0|^2$ gleichbleibt, aber die Übertragungsfunktion einen nichtminimalphasigen Phasengang aufweist, der in der Tat von 0 bis 2π dreht. Betrachten Sie dazu die Übertragungsfunktion eines Nullstellenpaares bei $z = e^{jwt}$.

Übung 3.16 – Eineindeutigkeit der Allpassbedingung.

Überzeugen Sie sich, dass die in Abschnitt 3.3.3 auf Seite 75 hergeleiteten Bedingungen

$$\frac{|z_{0q}|}{1+|z_{0q}|^2} = \frac{|z_{\infty q}|}{1+|z_{\infty q}|^2}$$

und

$$\Phi_{0q} = \Phi_{\infty q}$$

für den Zusammenhang zwischen Polen und Nullstellen der Übertragungsfunktion eines Allpasses tatsächlich von

$$z_{\infty q} = \frac{1}{z_{0q}^*}$$

und

$$z_{\infty q} = z_{0q}$$

erfüllt werden.

Übung 3.17 – Impulsantwort.

Die Sprungantwort eines zeitdiskreten Systems lautet
$a[n] = u[n](4 - 0.25^n)$.
Berechnen Sie die Impulsantwort des Systemes im Zeitbereich (d.h. direkt und ohne z-Transformation).

4

Signalverarbeitung mit zeitdiskreten Systemen

Im Bereich der Signalverarbeitung spielen analoge Systeme, wie wir sie im Kapitel 2 beschrieben haben, nach wie vor eine große Rolle. Immer häufiger werden jedoch zeitdiskrete Systeme (in Form von digitalen Systemen) verwendet, da diese gegenüber den analogen Schaltungen eine Reihe von Vorteilen aufweisen: sie sind genauer, liefern störunempfindliche und reproduzierbare Ergebnisse, lassen sich leichter entwickeln und werden bereits seit längerem von einer immensen Menge an Theorie, Bauelementen und Entwicklungssystemen begleitet. Dabei ist die digitale oder zumindest die zeitdiskrete Variante der Signalverarbeitung alles andere als naheliegend. Dem entsprechend ist diese Technik relativ „neu“ im Vergleich zur gesamten Entwicklung der technischen Signalverarbeitung.

Das Einsatzgebiet von Signalverarbeitungssystemen ist jedoch nach wie vor analog geblieben, da unsere Umwelt hauptsächlich aus kontinuierlichen analogen Signalen besteht. Wie kann jedoch ein digitales System die selben Aufgaben erfüllen, wie beispielsweise ein aktiver analoger Filter? Ermöglicht wurde dies dadurch, dass folgende Problemstellungen gelöst wurden:

- Ein analoges Signal kann ohne Verlust relevanter Information in ein digitales Signal umgewandelt werden.
- Das digitale Signal kann so verarbeitet werden, dass eine beliebige Übertragungsfunktion des Gesamtsystems entsteht.
- Aus jedem digitalen Signal kann wieder ein analoges Signal erzeugt werden.

In dieser Reihenfolge wollen wir auch die zeitdiskrete Verarbeitung von analogen Signalen erläutern. Unser Ausgangspunkt sei also ein kontinuierliches analoges Signal, das zur Verarbeitung mit einem zeitdiskreten System abgetastet, quantisiert und codiert wird. Anstelle von analogen Signalen werden nunmehr diskrete Zahlenfolgen verarbeitet, die bereits im ersten Kapitel dieses Buches erwähnt wurden. Wir werden sehen, dass auch im zeitdiskreten Bereich eine Signalverarbeitung möglich ist, die zu einer beliebigen Übertragungsfunktion des Gesamtsystems führt. Anschließend wird die verarbeitete Zahlenfolge wieder in ein analoges Signal umgewandelt – ein Schritt, den wir

„Rekonstruktion“ nennen. Damit ist das System komplett, um analoge Signale in beliebiger Weise zu verarbeiten.

Nicht jedes zeitdiskrete System benötigt all diese Komponenten, denn in manchen Fällen sind analoge Eingangssignale nicht nötig bzw. auch die Ausgabe analoger Signale ist nicht immer erforderlich. Trotzdem wird in diesem Kapitel exemplarisch der gesamte Signalpfad Schritt für Schritt durchlaufen, so dass der Leser eine systematische Orientierung erhält. Wir werden auf die mathematischen Beschreibungsmöglichkeiten der Einzelkomponenten eingehen, Fehlerbetrachtungen anstellen und einen Überblick über verschiedene Varianten geben, die in der heutigen Digitaltechnik von Bedeutung sind.

4.1 Grundlegende Begriffe und Zielstellungen

Zeitdiskrete Signale und Systeme sind ein rein theoretisches Konstrukt. Natürlich arbeiten auch zeitdiskrete Systeme mit kontinuierlichen Signalen auf kontinuierlichen Signalträgern wie beispielsweise Spannungen oder Strömen.

Der Unterschied besteht allein in der Interpretation dieser Signale. Während der Informationsgehalt eines analogen Signals innerhalb eines endlichen Zeitausschnittes theoretisch unendlich groß ist, wird ein diskretes Signal künstlich quantisiert, und zwar

1. durch diskrete zeitliche Schritte („Abtastung“) und
2. durch diskrete Amplitudenschritte („Quantisierung“)

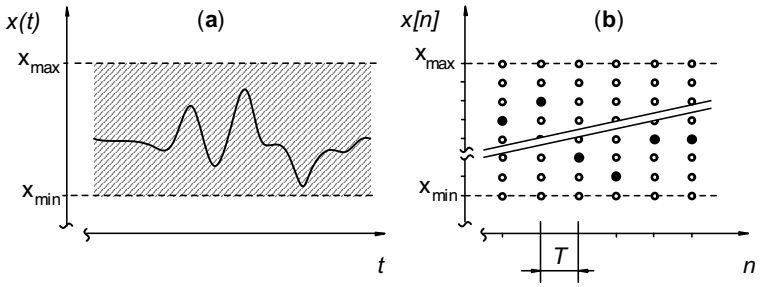
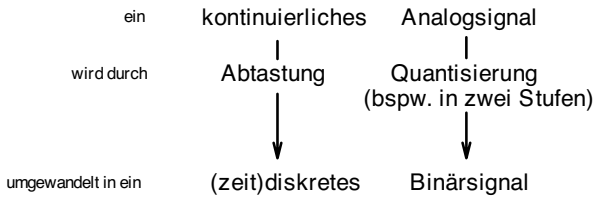


Abbildung 4.1. (**a**) Arbeitsbereich eines analogen Systems (schraffiert) und Beispiel-Signal (**b**) Arbeitsbereich eines diskreten Systems (Kreise) mit Beispiel-Signal (ausgefüllte Kreise)

Abbildung 4.1 soll dies verdeutlichen. Ein analoges Signal kann zu jedem Zeitpunkt t innerhalb der Grenzen x_{min} und x_{max} jeden beliebigen Wert einnehmen. Bei einem analogen Signalträger, der ein zeitdiskretes Signal überträgt, ist dies natürlich genauso. Jedoch wird dieser Signalträger im Bereich zeitdiskreter Signale dahingehend interpretiert, dass er innerhalb eines Abtastintervalls T jeweils nur genau einen Wert liefert. Dieser Wert ist außerdem – im

Gegensatz zu analogen Signalen – beschränkt auf einen abzählbaren und endlichen Wertebereich. Weit verbreitet sind beispielsweise die sog. Binär-Signale, deren Wertebereich auf zwei mögliche Werte beschränkt ist.

Wichtig für das weitere Verständnis ist die begriffliche Unterscheidung der zeitlichen Quantisierung und der Amplitudenquantisierung. Halten wir also zunächst noch einmal fest:



Ein (zeit)diskretes Signal ist also ein zeitlich quantisiertes (abgetastetes) Signal, während das Besondere an einem Binär-Signal dessen Amplitudenquantisierung (in diesem Falle mit zwei Quantisierungsstufen) ist. Die Vorteile dieser Quantisierungen sind die verbesserte Störsicherheit sowie eine vereinfachte Signalverarbeitung.

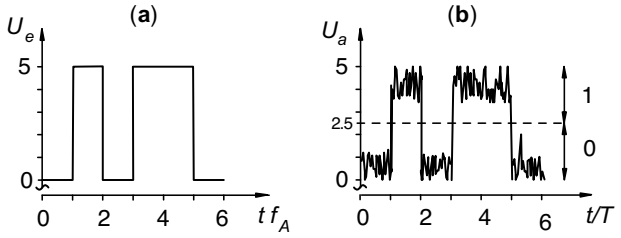


Abbildung 4.2. (**a**) Binärsignal „010110“ (**b**) Überlagerung mit Rauschen

Verbesserte Störsicherheit

In technischen Systemen sind Rauschprozesse allgegenwärtig. Problematisch wird dies, wenn Information möglichst unverfälscht über räumliche oder zeitliche Distanzen übertragen werden soll. Bei analoger Signalverarbeitung überlagert sich in jeder Systemkomponente die Nutzinformation mit unvermeidlichem Rauschen, was durch erhöhten technischen Aufwand begrenzt, jedoch nie ganz verhindert werden kann.

Handelt es sich hingegen um Zeit- und amplitudenquantisierten Signale, sind Rauscheinflüsse weniger störend. Am Beispiel von zeitdiskreten Binär-Signalen soll dies kurz veranschaulicht werden. Binäre Signale liefern pro Zeittakt genau einen von zwei möglichen Werten, die wir hier mit „0“ bzw.

„1“ bezeichnen wollen. Eine „0“ soll durch eine Spannung von 0V und eine „1“ durch eine Spannung von 5V repräsentiert werden. Abb. 4.2 zeigt, wie ein solches Signal zwar durch Rauschen beeinflusst wird, dies jedoch keinerlei Einfluß auf den Informationsgehalt des Signals hat, wenn man lediglich betrachtet, ob der Spannungswert im Intervall oberhalb bzw. unterhalb einer Schwelle (hier 2,5V) bleibt.

Vereinfachte Signalverarbeitung

Um analoge Signale auf arithmetische Weise zu verarbeiten, können Operationsverstärker-Schaltungen verwendet werden. Damit können Ströme oder Spannungen beispielsweise addiert, subtrahiert, multipliziert oder integriert werden. Der Aufwand, den ein solches analoges Rechenwerk erfordert, ist allerdings trotz fortschreitender Integrationsdichte enorm, denn jeder einzelne Operationsverstärker erfordert eine größere Zahl von Transistoren. Zusätzliche Probleme bereitet obendrein die Verlustleistung großer Transistor-Anordnungen, das eben bereits erwähnte Rauschen in solchen Schaltungen und die Frage, wie beispielsweise Werte über längere Zeit gespeichert werden können.

Nun muss jedoch nicht unbedingt die Spannung oder die Stromstärke direkt verwendet werden, um einen arithmetischen Wert zu codieren. Dies entspräche dem Versuch, Zahlen durch Striche unterschiedlicher Länge zu notieren. Wie aus der Geometrie bekannt, können damit auch arithmetische Berechnungen angestellt werden, jedoch sind die Grenzen gegenüber unserer heutigen Zahlendarstellung anhand der zehn verschiedenen Ziffernsymbole offensichtlich. In ähnlicher Weise können Zahlen auch dergestalt durch einen Signalträger codiert werden, dass sich arithmetische Operationen auf einfache Weise mit technischen Systemen erledigen lassen.

Die wesentliche Idee für die heutige Digitaltechnik erkannte bereits im Jahre 1605 – sicher nicht als erster – der damals 17-jährige Sir Francis Bacon (Bacon, 1605, 1640) mit seinem Code „omnia per omnia“. Zwar ging es ihm zunächst um Steganographie, also das Verbergen von geheimer Information in Texten, aber das damals verwendete Prinzip ist mit dem heutigen identisch: in einem Code, der aus einer Sequenz von zwei verschiedenen Symbolen besteht, kann beliebige Information gespeichert werden. Diese zwei Symbole lassen sich verhältnismäßig leicht mit allen Signalträgern codieren, beispielsweise durch positive und negative elektrische Spannung, durch Töne mit zwei unterschiedlichen Tonhöhen oder durch schwarzen und weißen Rauch.

Es könnten auch drei oder zehn Zustände sein, jedoch ist die Verwendung dieses binären Zustandsraumes die einfachste Variante, da beide Zustände lediglich durch eine einzige Entscheidungsschwelle voneinander getrennt werden können. Wie in Abb. 4.2 bereits deutlich wurde, lässt sich damit außerdem eine hohe Störsicherheit gegen Rauscheinflüsse erreichen. Der Inhalt einer beliebigen Information kann nun auf unterschiedliche Weise durch eine (zeitliche oder räumliche) Sequenz mehrerer dieser Binärsymbole codiert werden. Im Kapitel 4.3 werden wir auf übliche binäre Codierungsverfahren genauer eingehen.

4.2 Abtastung

Diskrete Signale sind dadurch gekennzeichnet, dass sie uns in einem bestimmten Raster, dem Abtastintervall T, jeweils nur genau einen Wert liefern. Üblicherweise sind diese Intervalle von gleicher Dauer, so dass wir von einer periodischen Abtastung mit der Abtastfrequenz f_A sprechen, wobei:

$$f_A = \frac{1}{T} \tag{4.1}$$

Von theoretischer Seite betrachtet reicht somit die Angabe der Abtastfrequenz, die Folge der Werte für alle Abtastintervalle sowie ein definierter Startzeitpunkt, um ein zeitdiskretes Signal komplett zu beschreiben. Wir verwenden hier die bereits im ersten Kapitel eingeführte Notation in Form einer Folge, bei der wir, falls erforderlich, das zu $t = 0$ gehörige Element durch Unterstreichen markieren:

$$x[n] = \{x_a, x_{a+1}, \ldots, \underline{x_0}, \ldots, x_{b-1}, x_b\} = \{x_n\}_{n=a}^{b} \tag{4.2}$$

Einen einzelnen Abtastwert – also ein Element dieser Folge – notieren wir unter Verwendung von eckigen Klammern oder Indices:

$$n \Rightarrow x[n] = x_n \tag{4.3}$$

Wie mag dies nun in einem konkreten System aussehen? Von einem kontinuierlichen Signal benötigen wir je Abtastintervall genau einen Wert. Dazu könnten wir beispielsweise die über das Intervall oder einen kürzeren Zeitraum gemittelte Amplitude des Signals heranziehen. Weiter verbreitet sind jedoch die Verfahren Track&Hold bzw. Sample&Hold. Beiden ist gemeinsam, dass das analoge Eingangssignal in jedem Abtastintervall lediglich zu einem definierten Zeitpunkt gemessen wird. Die restlichen Werte des Signals werden damit vollständig ignoriert. Dieser Analogwert wird anschließend für die Dauer des Abtastintervalls (Sample&Hold) bzw. für einen Teil des Intervalls (Track&Hold) durch ein analoges Speicherelement (i.Allg. ein Kondensator) festgehalten. Das Ausgangssignal eines Sample&Hold-Gliedes hat somit beispielsweise den in 4.3(a) idealisiert dargestellten treppenförmigen Verlauf. Die Kreise markieren in dieser Darstellung die Abtastzeitpunkte und -werte.

Das Halten des Signals ist zwar eine technische Notwendigkeit, jedoch können wir die theoretischen Betrachtungen vereinfachen, wenn wir das kontinuierliche Signal ausschließlich zu den Abtastzeitpunkten betrachten.

4.2.1 Mathematische Beschreibung des Abtastprozesses

Eine geeignete mathematische Beschreibung für den Abtastprozeß ist eine Folge von gewichteten Delta-Impulsen. Diese sind nur im Sinne von Distributionen definiert. Ihre Transformation in den Frequenzbereich ist jedoch

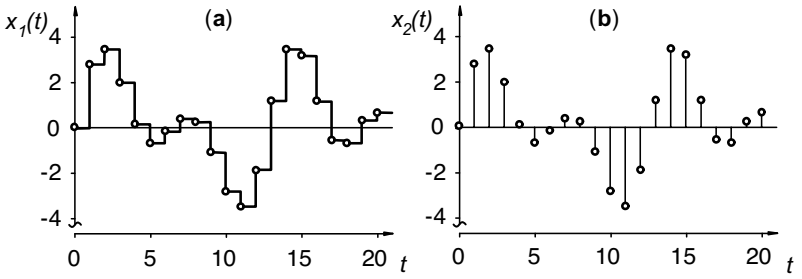


Abbildung 4.3. (**a**) Treppenförmiger Signalverlauf durch Sample&Hold (**b**) Zeitdiskretes Signal in Form einer symbolisch dargestellten Dirac-Impulsfolge

einfach und nützlich. Diese Transformation geschieht mit Hilfe der Fourier-Transformation. Wir betrachten daher zunächst einige Eigenschaften der Fourier-Transformation und der Delta-Funktion, mit deren Hilfe wir den Abtastprozeß anschliessend beschreiben können.

Eigenschaften der Fourier-Transformation

Die Fourier-Transformation hat folgende Symmetrieeigenschaft für negative Frequenzen. Sei $F(x,\omega)$ die Fouriertransformierte von $x(t)$ an der Stelle ω. Dann gilt

$$\begin{aligned} F(x,\omega) &= \int_{t=-\infty}^{\infty} x(t)e^{-j\omega t}\mathrm{d}t \\ &= \int_{t=-\infty}^{\infty} \left[x^*(t)e^{-j(-\omega)t}\right]^* \mathrm{d}t \\ &= \left[\int_{t=-\infty}^{\infty} x^*(t)e^{-j(-\omega)t}\mathrm{d}t\right]^* = [F(x^*,-\omega)]^* \end{aligned} \tag{4.4}$$

Mit Zeitumkehrung gilt auch

$$\begin{aligned} F(x(t),\omega) &= \int_{t=-\infty}^{\infty} x(t)e^{-j\omega t}\mathrm{d}t \\ &= \int_{t=-\infty}^{\infty} x(-t)e^{-j(-\omega)t}\mathrm{d}t = F(x(-t),-\omega) \end{aligned} \tag{4.5}$$

Anwendungen der Gl. (4.4) werden in Beispiel 4.1 und in den Übungen 4.1 und 4.2 diskutiert. Ist also $x(t)$ ein Signal mit zeitlich konstanter Phase oder mit Zeitumkehrinvarianz $x(t) = x(-t)$, so gilt speziell:
$|X(\omega)| = |F(x,\omega)| = |F(x,-\omega)| = |X(-\omega)|$.
Der Betrag der Fouriertransformierten eines solchen Signals ist also eine gerade Funktion in ω.

Beispiel 4.1 – Konjugiert gerade Fouriertransformierte.

Wir zeigen, dass die Fouriertransformierte eines reellen Signals eine konjugiert gerade Funktion von ω ist
a) unter Benutzung von Gl. (4.4)
b) durch Benutzen der Definition der Fouriertransformierten
c) Weiterhin zeigen wir an einem geeigneten Beispiel, dass für Signale mit zeitlich variabler Phase und ohne Zeitumkehrinvarianz im Allgemeinen keine Symmetrieeigenschaften gelten.

Lösung:
a) Da für reelle Signale $x^* = x$, gilt unter Benutzung von Gl. (4.4):
$X(\omega) = F(x, \omega) = [F(x^*, -\omega)]^* = [F(x, -\omega)]^* = X^*(-\omega)$.
b) Für reelle Signale gilt unter Benutzung der Definition der Fouriertransformierten

$$\begin{aligned} X(\omega) &= \int_{t=-\infty}^{\infty} x(t) e^{-j\omega t} \mathrm{d}t \\ &= \int_{t=-\infty}^{\infty} x(t) \cos(\omega t) \mathrm{d}t + j \int_{t=-\infty}^{\infty} x(t) \sin(\omega t) \mathrm{d}t \\ &= \int_{t=-\infty}^{\infty} x(t) \cos(-\omega t) \mathrm{d}t - j \int_{t=-\infty}^{\infty} x(t) \sin(-\omega t) \mathrm{d}t \\ &= X^*(-\omega) \end{aligned}$$

c) Als Beispiel für Signale mit zeitlich variabler Phase und ohne Zeitumkehrinvarianz verwenden wir $x(t) = \exp(-j\omega_0 t)$. Die Fouriertransformierte lautet $X(\omega) = \delta(\omega + \omega_0)$. Sie weist nur einen Anteil bei $(-\omega_0)$ und damit keine Symmetrieeigenschaften auf. □

Betrachtungen im analogen Bereich

Im *analogen* Bereich führen wir die Abtastung mit einer Delta-Folge ein, wofür wir wieder mit Distributionen arbeiten müssen. Die Abtastung des analogen Signals x(t) lautet dann:

$$x_n(t) = x(t) \sum_{n=-\infty}^{\infty} \delta(t - nT) \tag{4.6}$$

So entsteht beispielsweise der in Abb. 4.3(b) gezeigte Signalverlauf, wobei die Delta-Impulse nur schematisch gezeichnet werden können, da sie ja prinzipiell eine unendlich hohe Amplitude besitzen. Stattdessen symbolisiere die Länge der eingezeichneten Linie die Fläche unter dem gewichteten Delta-Impuls. Eine Linie der Länge eins symbolisiert also $\delta(t)$ und eine Linie der Länge a dementsprechend $a\delta(t)$.

Wie sieht das Spektrum des abgetasteten Signals aus? Dazu berechnen wir zunächst das Spektrum der reinen Delta-Folge.

Delta-Folge im Frequenzbereich

Wir betrachten die unendliche Folge von (analogen) Dirac-Pulsen

$$x(t) = \sum_{m=-\infty}^{\infty} \delta(t - mT) \tag{4.7}$$

Die Fouriertransformierte dieser Funktion ist zunächst

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt = \sum_{m=-\infty}^{\infty} e^{-j\omega mT} \tag{4.8}$$

Für jedes $\omega \neq k(2\pi/T)$ mit ganzzahligem k ist diese Summe 0, da sich alle Werte auf dem Einheitskreis zu Null addieren. Für $\omega = k \cdot (2\pi/T)$ hingegen wird der Betrag der Summe unendlich groß. Die Summe muss also den Wert haben:

$$X(\omega) = c \sum_{m=-\infty}^{\infty} \delta(\omega - m\Omega) \tag{4.9}$$

mit einer noch zu bestimmenden Konstanten c und der Abtast-Kreisfrequenz $\Omega = 2\pi/T$. Um den Wert der Konstanten c zu bestimmen, wenden wir die inverse Fourier-Transformation an und erhalten:

$$x(t) = \sum_{m=-\infty}^{\infty} \delta(t-mT) = \frac{c}{2\pi} \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} e^{j\omega t} \delta(\omega - m\Omega) d\omega = \frac{c}{2\pi} \sum_{m=-\infty}^{\infty} e^{jm\Omega t} \tag{4.10}$$

Anschließend integrieren wir über eine Breite von T:

$$\int_{-T/2}^{T/2} x(t) dt = 1 = \frac{c}{2\pi} \sum_{m=-\infty}^{\infty} \int_{-T/2}^{T/2} e^{jm\Omega t} dt = \frac{c}{2\pi} \sum_{m=-\infty}^{\infty} T\delta(m) = \frac{cT}{2\pi} = \frac{c}{\Omega} \tag{4.11}$$

und erhalten also letztlich $c = \Omega$.

Wir können aus diesen Herleitungen folgende drei wichtigen Gleichungen extrahieren:

Theorem 4.1 (Unendliche Summen von Distributionen).

$$\sum_{m=-\infty}^{\infty} \delta(t - mT) = \frac{1}{T} \sum_{m=-\infty}^{\infty} e^{jm\Omega t} \tag{4.12}$$

$$\sum_{m=-\infty}^{\infty} \delta(\omega - m\Omega) = \frac{1}{\Omega} \sum_{m=-\infty}^{\infty} e^{jm\omega T} \tag{4.13}$$

$$FT\left\{ \sum_{m=-\infty}^{\infty} \delta(t - mT) \right\} = \Omega \sum_{m=-\infty}^{\infty} \delta(\omega - m\Omega) \tag{4.14}$$

Abgetastetes Signal im Frequenzbereich

Ausgehend von diesen Ergebnissen, können wir nun das Spektrum eines abgetasteten Signals $x(t)$ ermitteln. Es gilt:

Theorem 4.2 (Periodisches Spektrum eines abgetasteten Signals). *Das Spektrum eines mit der Periode T (entspricht der Abtastfrequenz f_A) abgetasteten Signals ist die Überlagerung von um $\Omega = 2\pi/T = 2\pi f_A$ gegeneinander versetzter, periodisch fortgesetzter und normierter Spektren des zugehörigen kontinuierlichen Signals:*

$$X_n(\omega) = \frac{1}{T} \sum_{n=-\infty}^{\infty} X(\omega - n\Omega) \tag{4.15}$$

Zum Beweis transformieren wir beide Seiten der Gleichung (4.6) mittels Fouriertransformation, wodurch auf der rechten Seite eine Faltung entsteht:

$$\begin{aligned} X_n(\omega) &= X(\omega) * \frac{1}{T} \sum_{n=-\infty}^{\infty} \delta(\omega - n\Omega) \\ &= \int_{v=-\infty}^{\infty} X(v) \frac{1}{T} \sum_{n=-\infty}^{\infty} \delta(\omega - n\Omega - v) dv \\ &= \frac{1}{T} \sum_{n=-\infty}^{\infty} X(\omega - n\Omega) \end{aligned} \tag{4.16}$$

□

Betrachtungen im diskreten Bereich

Wir haben den Abtastprozeß bisher im analogen Bereich betrachtet. Sehen wir uns die Fouriertransformierte des abgetasteten Signals an

$$X_n(\omega) = \sum_{n=-\infty}^{\infty} x(nT) e^{-j\omega nT},$$

so können wir für das analoge Signal $x(nT)$ das entsprechende diskrete Signal $x[n]$ einsetzen (es wird ja nur über ganzzahlige n summiert) und sind somit im diskreten Bereich. Dann erhalten wir

$$X_n(\omega) = \sum_{n=-\infty}^{\infty} x[n] e^{-j\omega nT} \tag{4.17}$$

Der rechte Ausdruck von (4.17) stellt eine Fourierreihe mit Koeffizienten $x[n]$ dar. Offensichtlich ist diese Fouriertransformierte periodisch, denn es gilt für alle ganzzahligen k:

$$X_n(\omega) = X_n(\omega + k2\pi/T) = X_n(\omega + k\Omega) \tag{4.18}$$

Wir suchen jetzt nach einer Umkehrtransformation, mit der wir aus $X_n(\omega)$ direkt wieder $x[n]$ gewinnen, d.h. wir möchten im diskreten Bereich bleiben. Da $X_n(\omega)$ periodisch ist, ist diese Umkehrtransformation die inverse Fouriertransformation einer periodischen Fourierreihe. Wir können aber auch unsere Kenntnis nutzen, dass der Frequenzgang die z-Transformierte auf dem Einheitskreis ist. Mithin erhalten wir die Folge $x[n]$ als inverse z-Transformierte auf dem Einheitskreis. Es muss also gelten

$$x[n] = \frac{1}{2\pi j} \oint z^{n-1} X_n(z) dz \quad \text{mit} \quad z = e^{-j\omega T}, dz = -jTz d\omega \tag{4.19}$$

Daraus erhalten wir

$$x[n] = \frac{T}{2\pi} \int_{\omega=-\pi/T}^{\pi/T} e^{jn\omega T} X_n(\omega) d\omega \tag{4.20}$$

bzw.

$$x[n] = \frac{1}{2\pi} \int_{\Phi=-\pi}^{\pi} e^{jn\Phi} X_n(\Phi)\ d\Phi \quad \text{mit} \quad \Phi = \omega T = \omega / f_A \tag{4.21}$$

Dies ist in der Tat die inverse Fouriertransformation für eine periodische Funktion. Wir überzeugen uns durch Einsetzen von $X_n(\Phi)$, dass dies wirklich die gesuchte Umkehrtransformation ist:

$$\begin{aligned} x[n] &= \frac{1}{2\pi} \int_{\Phi=-\pi}^{\pi} e^{jn\Phi} X_n(\Phi)\ d\Phi \\ &= \frac{1}{2\pi} \int_{\Phi=-\pi}^{\pi} \sum_{k=-\infty}^{\infty} x[k] e^{j(n-k)\Phi}\ d\Phi \\ &= \sum_{k=-\infty}^{\infty} x[k] \frac{1}{2\pi} \int_{\Phi=-\pi}^{\pi} e^{j(n-k)\Phi}\ d\Phi \\ &= \sum_{k=-\infty}^{\infty} x[k] \delta(n-k) = x[n] \end{aligned} \tag{4.22}$$

Damit haben wir ein Transformationspaar im diskreten Bereich. Wir wiederholen noch einmal: Aus der Abtastgleichung (4.6)

$$x_n(t) = x(t) \sum_{n=-\infty}^{\infty} \delta(x - nT)$$

folgt die zeitdiskrete, frequenzkontinuierliche Fouriertransformation (**T**ime **D**iscrete **F**ourier **T**ransformation):

Theorem 4.3 (Zeitdiskrete, frequenzkontinuierliche Fouriertransformation (TDFT)).

$$X_n(\omega) = \sum_{n=-\infty}^{\infty} x[n] e^{-j\omega nT} \tag{4.23}$$

$$x[n] = \frac{T}{2\pi} \int_{\omega=-\pi/T}^{\pi/T} e^{jn\omega T} X_n(\omega) d\omega \tag{4.24}$$

Dieses können wir nun nutzen, um auf einem zweiten Weg das Spektrum ohne Benutzen von Distributionen herzuleiten.

Die Funktion im Zeitbereich ergibt sich allgemein aus dem Frequenzgang

$$x(t) = \frac{1}{2\pi} \int_{\omega=-\infty}^{\infty} X(\omega) e^{j\omega t} d\omega \tag{4.25}$$

Daraus erhalten wir bei Zeitpunkten $t = kT$ die Beziehung

$$\begin{aligned} x[k] = x(t = kT) &= \frac{1}{2\pi} \int_{\omega=-\infty}^{\infty} X(\omega) e^{j\omega kT} d\omega \\ &= \sum_{n=-\infty}^{\infty} \frac{1}{2\pi} \int_{\omega=(-1+2n)\pi/T}^{(1+2n)\pi/T} X(\omega) e^{j\omega kT} d\omega \\ &= \frac{T}{2\pi} \int_{\omega=-\pi/T}^{\pi/T} \left[\frac{1}{T} \sum_{n=-\infty}^{\infty} X(\omega + \frac{2\pi}{T} n) \right] e^{j\omega kT} d\omega \\ &= \frac{1}{2\pi} \int_{\Phi=-\pi}^{\pi} \left[\frac{1}{T} \sum_{n=-\infty}^{\infty} X((\Phi + 2\pi n)/T) \right] e^{j\Phi k} d\Phi \end{aligned} \tag{4.26}$$

Durch Vergleich mit den oben angegebenen Koeffizienten der zu diskreten Systemen gehörigen Fourierreihe (4.24) ergibt sich im Frequenzbereich

$$X_n(\omega) = \frac{1}{T} \sum_{n=-\infty}^{\infty} X(\omega + n\Omega) \tag{4.27}$$

also dasselbe Resultat wie oben im Theorem 4.2 unter Benutzung von Distributionen.

Frequenzgang eines diskreten Systems

Wie sieht die Übertragungsfunktion eines Systems im Frequenzbereich aus, dass Signale nur zu diskreten Zeitpunkten verarbeitet? Dazu verwenden wir die bereits erwähnten Zusammenhänge aus Kapitel 3 und betrachten die Anregung mit der Exponentialfolge z^k für $z = \exp(j\omega T)$ (Vgl. dazu Kapitel 3.2):

$$x[k] = e^{j\omega Tk} = e^{j\Phi k} \quad \text{mit} \quad \Phi = \omega T = \omega / f_A \tag{4.28}$$

wobei f_A weiterhin die Abtastfrequenz bezeichnet. Die Übertragungsfunktion im Frequenzbereich (der sog. *Frequenzgang*) ist dann

$$H(z = e^{j\Phi k}) = \sum_{k=-\infty}^{\infty} h[k]e^{-j\Phi k} = H(e^{jk(\Phi+2\pi n)}) \tag{4.29}$$

Als *Amplitudengang* wird der Betrag des Frequenzganges bezeichnet, als *Phasengang* seine Phase. Es ist folgendes festzustellen:

Theorem 4.4 (Periodizität und Symmetrie des Frequenzganges).

- *Der Frequenzgang ist periodisch in $\Phi = \omega T = \omega/f_A$ mit ganzzahligen Vielfachen von 2π.*
- *Wenn die Impulsantwort reell ist, so ist der Amplitudengang (Betrag des Frequenzganges) eine gerade, der Phasengang ϕ dagegen eine ungerade Funktion in Φ bezüglich $\Phi = 0$.*

Letzteres sieht man aus

$$\begin{aligned}|H(e^{j\Phi k})|^2 &= \sum_{k=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} h[k]h[n](\cos(\Phi k)\cos(\Phi n) + \sin(\Phi k)\sin(\Phi n)) \\ &= \sum_{k=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} h[k]h[n](\cos(-\Phi k)\cos(-\Phi n) + \sin(-\Phi k)\sin(-\Phi n)) \\ &= |H(e^{j(-\Phi)k})|^2 \end{aligned} \tag{4.30}$$

$$\tan\phi(\Phi k) = \frac{\sum_{-\infty}^{\infty} h[k](\sin(\Phi k))}{\sum_{-\infty}^{\infty} h[k](\cos(\Phi k))} = \frac{-\sum_{-\infty}^{\infty} h[k](\sin(-\Phi k))}{\sum_{-\infty}^{\infty} h[k](\cos(-\Phi k))} = -\tan\phi(-\Phi k) \tag{4.31}$$

und somit

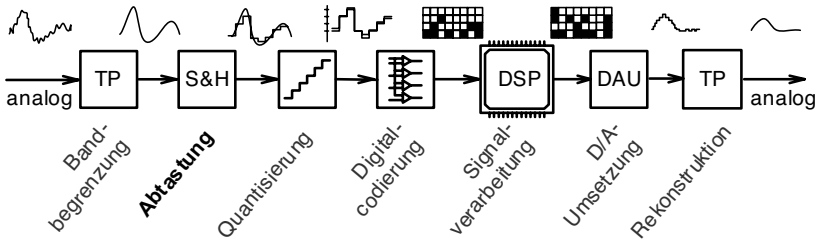
$$\phi(\Phi k) = -\phi(-\Phi k) \tag{4.32}$$

Weil der Frequenzgang diskreter Systeme periodisch ist, kann er als Fourierreihe interpretiert werden mit den Koeffizienten

$$h(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H(e^{j\Phi})e^{j\Phi k}d\Phi \tag{4.33}$$

□

4.2.2 Das Abtasttheorem



Wir haben nun bereits eine mathematische Beschreibung des Abtastprozesses erwähnt, ohne Sedoch darauf einzugehen, was dies für den Informationsgehalt unseres ursprünglichen Signals bedeutet. Offensichtlich reduziert man gleichermaßen gewaltsam die Information des kontinuierlichen Signals, so dass man sich berechtigterweise der Frage nach der Zulässigkeit eines solchen Vorgangs zu stellen hat.

Interessanter Weise stellt sich heraus, dass durch den Abtastvorgang trotz der Tatsache, dass man nur wenige Werte eines bestimmten Messintervalls berücksichtigt, unter bestimmten Voraussetzungen keinerlei Information über das analoge Signal verloren geht. Dieser Sachverhalt ist für die gesamte Digitaltechnik und damit praktisch für die gesamte technische Signalverarbeitung von fundamentaler Bedeutung. In der 7iteratur wird dies gemeinhin als „Abtasttheorem nach Shannon“ bezeichnet. Historisch wurde 1924 von Harry Ny uist 1 9-19 6 eine erste Formulierung der notwendigen Bandbreite für Informationsübertragung vorgestellt Ny uist, 1924 . Ny uist brachte auch 192 die erste Formulierung des Abtasttheorems, allerdings noch ohne Beweis Ny uist, 192 . Dieser wurde erst 21 *f*ahre später von Claude Elwood Shannon 1916-2001 erbracht in seiner klassischen Veröffentlichung von 1949, „Communication in the presence of noise“ Shannon, 1949 . Dieses Papier gilt als Grundlage der Informationstheorie. Andere Bezeichnungen des Abtasttheorems beziehen sich auf die Namen Whittaker, Kotelnikov oder Kramer. Wir nennen es an dieser Stelle kurz „*Abtasttheorem*“ , das folgendes aussagt:

445 (**b** /)4

Ein analoges j *ignal sei bandbegrenät, d.h. seine FouriertransJormierte ver schwindet identisch Jür* $|f| > f_{\max}$. *Dann lzsst sich aus dem abgetasteten* j *ig nal das ursprüngliche* j *ignal komplett und eindeutig wiedergewinnen. Daäu ist die AbtastJrequenä* f_A *mindestens doppelt so groß äu wzhlen wie* $f_{\max}$.

Das Abtasttheorem folgt unmittelbar aus Theorem 4.2 auf Seite 93 Gl. 4.15 über die Periodizität des Spektrums eines abgetasteten Signals. Wir stellen den Verlauf des Betrages des periodischen Spektrums grafisch dar. Dabei wird eine schematische Darstellung gewählt, bei der der Betrag des Spektrums eine gerade Funktion in ω ist. Dies gilt für die Funktionen der Gln. 4.4 oder 4.5 , speziell für reelle Funktionen. Dies ist aber für die Gültigkeit des Abtasttheorems keine notwendige Voraussetzung: Wie Beispiel 4.1 c zeigt, gibt es auch andere bandbegrenzte Signale. Die gerade Darstellung des Signals wurde hier also nur aus Gründen der übersichtlicheren Darstellung gewählt.

In Abb. 4.4 a ist schematisch ein Spektrum eines Signals gezeigt. $f_{\max}$ ist die größte vorkommende Fre uenz im Spektrum des Signals. Den Bereich von $0 \leq f \leq f_{\max}$ nennt man den *asisbandbereich* oder kurz das Basisband. Die periodisch folgenden Spektren mit höheren Fre uenzen heissen Seitenbänder.

In Abb. 4.4 b ist gezeigt, wie durch Abtastung dieses Signals mit der Abtastfre uenz f_A gemäß Theorem 4.2 ein mit f_A periodisches Spektrum entsteht. Man erkennt: Ist die Abtastfre uenz f_A mindestens doppelt so groß, wie die größte vorkommende Fre uenz $f_{\max}$ des Signalspektrums, so überlagern

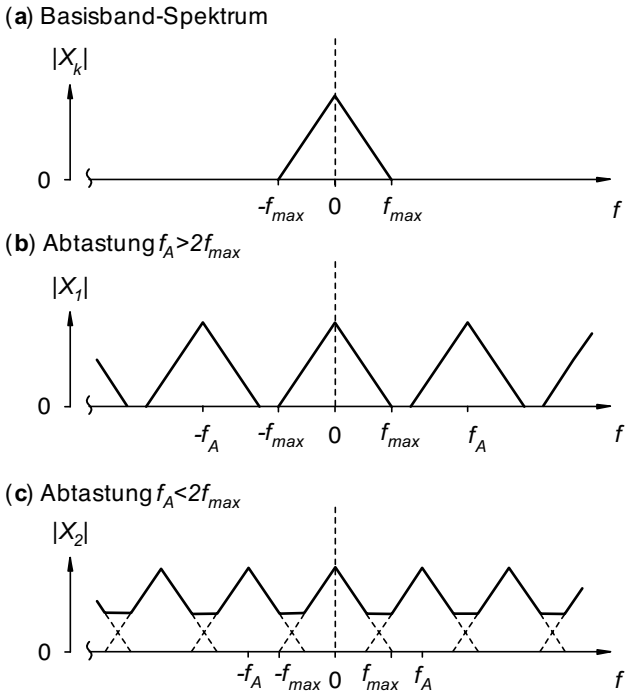


Abbildung 4.4. Spektren eines mit unterschiedlichen Frequenzen abgetasteten Signals

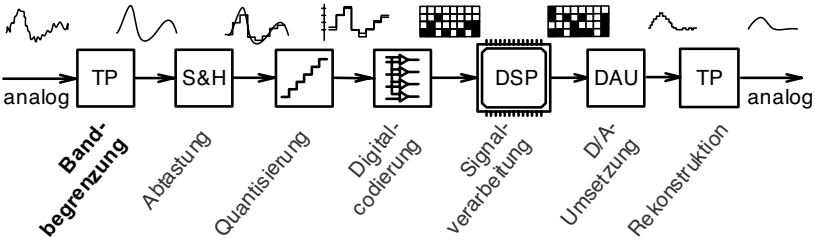
sich die periodischen Anteile nicht und das Signal könnte exakt rekonstruiert werden – beispielsweise einfach dadurch, dass durch ein Tiefpassfilter mit der Grenzfrequenz $f_{\max}$ wieder das Basisband vom restlichen Spektrum separiert wird.

Ist die Abtastfrequenz jedoch kleiner ($f_A < 2\,f_{\max}$, Abb. 4.4(c)), so führt dies zu Überlappungen, in denen ein Teil des Signalspektrums in den Basisbandbereich gespiegelt und dort addiert wird. Dies wiederholt sich auch für alle höheren Seitenbänder. Dieses Phänomen wird sinnfälligerweise auch *aliasing* genannt.

Da sich in diesen Bereichen die zu verschiedenen Frequenzen gehörenden Spektralkoeffizienten addieren, kann das ursprüngliche Spektrum daraus ohne weitere Zusatzinformation nicht wieder ermittelt werden. □

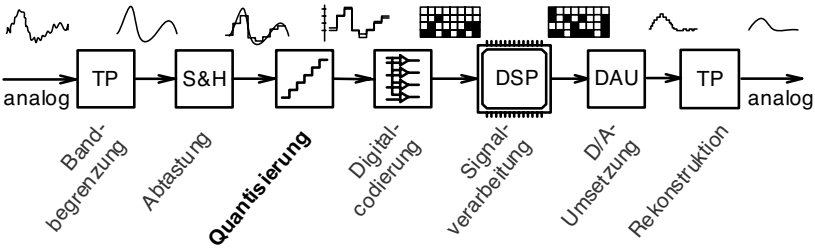
Zum weiteren Verständnis des Abtasttheorems stehen die Übungen 4.5 und 4.6 zur Verfügung.

4.2.3 Tiefpassfilterung



Wie eben beschrieben gehen bei zu großer Bandbreite des abgetasteten Signals nicht nur die höherfrequenten Signalanteile verloren, sondern auch Anteile unterhalb $f_A/2$ werden unbrauchbar. Um dies zu vermeiden, sollte ein Signal vor seiner Abtastung mit einem „anti-aliasing-Filter“ gefiltert werden. Ein solcher Tiefpassfilter begrenzt dementsprechend die Signalbandbreite auf die Hälfte der Abtastfrequenz. Noch stärkere Begrenzungen sind natürlich ebenfalls möglich.

4.2.4 Quantisierung



Nach Tiefpassfilterung und Abtastung wird das Signal quantisiert. Eine naheliegende Methode der Amplitudenquantisierung ist die in Abb. 4.5 dargestellte *gleichmäßige Quantisierung*. Der gesamte Wertebereich des zu quantisierenden Signals x wird in n gleiche Teilintervalle (Quantisierungsintervalle) der Größe q unterteilt. Jedem Intervall kann daraufhin ein Symbol zugeordnet werden, das in Anbetracht der hier verwendeten digitalen Systeme binär zu codieren ist. Die Quantisierung kann zwar prinzipiell beliebig fein erfolgen, jedoch steigt der Aufwand mit der Zahl der gewählten Quantisierungsstufen. In der Praxis stößt man bei diesem Schritt sogar recht schnell an Grenzen.

Durch die begrenzte Zahl von Quantisierungsstufen ist ein Kompromiss zwischen der Auflösung (reziprok zum Quantisierungsfehler) und dem Wertebereich zu finden. Nicht immer ist dabei die gleichmäßige Quantisierung optimal und für einige Anwendungen existieren entsprechend angepasste Quantisierungskennlinien. Dabei werden selten die Quantisierungsstufen selbst ungleichmäßig verteilt, sondern das analoge Eingangssignal wird üblicherweise zunächst mit Hilfe eines nichtlinearen Systems geeignet angepasst.

Im Audio-Bereich verwendet man dazu einen sog. Kompander, bestehend aus einem Kompressor, der das hochdynamische Audio-Signal in ein Signal

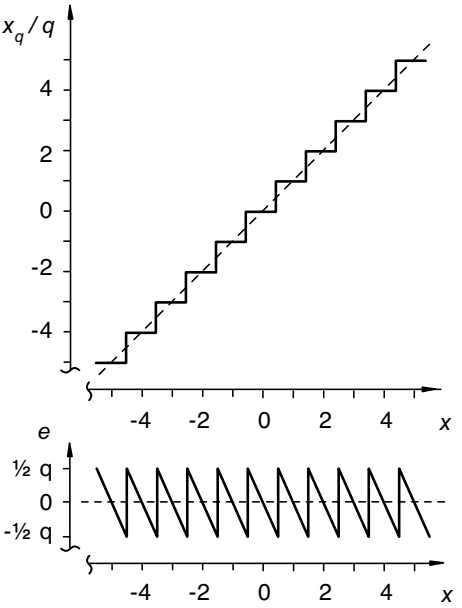


Abbildung 4.5. Quantisierungskennlinie und Quantisierungsfehler bei gleichmäßiger Quantisierung

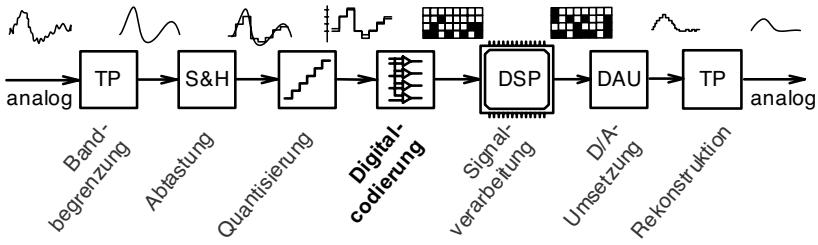
mit geringer Dynamik umwandelt sowie einen Expander, der eine dazu inverse Kennlinie besitzt und die Kompression wieder rückgängig macht. Das Signal mit geringer Dynamik eignet sich dann für eine Quantisierung mit gleichmäßig verteilten Quantisierungsstufen.

Auf eine Diskussion der Auswirkung der Quantisierung werden wir weiter unten eingehen, da dieser Schritt im eigentlichen Sinne nicht unabhängig vom nächsten Verarbeitungsschritt – der Codierung – zu betrachten ist.

4.3 Codierung

Wie bereits erwähnt wurde, verwendet die Digitaltechnik binäre Signale, um ein bestimmte Information zu codieren. Da ein einzelnes Binär-Signal nur einen von zwei verschiedenen Zuständen pro Zeitschritt einnehmen kann, benötigt man entweder mehrere Zeitschritte, um einen Wert aus einem größeren Zahlenbereich zu codieren, oder man verwendet parallel mehrere Binär-Signale. Eine solche Kombination von Binär-Signalen nennt man ein digitales Signal. Wir werden hier hauptsächlich auf diese Art von Signalen eingehen, da sie am weitesten verbreitet sind. Man könnte jedoch alle hier angestellten Überlegungen sinngemäß ebenfalls anwenden, wenn die Zahl der möglichen Signalzustände größer als zwei wäre.

4.3.1 Grundbegriffe



Bit, Wörter und Wortbreite

Codieren wir binär – wovon im Folgenden generell ausgegangen werden soll – stehen uns zunächst nur zwei mögliche Symbole (hier mit „0“ und „1“ bezeichnet) zur Verfügung. Erst durch eine geeignete Kombination von mehreren solcher „Bits“ (*bi*nary digi*ts*) lässt sich eine größere Anzahl von Symbolen codieren.

Ein Block von mehreren Bit, die zusammengenommen zur Codierung eines Symbols herangezogen werden, nennt man Wort; die Anzahl der verwendeten Bit wird als Wortbreite bezeichnet. Die Anzahl der mit einem Codewort darstellbaren Zeichen ist abhängig von der Wortbreite.

Zur Codierung N unterschiedlicher Symbole (bspw. Quantisierungsstufen) benötigen wir eine Wortbreite von mindestens $W = \log_2 N$ bit. Dies gilt allerdings nur, wenn für jedes Symbol die gleiche Wortlänge verwendet wird. In der Nachrichtentechnik kommen jedoch auch dynamische Wortlängen zum Einsatz. (Bsp: „0“ $= 0_b$, „1“ $= 10_b$, „2“ $= 11_b$)

Entropie und Redundanz

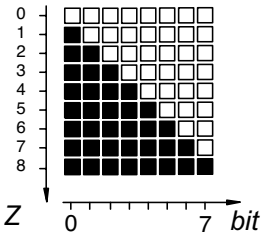
Die Übertragung oder Speicherung von codierten Daten verlangt in vielen Fällen eine Reduktion von Redundanz. Ziel ist es, eine rekonstruierbare Datencodierung zu finden, die die gewünschte Information unter Verwendung möglichst kurzer Codesequenzen enthält. Dazu ist Redundanz (irrelevante Information) zu entfernen und die Entropie zu maximieren. Redundanz kann sowohl auf Signalebene vermindert werden (Bsp.: Fano-Shannon- und Huffman-Codierung), als auch auf semantischer Ebene (Bsp.: Psychoakustische, -visuelle Modelle). Anwendung finden hierbei Prädiktions-Codierer und Kompressionsalgorithmen.

4.3.2 Codierungsarten

Die Codierung von Quantisierungsstufen bzw. die Darstellung von Zahlen allgemein kann auf unterschiedliche Weise erfolgen. In der Praxis haben sich eine ganze Reihe von Codierungsarten (Cover und Thomas, 1991) etabliert – wir gehen hier auf einige ein, die in digitalen Systemen gebräuchlich sind.

Unärcode

Bei diesem Code entscheidet die Anzahl der gesetzten Bit über das Codewort. Bei einer Wortlänge W sind demnach $W+1$ unterschiedliche Zeichen codierbar. Jede Codestelle besitzt die gleiche Wertigkeit. Die Unärcodierung findet in unterschiedlicher Form relativ häufig Anwendung.



Binärcode

Der Binärcode verwendet ein Stellenwertsystem, wie es auch für die gebräuchliche dezimale Zahlendarstellung verwendet wird:

$$Z_p = \sum_{i=-n}^{v} z_i r^i \tag{4.34}$$

Dabei bezeichnet v die Anzahl der Vorkommastellen und n die Anzahl der Nachkommastellen. Die Basis r bezeichnet man als Radix. Zur Codierung N unterschiedlicher Symbole (bspw. Quantisierungsstufen) benötigen wir eine Wortbreite von mindestens $W = \log_r N$. Im Falle einer Binärzahl ist die Radix $r = 2$. Diese Summenschreibweise erfasst somit auch die übliche dezimale Zahlendarstellung ($r = 10$) sowie weitere Zahlensysteme. Man beachte, dass die Summe (4.34) keinen „Überlauf“ und auch keinen „Rest“ besitzt – bei fortschreitender Addition oder fortschreitender Division in solchen Zahlensystemen entsteht also irgendwann ein Fehler. Wir werden diese Phänomene in Abschnitt 4.3.3 betrachten.

Beispiel 4.2 – Dezimalwert einer vorzeichenlosen Binärzahl.

0100.10_{b+} (Punkt kennzeichnet Kommastelle) $z_{-1} = 1,\ z_2 = 1$,
$Z_p = 0 \cdot 2^3 + 1 \cdot 2^2 + 0 \cdot 2^1 + 0 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} = 4,5_d$ □

Sollen auch negative Zahlen erfasst werden, bietet sich der Einsatz eines Vorzeichen-Bits an, was jedoch den Nachteil aufweisen würde, dass die Zahlen „-0“ und „+0“ durch zwei unterschiedliche Codes repräsentiert werden würden. Dies lässt sich mit einer *Zweierkomplement*-Darstellung umgehen. Um also eine positive normierten Binärzahl bestehend aus den Koeffizienten k_i in eine negative Binärzahl in Zweierkomplementdarstellung umzuwandeln, sind sämtliche Bits einschließlich des Vorzeichenbits zu negieren und anschließend zu dem so gewonnenen Ergebnis eine 1 zu addieren.

$$-z_n = 1 - \sum_{i=-n}^{-1} k_i 2^i \tag{4.35}$$

Beispiel 4.3 – Zweierkomplement-Darstellung.

Umwandlung der negativen Dezimalzahl -5_d in Zweierkomplement-Darstellung:
$+5_d$ binär codiert: 00101_b
Komplementbildung: 11010_b
Addition von 1: 00001_b
Ergebnis -5_d: 11011_b
Es ist zu beachten, dass das erste Bit von links (HSB – highest significant bit) zwar als „Vorzeichenindikator" dient, der Rest der im Zweierkomplement dargestellten Zahl jedoch nicht mit dem Betrag übereinstimmt. □

Ein wesentlicher Vorteil des Zweierkomplement-Codes gegenüber dem Vorzeichen-Betrag-Coden liegt darin, dass eine Subtraktion auf eine Addition einer negativen Zahl zurückgeführt werden kann. Das vereinfacht die Implementierung des Additions- und Subtraktionsalgorithmus.

Weniger gebräuchlich ist die Einerkomplement-Codierung. Hierbei werden negative Zahlen dargestellt, indem das Komplement ohne Addition einer Eins gebildet wird. Wie in folgender Tabelle 4.1 zusammengefasst, erhält man jedoch auch hierfür wiederum unterschiedliche Codes für $+0_d$ und -0_d.

	Kleinster Wert	-0	+0	Größter Wert
Vorzeichen+Betrag	1111...1	1000...0	0000...0	0111...1
Einerkomplement	1000...0	1111...1	0000...0	0111...1
Zweierkomplement	1000...0	0000...0	0000...0	0111...1

Tabelle 4.1. Kleinste und grösste Werte für unterschiedliche Codes

Der kleinste und größte Wert für eine Wortbreite von $W+1$ bits, d.h. W bits für das Wort und ein Bit für das Vorzeichen, sind für Vorzeichen–Betrag–Code und Einerkomplement–Code jeweils (2^W-1) und $-(2^W-1)$. Es wird also ein Zahlenbereich der Breite $(2^{W+1}-2)$ abgedeckt, dies ist um Eins weniger als die theoretisch mögliche Breite $(2^{W+1}-1)$. Grund ist natürlich das doppelte Repräsentieren der Null. Im Zweierkomplement–Code wird dieser Mangel beseitigt, der kleinste und größte Wert sind jeweils (2^W-1) und $-(2^W)$. Man beachte also in obiger Tabelle, dass der kleinste Wert für Einerkomplement– und Zweierkomplement–Code zwar binär gleich lauten, aber zwei verschiedene Zahlen darstellen.

Der in Abb. 4.6(**a**) gezeigte 8-4-2-1-Code stellt eine in der digitalen Signalverarbeitung weit verbreitete Codierung dar. Er ergibt sich aus der binären Zahlendarstellung, wie sie in Formel (4.34) angegeben ist. Variationen des Binärcodes erfassen wie oben dargestellt auch Kommazahlen und negative Werte.

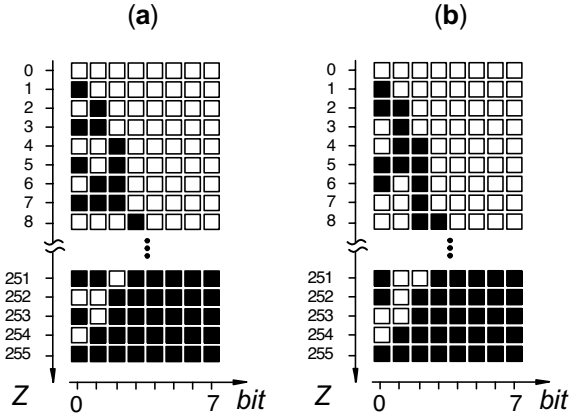


Abbildung 4.6. (**a**) 8-4-2-1-Code (**b**) Gray-Code

Fließkomma-Codes

Unpraktisch bei den bisher vorgestellten Binärcodes ist, dass beispielsweise beim Übergang von der betragsmäßig kleinsten darstellbaren Zahl zur darauf folgenden Zahl ein Unterschied im Faktor 2 besteht, während zwei benachbarte größere Zahlen mit gemeinhin unnötiger relativer Genauigkeit aufgelöst werden. Hierzu ein Beispiel. Wir nehmen an, eine Längenmessung soll im Bereich $[0m, 1000m]$ durchgeführt werden und es steht ein Code mit 1001 Worten zur Verfügung. Man wird diese Worte kaum gleichmässig in $1m$-Intervallen verteilen, wenn es um relative Genauigkeit geht: Diese betrüge dann bei einer Messung von $1m$ $\pm 100\%$ und bei einer Messung von $1000m$ $\pm 0{,}1\%$.

Sinnvoller ist für viele technische Anwendungen eine logarithmische Verteilung der Quantisierungsstufen, wie sie annähernd durch Fließkomma-Zahlen realisiert wird. Eine Fließkomma-Zahl wird dargestellt als

$$Z_f = M \cdot R^E \text{ mit } \frac{1}{R} \leq |M| < 1. \tag{4.36}$$

Dabei bezeichnen M die Mantisse, R die Radix und E den Exponenten dieser Zahlendarstellung. An dieser Stelle soll wiederum von einer Radix $R = 2$ ausgegangen werden, da wir weiterhin Binär-Signale verwenden wollen. Die Mantisse liegt damit im Bereich $0.5 \leq |M| < 1$ und könnte – ebenso wie der Exponent E – in Zweierkomplement-Codierung abgespeichert werden. Der Exponent ist ganzzahlig; kann jedoch durchaus auch negative Werte enthalten.

Beispiel 4.4 – Wertebereich einer 4-bit Fließkommazahl.

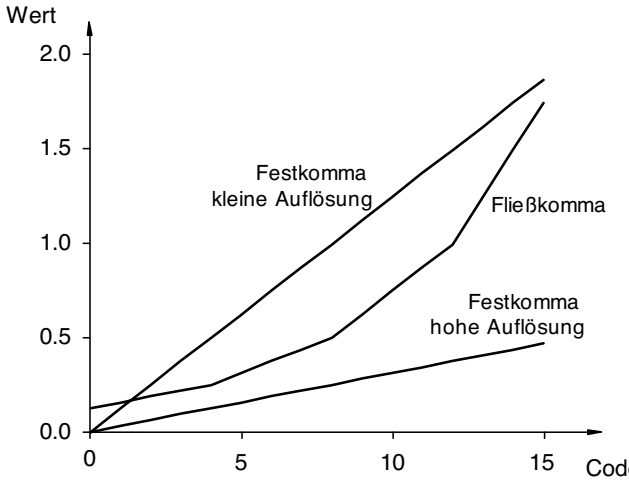
Berechnen Sie die Wertemenge, die mit einer positiven 4-Bit-Fließkomma-zahl darstellbar ist und vergleichen Sie diese mit einer Festkommadarstellung, die a) etwa den gleichen Wertebereich überdeckt, b) die gleiche maximale Auflösung besitzt.

Lösung:
Die verfügbare Wortbreite kann beliebig in Mantisse und Exponent unterteilt werden. Von der 4-bit-Gesamtwortbreite wählen wir beispielhaft die ersten zwei Bit als Mantisse und die restlichen beiden Bit als Exponent. Die Mantisse teilen wir gleichmäßig im Wertebereich $0.5 \leq M < 1$ auf, wobei wir zunächst auf negative Werte verzichten. Der Exponent ist ganzzahlig zu wählen. Wir wählen in diesem Beispiel die Exponenten -2, -1, 0 und 1 und codieren diesen im Zweierkomplement-Code.

Mantisse:	00_b:0.500_d	01_b:0.625_d	10_b:0.750_d	11_b:0.875_d
Exponent:	10_b:-2_d	11_b:-1_d	00_b:$+0_d$	01_b:$+1_d$

Mit diesen Mantisse-Exponent-Kombinationen erhalten wir nach Gl. (4.36) folgende Werte:

$E = -2:$	0.1250	0.1563	0.1875	0.2188
$E = -1:$	0.2500	0.3125	0.3750	0.4375
$E = +0:$	0.5000	0.6250	0.7500	0.8750
$E = +1:$	1.0000	1.2500	1.5000	1.7500



Es wird somit ein Wertebereich von ca. 2 überdeckt. In einer Festkomma-Codierung wäre somit ein Quantisierungsintervall von 0.125 notwendig, um mit 16 Symbolen den selben Wertebereich zu überdecken. Die Fließkommadarstellung bietet jedoch eine maximale Auflösung von $Z_1 - Z_0 = 0.0313$. Eine Festkommadarstellung mit dieser Auflösung würde nur einen Wertebereich von 0.4695 überdecken. □

Spezielle Codes

Für verschiedene Anwendungen haben sich weitere Codierungsarten etabliert, von denen wir an dieser Stelle nur einige nennen wollen.

Der in Abb. 4.6(**b**) gezeigte Gray-Code ist ein Beispiel für einen so genannten „einschrittigen Code". Zählt man in diesem Code vor- oder rückwärts, ändert sich jeweils nur ein Bit. Damit kann beispielsweise in Zählschaltungen vermieden werden, dass beim Übergang von einem Wert zum nächsten zwischenzeitlich falsche Werte dadurch entstehen, dass nicht alle zu verändernden Bitstellen exakt gleichzeitig umschalten.

„Differentielle Codes" codieren die Signal-Information nicht in den *Werten* eines analogen Signalträgers, sondern vielmehr in dessen Zustands*änderung*. Man stelle sich beispielsweise die Spule im Schreib-/Lesekopf eines Festplattenlaufwerks vor, die mit einer längeren Sequenz von gewöhnlich codierten „Low"-Bits angesteuert wird. Offensichtlich entsteht beim Schreiben dieser Bits eine konstante Magnetisierung auf dem Datenträger. Soll diese Sequenz wieder eingelesen werden, so steht man vor der Problematik, dass lediglich ein *Wechsel* des Magnetisierungszustands im Lesekopf einen Stromimpuls induziert. Eine längere Reihe von „High"-Bits wäre hingegen nicht von einer „Low"-Bit-Reihe unterscheidbar, da in beiden Fällen kein Strom induziert wird. Ferner hat man in diesem Fall keinerlei Anhaltspunkte über die Dauer eines Bits, wodurch nur schwerlich eine Synchronisation mit der Lese-Logik erreichbar ist.

Einen Ausweg bietet eine „lauflängenbegrenzte" Codierung (run length limited – RLL). Diese Codierung stellt sicher, dass in einem Codewort nur eine begrenzte Anzahl von aufeinander folgenden Bits gleich ist. Dies wiederum bedeutet, dass genügend oft Signalwechsel stattfinden, an denen eine Synchronisation stattfinden kann. (Selbstsynchronisierende Codes.)

In Signalübertragungssystemen muss oftmals eine Potentialtrennung vorgenommen werden. Ein populäres Beispiel dafür sind Datenleitungen in Weitverkehrsnetzen oder in industrieller Umgebung. Um Störspannungen von Datenverarbeitungsanlagen fernzuhalten, wird das analoge Trägersignal beispielsweise durch einen Übertrager (Transformator) vom Gleichspannungsanteil befreit. Dadurch geht jedoch auch jegliche Information über den ursprünglichen Pegel des Nutzsignals verloren, da das Signal im Übertrager gewissermaßen differenziert wird. Eine lauflängenbegrenzte Codierung (beispielsweise Manchester-Code) schafft auch hier Abhilfe.

4.3.3 Quantisierungsfehler

Nicht nur im Zuge der Quantisierung beim Umsetzen eines Analogwertes in einen digitalen Wert kommt es zu Ungenauigkeiten, sondern auch bei der Verarbeitung von digitalen Werten treten unter Umständen aufgrund der oben beschriebenen begrenzten Anzahl von zur Verfügung stehenden Codes Fehler auf.

Wir unterscheiden dabei zwischen zwei verschiedenen Fällen. Auf der einen Seite ist das Digitalwort anzupassen auf die zur Verfügung stehende Registerlänge, die in gebräuchlichen digitalen Systemen zwischen 8 und 128 Bit liegt. Überschreitet ein Wert, der beispielsweise im Verlaufe eines Algorithmus aus einer Multiplikation entstanden sein könnte, die zur Verfügung stehende Registerwortlänge, kommt es zu Überlauffehlern (Overflow). Verwendet man beispielsweise eine vorzeichenlose 8-bit Ganzzahldarstellung, so resultiert die Multiplikation von 128_d mit 2_d in der Zahl 0_d, was natürlich einen extremen Fehler darstellt. Auf der anderen Seite sind Binärzahlen in einigen Fällen (beispielsweise nach einer Division) zu Runden. Auf beide Arten von Fehlern – *Fehler durch Überlauf* und *Fehler durch Runden* soll im Folgenden kurz eingegangen werden.

Einen Overflow vorab zu berechnen und vermeiden zu wollen stellt in realen Digitalsystemen ein mit wachsender Komplexität des zugrundeliegenden Algorithmus zunehmendes Problem dar. Deshalb wird bei der Programmierung digitaler Algorithmen üblicherweise anders vorgegangen. Gängige Arithmetik-Einheiten verfügen über umfangreiche Kontrollmechanismen, die einen Überlauffehler im Nachhinein feststellen können. Zwar ist das restliche Ergebnis dann nicht mehr brauchbar, jedoch dient diese Überlauf-Detektion dazu, den Algorithmus abzubrechen und das Ergebnis zu verwerfen. Aus diesem Grunde stellt ein Überlauf keinen Fehler im Sinne eines unvermeidlichen Rauschens dar. Schließlich entstehen keinerlei Fehler durch Überlauf, solange kein Überlauf stattfindet.

Trotzdem hat diese Methode den Nachteil, dass die Zuverlässigkeit eines Digitalsystems damit ungewiss wird. Um dem empirischen Auftreten von Überlauffehlern entgegenzuwirken, kann auch anders vorgegangen werden: Vor jeder kritischen Operation wird eine Normierung der digital dargestellten Zahl vorgenommen. Normieren wir beispielsweise vor oder nach jeder Multiplikation sämtliche Faktoren auf den Wertebereich $-1 \leq z < +1$, so ist das zu erwartende Ergebnis unabhängig von den tatsächlichen Werten sicher zu prognostizieren.

Welchen Nachteil besitzt diese Normierung (neben dem zusätzlichen Ressourcenbedarf)? Will man eine Division vermeiden, lässt sich die Normierung größerer Zahlen auf $-1 \leq z < +1$ durch Verschieben der Kommastelle nach links realisieren, was bei Binärzahlen einer Division durch zwei bzw. einer SHR (shift-right)-Operation gleichkommt. Offensichtlich werden dabei die LSB (die niederwertigsten Bits - least significant bit) gelöscht. Man erkauft sich somit die Determinierbarkeit des Algorithmus (Vermeidung eines Überlauf-Fehlers) durch ein permanentes statistisch verteiltes Rauschen. Anhand der Zweierkomplement-Darstellung soll dessen Größe abgeschätzt werden.

Gehen wir von einer beliebigen Zahl z in Zweierkomplement-Darstellung aus, dessen Stellenzahl b jedoch um a Bits (plus Vorzeichenbit) größer ist, als uns in den r Bits (plus Vorzeichenbit) eines Registers zur Verfügung steht

($b = a + r$). Nun haben wir drei Möglichkeiten, um diese Zahl im Register abzulegen:

1. Wir schneiden die a höchstwertigen Bits ab. Das kommt einem Überlauf-Fehler gleich und beschert uns im Extremfall den maximal erreichbaren Fehler.
2. Wir normieren z auf $-1 \leq z < +1$, schieben dabei das Komma direkt hinter das Vorzeichenbit und speichern das normierte z linksbündig. Dabei *schneiden* wir die a niederwertigsten Bits ab.
3. Wir normieren wie im 2. Fall jedoch *runden* wir anschließend auf die Nachkomma Stellenzahl $r - 1$, wodurch wiederum a Bits verlorengehen.

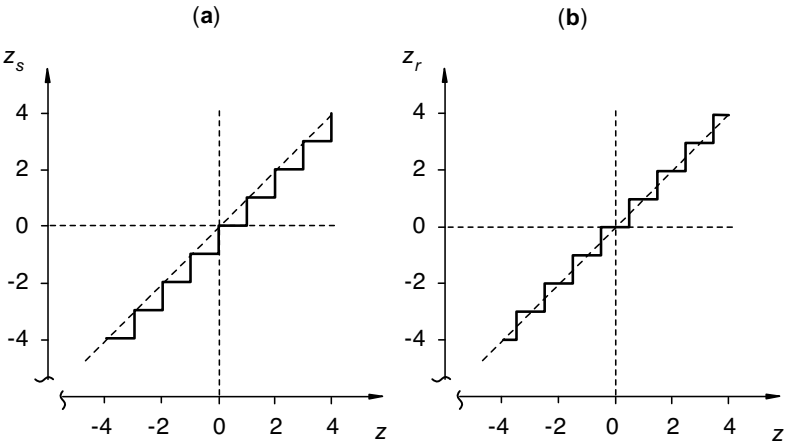


Abbildung 4.7. Kennlinien zum Schneiden (**a**) bzw. Runden (**b**) einer Zweierkomplementzahl

Beispiel 4.5 – Schneiden und Runden einer normierten Binärzahl in Zweierkomplement-Codierung.

Berechnen Sie den Betrag des Fehlers, der entsteht, wenn die als Zweierkomplement-Zahl interpretierten 5-bit Ganzzahlen 01010_b, 11010_b und 00111_b in einem normierten 3-bit-Register a) mittels Abschneiden der LSB und b) mittels Runden abgespeichert werden sollen.

Lösung:

	binär	dezimal	binär	dezimal	binär	dezimal
Wert:	01001	9	11010	-6	00111	7
normieren:	0.1001	9/16	1.1010	-6/16	0.0111	7/16
schneiden:	0.10--	8/16	1.10--	-8/16	0.01--	4/16
Fehlerbetrag:	0.0001	1/16	0.0010	2/16	0.0011	3/16
runden$^{(1)}$:	0.10--	8/16	1.11--	-4/16	0.10--	8/16
Fehlerbetrag:	0.0001	1/16	0.0010	2/16	0.0001	1/16

$^{(1)}$: Die Regel zum Runden einer Zahl z in einer Basis mit Radix R auf eine ganze Zahl lautet:

$$z_r = \text{trunc}_0\left(z + \frac{1}{2}\right), \tag{4.37}$$

wobei „trunc_0“ das Abschneiden nach der nullten Nachkommastelle bezeichne, d.h. sämtliche Nachkommastellen werden entfernt. Das Runden auf n Nachkommastellen könnte man somit folgendermaßen formulieren:

$$z_r = \text{trunc}_n\left(z + \frac{1}{2}R^{-n}\right) \tag{4.38}$$

In der selben Form geschieht auch das Runden einer binärcodierten Zahl in Zweierkomplementdarstellung auf n Nachkommastellen durch eine Addition von $0.5 \cdot 2^{-n} = 2^{-(n+1)}$ und einem anschließenden Abschneiden aller Nachkommastellen hinter n. Auch hier werden dadurch negative Zahlen in positive Richtung aufgerundet. □

Nach diesem Beispiel soll nun allgemein berechnet werden, welcher Fehler beim Schneiden oder Runden einer binär codierten Zahl in Zweierkomplement-Darstellung maximal zu erwarten ist.

Schneiden einer (verlustfrei) normierten Zahl z

Durch *Abschneiden* der a niederwertigsten Bits erhalten wir z_s mit einer Abweichung vom ursprünglichen z von

$$e_{s+} = z_s - z \tag{4.39}$$

e_{s+} ist in allen Fällen negativ (oder null), denn

- ist z positiv, entfallen beim Schneiden u.U. einige Nachkommastellen, wodurch z_s kleiner wird als z
- Ist z negativ, wird diese Zahl in Zweierkomplementdarstellung durch das Schneiden ebenfalls kleiner, denn das Ersetzen von Einsen durch Nullen bewirkt hier eine Vergrößerung ihres Betrages. Die Zahl $1.00...0_b$ ist in Zweierkomplementdarstellung die kleinste darstellbare Zahl und $1.11...1_b$ die größte negative Zahl

Sind die abgeschnittenen Stellen Nullen, entsteht kein Fehler, wohingegen der maximale Fehler auftritt, wenn alle a abgeschnittenen Bits Einsen sind. Da

$$\sum_{i=-\infty}^{n} 2^i = 2^{n+1} \tag{4.40}$$

(Bsp: $2+1+1/2+1/4+1/8+... = 4$) übersteigt der Fehler durch Schneiden auch bei beliebig hoher ursprünglicher Auflösung der Zahl z nie die Wertigkeit $z_{LSB(r)}$ des LSB des Registers. Da letzteres eine Wortbreite r (zzgl. Vorzeichenbit) besitzt, können wir den maximalen Fehler folgendermaßen eingrenzen.

$$-z_{LSB(r)} = -2^{-r} < e_s \leq 0 \tag{4.41}$$

(Bsp: $r = 4 \rightarrow z_{LSB(r)} = 1/16$). Allerdings besaß auch die ursprüngliche Zahl z eine begrenzte Stellenzahl und war somit bereits mit einem ungewissen Fehler behaftet. Soll der Fehler berechnet werden, der maximal durch das Schneiden entstanden sein kann, muss von (4.41) die ursprüngliche Auflösungsgrenze abgezogen werden. Man könnte nun unterscheiden, ob das ursprüngliche z durch sorgfältiges Runden oder aber ebenfalls durch Schneiden entstanden ist. Beschränken wir uns auf den letzten Fall, ist das z ebenfalls mit einer Unschärfe behaftet, die der Wertigkeit des LSB (Quantisierungsstufe) entspricht:

$$z_{LSB(z)} = 2^{-b} = 2^{-(r+a)} \tag{4.42}$$

Der mögliche Fehlerbereich durch Schneiden einer normierten und ebenfalls durch Schneiden entstandenen Zweierkomplementzahl mit $b+1$ Bits auf die Länge von $r+1$ Bits beträgt also:

$$2^{-b} - 2^{-r} < e_s \leq 0 \tag{4.43}$$

Runden einer (verlustfrei) normierten Zahl z

Unter der Voraussetzung, dass *gerundet* wird, halbiert sich der maximale Fehler betragsmäßig auf eine halbe Quantisierungseinheit. Er kann jetzt jedoch positiv und negativ sein:

$$-0.5z_{LSB(r)} = -2^{-(r+1)} < e_{r+} \leq 2^{-(r+1)} = 0.5z_{LSB(r)} \tag{4.44}$$

Durch das Runden einer normierten Zweierkomplementzahl mit $b+1$ Bits auf die Länge von $r+1$ Bits entsteht demzufolge ein zusätzlicher Fehler von

$$2^{-(b+1)} - 2^{-(r+1)} < e_{r+} \leq 2^{-(r+1)} - 2^{-(b+1)} \tag{4.45}$$

In diesem Fall ist $2^{-(b+1)}$ die Genauigkeit der ursprünglichen Zahl z – diesmal nehmen wir also an, dass die ursprüngliche Zahl z ebenfalls durch Runden

entstanden ist, wodurch der maximale Fehler nur noch einer halben Quantisierungseinheit entspricht. Andernfalls muss entsprechend wieder 2^{-b} angesetzt werden.

4.3.4 Quantisierungsfehler als stochastisches Signal

Wir nehmen den Quantisierungsfehler als Rauschsignal an. Dies ist in guter Näherung möglich, da die Quantisierung der einzelnen Daten unkorreliert und zufällig auftritt. Das Quantisierungsrauschen ist dann eine Signalfolge $e_r[n]$ mit Mittelwert m und Varianz σ. Auf derartige stochastische Signale werden wir in Kap. 7 noch ausführlicher eingehen.

Wir beschreiben diese Größen (Mittelwert und Varianz) als Erwartungswerte. Der Erwartungswert $E[f(x)]$ einer Größe $f(x)$ mit Wahrscheinlichkeitsdichte $p(x)$ ist allgemein definiert als

$$E\left[f(x)\right] = \sum_{\forall x} f(x)p(x) = \int_{\forall x} f(x)p(x)\mathrm{d}x \tag{4.46}$$

wobei die Summe bei diskreten x über alle auftretenden x bzw. das Integral bei kontinuierlichen x über den gesamten x-Raum läuft, also bei eindimensionalem x entsprechend über die x-Achse.

Das Bilden von Erwartungswerten ist eine lineare Operation, die mit anderen linearen Operationen vertauscht werden kann: z.B. Summenbildung, Mittelung, Antwort eines linearen Systems, Faltung, Integration, etc. Wir werden von dieser Möglichkeit reichlich Gebrauch machen.

Weiterhin ist der Erwartungswert des Produkts zweier unkorrelierter Größen gleich dem Produkt der Erwartungswerte dieser Größen. Beispielsweise bedeutet dies, dass der Erwartungswert

$$E\left[f(x)g(x)\right] \;=\; E\left[f(x)\right]\; E\left[g(x)\right] \tag{4.47}$$

für unkorrelierte Funktionen $f(x)$ und $g(x)$ ist. Für miteinander korrelierte Größen gilt dieser Zusammenhang jedoch nicht. Hier ist also spezielle Vorsicht angebracht. Wir werden dies bei der Bestimmung der Varianz am Ausgang eines Systems zu beachten haben.

Spezielle Funktionen von x sind $f(x) = x$, in diesem Fall erhalten wir den linearen *Mittelwert* μ:

$$\mu = E[x] \tag{4.48}$$

sowie $f(x) = (x - \mu)^2$. In diesem Fall erhalten wir die *Varianz* σ:

$$\sigma = E[(x - \mu)^2] \tag{4.49}$$

4.3.5 Transformation von Zufallsgrößen durch Systeme

Es wurde festgestellt, dass aufgrund der begrenzten Wortbreite der digitalen Zahlendarstellung jedes Signal mit einem Quantisierungsrauschen behaftet

ist. Dies gilt natürlich auch für Signale, die am Eingang eines digitalen Signalverarbeitungssystems darstellen. In diesem und im nächsten Kapitel soll nun untersucht werden, wie sich das Rauschsignal innerhalb eines digitalen Systems fortpflanzt.

Das Quantisierungsrauschen addiert sich zu der Eingangsfolge $x[n]$ additiv. Es gilt also am Systemausgang

$$y_{\mathrm{x\&e}}[n] = S\{x[n]+e_r[n]\} = h[n]*(x[n]+e_r[n]) = h[n]*x[n]+h[n]*e_r[n] \quad (4.50)$$

Die Reaktion des Systems auf das Rauschen ist somit ebenfalls additiv.

Der Mittelwert μ_{ein} der Rauschfolge am Eingang des Systems ist gegeben als Erwartungswert

$$\mu_{\mathrm{ein}} = E[e_r[n]] \quad (4.51)$$

Am Ausgang das Systems erzeugt das Eingangsrauschen ein Signal mit Mittelwert

$$\begin{aligned}\mu_{\mathrm{aus}} &= E[h[n] * e_r[n]] = h[n] * E[e_r[n]] = h[n] * \mu_{\mathrm{ein}} = \mu_{\mathrm{ein}}(h[n] * 1) \\ &= \mu_{\mathrm{ein}} \sum_{n=-\infty}^{\infty} h[n] = \mu_{\mathrm{ein}} H(e^{j0}) = \mu_{\mathrm{ein}} H(z=1) \end{aligned} \quad (4.52)$$

Um die Varianz am Ausgang zu bestimmen, betrachten wir den Erwartungswert von $y[n]^2$. Zum einen gilt

$$\begin{aligned} E\left[(y[n])^2\right] &= E\left[(y[n] - \mu_{\mathrm{aus}} + \mu_{\mathrm{aus}})^2\right] \\ &= E\left[(y[n] - \mu_{\mathrm{aus}})^2\right] + 2\mu_{\mathrm{aus}} E\left[y[n] - \mu_{\mathrm{aus}}\right] + \mu_{\mathrm{aus}}^2 \\ &= \sigma_{\mathrm{aus}} + 2\mu_{\mathrm{aus}}(\mu_{\mathrm{aus}} - \mu_{\mathrm{aus}}) + \mu_{\mathrm{aus}}^2 \\ &= \sigma_{\mathrm{aus}} + \mu_{\mathrm{aus}}^2 \end{aligned} \quad (4.53)$$

Man beachte, dass in der Herleitung dieser Relation an keiner Stelle Systemeigenschaften benutzt wurden.

Zum anderen können wir mit Hilfe der Systemeigenschaften schreiben

$$\begin{aligned} E[(y[n])^2] &= E\left[(h[n] * e_r[n])^2\right] \\ &= E\left[\sum_{k=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} h[k]h[m]e_r[n-k]e_r[n-m]\right] \\ &= \sum_{k=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} h[k]h[m] E\left[e_r[n-k]e_r[n-m]\right] \\ &= \sum_{k=-\infty}^{\infty} h[k]^2 E\left[e_r[n-k]^2\right] \\ &\quad + \sum_{k=-\infty}^{\infty} \sum_{m\neq k} h[k]h[m] E\left[e_r[n-k]\right] E\left[e_r[n-m]\right] \end{aligned} \quad (4.54)$$

Die letzte Zerlegung ist notwendig, da bei gleichem Index $k = m$ die Größen $e[n-k]$ und $e[n-m]$ nicht mehr unkorreliert sind und daher der Erwartungswert des Quadrates gebildet werden muss. Für den Erwartungswert dieses Quadrats der Eingangsrauschfolge gilt analog zu der obigen Betrachtung

$$E\left[(e_r[n])^2\right] = \sigma_{\text{ein}} + \mu_{\text{ein}}^2 \tag{4.55}$$

so dass weiterhin gilt:

$$\begin{aligned} E\left[(y[n])^2\right] &= \left(\sigma_{\text{ein}} + \mu_{\text{ein}}^2\right) \sum_{k=-\infty}^{\infty} h[k]^2 + \mu_{\text{ein}}^2 \sum_{k=-\infty}^{\infty} \sum_{m \neq k} h[k]h[m] \\ &= \left(\sigma_{\text{ein}} + \mu_{\text{ein}}^2\right) \sum_{k=-\infty}^{\infty} h[k]^2 + \mu_{\text{ein}}^2 \sum_{k=-\infty}^{\infty} h[k] \left(-h[k] + \sum_{m=-\infty}^{\infty} h[m]\right) \\ &= \sigma_{\text{ein}} \sum_{k=-\infty}^{\infty} h[k]^2 + \mu_{\text{ein}}^2 \left(\sum_{k=-\infty}^{\infty} h[k]\right)^2 \\ &= \sigma_{\text{ein}} \sum_{k=-\infty}^{\infty} h[k]^2 + \mu_{\text{aus}}^2 \end{aligned} \tag{4.56}$$

Mithin gilt für die Varianz des Ausgangssignals

$$\sigma_{\text{aus}} = \sigma_{\text{ein}} \sum_{k=-\infty}^{\infty} h[k]^2 \tag{4.57}$$

und mit der Parseval'schen Gleichung (Theorem 3.9 auf Seite 56) können wir auch schreiben

$$\sigma_{\text{aus}} = \sigma_{\text{ein}} \frac{1}{2\pi j} \oint H(z)H(1/z)z^{-1}dz \tag{4.58}$$

Die Varianz am Ausgang entsteht also aus der Varianz am Eingang durch Multiplikation mit der Summe der Quadrate der Impulsantwortfolge.

Der Mittelwert des Eingangs hat keinen Einfluss auf die Varianz des Ausgangs. Dies ist offensichtlich eine Konsequenz aus der Linearität des Systems. Denn wir haben gesehen, dass der Erwartungswert des Ausgangsquadrats immer durch Varianz minus Quadrat des Ausgangsmittelwertes gegeben ist – unabhängig von den Systemeigenschaften. Hingegen entstand dasselbe Quadrat des Ausgangsmittelwertes aus den Systemeigenschaften in der zweiten Herleitung nur durch die Linearität des Systems. Bei nichtlinearen Systemen müssen wir also einen Einfluss des Mittelwertes auf die Ausgangsvarianz erwarten (Kay, 1993). Dieser Einfluss wir allgemein als *bias* (engl. Verschiebung) bezeichnet.

Diese Ergebnisse gelten allgemein für die Transformation von Zufallsgrößen durch LTI-Systeme.

Wir fassen noch einmal zusammen in folgendem Theorem:

Theorem 4.6 (Transformation von Zufallsgrößen durch LTI-Systeme).

$$\mu_{\text{aus}} = \mu_{\text{ein}} \sum_{n=-\infty}^{\infty} h[n] = \mu_{\text{ein}} H(e^{j0}) = \mu_{\text{ein}} H(z=1) \tag{4.59}$$

$$\sigma_{\text{aus}} = \sigma_{\text{ein}} \sum_{k=-\infty}^{\infty} h[k]^2 = \sigma_{\text{ein}} \frac{1}{2\pi j} \oint H(z) H(1/z) z^{-1} dz \tag{4.60}$$

4.3.6 Transformation des Quantisierungsrauschens durch LTI-Systeme

Wir betrachten jetzt eine Quantisierung des Signals x. Es werde repräsentiert durch normierte $(-1 \leq x < 1)$ binärcodierte Werte in Zweierkomplementdarstellung mit b Bits zzgl. Vorzeichenbit. Es stehe jedoch nur eine Register-Wortlänge von c Bit plus Vorzeichenbit zur Verfügung $(c < b)$. Das Signal werde nicht gerundet, sondern abgeschnitten. Wie in Kapitel 4.3.3 erläutert, entsteht dabei ein betragsmäßig maximaler Fehler von

$$Q = 2^{-b} - 2^{-c} \tag{4.61}$$

Da der Fehler gleichverteilt ist, entsteht durch Schneiden ein Quantisierungsrauschen mit dem Mittelwert

$$\mu_{\text{ein}} = Q/2 = \frac{1}{2}(2^{-b} - 2^{-c}) \tag{4.62}$$

Die Varianz des Quantisierungsrauschen erhalten wir unter Gleichverteilungsannahme

$$\begin{aligned} \sigma_{\text{ein}} &= \int_{e=-Q/2}^{Q/2} \frac{1}{Q} e^2 de = Q^2/12 \\ &= \frac{1}{12}(2^{-b} - 2^{-c})^2 \end{aligned} \tag{4.63}$$

Nach dem oben aufgeführten Satz über die Transformation von Zufallsgrössen durch LTI-Systeme entsteht daraus am Ausgang ebenfalls ein Rauschen mit

$$\begin{aligned} \mu_{\text{aus}} &= \frac{1}{2}(2^{-b} - 2^{-c}) H(z=1) \\ \sigma_{\text{aus}} &= \frac{1}{12}(2^{-b} - 2^{-c})^2 \frac{1}{2\pi j} \oint H(z) H(1/z) z^{-1} dz \end{aligned} \tag{4.64}$$

Beispiel 4.6 – Rauschen durch Schneiden an einem LTI-System.

Gegeben sei ein LTI-System mit

$$H(z) = \frac{z-1}{z-a} \quad , \quad 0 < |a| < 1$$

und ein Prozeß, der Zahlen mit $b = 8$ Bit auf $c = 4$ Bit beschneidet. Wie sieht der Amplitudengang des Systems für $a = 1/2$ aus? Welche Parameter kennzeichnen das Quantisierungsrauschen am Ausgang? Wie groß sind diese für $a = 1/2$?

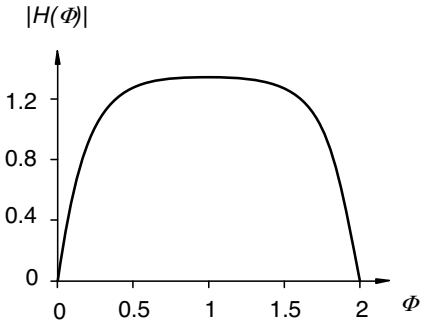
Lösung:
Der Amplitudengang ist der Betrag des Frequenzganges, also der Betrag der Übertragungsfunktion bei $z = e^{j\omega T}$. Wir schreiben

$$H(\omega T) = \frac{e^{j\omega T} - 1}{e^{j\omega T} - a} = \frac{\cos(\omega T) - 1 + j\sin(\omega T)}{\cos(\omega T) - a + j\sin(\omega T)}$$

und mit $\Phi = \omega T$ bei $a = 1/2$:

$$|H(\Phi)|^2 = \frac{(\cos(\Phi) - 1)^2 + \sin^2(\Phi)}{(\cos(\Phi) - 0.5)^2 + \sin^2(\Phi)}.$$

Dies hat den im folgenden gezeigten Verlauf (periodisch mit 2π).



Nach der Parseval'schen Gleichung ist das Integral

$$\frac{1}{2\pi a j} \oint \frac{(z-1)^2}{z(z-a)(z-1/a)} dz$$

zu bestimmen, wobei z im gemeinsamen Konvergenzgebiet von $H(z)$ und $H(1/z)$ liegen muss. Damit muss ein Kreisring vom Radius r mit $a < r < 1/a$ als Integrationsweg gewählt werden. Innerhalb dieses Kreises liegen zwei Residuen an den Polen $z = 0$ und $z = a$. Das Integral nimmt damit den Wert

$$\frac{1}{a}\left[1+\frac{(a-1)^2}{a(a-1/a)}\right]=\frac{2}{1+a}$$

an. Das Quantisierungsrauschen am Ausgang ist dadurch von der Form

$$\mu_{\text{aus}}=\frac{1}{2}(2^{-8}-2^{-4})H(z=1)=0$$
$$\sigma_{\text{aus}}=\frac{1}{12}(2^{-8}-2^{-4})^2\frac{2}{1+a}=\frac{75}{2^{17}(1+a)}$$

Für a=1/2 nimmt die Varianz am Eingang den Wert 0,00029, am Ausgang 0,00038 an. Die Standardabweichung einer Größe am Ausgang ist damit $\sigma^{1/2}=0.0195$, während die kleinste auflösbare Zahl am Ausgang bei einer 4-Bit-Genauigkeit $1/16=0.0625$ beträgt. Die Standardabweichung am Ausgang beträgt damit 31% der Größe der kleinsten auftretenden Zahl! Dies ist eine Konsequenz des sehr groben Schneidens von 8 auf 4 Bit. □

4.3.7 Grenzzyklusschwingungen

Bisher haben wir gesehen, dass die Quantisierung zu Rauschen führt, welches sich am Ausgang eines Signalverarbeitungssystems ggf. verstärkt. Ebenfalls wurde festgestellt, dass das Rauschen durch zweierlei Effekte hervorgerufen wird: Kleinsignalrauschen durch Runden oder Abschneiden von niederwertigen Bits sowie Großsignalrauschen durch Überlauffehler in den höherwertigen Bits. Wir wollen nun allgemeine Filterstrukturen hinsichtlich der quantisierungsbedingten Rauscheffekte untersuchen.

4.3.8 Kleinsignalrauschen - Unterschreiten von Quantisierungsstufen

Betrachten wir zunächst das einfache rückgekoppelte System aus Abb. 4.8. Es enthält einen digitalen Multiplizierer, den wir der besseren Beschreibung wegen in einen idealen Multiplizierer mit dem Verstärkungsfaktor A mit unbeschränkter Auflösung und einen nachgeschalteten Quantisierer Q aufteilen.

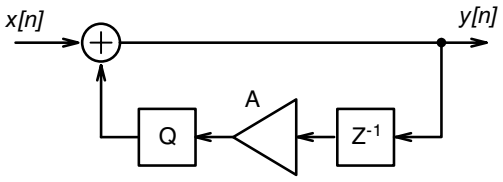


Abbildung 4.8. Rückgekoppeltes System erster Ordnung

Das System lässt sich beschreiben, indem wir die einzelnen Elemente in der Reihenfolge entgegen der Pfeilrichtungen ablesen:

$$y(n) = x(n) + Q[ay(n-1)] \tag{4.65}$$

Dabei möge $|a| < 1$ gelten, damit das System begrenzte Amplituden behält.

Wir wollen nun zunächst an einem Beispiel untersuchen, wie sich dieses System bei Erreichen der Quantisierungsstufe verhält. Dazu setzen wir $x[n] = 0$, laden aber die Speicher des Systems vor Beginn der zeitlichen Betrachtung. Es sei $a = -1/2$, und Q ein Rundungsquantisierer mit $b = 5 + 1$ Bit im Vorzeichen-Betrag-Code. Wir setzen $y[-1] = 5/64$. Dann ergibt sich der folgende Zeitverlauf:

n	$ay(n-1)$ dezimal	binär	$Q[ay(n-1)]$ binär	dezimal
0	-5/64	1.000101	1.00011	-3/32
1	+3/64	0.000011	0.00010	+1/16
2	-1/32	1.000010	1.00001	-1/32
3	+1/64	0.000001	0.00001	+1/32
4	-1/64	1.000001	1.00001	-1/32
5	...	...	...	...

Es ergibt sich also eine Schwingung der Periode 2 in der Höhe der Quantisierungsstufe $Q = 2^{-5}$. Das Verhalten weiterer Systeme vom Typ der Abb. 4.8 wird in Übung 4.3 untersucht.

Wir können jetzt bezüglich des Systemverhaltens folgende Fälle unterscheiden. Dabei möge nach wie vor $|a| < 1$ gelten, damit das System begrenzte Amplituden behält.

Theorem 4.7 (Kleinsignalrauschen).
Systeme vom Typ der Abb. 4.8 zeigen folgendes Verhalten:

a) *Der Multiplikationsfaktor ist $|a| \geq 1/2$. Dann kommt es zu Grenzzyklusschwingungen. Dies liegt daran, dass der letzte noch exakt darstellbare Wert (in Höhe der Quantisierungsstufe), mit a multipliziert und wieder gerundet, vom Betrag her keine Änderung erfährt. Es ergibt sich also*
   - *für $a > 1/2$ ein konstanter Endwert, also eine Grenzzyklusschwingung der Periode 1.*
   - *für $a < -1/2$ eine Grenzzyklusschwingung der Periode 2.*

b) *Der Multiplikationsfaktor ist $|a| < 1/2$. Dann sinkt der Ausgabewert auf Null und verbleibt dort. Dies liegt daran, dass der Betrag des letzten noch exakt darstellbaren Wertes (in Höhe der Quantisierungsstufe) durch die Multiplikation auf (exakt) weniger als seine Hälfte reduziert wird, was nach Runden Null ergibt.*

Gegenmaßnahmen zu diesen Effekten sind nur in beschränktem Maße möglich. Durch Erhöhen der verfügbaren Wortbreite wird das Problem zunächst vermieden, taucht dann aber bei entsprechend höherer Auflösungsstufe wieder auf. Falls möglich, sollte der Faktor a geeignet gewählt werden. Am Beispiel einer gedämpften Schwingung sei dies illustriert:

Ist man an der Amplitude, also an der Einhüllenden der Schwingung interessiert, so soll diese für lange Zeiten gegen Null laufen. Dies wird mit einem Faktor $|a| < 1/2$ erreicht, man erkauft sich dies allerdings durch das abrupte Abbrechen der Schwingung ab der Quantisierungsstufe.

Ist man an der Schwingungseigenschaft interessiert, ist ein Faktor $a < -1/2$ zu wählen. Das System schwingt dann auch für lange Zeiten, allerdings wird das Abklingen der Amplitude ab der Quantisierungsstufe nicht mehr sichtbar.

4.3.9 Großsignalrauschen - Überlaufeffekte

Wir wollen nun das umgekehrte Problem zu dem gerade dargestellten, nämlich das *Überschreiten* des Zahlenbereichs, untersuchen. Dazu betrachten wir als Beispiel einen Addierer. Er möge sich darstellen lassen als idealer Addierer (wir könnten auch ein anderes Element betrachten) in Festkommadarstellung, der allerdings keinen Überlauf hat. Der Überlaufswert wird also einfach abgeschnitten, wodurch der Addierer eine sägezahnförmige Kennlinie f(S) als Funktion der Summe S hat. Dazu geben wir ein Beispiel für den Überlauf im Vorzeichen-Betrag-Code:

$$\begin{array}{r} 0.\,1\,1\,1\,1\,1 \\ +\,0.\,0\,0\,0\,0\,1 \\ \hline =\,1.\,0\,0\,0\,0\,0 \end{array}$$

Der Addierer springt also von $1-Q$ bei Addition von Q auf -1. Dies stellt sich wie Abb. 4.9(b) abgebildet dar.

Wir betrachten nun das System 2. Ordnung aus Abb. 4.9(a) und wollen untersuchen, ob diese sägezahnförmige Kennlinie $f(S)$ zu Schwingungen oder ähnlichen Effekten führen kann.

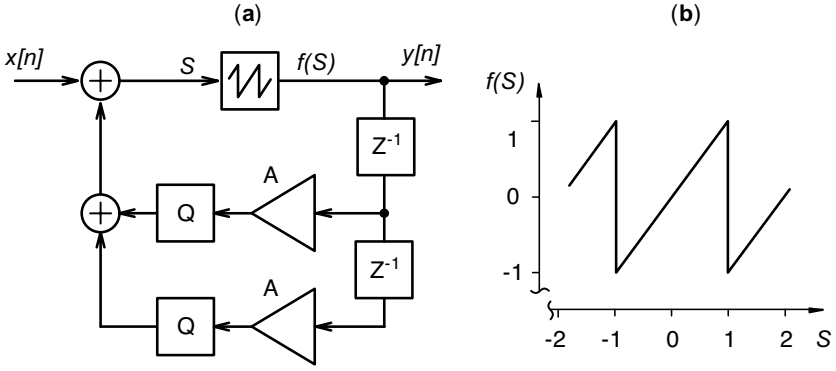


Abbildung 4.9. Rückgekoppeltes System zweiter Ordnung und die Kennlinie eines Elementes mit Überlauf

Dieses System wird beschrieben durch die Gleichung

$$y(n) = f(x[n] + ay[n-1] + by[n-2]) \tag{4.66}$$

Wir nehmen zur Vereinfachung wieder ein System mit $x[n] = 0$ an, dessen Speicher aber zu Beginn der Betrachtung ($n = 0$) von Null verschieden sein mögen. Dann gilt:

$$y(n) = f(ay[n-1] + by[n-2]) \tag{4.67}$$

Da $|y| < 1$, folgt für das Argument der Funktion f:

$$|ay[n-1] + by[n-2]| \leq |a| + |b| \tag{4.68}$$

Eine hinreichende Bedingung für das Vermeiden von Addierer-Überläufen ist es also, das Sägezahnverhalten von f völlig zu vermeiden. Dies gelingt mit

$$|a| + |b| < 1 \tag{4.69}$$

Wir wollen nun untersuchen, ob es zu Schwingungen kommen kann. Dazu machen wir den Ansatz

$$y[n] = (-1)y[n-1] \tag{4.70}$$

und untersuchen, ob diese Bedingung erfüllt werden kann. Einsetzen in die Systemgleichung liefert

$$y(n) = f((-a+b)y[n]) \tag{4.71}$$

Die Sägezahnkennlinie liefert darüber hinaus den selben Wert für alle um ganzzahlige Vielfache von 2 verschobene Eingangswerte, so dass auch gilt:

$$y(n) = f((-a+b)y[n] + 2k) \tag{4.72}$$

Es kann nun immer ein k^* so gewählt werden, dass das Argument der Kennlinie zwischen -1 und 1 liegt, in diesem Fall gilt $f(x) = x$ und damit

$$y(n) = (-a+b)y[n] + 2k^* \tag{4.73}$$

also

$$y(n) = \frac{2k^*}{1+a-b} \tag{4.74}$$

wobei $k^* \neq 0$ sein muss, damit $y[n]$ nicht identisch Null ist. Unser Ansatz, mit dem $y[n]$ eine Schwingung durchführt ist also erfüllbar.

Wir wollen nun wissen, welche Bedingungen die Koeffizienten erfüllen müssen. Da $|y[n]| < 1$ gilt, folgt

$$|1 + a - b| > 2|k^*| \tag{4.75}$$

bzw.

$$a - b > 2|k^*| - 1 \text{ oder } a - b < -2|k^*| - 1 \tag{4.76}$$

und da $k^* \neq 0$, gilt ferner

$$a - b > 1 \text{ oder } a - b < -3 \tag{4.77}$$

Übung 4.4 verifiziert, dass die letztere Bedingung auf ein instabiles System führt. Es bleibt die erste Bedingung, und wir fassen noch einmal zusammen:

Theorem 4.8 (Addierer-Überlaufschwingungen).
Addierer-Überlaufschwingungen bei Systemen 2. Ordnung treten nicht auf bei $|a| + |b| < 1$. *Bei* $a - b > 1$ *treten Addierer-Überlaufschwingungen der Periode 2 auf.*

Beispiel 4.7 – Addierer-Überlaufschwingungen.

Sei $a = 5/4$ und $b = -3/4$, und ein System mit auf Null gesetztem Eingang gegeben mit den Anfangswerten $y[-1] = 2/3$ und $y[-2] = -2/3$. Betrachten Sie die Möglichkeit von Addierer-Überlaufschwingungen!
Lösung:
Wir untersuchen die Bedingung $|a| + |b| < 1$ und stellen fest, dass diese wegen $|a| + |b| = 2 > 1$ verletzt ist. Es können also Addierer-Überlaufschwingungen auftreten. Wir untersuchen, ob für diese unser Ansatz gilt. Mit $a - b = 2 > 1$ ist die Bedingung für Addierer-Überlaufschwingungen der Periode 2 erfüllt mit $k^* = 1$. Gemäß

$$y[n] = \frac{2k^*}{1 + a - b} = 2/3$$

stellen sich also Grenzzyklusschwingungen der Amplitude $2/3$ ein.
Wir verifizieren dies in der folgenden Tabelle:

n	$y[n-1]$	$y[n-2]$	$S = \frac{5}{4}y[n-1] - \frac{3}{4}y[n-2]$	$y[n] = f(S)$
0	2/3	- 2/3	4/3	- 2/3
1	- 2/3	2/3	- 4/3	2/3
2	2/3	- 2/3	4/3	- 2/3
...	...	...	...	...

In der Tat stellen sich also Grenzzyklusschwingungen der Periode 2 mit Amplitude $2/3$ ein. □

Wir fragen wieder, ob sich diese Addierer-Überlaufschwingungen vermeiden lassen. Wie aus der Herleitung ersichtlich, treten sie offensichtlich wegen der Periodizität der Kennlinie auf. Es liegt also nahe, die Kennlinie so zu modifizieren, dass Periodizität nicht mehr gegeben ist. Dies wird beispielsweise durch Sättigungs-Kennlinien (Abb. 4.10) erreicht. Man bezeichnet einen solchen Addierer üblicherweise als „Sättigungs-Addierer" (Saturation-Adder).

Anstelle eines Overflow-Fehlers, der – wie oben gezeigt – zu Großsignalrauschen führt, findet eine Begrenzung auf einen der beiden Sättigungswerte statt (clipping). Damit werden die Eingabewerte der Multiplikatoren ständig kleiner, als sie bei einem unbegrenzten, aber BIBO-stabilen System sein sollten. Die ständige Reduktion der Eingabewerte sorgt dafür, dass der Grenzwert $y[n] = 0$ für große n erreicht wird.

4.3.10 Abklingen von Grenzzyklus-Schwingungen

Wir geben hier eine anschaulich-analytische Behandlung des Verhaltens von Systemen mit der Kennlinie aus Abb. 4.10.

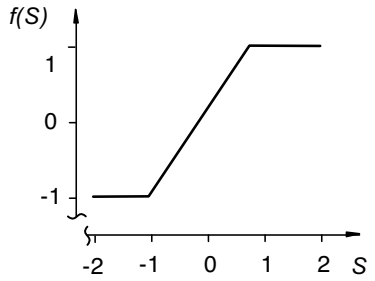


Abbildung 4.10. Sättigungskennlinie zur Vermeidung von Überlaufschwingungen

Die gesättigten Enden der Kennlinie zwingen die Summe S in Beträge kleiner als 1. Denn weit im gesättigten positiven Teil wären alle $y[n] \leq 1$ und damit $S < 1/2$, so dass der lineare Teil der Kennlinie in wenigen Schritten erreicht wird. Gleiches gilt weit im gesättigten negativen Teil.

Wir können das System 2. Ordnung nun für den Fall, dass nur der lineare Teil der Kennlinie benutzt wird, wie folgt beschreiben. Dazu definieren wir zwei Hilfsgrößen

$$v[n] = y[n-1] \quad \text{und} \quad w[n] = y[n-2] \tag{4.78}$$

Dann gilt:

$$\begin{aligned} v[n+1] &= av[n] + bw[n] \\ w[n+1] &= v[n] \end{aligned} \tag{4.79}$$

Der Vektor $[v[n+1], w[n+1]]^T$ ergibt sich also aus dem Vektor $[v[n], w[n]]^T$ durch Multiplikation mit der Matrix

$$A = \begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix} \tag{4.80}$$

Nach n Zeitschritten entspricht das einer Multiplikation mit A^n. Damit die Beträge der dann entstehenden Vektoren gegen 0 streben, muss gelten, dass alle Eigenwerte der Matrix A vom Betrag kleiner als 1 sind. Diese Eigenwerte lauten

$$\lambda = \frac{a}{2} \pm \sqrt{\frac{a^2}{4} + b} \tag{4.81}$$

Wir untersuchen die Beträge der Eigenwerte.

a) Für $a^2/4 + b > 0$ ist die Wurzel reell. Der Betrag des grössten Eigenwertes ist dann

$$|\lambda|_{\max} = \frac{|a|}{2} + \sqrt{\frac{a^2}{4} + b} \tag{4.82}$$

und damit muss gelten:

$$\frac{|a|}{2} + \sqrt{\frac{a^2}{4} + b} < 1 \tag{4.83}$$

woraus folgt $|a| + b < 1$. Die beiden Bedingungen für diesen Fall lauten also zusammengefasst: $-a^2/4 < b < 1 - |a|$. Da für alle $|a| \neq 2$ gilt: $-a^2/4 < 1 - |a|$, ergibt sich aus den beiden Schranken nie eine leere Menge.

b) Für $a^2/4 + b \leq 0$ ist die Wurzel imaginär. Beide Eigenwerte haben dann denselben Betrag,

$$|\lambda|^2_{\max} = -b \tag{4.84}$$

und damit muss gelten:

$$b > -1 \tag{4.85}$$

Die beiden Bedingungen für diesen Fall lauten also zusammengefasst: $-1 < b \leq -a^2/4$. Diese Schranken sind nur erfüllbar für $|a| < 2$.

Wir betrachten die Grenzen in den beiden o.a. Fällen und stellen fest: Falls $|a| < 2$, können beide Gebiete zusammengefasst werden: $-1 < b < 1 - |a|$. Falls $|a| \geq 2$, gilt nur der Fall a): $-a^2/4 < b < 1 - |a|$. Auch diese beiden Fälle können wir nun zusammenfassen in folgendem

Theorem 4.9 (Konvergenz bei Sättigungs-Addierern).
Bei Systemen 2. Ordnung mit Sättigungs-Addierern klingen Schwingungen bei beliebiger Anregung ab, und das System konvergiert für $n \to \infty$ zu $y[n] = 0$, wenn gilt:

$$\min(-1, -a^2/4) < b < 1 - |a| \tag{4.86}$$

Beispiel 4.8 – Konvergenz bei Sättigungs-Addierern.

Wir betrachten erneut das System aus Beispiel 4.7, diesmal mit der Kennlinie aus Abb. 4.10. Es ergeben sich folgende Werte:

n	$y[n-1]$	$y[n-2]$	$S = 5/4y[n-1] - 3/4y[n-2]$	$y[n] = f(S)$
0	2/3	- 2/3	4/3	1
1	1	2/3	3/4	3/4
2	3/4	1	4/3	3/16
3	3/16	3/4	- 21/64	- 21/64
...	...	...	...	...

Im weiteren Verlauf wird das Vorzeichen von $y[n]$ alternieren, der Betrag aber endgültig gegen 0 streben. Um dies zu zeigen, untersuchen wir die Verhältnisse des Theorems 4.9 für die Werte $a = 5/4$ und $b = -3/4$. Es ist zu zeigen
$\min(-1, -a^2/4) = -1 < b = -3/4 < 1 - |a| = -1/4$.
Dies ist offensichtlich erfüllt. Damit ist gezeigt, dass, von endlichem „Anstoßen" an den Sättigungsbereich zu frühen Zeitpunkten abgesehen, $y[n] = 0$ für große n erreicht wird.

□

Eine weiterführende Betrachtung ist mit linearen Differenzengleichungen möglich, auf die wir im Kapitel 5 noch ausführlich eingehen werden.

4.4 Rekonstruktion

Das bereits erwähnte Abtasttheorem behauptet, dass sich aus der diskreten Folge von Abtastwerten das zugrundeliegende kontinuierliche Signal unter bestimmten Voraussetzungen verlustfrei wiedergewinnen, sprich in ein identisches analoges Signal zurück wandeln lässt. Auch wenn dies in den seltensten Fällen beabsichtigt ist, wird diese Umwandlung häufig als „Rekonstruktion“ bezeichnet. Man meint damit den Prozess, der das diskrete Signal anhand seiner diskreten Werte in den kontinuierlichen Signalverlauf eines analogen Signals umwandelt. Mit diesem Schritt schließt sich die Kette des von uns betrachteten digitalen Systems zur Signalverarbeitung in einer analogen Umgebung, womit wir gleichzeitig zum letzten Thema dieses Kapitels kommen.

Um ein digitales Signal in ein analoges Signal umzusetzen, verdeutliche man sich wiederum, dass das digitale Signal gegenüber dem analogen Signal in zweierlei Hinsicht quantisiert wurde: erstens in der Amplitude und zweitens in der Zeit.

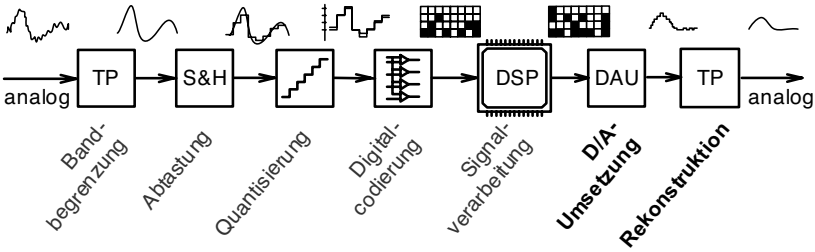
Amplitudenquantisierung

Weiter oben haben wir bereits festgestellt, dass die Amplitudenquantisierung als prinzipbedingtes Rauschen anzusehen ist und die dabei entfernte Signalinformation daher nicht in irgend einer Weise rekonstruiert werden kann. Man kann dem dabei entstehenden Fehler wie gezeigt lediglich mit größeren Wortbreiten oder geeigneten Codierungsverfahren begegnen. Dieser Fehler überträgt sich unvermeidlich auf das analoge Signal, das aus dem digitalen erzeugt werden soll. Durch ausreichenden Signal-Störabstand oder geeignete Modulationsverfahren kann vermieden werden, dass dabei Nutzinformation verloren geht.

Zeitliche Quantisierung

Wie sieht es jedoch mit der zeitlichen Quantisierung aus? Auch hier ging Information verloren, wobei sich das dadurch entstehende Rauschen an anderer Stelle zeigt, nämlich gemäß dem Abtasttheorem in den höherfrequenten Signal-Anteilen. Es kann daher auch hier vermieden werden, dass eine üblicherweise bandbegrenzte Nutzinformation davon beeinflusst wird.

Wir werden nun darauf eingehen, wie die Rekonstruktion eines diskreten Signals realisiert werden kann und gehen dabei auch hier sowohl auf die mathematische Beschreibungsmöglichkeit, als auch auf die praktischen Belange dieses Prozesses ein.



4.4.1 Analogwerte aus digitalen Codeworten

Am Beginn dieses Kapitels wurden die Verarbeitungsschritte erläutert, die vom analogen Signal zum codierten diskreten Signal führen. Kehrt man die Reihenfolge dieser Schritte um, erhält man – wer hätte es gedacht – den theoretischen Weg zurück zum analogen Signal. Allerdings sind die dazu verwendeten Komponenten wiederum anders gestaltet, weswegen der Rekonstruktion ein eigener Abschnitt gewidmet wurde.

Zunächst ist die Codierung des diskreten Signals aufzuheben. Aus dem Code ist dafür wieder der entsprechende „Wert“ zu ermitteln, der ursprünglich damit codiert werden sollte.

Weiter geht es mit einer De-Quantisierung (dies ist aufgrund des eher theoretischen Konstrukts kein gängiger Fachbegriff). Der „Wert“, der lediglich als Index einer Quantisierungsstufe aufzufassen ist, wird dabei wieder in die zugeordnete physikalische Größe auf dem Signalträger gewandelt.

Ist der Signalträger des analogen Signals beispielsweise eine Spannung und das digitale Signal binär codiert, so können die beiden eben genannten Schritte durch selektive Addition von Referenzspannungen oder -strömen erreicht werden. Dadurch, dass in Digitalsystemen üblicher Weise auch die Binärwerte durch Spannungen repräsentiert werden, besteht ein einfaches Verfahren darin, die einzelnen Bitleitungen b_N durch je einen einen Widerstand der Größe $R_N = R_q \cdot 2^N$ einen Strom erzeugen zu lassen und alle Ströme über einen Operationsverstärker o.ä. zu addieren.

Soweit zur technischen Lösung. Damit sind wir nun in der Lage, aus einem digitalen Codewort in beliebiger Weise einen analogen Signalwert zu erzeugen. Der nun folgende Schritt ist die Umkehrung des Abtastprozesses. Aus einer diskreten Folge analoger Werte ist also ein kontinuierliches Signal zu erzeugen, das insbesondere die spektrale Information erhalten soll. Hierfür existieren verschiedene Verfahren, die im Folgenden genauer untersucht werden.

4.4.2 Rekonstruktion durch Tiefpass

Im Kapitel 4.2 haben festgestellt, dass das Spektrum eines Signals durch den Abtastprozeß periodisch wird. Weiterhin konnten wir konstatieren, dass das Basisbandspektrum des Originalsignals bei ausreichend großer Abtastfrequenz

identisch ist mit dem Spektrum des abgetastete Signals. Ein Verfahren zur Rekonstruktion wäre somit der Einsatz eines Tiefpasses, der Frequenzen unterhalb der höchsten im Signal vorkommenden Frequenz unverändert überträgt und oberhalb der halben Abtastfrequenz vollständig sperrt.

Je näher diese beiden Eckfrequenzen beieinander liegen, desto schwieriger wird die technische Realisierung dieses analogen Filters. Oftmals wird deshalb im diskreten Bereich ein Ausgangssignal mit höherer Abtastfrequenz erzeugt. Fehlende Werte werden einfach durch Wiederholung oder Interpolation ergänzt.

Aufgrund der Einfachheit des Verfahrens wird es von nahezu allen praktischen Digital–Analog-Umsetzern verwendet. Mit anderen Verfahren, so auch mit dem im Folgenden beschriebenen, können die Rekonstruktionsfehler vermindert werden. Dafür wird jedoch zwischen zwei oder mehreren Abtastwerten interpoliert, was im Allgemeinen einen höheren Aufwand erfordert und zu einem größeren zeitlichen Versatz zwischen digitalem und analogen Signal führt.

4.4.3 Rekonstruktion bei unendlicher Folgenlänge

Zur Rekonstruktion eines Signals aus den Abtastwerten gehen wir von der inversen Fouriertransformierten

$$x(t) = \frac{1}{2\pi} \int_{\omega=-\infty}^{\infty} X(\omega) e^{j\omega t} d\omega \tag{4.87}$$

aus. Nach Voraussetzung des Abtasttheorems ist die Fouriertransformierte bandbegrenzt, und innerhalb des Bandes sind Fouriertransformierte des analogen und des abgetasteten Signals bis auf einen Faktor T identisch, so dass auch gilt

$$\begin{aligned} x(t) &= \frac{T}{2\pi} \int_{\omega=-\Omega/2}^{\Omega/2} X_n(\omega) e^{j\omega t} d\omega \\ &= \frac{1}{\Omega} \int_{\omega=-\Omega/2}^{\Omega/2} \sum_{m=-\infty}^{\infty} x[m] e^{-jm\omega T} e^{j\omega t} d\omega \end{aligned} \tag{4.88}$$

Nach Ausführung der Integration erhalten wir

$$\begin{aligned} x(t) &= \frac{1}{\Omega} \sum_{m=-\infty}^{\infty} x[m] \, \frac{1}{j(t-mT)} e^{j\omega(t-mT)} \Big|_{-\Omega/2}^{\Omega/2} \\ &= \sum_{m=-\infty}^{\infty} x[m] \frac{\sin((t-mT)\pi/T)}{(t-mT)\pi/T} \end{aligned} \tag{4.89}$$

und unter Verwendung von $\mathrm{sinc}(x) = \sin(x)/x$ den Rekonstruktionssatz:

Theorem 4.10 (Rekonstruktionssatz).

$$x(t) = \sum_{m=-\infty}^{\infty} x[m] \operatorname{sinc}\left(\frac{\pi(t - mT)}{T}\right) \tag{4.90}$$

Die Funktion $\operatorname{sinc}(x)$ ist bekannt als Fouriertransformierte eines Rechteckfensters (dort wegen der Bandbeschränkung). Sie nimmt dabei den Wert 1 bei $x = 0$ an und hat das in Abb. 4.11 gezeigte Aussehen. Insbesondere bei $x(t) = 1$ für alle t erhalten wir aus dem Rekonstruktionssatz für alle t die Beziehung

$$1 = \sum_{m=-\infty}^{\infty} \operatorname{sinc}\left(\frac{\pi(t - mT)}{T}\right) \tag{4.91}$$

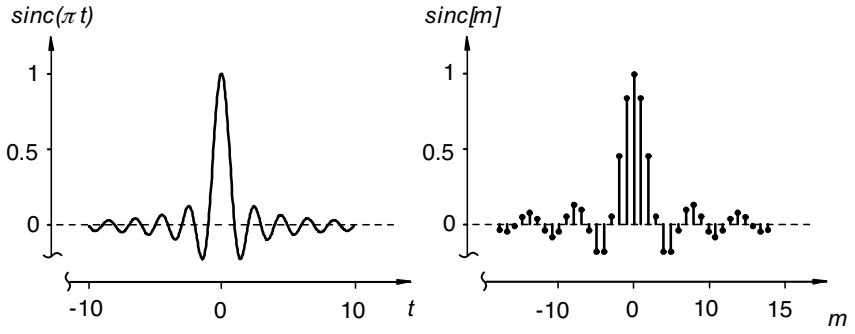


Abbildung 4.11. Verlauf der Funktion $\operatorname{sinc}(x) = \sin(x)/x$

Wir erkennen daraus anschaulich, dass eine Überlagerung unendlich vieler sinc-Funktionen, zentriert zu beliebigen Zeitpunkten t, eine Konstante ergibt. So ist es verständlich, dass der Rekonstruktionssatz tatsächlich ein „glattes" Aussehen der ursprünglichen Funktion wieder herstellt.

Wir wollen uns an einem Beispiel überlegen, dass diese Rekonstruktion bei Einhaltung der Bandbeschränkung das richtige Resultat liefert bzw. sonst zu aliasing führt.

Beispiel 4.9 – Effekte des Aliasing.

Wir betrachten die abgetastete Folge $x[m] = \exp(j\,\nu\,mT)$ mit einer noch unbestimmten Frequenz ν. In der Integralformulierung des Rekonstruktionssatzes wird daraus eine kontinuierliche Funktion

$$\begin{aligned} x(t) &= \frac{1}{\Omega} \int_{\omega=-\Omega/2}^{\Omega/2} \sum_{m=-\infty}^{\infty} x[m] e^{-jm\omega T} e^{j\omega t} d\omega \\ &= \frac{1}{\Omega} \int_{\omega=-\Omega/2}^{\Omega/2} \sum_{m=-\infty}^{\infty} e^{jm(v-\omega)T} e^{j\omega t} d\omega \end{aligned}$$

Zur Ausführung der Summe benutzen wir die bereits hergeleitete Identität (4.13) (siehe Theorem 4.1 auf Seite 92)

$$\begin{aligned} x(t) &= \int_{\omega=-\Omega/2}^{\Omega/2} \sum_{m=-\infty}^{\infty} \delta(v-\omega-m\Omega) e^{j\omega t} d\omega \\ &= e^{jvt} e^{-jk\Omega t} \end{aligned}$$

wobei $\Omega/2 < \nu - k\Omega < \Omega/2$ wegen der Impuls-Funktion. Liegt also ν in dem durch die Abtastung vorgegebenen Band, so erhalten wir in der Tat $k = 0$ und damit die zu der Folge $x[m] = \exp(j\,\nu\,mT)$ gehörige Funktion $x(t) = \exp(j\,\nu\,t)$. Liegt ν nicht in diesem Band, so kommt es zu einer Frequenzverschiebung in das durch die Abtastung vorgegebene Band hinein, also genau zu dem so genannten „aliasing". Die Frequenzverschiebung erfolgt dabei immer um Vielfache von Ω, woran man den Einfluss der Abtastung auf das Aliasing deutlich sieht. □

4.4.4 Rekonstruktion bei endlicher Folgenlänge

Nach dem Rekonstruktionssatz (Theorem 4.10) müssen alle Folgenglieder von $-\infty < t < \infty$ vorhanden sein. In der Praxis kennt man jedoch nur eine endliche Folge von Messwerten oder möchte sich aus rechentechnischen Gründen auf endliche Folgenlänge beschränken. Wir führen daher ein *Fenster* (engl. „window") p(k) ein. Dies ist eine Folge, die die Folge beschneidet und ggf. die Folgenglieder gewichtet. Das einfachste Fenster ist ein Rechteckfenster mit

$$p(k) = \begin{cases} 1 & 0 \le k < N \\ 0 & \text{sonst} \end{cases} \tag{4.92}$$

Eine Multiplikation von $x[k]$ mit dieser Fensterfunktion schneidet also N Folgenglieder aus.

Wir interessieren uns für den Zusammenhang zwischen dem Spektrum der unendlich langen und dem der gefensterten Folge. Aus der unendlich langen Folge

$$X_n(\omega) = \sum_{n=-\infty}^{\infty} x[n] e^{-j\omega nT} \tag{4.93}$$

erhalten wir nach Fensterung

$$X_n^p(\omega) = \sum_{n=-\infty}^{\infty} x[n] p[n] e^{-j\omega nT} \tag{4.94}$$

Wir ersetzen die Folge $x[n]$ mit Hilfe der inversen Fouriertransformation für $t = nT$

$$X_n^p(\omega) = \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} X_n(\nu) e^{j\nu nT} p[n] e^{-j\omega nT} d\nu$$
$$= \int_{-\infty}^{\infty} X_n(\nu) \sum_{n=-\infty}^{\infty} p[n] e^{-j(\omega-\nu)nT} d\nu \tag{4.95}$$

Darin ist die letzte Summe der Frequenzgang der Fensterfolge, den wir mit $P((\omega - \nu))$ bezeichnen. Es ergibt sich

$$X_n^p(\omega) = \int_{-\infty}^{\infty} X_n(\nu) P((w - \nu)) d\nu = X_n(\omega) * P(\omega) \tag{4.96}$$

d.h. der Frequenzgang der gefensterten Folge entsteht aus der Faltung der unendlichen Folge mit dem Frequenzgang der Fensterfunktion. Dies entspricht dem

Theorem 4.11 (Rekonstruktionssatz bei gefensterter Folgenlänge).

$$x(t) = \sum_{m=-\infty}^{\infty} x[m] p[m] \operatorname{sinc}\left(\frac{\pi(\mathrm{t} - \mathrm{mT})}{\mathrm{T}}\right) \tag{4.97}$$

Für eine Rechteckfolge ergibt sich[1]

$$P(\omega) = \sum_{n=-\infty}^{\infty} p[n] e^{-j\omega nT} = \sum_{n=0}^{N-1} e^{-j\omega nT} = e^{-j\omega(N-1)T/2} \frac{\sin(N\omega T/2)}{\sin(\omega T/2)} \tag{4.98}$$

und damit

$$X_n^p(\omega) = X_n(\omega) * \left(e^{-j\omega(N-1)T/2} \frac{\sin(N\omega T/2)}{\sin(\omega T/2)}\right) \tag{4.99}$$

sowie

$$x(t) = \sum_{m=0}^{N-1} x[m] \operatorname{sinc}\left(\frac{\pi(t - mT)}{T}\right) \tag{4.100}$$

Der Betrag der Funktion $P(\omega)$ für die Rechteckfolge ist in Abb. 4.12 dargestellt. Alle Maxima der Beträge der Funktion $\sin(Nx)/\sin(x)$ haben allgemein den Wert N.

Für große N nähert sich der Betrag der Funktion $P(\omega)$ für die Rechteckfolge immer mehr einer Summe von Delta-Funktionen für $\omega = k\Omega$ an, so dass das Spektrum immer besser rekonstruiert werden kann.

[1] Zur Berechnung der Summe der Exponentiale und der Beträge der Maxima in Abb. 4.12 siehe die Rechnung beginnend mit Gl. 6.24 auf Seite 157.

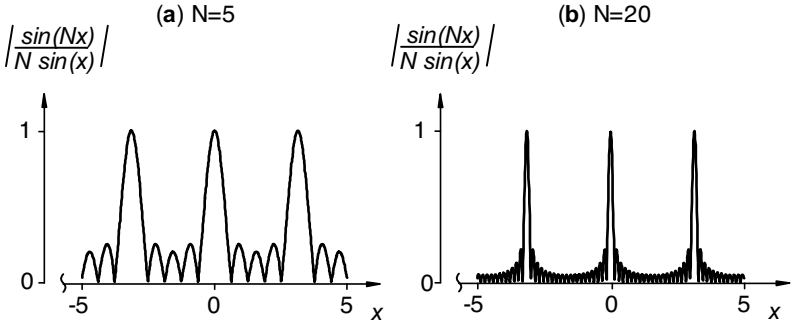


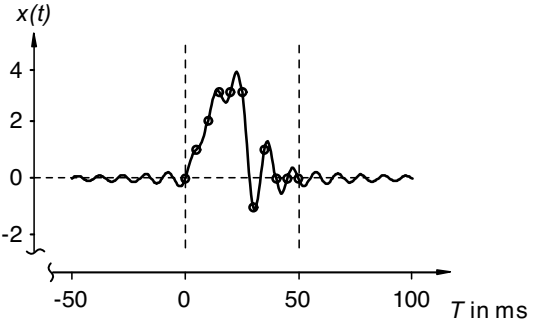
Abbildung 4.12. Verlauf des Betrages der Funktion $sin(Nx)/N \sin(x)$ mit (**a**) $N = 5$ und (**b**) $N = 20$

Beispiel 4.10 – Rekonstruktion endlicher Folgen.

Die Folge $x[m] = \{\underline{0}; 1; 2; 3; 3; 3; -1; 1; 0; 0; 0\}$ soll mit Hilfe des Rekonstruktionssatzes in ein kontinuierliches Signal gewandelt werden. Das Abtastintervall betrage dabei pro Folgenwert $5ms$.

Lösung:

Die folgende Abbildung zeigt das Ergebnis. Die Kreise kennzeichnen die durch die diskrete Folge vorgegebenen Stützwerte, durch die das rekonstruierte Signal verläuft. Die senkrechten Strich-Linien bezeichnen das Rechteckfenster, ausserhalb dessen die wahren Funktionswerte Null werden. Die Interpolationseigenschaft der sinc-Funktion bewirkt auch für die rekonstruierte Funktion dort einen Abfall auf Null.



□

4.4.5 Andere Rekonstruktionen des analogen Signals

Es wurde bereits festgestellt, dass die Wandlung eines diskreten Signals in ein kontinuierliches Signal einer Interpolation zwischen Stützstellen gleichkommt. Je nach Anwendungsfall eignen sich dafür verschiedene Verfahren. Wir geben hier nur ein Beispiel für eine andere, näherungsweise Rekonstruktion des Abtastsignals (B: Bandbreite des Signals $x(t)$):

$$x(t) = \sum_{n=-\infty}^{\infty} x\left(\frac{n}{2B}\right) \operatorname{sinc}[\pi(2Bt - n)] \tag{4.101}$$

Dieses Ergbnis folgt aus dem Rekonstruktionssatz (Theorem 4.10) unter Beachtung von $x[n] = x(nT)$ und nach Setzen von $T = 1/(2B)$.

Übungen

Übung 4.1 – Symmetrie der Fouriertransformation.

Zeigen Sie, dass die Fouriertransformierte eines rein imaginären Signals $jx(t)$ eine konjugiert ungerade Funktion von ω ist
a) unter Benutzung von Gl. (4.4)
b) durch Benutzen der Definition der Fouriertransformierten
Orientieren Sie sich an der Vorgehensweise in Beispiel 4.1.

Übung 4.2 – Symmetrie des Betrages der Fouriertransformierten.

Zeigen Sie, dass der Betrag der Fouriertransformierten eines Signals mit konstanter Phase eine gerade Funktion von ω ist
a) unter Benutzung von Gl. (4.4)
b) durch Benutzen der Definition der Fouriertransformierten
Orientieren Sie sich an der Vorgehensweise in Beispiel 4.1.

Übung 4.3 – Rundungseffekte.

a) In dem System aus Abb. 4.8 sei der Multiplikationsfaktor $a = 1/2$. Erstellen Sie die o.a. Tabelle für diesen Fall. Wie verhält sich das System bei Erreichen der Quantisierungsstufe?
b) In dem selben System sei der Multiplikationsfaktor $a = 1/4$. Erstellen Sie die o.a. Tabelle für diesen Fall. Wie verhält sich das System bei Erreichen der Quantisierungsstufe?

Übung 4.4 – Bedingung für Schwingungen.

Verifizieren Sie, dass die in Gleichung (4.77) aufgeführte Bedingung $a - b > 1$ zu einem stabilen System führen kann. Unter welchen weiteren Randbedingungen ist das System stabil? Verifizieren Sie, dass die zweite hergeleitete Bedingung $a - b < -3$ zu einem instabilen System führt.

Übung 4.5 – Abtastung.

Beantworten Sie folgende vier Fragen:

1. Durch welche Formel lässt sich die ideale Abtastung beschreiben? Warum?
2. Was wird durch die Abtastung diskretisiert?

3. Tasten Sie ein selbst gewähltes Signal $x(t)$ mit der Abtastfrequenz f_A ab und transformieren Sie das Ergebnis in den Frequenzbereich.
4. Erstellen Sie eine Skizze von $X(j\omega)$ und zeigen Sie anhand ihres Graphen mit welcher Frequenz f_A mindestens abgetastet werden muss, damit $X(j\omega)$ eindeutig rekonstruierbar ist.

Übung 4.6 – Abtastung und Aliasing.

Berechnen Sie (oder schlagen Sie sie nach) die Fouriertransformierte der zeitlichen Normalverteilung

$$N(t) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{t^2}{2\sigma}\right)$$

Können Sie eine solche zeitliche Funktion abtasten, mit Hilfe der digitalen Signalverarbeitung weiter verarbeiten (oder auch einfach nur ohne Verarbeitung übertragen) und das zeitanaloge Verhalten rekonstruieren?

Welchen Effekt beobachten Sie, wenn Sie abtasten, und warum? Liesse sich dieser Effekt vermeiden? Liesse er sich wenigstens näherungsweise vermeiden, und worin besteht die Näherung? Welche Verfälschung ergibt sich damit?

Übung 4.7 – Zahlendarstellung.

Schreiben Sie die folgenden Dezimalzahlen in Einer- sowie in Zweier-Komplement-Darstellung: 3,625, -3,625

Übung 4.8 – Quantisierung.

1. Wie lautet die Definition der Wortbreite W?
2. Bestimmen Sie die Anzahl der Quantisierungsstufen.
3. Geben Sie zwei mögliche Formen der Quantisierung an.
4. Was ist der Quantisierungsfehler und wie ist er begrenzt?

Übung 4.9 – Quantisierungsfehler.

Stellen Sie die Wahrscheinlichkeitsdichtefunktion $p(e)$ für den Quantisierungsfehler e (weiße Rauschquelle) beim Runden und beim Abschneiden grafisch dar. Berechnen Sie den Mittelwert und die Varianz des Quantisierungsrauschens für diese beiden Fälle.

5

Differenzengleichungen

Bereits in Kap. 3.2 sind wir auf die Differenzengleichung als eine Beschreibungsform eines zeitdiskreten Systems eingegangen. Wenn man bei gegebener Anregung eine Systemreaktion berechnen will, muss die Differenzengleichung gelöst werden. Dies entspricht dem Ziel des Lösens einer Differentialgleichung im Bereich von zeitkontinuierlichen Systemen.

Um aus einer Differenzengleichung und gegebenen Anfangswerten eine Lösung für die Systemreaktion zu berechnen, können verschiedene Verfahren angewendet werden. Wir wollen hier

- eine direkte Lösung (in Analogie zu Differentialgleichungen),
- die Lösung über die einseitige Z-Transformation,
- die Lösung über Systemmatrizen

behandeln.

Zunächst sei auf folgende Merkmale von Differenzengleichungen hingewiesen:

1. Durch Abtastung eines kontinuierlichen Signals entstehen zeitdiskrete Folgen. Wir haben es also mit Differenzen (von Folgentermen), statt mit Differentialen (von Funktionen im analogen Bereich) zu tun.
2. Durch die LTI-Eigenschaften des Systems sind die Differenzengleichungen linear und zeitinvariant (invariant unter Verschiebung des Folgenindex). Insbesondere kommen Operationen wie die Multiplikation von Signalen untereinander nicht vor (im Gegensatz zu skalaren Multiplikationen mit einem konstanten Faktor, die der Verstärkung dienen).
3. Die „physikalischen“ Elemente von Differenzengleichungen sind Addierer, Multiplizierer, und Zeitverzögerer. Andere lineare Elemente wie Differentiatoren und Integratoren wurden hier nicht behandelt, für sie stehen Hilbert-Transformationen zur Verfügung
4. Die physikalischen Elemente der Differenzengleichungen lassen sich in Rechnern hervorragend abbilden.

Wir verwenden wie schon zuvor für die Blockschaltbild-Darstellung die Symbole in Abb. 5.1.

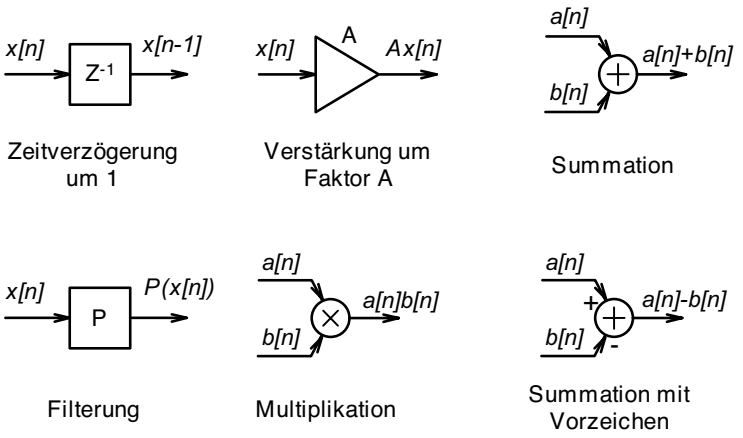


Abbildung 5.1. Blockschaltbilder in Differenzengleichungen

5.1 Direkte Lösung der Differenzengleichung

Als rückgekoppeltes System mit Verzögerungsglied und Multiplizierer sei das Beispiel in Abb. 5.2 betrachtet. Das System lässt sich beschreiben, indem

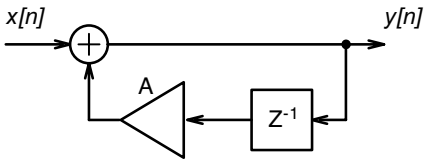


Abbildung 5.2. Rückgekoppeltes System

wir die einzelnen Elemente in der Reihenfolge entgegen den Pfeilrichtungen ablesen:

$$y[n] = x[n] + Ay[n-1]. \tag{5.1}$$

Damit das System begrenzte Amplituden behält, gelte $|A| < 1$. Die Anregung möge mit der Sprungfolge $x[n] = u[n]$ erfolgen.

Um eine direkte Lösung zu erreichen, bilden wir analog zur Differentialgleichung die homogene und partikuläre Lösung und verlangen wegen der Kausalität $y[n] = 0$ für alle $n < 0$. Die homogene Differenzengleichung lautet:

$$y_h[n] - Ay_h[n-1] = 0. \tag{5.2}$$

Analog zu Differentialgleichungen machen wir den Ansatz

$$y_h[n] = q^n. \tag{5.3}$$

Einsetzen von (5.3) in die homogene Differenzengleichung (5.2) liefert $q_h = A$ und damit die homogene Lösung:

$$y_h[n] = kA^n, \quad k = konst. \tag{5.4}$$

wobei die Konstante k noch zu bestimmen ist.

Die partikuläre Differenzengleichung lautet

$$y_p[n] - Ay_p[n-1] = u[n]. \tag{5.5}$$

Für die partikuläre Lösung lassen wir uns von der Beobachtung leiten, dass das System mit der Sprungfolge angeregt wird. Nach einer sehr langen Zeitdauer sollte die Lösung der Differenzengleichung eine Konstante sein. Wir machen also den Ansatz

$$y_p[n] = C, \tag{5.6}$$

setzen in die partikuläre Differenzengleichung ein und erhalten $C = \frac{1}{1-A}$.

Wir müssen dabei sicherstellen, dass $y[n] = 0$ für alle $n < 0$ ist. Speziell geben wir $y[-1] = 0$ vor und bilden die vollständige Lösung als Summe aus homogener und partikulärer Lösung

$$y[n] = kA^n + \frac{1}{1-A}. \tag{5.7}$$

Die offene Konstante k aus (5.5) bestimmen wir mit $y[-1] = 0$ zu $k = \frac{A}{A-1}$ und erhalten die Lösung:

$$y[n] = \frac{A^{n+1} - 1}{A - 1} u[n]. \tag{5.8}$$

Zur direkten Lösung von Differenzengleichungen siehe auch Übung 5.1.

5.2 Die einseitige Z-Transformation

Eine andere Möglichkeit, Differenzengleichungen zu lösen, ist die Verwendung der einseitigen Z-Transformation. Diese stellt nichts Neues im Vergleich zur normalen Z-Transformation dar, erlaubt uns jedoch, die Kausalität von Systemen direkt in die Transformationsgleichungen einzubeziehen.

Sei $f[n]$ eine kausale Folge, also $f[n] = 0$ für $n < 0$. Das System soll natürlich auch erst für $n \geq 0$ reagieren. Wir wollen dies in die Z-Transformation

einbauen. Im kausalen Bereich können wir dann allgemein die einseitige Z-Transformation als die (normale) Z-Transformation definieren:

$$F_e(z) = Z_e\{f[n]\} = \sum_{n=0}^{\infty} f[n]z^{-n}. \tag{5.9}$$

Bei einer Verschiebung der Eingangsfolge ergibt sich nun

$$Z_e\{f[n+m]\} = \sum_{n=0}^{\infty} f[n+m]z^{-n} = z^m \sum_{n=m}^{\infty} f[n]z^{-n}. \tag{5.10}$$

Bei der normalen Z-Transformation liegt die untere Grenze der Summe bei $-\infty$ und es hätte sich somit nichts geändert. Hier jedoch müssen wir die untere Grenze wieder separat auf Null transformieren, um die einseitige Z-Transformation auch auf der rechten Seite zu erhalten:

$$Z_e\{f[n+m]\} = z^m \sum_{n=m}^{\infty} f[n]z^{-n} = z^m \sum_{n=0}^{\infty} f[n]z^{-n} - \sum_{n=0}^{m-1} f[n]z^{m-n}$$

$$Z_e\{f[n+m]\} = z^m F_e(z) - \sum_{n=0}^{m-1} f[n]z^{m-n}. \tag{5.11}$$

Die einseitige Z-Transformation enthält also die Anfangswerte explizit.

5.3 Lösung der Differenzengleichung über einseitige Z-Transformation

Das Lösen der Differenzengleichung über die einseitige Z-Transformation betrachten wir an einem Beispiel, wodurch die Vorgehensweise klar werden soll. Wir wählen wieder wie oben mit $|A| < 1$:

$$y[n] - Ay[n-1] = u[n]. \tag{5.12}$$

Nach Anwendung der einseitigen Z-Transformation ergibt sich aus (5.12):

$$Y_e(z) - Az^{-1}Y_e(z) - Ay[-1] = \frac{z}{z-1} \quad \text{für} \quad |z| > 1. \tag{5.13}$$

Da $y[-1] = 0$ ist, erhalten wir

$$Y_e(z) = \frac{z^2}{(z-A)(z-1)}. \tag{5.14}$$

Die Rücktransformation führen wir z.B. mit Hilfe des Residuensatzes aus, wobei wir gleich sehen werden, dass keine komplizierten Rechnungen durchzuführen sind. Wir benötigen die Residuen der Funktion

$$z^{n-1}Y_e(z) = \frac{z^{n+1}}{(z-A)(z-1)}. \tag{5.15}$$

Da eine Kontur mit $|z| > 1$ zu wählen ist, und da $|A| < 1$, sind alle Pole in der Kontur enthalten. Für $n \geq -1$ erhalten wir einfache Pole bei 1 und A:

$$y[n] = \frac{A^{n+1}-1}{A-1}\,, \quad \text{speziell} \quad y[-1] = 0. \tag{5.16}$$

Für $n < -1$ erhalten wir einen zusätzlichen, $-(n+1)$-fachen Pol bei $z = 0$. Da dessen Residuum kompliziert zu berechnen ist, bilden wir gemäß dem Inversionssatz im $1/z$–Bereich (3.51) die Funktion

$$z^{-n-1}Y_e(1/z) = z^{-n-1}\frac{z^{-2}}{(1/z-A)(1/z-1)} = \frac{z^{-n-1}}{(1-Az)(1-z)}. \tag{5.17}$$

Die Residuen dieser Funktion sind nun für eine Kontur mit $|z| < 1$ zu bestimmen. Innerhalb solcher Konturen verbleibt der Pol bei Null. Diese Polstelle existiert lediglich für $n \geq 0$, bei allen anderen Werten mit $n < 0$ und Kontur mit $|z| < 1$ treten keine Pole mehr auf. In diesen Fällen ist $y[n] = 0$.

Fassen wir die Teillösungen für $n \geq -1$ und $n < -1$ zusammen, so erhalten wir nur von Null verschiedene Beiträge für $n \geq 0$:

$$y[n] = \frac{A^{n+1}-1}{A-1}u[n] \tag{5.18}$$

wie bereits zuvor in der direkten Herleitung aus Formel (5.1). Wir sehen, dass dieser Lösungsweg über die einseitige Z-Transformation die Anfangsbedingungen bereits richtig enthält. Darüber hinaus wird die Differenzengleichung auf eine systematische Art und Weise gelöst.

5.4 Lösung von Differenzengleichungssystemen

Wir betrachten nun das in Abb. 5.3 angegebene Beispiel, wobei $u(n)$ eine allgemeine Anregung sein soll (nicht die Sprungfolge).

Da es sich um ein System mit 2 Verzögerungsgliedern handelt, definieren wir die *Zustandsgrößen* (state variables) x_1 und x_2 zur Beschreibung des internen Systemzustandes. Diese Zustandsgrößen können 2 Funktionen haben:

1. bei Systemen, die schon unendlich lange Zeit laufen, beschreiben sie interne Systemzustände lediglich als Hilfsgrößen. Die Systemgleichung lässt sich dann – analog zu Differentialgleichungssystemen – entweder als Gleichung höherer Ordnung aufschreiben oder unter Verwendung der Hilfsgrößen als ein System von Gleichungen erster Ordnung.

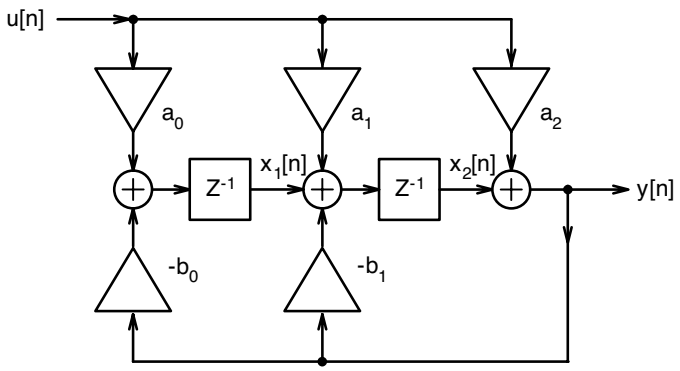


Abbildung 5.3. Beispielsystem 2. Ordnung

2. Bei (kausalen) Systemen, die ab dem Zeitpunkt $t = 0$ eingeschaltet werden, können zusätzlich zur Eingangsgröße $u[0]$ auch die Zustandsgrößen $\mathbf{x}[0]$ (die wir zusammengefasst als fett gedruckten Vektor schreiben) zum Zeitpunkt $t = 0$ von außen *gesetzt* werden. Wegen der Linearität des Systems sind die Zustandsgrößen im weiteren Verlauf Linearkombinationen aus der zeitlichen Entwicklung der Anfangsgrößen $\mathbf{x}[0]$ sowie der Systemreaktion auf die Eingangsfolge $u[n]$.

Wir werden beide Funktionen der Zustandsgrößen kennen lernen und mathematisch formulieren.

5.4.1 Systemgleichungen mit Zustandsgrößen

Die Systemgleichungen erhält man direkt durch Ablesen aus der Abb. 5.3. Wir erinnern uns, dass wir bei der Betrachtung der Addierer-Sättigung bereits ein ähnliches Gleichungssystem (Gln. 4.80 auf Seite 121) erhalten hatten:

$$\begin{pmatrix} x_1(n+1) \\ x_2(n+1) \end{pmatrix} = \begin{pmatrix} 0 & -b_0 \\ 1 & -b_1 \end{pmatrix} \begin{pmatrix} x_1(n) \\ x_2(n) \end{pmatrix} + \begin{pmatrix} a_0 - a_2 b_0 \\ a_1 - a_2 b_1 \end{pmatrix} u(n), \tag{5.19}$$

bzw. für $y(n)$:

$$y(n) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T \begin{pmatrix} x_1(n) \\ x_2(n) \end{pmatrix} + a_2 u(n). \tag{5.20}$$

Wir können nun Lösungswege für solche Gleichungssysteme angeben. Zunächst aber verallgemeinern wir das obige Beispiel auf Systeme höherer Ordnung. Für sie ergeben sich durch die höhere Ordnung M weitere Gleichungen, wobei die im folgenden benutzten Matrizen von der Größe $M \times M$ und die Vektoren von der Dimension M sind:

$$\mathbf{x}(n+1) = \mathbf{A}\mathbf{x}(n) + \mathbf{b}u(n) \tag{5.21}$$

$$y(n) = \mathbf{c}^T \mathbf{x}(n) + du(n). \tag{5.22}$$

Dabei ist **A** die sogenannte **Systemmatrix**. Alle fett gedruckten Kleinbuchstaben sind Vektoren.

Setzen wir die erste Gl. (5.21) wiederholt (n–mal) in sich selbst ein, erhalten wir:

$$\mathbf{x}(n) = \mathbf{A}^n \mathbf{x}(0) + \sum_{m=0}^{n-1} \mathbf{A}^{n-m-1} \mathbf{b} u(m) \tag{5.23}$$

und damit für den Ausgang eine geschlossene Lösung:

Theorem 5.1 (Ausgangsverhalten einer Differenzengleichung höherer Ordnung).

$$y(n) = \mathbf{c}^T \mathbf{A}^n \mathbf{x}(0) + du(n) + \mathbf{c}^T \sum_{m=0}^{n-1} \mathbf{A}^{n-m-1} \mathbf{b} u(m) \tag{5.24}$$

Der erste Summand beschreibt die Abhängigkeit vom Anfangszustand, während die folgenden Summanden den Einfluss der Eingangsfolge berücksichtigen. Die Eingangsfolge $u[m]$ wird dabei nach n Zeitschritten 0– bis $(n-1)$–mal mit der Systemmatrix multipliziert, wobei die am weitesten in der Vergangenheit liegenden Glieder der Eingangsfolge mit den höchsten Potenzen von A multipliziert werden.

Diese geschlossene Lösung für die Ausgangsfolge hat also den Nachteil, dass alle Werte der Eingangsfolge berücksichtigt werden müssen. Weiter unten werden wir aus der Übertragungsfunktion eine (nicht geschlossene) Differenzengleichung M-ter Ordnung herleiten, wie sie auch aus der Funktionsskizze direkt abgelesen werden kann. Diese Differenzengleichung enthält Rückkopplungen des Ausgangs zum Eingang, wodurch die Matrixpotenzierungen und das Verwenden aller Werte der Eingangsfolge beseitigt wird.

Aus der geschlossenen Lösung sieht man, dass Potenzen von Matrizen berechnet werden müssen. Dies wollen wir auf zwei Wegen tun.

5.4.2 Matrixpotenzierung über Eigenwerte

Wir benutzen eine Zerlegung der (M, M)-Matrix **A** mit Hilfe der *Eigenvektorgleichung*

$$\mathbf{A}\mathbf{v} - \lambda \mathbf{E}\mathbf{v} = \mathbf{0}. \tag{5.25}$$

Dabei ist **E** die **Einheitsmatrix** und **v Eigenvektor** zum zugehörigen **Eigenwert** λ. Dieses Gleichungssystem hat M Lösungen $(\lambda, \mathbf{v})$, wobei für Matrizen vom Rang $R < M$ eine freie Wahl der Basis in einem Untervektorraum vom Rang $R - M$ besteht. Die Eigenwerte λ ergeben sich aus der *charakteristischen Gleichung*

$$\det(\mathbf{A} - \lambda \mathbf{E}) = \mathbf{0}. \tag{5.26}$$

Die dazugehörigen normierten Eigenvektoren $\mathbf{v}$ ergeben sich nach Einsetzen der Eigenwerte in die Eigenvektorgleichung unter zusätzlicher Berücksichtigung von $|\mathbf{v}| = 1$. Auch andere Normierungen sind möglich, allerdings müssen alle Eigenvektoren auf denselben Wert normiert werden.

Wir schreiben nun die Eigenvektoren spaltenweise in eine Matrix $\mathbf{V}$ und die Eigenwerte in eine Diagonalmatrix $\mathbf{\Lambda}$. Dann gilt

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}; \quad \mathbf{V}\mathbf{V}^{-1} = \mathbf{E}. \tag{5.27}$$

Demzufolge bestimmen wir

Theorem 5.2 (Matrixpotenzierung über Eigenwerte).

$$\mathbf{A}^n = (\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1})^n = \mathbf{V}\mathbf{\Lambda}^n\mathbf{V}^{-1}. \tag{5.28}$$

$\mathbf{V}$ *ist dabei die Matrix der normierten Eigenvektoren (spaltenweise), und die Diagonalmatrix* $\mathbf{\Lambda} = \mathbf{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M)$ *enthält die Eigenwerte. Diese ergeben sich aus der Lösung der charakteristischen Gleichung* (5.26) $\det(\mathbf{A} - \lambda\mathbf{E}) = \mathbf{0}$. *Die Eigenvektoren ergeben sich aus der Lösung der* Eigenvektorgleichung (5.25)*:* $(\mathbf{A} - \lambda\mathbf{E})\mathbf{v} = \mathbf{0}$.

Wir können damit folgende Aufwandsabschätzung geben: Eine Matrix n-mal mit sich selbst zu potenzieren erfordert nM^2 Multiplikationen und nM^2 Additionen.

Im Gegensatz dazu benötigen wir für die Lösung einer Eigenwertaufgabe ca. M^2 Multiplikationen (sowie ggf. die aufwändige Lösung eines algebraischen Gleichungssystems M-ter Ordnung). Für die Potenzierung der Eigenwerte M Operationen, sowie zwei (!) Matrixmultiplikationen, also $2M^2$ Multiplikationen und $2M^2$ Additionen.

Insgesamt sind dies ca. $3M^2$ Multiplikationen und $3M^2$ Additionen sowie die Lösung eines algebraischen Gleichungssystems M-ter Ordnung. Man erkennt also, speziell wenn M klein ist (geschlossene Lösungen eines algebraischen Gleichungssystems M-ter Ordnung sind bis $M = 5$ möglich) und n groß, dass diese Methode eine enorme Aufwandsersparnis bringt.

Beispiel 5.1 – Matrixpotenzierung über Eigenwerte.

Wir betrachten erneut das in Abb. 5.3 skizzierte System. Die Systemmatrix $\mathbf{A}$ lautet:

$$\mathbf{A} = \begin{pmatrix} 0 & -b_0 \\ 1 & -b_1 \end{pmatrix}. \tag{5.29}$$

Wir wählen als Beispiel $b_0 = -3/4$ und $b_1 = 1$ und erhalten

$$\mathbf{A} = \begin{pmatrix} 0 & 3/4 \\ 1 & -1 \end{pmatrix}. \tag{5.30}$$

$\mathbf{A}^n$ wird über die Eigenwerte bestimmt. Dazu lösen wir

$$0 = \det(\mathbf{A} - \lambda\mathbf{E}) = \det\begin{pmatrix} -\lambda & 3/4 \\ 1 & -1-\lambda \end{pmatrix} = \lambda^2 + \lambda - 3/4 \qquad (5.31)$$

und erhalten $\lambda_1 = 1/2$ und $\lambda_2 = -3/2$. Die Eigenvektoren erhalten wir aus:

$$\mathbf{0} = (\mathbf{A} - \lambda\mathbf{E})\mathbf{v} = \begin{pmatrix} -\lambda & 3/4 \\ 1 & -1-\lambda \end{pmatrix},$$
$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -\lambda x + 3/4y \\ x + (-1-\lambda)y \end{pmatrix}. \qquad (5.32)$$

Aus dieser Vektorgleichung reicht es, eine (!) Zeile zu nehmen, um x und y (bis auf einen Normierungsfaktor) zu bestimmen. Die andere Zeile ist nämlich wegen der Nulldeterminante linear abhängig von der gewählten. (Im allgemeinen Fall von (M, M)-Matrizen reicht es, $(M-1)$ Zeilen zu benutzen.) Für die beiden Eigenwerte ergeben sich auf 1 normierte Eigenvektoren zu

$$\lambda_1 = 1/2: \quad \mathbf{v}_1 = \begin{pmatrix} \frac{3}{\sqrt{13}} \\ \frac{2}{\sqrt{13}} \end{pmatrix}; \quad \lambda_2 = -3/2: \quad \mathbf{v}_2 = \begin{pmatrix} \frac{1}{\sqrt{5}} \\ \frac{-2}{\sqrt{5}} \end{pmatrix}. \qquad (5.33)$$

Wir fassen Eigenwerte und Eigenvektoren zu Matrizen zusammen:

$$\mathbf{\Lambda} = \mathbf{diag}(\lambda_1, \lambda_2) = \begin{pmatrix} 1/2 & 0 \\ 0 & -3/2 \end{pmatrix}, \quad \mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2] = \begin{pmatrix} \frac{3}{\sqrt{13}} & \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{13}} & \frac{-2}{\sqrt{5}} \end{pmatrix}. \qquad (5.34)$$

Es gilt nun nach den Regeln der linearen Algebra:

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}, \qquad (5.35)$$

was genau die gesuchte Diagonalisierung der Matrix $\mathbf{A}$ ist. Im Falle symmetrischer Matrizen $\mathbf{A}$ kann die Inverse $\mathbf{V}^{-1}$ durch die Transponierte $\mathbf{V}^T$ ersetzt werden. Für $(2,2)$-Matrizen geben wir hier der Praktikabilität halber eine allgemeine, geschlossene Lösung zur Inversion an:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \frac{d}{\Delta} & \frac{-b}{\Delta} \\ \frac{-c}{\Delta} & \frac{a}{\Delta} \end{pmatrix} = \mathbf{E}\,, \quad \text{mit} \quad \Delta = \det(\mathbf{A}) = ad - bc. \qquad (5.36)$$

Wir erhalten

$$\mathbf{V}^{-1} = \begin{pmatrix} \frac{\sqrt{13}}{4} & \frac{\sqrt{13}}{8} \\ \frac{\sqrt{5}}{4} & \frac{-3\sqrt{5}}{8} \end{pmatrix} \qquad (5.37)$$

und damit

$$\mathbf{A}^n = \mathbf{V}\mathbf{\Lambda}^n\mathbf{V}^{-1} = \begin{pmatrix} \frac{3}{\sqrt{13}} & \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{13}} & \frac{-2}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} (1/2)^n & 0 \\ 0 & (-3/2)^n \end{pmatrix} \begin{pmatrix} \frac{\sqrt{13}}{4} & \frac{\sqrt{13}}{8} \\ \frac{\sqrt{5}}{4} & \frac{-3\sqrt{5}}{8} \end{pmatrix} \qquad (5.38)$$

und nach Ausführung der Matrixmultiplikationen:

$$\mathbf{A}^n = \begin{pmatrix} 3/4(1/2)^n + 1/4(-3/2)^n & 3/8(1/2)^n - 3/8(-3/2)^n \\ 1/2(1/2)^n - 1/2(-3/2)^n & 1/4(1/2)^n + 3/4(-3/2)^n \end{pmatrix}. \tag{5.39}$$

Man überzeugt sich, dass sich z.B. für $n = 0$ die Einheitsmatrix, für $n = 1$ wieder die Matrix $\mathbf{A}$ ergibt. Wird das System z.B. mit $x_0 = [0, 1]$ und ohne äußere Anregung angefahren, so ergibt sich

$$\mathbf{x}(n+1) = \mathbf{A}^n\mathbf{x}(0) = \begin{pmatrix} \frac{3}{\sqrt{13}} & \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{13}} & \frac{-2}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} (1/2)^n & 0 \\ 0 & (-3/2)^n \end{pmatrix} \begin{pmatrix} \frac{\sqrt{13}}{4} & \frac{\sqrt{13}}{8} \\ \frac{\sqrt{5}}{4} & \frac{-3\sqrt{5}}{8} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{5.40}$$

Wir können die Matrixmultiplikationen (von links nach rechts), die wir oben schon durchgeführt haben, benutzen. Wir können aber auch, falls wir nur am Endresultat interessiert sind, die im folgenden einmal detailliert durchgeführte Multiplikation von Matrizen mit Vektoren von rechts nach links durchführen, was den Rechenaufwand im ganzen verringert.

$$\begin{aligned} \mathbf{x}(n+1) &= \begin{pmatrix} \frac{3}{\sqrt{13}} & \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{13}} & \frac{-2}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} (1/2)^n & 0 \\ 0 & (-3/2)^n \end{pmatrix} \begin{pmatrix} \frac{\sqrt{13}}{8} \\ \frac{-3\sqrt{5}}{8} \end{pmatrix} \\ &= \begin{pmatrix} \frac{3}{\sqrt{13}} & \frac{1}{\sqrt{5}} \\ \frac{2}{\sqrt{13}} & \frac{-2}{\sqrt{5}} \end{pmatrix} \begin{pmatrix} \frac{\sqrt{13}}{8}(1/2)^n \\ \frac{-3\sqrt{5}}{8}(-3/2)^n \end{pmatrix} = \begin{pmatrix} \frac{3}{8}\{(1/2)^n - (-3/2)^n\} \\ \frac{1}{4}\{(1/2)^n + 3(-3/2)^n\} \end{pmatrix} \end{aligned} \tag{5.41}$$

□

5.4.3 Matrixpotenzierung über Z-Transformation

Das Systemverhalten ohne äußere Anregung ist durch den Zusammenhang $\mathbf{x}[n+1] = \mathbf{A}x[n]$ gegeben. Nach der einseitigen Z-Transformation erhalten wir

$$z\mathbf{X}(z) - z\mathbf{x}[0] = \mathbf{A}\mathbf{X}(z) \tag{5.42}$$

oder

$$\mathbf{X}(z) = z(z\mathbf{E} - \mathbf{A})^{-1}\mathbf{x}[0] \tag{5.43}$$

bzw. rücktransformiert:

$$\mathbf{x}[n] = Z^{-1}\{z(z\mathbf{E} - \mathbf{A})^{-1}\}\mathbf{x}[0]. \tag{5.44}$$

Andererseits gilt nach Gl. (5.23) aber auch $\mathbf{x}[\mathbf{n}] = \mathbf{A^n}\,\mathbf{x}[\mathbf{0}]$ ohne äußere Anregung. Vergleich der beiden Gleichungen liefert die nichttriviale Lösung

$$\mathbf{A}^n = Z^{-1}\{z(z\mathbf{E} - \mathbf{A})^{-1}\} \tag{5.45}$$

oder

Theorem 5.3 (Matrixpotenzierung über Z-Transformation).

$$\mathbf{A}^n = Z^{-1}\{(\mathbf{E} - z^{-1}\mathbf{A})^{-1}\}. \tag{5.46}$$

Die inverse Z-Tansformation ist dabei für jeden Eintrag der Matrix vorzunehmen.

Dieser Zusammenhang liefert (alternativ zu der Behandlung über charakteristische Gleichungen) eine weitere Möglichkeit zur Berechnung von Potenzen von Matrizen. Die Rücktransformation versteht sich für jeden Eintrag der Matrix. Man vergleiche mit dem Fall, dass A ein Skalar ist: dann ergibt sich gerade der schon in der Einführung zur Z-Transformation genannte Fall

$$a^n = Z^{-1}\{\frac{1}{1 - a/z}\} \tag{5.47}$$

für die Folge a^n $(n \geq 0)$.

Beispiel 5.2 – Matrixpotenzierung über Z-Transformation.

Wie schon in Beispiel 5.1 betrachten wir die Systemmatrix

$$\mathbf{A} = \begin{pmatrix} 0 & 3/4 \\ 1 & -1 \end{pmatrix} \tag{5.48}$$

Wir wollen nun die Matrix–Potenzierung über die einseitige Z-Transformation durchführen. Wir setzen die Matrix $\mathbf{A}$ in die Gleichung:

$$\mathbf{A}^n = Z^{-1}\{(\mathbf{E} - z^{-1}\mathbf{A})^{-1}\} \tag{5.49}$$

ein und erhalten so für die rechte Seite:

$$(\mathbf{E} - \mathbf{z^{-1}A}) = \begin{pmatrix} 1 & -3/4z^{-1} \\ -z^{-1} & 1 + z^{-1} \end{pmatrix} \tag{5.50}$$

Die Berechnung der inversen Matrix liefert uns:

$$\begin{aligned} (\mathbf{E} - \mathbf{z^{-1}A})^{-1} &= \begin{pmatrix} \frac{4(z+1)z}{4z^2+4z-3} & \frac{3z}{4z^2+4z-3} \\ \frac{4z}{4z^2+4z-3} & \frac{4z^2}{4z^2+4z-3} \end{pmatrix} \\ &= \frac{1}{4(z-1/2)(z+3/2)} \begin{pmatrix} 4(z+1)z & 3z \\ 4z & 4z^2 \end{pmatrix}. \end{aligned} \tag{5.51}$$

Wir werden später noch sehen, dass die Produktzerlegung des Nenners (nicht zufällig) die Eigenwerte der Systemmatrix ergibt.
Zur Durchführung der inversen Z-Transformation wird nun jedes Element einzeln rücktransformiert, es ergibt sich (ggf. nach Partialbruchzerlegung) wieder das Ergebnis

$$\begin{aligned} \mathbf{A}^n &= Z^{-1}\{(\mathbf{E} - z^{-1}\mathbf{A})^{-1}\} \\ &= \begin{pmatrix} 3/4(1/2)^n + 1/4(-3/2)^n & 3/8(1/2)^n - 3/8(-3/2)^n \\ 1/2(1/2)^n - 1/2(-3/2)^n & 1/4(1/2)^n + 3/4(-3/2)^n \end{pmatrix}. \end{aligned} \tag{5.52}$$

Auch hier hätten wir zur Berechnung von $\mathbf{A}^n\mathbf{x}[0]$ die inverse Z-Transformation nicht über alle 4 Komponenten durchführen müssen. Wegen der Linearität der inversen Z-Transformation gilt auch

$$\mathbf{A}^n\mathbf{x}[0] = Z^{-1}\{(\mathbf{E} - z^{-1}\mathbf{A})^{-1}\mathbf{x}[0]\} \tag{5.53}$$

und damit, mit $\mathbf{x}[0] = [0, 1]$ wie oben,

$$\mathbf{A}^n\mathbf{x}[0] = Z^{-1}\{\begin{pmatrix} \frac{3z}{4z^2+4z-3} \\ \frac{4z^2}{4z^2+4z-3} \end{pmatrix}\} = \begin{pmatrix} \frac{3}{8}\{(1/2)^n - (-3/2)^n\} \\ \frac{1}{4}\{(1/2)^n + 3(-3/2)^n\} \end{pmatrix}. \tag{5.54}$$

Dabei waren nur die 2 Vektorkomponenten rückzutransformieren. □

Diese kurze Herleitung der Berechnung über Z-Transformationen darf nicht darüber hinweg täuschen, dass das Ergebnis der Bildung einer inversen Matrix und einer inversen Z-Transformation aller Elemente bedarf. Es gilt der Murphy'sche Satz von der Erhaltung der Komplexität: sie wird lediglich an einen anderen Punkt verlagert.

Zur Lösung von Differenzengleichungssystemen durch Z-Transformation oder Eigenwertmethode siehe auch Übung 5.3.

5.4.4 Überführen in Differenzengleichung höherer Ordnung

Ein System wie in Abb. 5.3 führt im Allgemeinen zu einem System von k Differenzengleichungen für die Hilfsgrößen $x_i(n)$,

$$\mathbf{x}(n+1) = \mathbf{A}\mathbf{x}(n) + \mathbf{b}u(n) \tag{5.55}$$

$$y(n) = x_k(n) + a_k u(n) \tag{5.56}$$

wobei

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & -b_o \\ 1 & 0 & 0 & \dots & 0 & -b_1 \\ 0 & 1 & 0 & \dots & 0 & -b_2 \\ \vdots & & & & & \\ 0 & \dots & \dots & 0 & 1 & -b_{k-1} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} a_0 - a_k b_0 \\ \vdots \\ a_{k-1} - a_k b_{k-1} \end{pmatrix}, \quad \mathbf{x}(n) = \begin{pmatrix} x_1(n) \\ \vdots \\ x_k(n) \end{pmatrix} \tag{5.57}$$

gilt. Dieses System kann man durch

$$
\begin{aligned}
y(n+k) &= x_k(n+k) + a_k u(n+k) \\
&= x_{k-1}(n+k-1) - b_{k-1}x_k(n+k-1) \\
&\quad +(a_{k-1} - a_k b_{k-1})u(n+k-1) + a_k u(n+k) \\
&= x_{k-1}(n+k-1) - b_{k-1}[y(n+k-1) - a_k u(n+k-1)] \\
&\quad +a_{k-1}u(n+k-1) + a_k u(n+k) \\
&= x_{k-2}(n+k-2) - b_{k-2}x_k(n+k-2) \\
&\quad +(a_{k-2} - a_k b_{k-2})u(n+k-2) \\
&\quad -b_{k-1}[\ldots] + a_{k-1}\ldots \\
&\vdots \quad \text{(analoge Schritte bis } x_1(n+1)\text{)} \\
y(n+k) &= -\sum_{i=0}^{k-1} b_i y(n+i) + \sum_{i=0}^{k} a_i u(n+i) \qquad (5.58)
\end{aligned}
$$

schrittweise in eine Differenzengleichung k-ter Ordnung für $y(n)$ überführen.

Der homogene Teil dieser Gleichung (mit $b_k = 1$)

$$\sum_{i=0}^{k} b_i y(n+i) = 0 \qquad (5.59)$$

führt mit dem Ansatz

$$y(n) = \lambda^n \qquad (5.60)$$

zu der charakteristischen Gleichung

$$\sum_{i=0}^{k} b_i \lambda^i = 0. \qquad (5.61)$$

Wir zeigen nun, dass die Lösungen dieser Gleichung mit den Eigenwerten der Matrix $\mathbf{A}$ identisch sind. Das heisst, $\det(\mathbf{A} - \lambda\mathbf{E}) = 0$ muss zur gleichen charakteristischen Gleichung führen. Dies kann durch folgende explizite Konstruktionsvorschrift geschehen.

1. Benutze den Determinanten-Entwicklungssatz für $\det(\mathbf{A} - \lambda\mathbf{E})$ wie folgt:
2. Entwickle $\mathbf{A} - \lambda E$ nach der letzten Spalte.
3. In jeder dabei entstehenden Unterdeterminante trägt **nur** die Diagonale zum Wert bei.
4. Das Verfahren unter 2) ergibt als Entwicklung der Unterdeterminanten der letzten Spalte genau Produkte (aus der Diagonalen) mit so vielen Faktoren λ, wie der Index des b angibt (siehe Gleichung unten). Alle restlichen Faktoren haben den Wert 1.
5. Nach dieser Konstruktionsvorschrift entsteht der gesuchte Zusammenhang

$$\det(A - \lambda E) = \sum_{i=0}^{k} b_i \lambda^i \quad \text{(beachte} \quad b_k = 1) \qquad (5.62)$$

Die durchgeführte Entwicklung nach der letzten Spalte verdeutlicht das vorgestellte Schema (beachte $b_k = 1$):

$$
\begin{aligned}
\det(\mathbf{A} - \lambda\mathbf{E}) &= \det \begin{pmatrix} -\lambda & 0 & 0 & \dots & 0 & -b_o \\ 1 & -\lambda & 0 & \dots & 0 & -b_1 \\ 0 & 1 & -\lambda & \dots & 0 & -b_2 \\ \vdots & & & & & \\ 0 & \dots & \dots & 1 & -\lambda & -b_{k-2} \\ 0 & \dots & \dots & 0 & 1 & -b_{k-1} - \lambda \end{pmatrix} \\
&= -b_0 \det \begin{pmatrix} 1 & -\lambda & 0 & \dots & 0 \\ 0 & 1 & -\lambda & \dots & 0 \\ \vdots & & & & \\ 0 & \dots & \dots & 1 & -\lambda \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix} + b_1 \det \begin{pmatrix} -\lambda & 0 & 0 & \dots & 0 \\ 0 & 1 & -\lambda & \dots & 0 \\ \vdots & & & & \\ 0 & \dots & \dots & 1 & -\lambda \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix} \\
&\quad - b_2 \det \begin{pmatrix} -\lambda & 0 & 0 & \dots & 0 \\ 1 & -\lambda & 0 & \dots & 0 \\ 0 & 0 & 1 & -\lambda & \dots \\ \vdots & & & & \\ 0 & \dots & \dots & 1 & -\lambda \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix} \\
&\quad + \dots + \\
&\quad + b_{k-2} \det \begin{pmatrix} -\lambda & 0 & 0 & \dots & 0 \\ 1 & -\lambda & 0 & \dots & 0 \\ 0 & 1 & -\lambda & \dots & 0 \\ \vdots & & & & \\ 0 & \dots & 1 & -\lambda & 0 \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix} \\
&\quad - (b_{k-1} + \lambda) \det \begin{pmatrix} -\lambda & 0 & 0 & \dots & 0 \\ 1 & -\lambda & 0 & \dots & 0 \\ 0 & 1 & -\lambda & \dots & 0 \\ \vdots & & & & \\ 0 & \dots & \dots & 1 & -\lambda \end{pmatrix} \\
&= 0
\end{aligned}
\tag{5.63}
$$

Man erkennt, dass in allen entstehenden Unterdeterminanten mit der Größe $(k-1)\times(k-1)$ nur die Diagonale einen Beitrag leistet. Diese enthält genau so viele Faktoren λ, wie der Index von b angibt; die restlichen Faktoren sind 1. Es ergeben sich genau die gesuchten Potenzen von λ.

Das direkte Lösen eines Systems von k Differenzengleichungen ist also vom gleichen Schwierigkeitsgrad (Eigenwertaufgabe für eine $k \times k$ Matrix), wie das Lösen der entsprechenden Differenzengleichung k-ter Ordnung.

Mit den Lösungen λ_i der charakteristischen Gleichung ergibt sich nun die allgemeine Lösung des homogenen Teils der Differenzengleichung zu

$$y(n) = \sum_{i=1}^{k} c_i \lambda_i^n, \tag{5.64}$$

mit (zur Zeit noch) frei wählbaren Konstanten c_i. Diese müssen aus den Anfangsbedingungen bestimmt werden. Die Lösung einer Differenzengleichung ist gleich der Summe aus der allgemeinen homogenen Lösung plus einer partikulären (speziellen) Lösung der inhomogenen Gleichung. Wir müssen also noch eine partikuläre Lösung finden. Es ist leicht zu sehen, dass eine konstante Lösung

$$y(n) = C, \quad \text{mit} \quad C = \frac{\sum_{i=0}^{k} a_i u(n+i)}{\sum_{i=0}^{k} b_i} \tag{5.65}$$

die inhomogene Differenzengleichung erfüllt. Es müssen nun noch die c_i bestimmt werden. Das geschieht durch Lösen eines linearen Gleichungssystems mit den gegebenen k Anfangsbedingungen $y(0)$ bis $y(k-1)$:

$$\begin{pmatrix} y(0) \\ \vdots \\ y(k-1) \end{pmatrix} = \begin{pmatrix} \lambda_1^0 & \dots & \lambda_k^0 \\ \vdots & & \\ \lambda_1^{k-1} & \dots & \lambda_k^{k-1} \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} + \begin{pmatrix} C \\ \vdots \\ C \end{pmatrix}. \tag{5.66}$$

Beispiel 5.3 – Lösung des Differenzengleichungssystems über charakteristische Gleichung.

Wir wählen (siehe Abb. 5.3) als Beispiel $b_0 = -3/4$ und $b_1 = 1$ und erhalten

$$\mathbf{A} = \begin{pmatrix} 0 & 3/4 \\ 1 & -1 \end{pmatrix}. \tag{5.67}$$

Es gebe keine äußere Anregung ($u(n) = 0$). Somit erhalten wir die folgende Differenzengleichung 2.Ordnung:

$$-3/4 y(n) + y(n+1) + y(n+2) = 0. \tag{5.68}$$

Diese führt zu der charakteristischen Gleichung

$$-3/4 + \lambda + \lambda^2 = 0 \tag{5.69}$$

mit den Lösungen $\lambda_1 = 1/2$ und $\lambda_2 = -3/2$.
In Beispiel 5.1 wurden die Anfangsbedingungen $x_1(0) = 0$ und $x_2(0) = 1$ verwendet. Das entspricht hier den Anfangsbedingungen

$$y(0) = x_2(0) = 1 \quad \text{und} \quad y(1) = x_2(1) = x_1(0) - b_1 x_2(0) = -1. \tag{5.70}$$

Damit können nun die c_i der allgemeinen Lösung $y(n) = c_1 \lambda_1^n + c_2 \lambda_2^n$ mit

$$\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1/2 & -3/2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = 0 \quad \text{zu} \quad c_1 = 1/4 \quad \text{und} \quad c_2 = 3/4 \tag{5.71}$$

bestimmt werden. Diese allgemeine Lösung stimmt mit den Ergebnissen aus Beispiel 5.1 überein. □

Zur Lösung von Differenzengleichungen höherer Ordnung auf direktem Weg und über Z–Transformation siehe auch Übung 5.2.

5.4.5 Allgemeine Systembeschreibung im Z-Bereich

Wir wenden uns wieder der Systembeschreibung zu. Dazu wiederholen wir noch einmal die Gln. (5.21) und (5.22):

$$\mathbf{x}(n+1) = \mathbf{A}\mathbf{x}(n) + \mathbf{b}u(n) \tag{5.72}$$

$$y(n) = \mathbf{c}^T\mathbf{x}(n) + du(n). \tag{5.73}$$

Nach einseitiger Z-Transformation ergibt sich:

$$z\mathbf{X}(z) - z\mathbf{x}[0] = \mathbf{A}\mathbf{X}(z) + \mathbf{b}U(z) \tag{5.74}$$

$$Y(z) = \mathbf{c}^T\mathbf{X}(z) + dU(z). \tag{5.75}$$

Auflösen dieser Gleichungen nach $Y(z)$ liefert

$$Y_e(z) = \mathbf{c}^T(z\mathbf{E} - \mathbf{A})^{-1}\mathbf{x}[0]z + [\mathbf{c}^T(z\mathbf{E} - \mathbf{A})^{-1}\mathbf{b} + d]U_e(z). \tag{5.76}$$

Wir erkennen die bereits angesprochenen 2 linear überlagerten Komponenten in $Y(z)$:

- der erste Summand ist der Systemausgang aufgrund der eingeprägten Anfangsgrößen $x[0]$.
- der zweite Summand beschreibt das Ausgangsverhalten aufgrund der Anregung $U(z)$.

Falls das System seit unendlich langer Zeit läuft, ist der erste Summand nicht möglich, da die Werte $x[0]$ durch die Systemvergangenheit festgelegt sind und nicht eingestellt werden können. In diesem Fall benutzen wir die normale Z-Transformation und erhalten

$$Y(z) = [\mathbf{c}^T(z\mathbf{E} - \mathbf{A})^{-1}\mathbf{b} + d]U(z) \tag{5.77}$$

bzw. für die Übertragungsfunktion

$$H(z) = \mathbf{c}^T(z\mathbf{E} - \mathbf{A})^{-1}\mathbf{b} + d. \tag{5.78}$$

Wir bilden die inverse Matrix in dieser Gleichung. Dazu benötigen wir das algebraische Komplement der Matrix $(z\mathbf{E} - \mathbf{A})$, welches wir mit $(z\mathbf{E} - \mathbf{A})'$ bezeichnen. Es gilt:

$$H(z) = \frac{\mathbf{c}^T(z\mathbf{E} - \mathbf{A})'\mathbf{b}}{\det(z\mathbf{E} - \mathbf{A})} + d. \tag{5.79}$$

Die Pole von $H(z)$ finden wir damit bei denjenigen z, die $\det(z\mathbf{E} - \mathbf{A}) = 0$ erfüllen. Dies ist aber – wie gezeigt – gerade die Gleichung zur Bestimmung der Eigenwerte von $\mathbf{A}$. Damit erhalten wir folgenden Satz:

Theorem 5.4 (Pole der Übertragungsfunktion eines Differenzengleichungssystems).
Die Pole der Übertragungsfunktion eines Differenzengleichungssystems sind identisch mit den Eigenwerten seiner Systemmatrix.

5.4.6 Zusammenhang zwischen Struktur und Z-Transformierter

Wir wenden uns nun noch einmal der Darstellung der Ausgangsgröße $y(n)$ in Form einer Differenzengleichung zu. Oben haben wir gesehen, dass die geschlossene Lösung für $y(n)$ alle Werte der Eingangsfolge $u(0) \dots u(n)$ sowie Matrixpotenzierungen verlangt. Dies liegt daran, dass eine reine *Vorwärtslösung* erzeugt wurde, die auf vergangene Werte des Ausgangs $y(n-k)$ nicht zugreift. Lassen wir solch einen Zugriff zu (damit ist das System z.T. rückgekoppelt), so können wir die Übertragungsfunktion

$$H(z) = \frac{Y(z)}{U(z)} = \frac{\mathbf{c}^T(z\mathbf{E} - \mathbf{A})'\mathbf{b}}{\det(z\mathbf{E} - \mathbf{A})} + d \tag{5.80}$$

nach leichter Umformung in

$$\det(z\mathbf{E} - \mathbf{A})Y(z) = [\mathbf{c}^T(z\mathbf{E} - \mathbf{A})'\mathbf{b} + \det(z\mathbf{E} - \mathbf{A})d]U(z) \tag{5.81}$$

auch zu einer Differenzengleichung M-ter Ordnung ändern – die Beschreibung als System muss ja äquivalent zu einer Differenzengleichung höherer Ordnung sein.

Durch Rücktransformation erhalten wir:

$$Z^{-1}\{\det(z\mathbf{E}-\mathbf{A})Y(z)\} = Z^{-1}\{[\mathbf{c}^T(z\mathbf{E} - \mathbf{A})'\mathbf{b}+\det(z\mathbf{E}-\mathbf{A})d]U(z)\}. \tag{5.82}$$

Die Determinanten, wie auch der Teil in eckigen Klammern, stellen Polynome M-ten Grades in z dar, so dass wir nach Rücktransformation (wegen des Verschiebungssatzes) auf beiden Seiten eine Summe erhalten, die im allgemeinen alle Werte von $y[n-M]$ bis $y[n]$ bzw. $u[n-M]$ bis $u[n]$ enthält. Vergleicht man das mit der Differenzengleichung, die man z.B. direkt aus Bild 5.3 abliest:

$$b_0 y[n-2] + b_1 y[n-1] + y[n] = a_0 x[n-2] + a_1 x[n-1] + a_2 x[n] \tag{5.83}$$

so entspricht es offensichtlich genau dieser Struktur.

Übungen

Übung 5.1 – Direkte Lösung einer Differenzengl. 1. Ordnung.

Lösen Sie die Differenzengleichung $y(n) = u(n) + ay(n-1)$ mit der Kausalitätsbedingung $y(-1) = 0$ auf direktem Weg!

Übung 5.2 – Fibonacci-Folge.

Die Differenzengleichung 2. Ordnung $y(n) = y(n-1) + y(n-2)$ mit Anfangswerten $y(0) = 1,\ y(1) = 1$ ist die rekursive Vorschrift für die so genannte Fibonacci-Folge.

a) Lösen Sie die Differenzengleichung auf direktem Weg! Verwenden Sie den Ansatz $y(n) = a^n$.
b) Zeigen Sie, dass die Differenzengleichung $y(n) = y(n-1) + y(n-2) + \delta(n)$ (also mit der Impulsfolge als Eingang) als *kausale* Lösung ebenfalls die Fibonacci-Folge liefert. Lösen Sie diese Differenzengleichung nun mittels normaler Z-Transformation! Vergleichen Sie ihr Ergebnis mit a)!
c) Lösen Sie nun $y(n) = y(n-1) + y(n-2)$ mittels *einseitiger* Z-Transformation, wodurch Sie die passenden Anfangsbedingungen für die Fibonacci-Folge direkt einbauen können.

Übung 5.3 – System von Differenzengleichungen 1. Ordnung.

Gegeben sei ein System von Differenzengleichungen erster Ordnung

$$\begin{aligned} y_1[n+1] &= y_2[n] \\ y_2[n+1] &= y_2[n] + y_1[n] \end{aligned}, \quad y_1[0] = y_2[0] = 1.$$

a) Wandeln Sie das System in die Matrixform $\mathbf{y}[n+1] = \mathbf{A}\,\mathbf{y}[n]$ um!
b) Lösen Sie das System mittels Matrixpotenzierung über Eigenwerte!
c) Lösen Sie es über die *einseitige* Z-Transformation! Was wäre bei Verwendung der normalen Z-Transformation zu beachten?

6

Die diskrete Fouriertransformation

Das nun folgende Kapitel beschäftigt sich mit der diskreten Fouriertransformation (DFT). Diese stellt eine Anpassung der von den analogen Signalen her bekannten Fouriertransformation (FT) an die Problematik der diskreten Signale und ihrer digitalen Verarbeitung dar.

Im ersten Abschnitt werden die Definitionen der diskreten Fouriertransformation (DFT) ausgehend von der analogen Fouriertransformation (FT) hergeleitet. Der zweite Abschnitt geht kurz auf die Eigenschaften der DFT ein. Dieser Teil ist bewusst knapp gehalten, da sich für die DFT nichts qualitativ Neues ergibt und daher nur die bereits bekannten Eigenschaften der Fouriertransformation in leicht modifizierter Form zu nennen sind. Der dritte Abschnitt befasst sich mit dem Zusammenhang der DFT zu anderen Transformationen, speziell zur Z-Transformation. Im vierten Abschnitt werden verschiedene Fensterfolgen vorgestellt. Abschließend wird im fünften Abschnitt die schnelle Fouriertransformation (FFT) als eine mögliche Implementierung der DFT behandelt. Hierbei wird besonders auf den Radix-2-Algorithmus mit Reduktion im Zeitbereich eingegangen.

6.1 Herleitung und Definition

Für analoge Signale ist die Fouriertransformation wie folgt definiert:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega} dt \tag{6.1}$$

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega} d\omega \tag{6.2}$$

Für die Bearbeitung durch ein digitales System ergeben sich hieraus einige Probleme. Die Variablen t und ω sind kontinuierlich, die Grenzen liegen im Unendlichen, die Integration führt zu einer sehr aufwändigen Berechnung.

Daher ist es angebracht, die Fouriertransformation an die Gegebenheiten in digitalen Systemen anzupassen.

Für eine gewisse Vereinfachung sorgt die zeitdiskrete Fouriertransformation (TDFT). Sie kann auf zeitdiskrete, bandbegrenzte Signale angewendet werden und ist folgendermaßen definiert:

$$X(\omega) = \sum_{k=-\infty}^{\infty} x[k] e^{-j\omega kT} \tag{6.3}$$

$$x[k] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(\omega) e^{j\omega kT} d(\omega T) \tag{6.4}$$

Aufgrund der diskreten Zeitachse kann bei der Hintransformation die Integration durch eine Summation ersetzt werden, bei der Rücktransformation kann die Integration auf die begrenzte Bandbreite beschränkt werden. Damit sind aber nur einige der Probleme der analogen FT gelöst.

Zur weiteren Vereinfachung soll nun auch die Frequenz ω diskretisiert werden. Dazu definiert man N diskrete Frequenzen innerhalb der Bandbreite Ω des Signals:

$$\omega = n\omega_0 = n\frac{\Omega}{N} = n\frac{2\pi}{NT}; \quad \text{beachte} \quad \Omega = \frac{2\pi}{T}. \tag{6.5}$$

Setzt man dies in die Gleichung der zeitdiskreten FT ein, so erhält man:

$$X(\omega) = \sum_{k=-\infty}^{\infty} x[k] e^{-j\omega kT} \tag{6.6}$$

$$X(n) = \sum_{k=-\infty}^{\infty} x[k] e^{-jn(\frac{2\pi}{N})k} = \sum_{k=-\infty}^{\infty} x[k] W_N^{nk}. \tag{6.7}$$

Dabei ist

$$W_N = e^{-j\frac{2\pi}{N}} \tag{6.8}$$

der komplexe Drehoperator. Seine Bedeutung lässt sich anschaulich in der komplexen Ebene erklären. Man stelle sich vor, der Einheitskreis wird von N Zeigern in gleich große Sektoren geteilt. Dann repräsentiert jeder dieser Zeiger eine Potenz von W_N^k. Mit wachsender Potenz durchläuft der Drehoperator den Einheitskreis ähnlich einem Uhrzeiger. Ist der Umlauf bei W_N^{N-1} beendet, beginnt er für W_N^N von neuem. W_N^k ist also periodisch mit der Grundperiode N:

$$W_N^k = W_N^{k+iN} = W_N^{(k \mod N)}, \quad i \in \mathbb{Z}. \tag{6.9}$$

Dabei steht **mod** für den Modulo-Operator, dessen Ergebnis dem Rest bei der ganzzahligen Division entspricht.

Um die Periodizität von W_N ausnutzen zu können, teilen wir nun die unendliche Summe in Blöcke von je N Summanden. Danach erfolgt eine Indexverschiebung, die alle Teilsummen in den Bereich $k = 0 \ldots N-1$ überführt. Damit ergibt sich

$$
\begin{aligned}
X(n) &= \sum_{k=-\infty}^{\infty} x[k] W_N^{nk} \\
&= \ldots + \sum_{k=0}^{N-1} x[k] W_N^{nk} + \sum_{k=N}^{2N-1} x[k] W_N^{nk} + \sum_{k=2N}^{3N-1} x[k] W_N^{nk} + \ldots \\
&= \ldots + \sum_{k=0}^{N-1} x[k]\ W_N^{nk} + \sum_{k=0}^{N-1} x[k+N] W_N^{n(k+N)} \\
&\quad + \sum_{k=0}^{N-1} x[k+2N] W_N^{n(k+2N)} + \ldots
\end{aligned} \tag{6.10}
$$

Die Periodizität (Gl. 6.9) des komplexen Drehoperators W_N^k mit Vielfachen von N wird nun ausgenutzt, um die Folgegliedern $x[k]$ periodenweise zu summieren.

$$
X(n) = \sum_{k=0}^{N-1} \left[\sum_{r=-\infty}^{\infty} x[k+rN] \right] W_N^{nk} \tag{6.11}
$$

Definiert man nun $\tilde{x}[k]$ wie folgt

$$
\tilde{x}[k] = \sum_{r=-\infty}^{\infty} x[k+rN], \tag{6.12}
$$

so erhält man:

$$
X(n) = \sum_{k=0}^{N-1} \tilde{x}[k] W_N^{nk}. \tag{6.13}
$$

Damit ist die Transformationsformel vom Zeit- in den Frequenzbereich gefunden.

Als nächstes muss eine Möglichkeit definiert werden, aus dem Frequenzbereich zurück in den Zeitbereich zu gelangen, das ist die inverse diskrete Fouriertransformation (IDFT). Dazu wird von einer weiteren Eigenschaft des komplexen Drehoperators Gebrauch gemacht. Summiert man diesen nämlich über eine Periode, ergibt sich die Summe null:

$$
\sum_{k=0}^{N-1} W_N^k = 0. \tag{6.14}
$$

Auch dies kann man sich leicht in der komplexen Ebene veranschaulichen. Potenziert man den Drehoperator mit einer weiteren ganzen Zahl n, so wird der Einheitskreis nicht nur einmal, sondern n mal durchlaufen. Es ergibt sich aber weiterhin die Summe null, es sei denn n nimmt den Wert null an. In diesem Fall gilt:

$$
\sum_{k=0}^{N-1} W_N^{k0} = \sum_{k=0}^{N-1} e^{-j\frac{2\pi}{N}k\cdot 0} = \sum_{k=0}^{N-1} e^0 = N \tag{6.15}
$$

Mit Hilfe der Diracfolge lässt sich dieser Zusammenhang wie folgt beschreiben:

$$\sum_{k=0}^{N-1} W_N^{kn} = N\delta[n]. \tag{6.16}$$

Für die inverse DFT kann man nun vermuten, dass wie schon bei der analogen Fouriertransformation eine starke Ähnlichkeit zwischen Hin- und Rücktransformation besteht. Daher liegt der folgende Ansatz nahe:

$$\begin{aligned}\sum_{n=0}^{N-1} X(n)W_N^{-(mn)} &= \sum_{n=0}^{N-1}\left[\sum_{k=0}^{N-1}\tilde{x}[k]W_N^{nk}\right]W_N^{-(mn)} = \sum_{k=0}^{N-1}\tilde{x}[k]\sum_{n=0}^{N-1}W_N^{n(k-m)} \\ &= \sum_{k=0}^{N-1}\tilde{x}[k]\left(N\delta[k-m]\right) \\ &= N\tilde{x}[m] \end{aligned} \tag{6.17}$$

Somit ist auch die inverse DFT gefunden.

Es bleibt das Problem, dass es sich bei $\tilde{x}[k]$ um eine Summe von Folgegliedern der eigentlich gesuchten Folge $x[k]$ handelt. Dieses Problem tritt aber nicht auf, wenn es sich bei $x[k]$ um eine endliche Folge handelt, die weniger als N aufeinander folgende Glieder hat. In diesem Fall gilt $x[k] = \tilde{x}[k]$. Somit ist die DFT für eine endliche Folge $x[k]$ der Länge $L \leq N$ eine eineindeutige Transformation.

Weiterhin ist zu beachten, dass (vgl. Gl. 6.9)

$$W_N^{(k+mN)n} = W_N^{k(n+mN)} = W_N^{kn} \tag{6.18}$$

$$W_N^{(-k+mN)n} = W_N^{k(-n+mN)} = W_N^{-kn} \tag{6.19}$$

für alle $m \in \mathbb{Z}$. Sowohl $X(n)$ wie auch $x[k]$ erscheinen in der DFT also periodisch (in N) nach beiden Seiten fortgesetzt. Ggf. sind dabei aliasing-Effekte zu beachten. In der Interpretation der DFT als Stützwerte der analogen FT bedeutet das folgendes: Werden die N zur Bestimmung der (I)DFT benutzten Werte als Abtastwerte einer analogen Funktion im Intervall $t = [0, (N-1)T]$ bzw. $\omega = [0, (N-1)\omega_0]$ interpretiert, so ist diese analoge Funktion periodisch fortzusetzen, damit DFT und analoge FT dieselben Werte ergeben. Ist also insbesondere die analoge Funktion selbst bereits periodisch, z.B. im Zeitbereich mit der Periode T', so ist für die Abtastung und Weiterverarbeitung mit der DFT $NT = mT'$ mit $m \in \mathbb{N}$ zu wählen. Wird diese Regel verletzt, so kommt es zum so genannten *Leckeffekt* (engl. leakage). Diese Bezeichnung rührt davon her, dass anschaulich die Hauptspektrallinie in die benachbarten diskreten DFT-Werte „leckt“ und damit das linienförmige spektrale Verhalten verschmiert wird. Siehe auch Beispiel 6.1(c).

Wir fassen unsere Erkenntnisse über die diskrete Fourier-Transformation wie folgt zusammen:

Theorem 6.1 (Diskrete Fourier-Transformation (DFT)).

$$DFT: \quad X(n) = \sum_{k=0}^{N-1} x[k] W_N^{nk} = \sum_{k=0}^{N-1} x[k] e^{-j\frac{2\pi}{N}kn} \tag{6.20}$$

$$IDFT: \quad x[k] = \frac{1}{N} \sum_{k=0}^{N-1} X(n) W_N^{-nk} = \frac{1}{N} \sum_{k=0}^{N-1} X(n) e^{j\frac{2\pi}{N}kn} \tag{6.21}$$

Die Werte der (I)DFT sind periodisch in N. Für endliche Folgen $x[k]$ bzw. $X(n)$ der Länge $L \leq N$ sind die Werte der (I)DFT an den Stellen $t = kT$ bzw. $\omega = n\omega_0 = n\frac{\Omega}{N} = n\frac{2\pi}{NT}$ identisch zur analogen (inversen) Fourier-Transformation einer periodisch fortgesetzten (mit Perioden NT für die DFT bzw. $\Omega = N\omega_0$ für die IDFT) analogen Funktion. Ist diese analoge Funktion selbst im Zeitbereich periodisch in T', und soll die DFT das Spektrum dieser Funktion liefern, so ist für die DFT das Zeitfenster $NT = mT'$ mit $m \in \mathbb{N}$ zu wählen. Wird diese Relation verletzt, so kommt es zum Leckeffekt. Dasselbe gilt äquivalent für die IDFT bei Vertauschung von Zeit- und Frequenzbereich.

Beispiel 6.1 – DFT für eine Winkelfunktion.

Die Funktion $x(t) = cos(t)$ hat nur Spektrallinien bei $|\omega| = 1$.
a) Sie werde an N Stellen $t = k \cdot 2\pi/N$ abgetastet, also $T = 2\pi/N$. Wir ermitteln die Spektralwerte bei den Frequenzen $\omega = n\frac{2\pi}{NT} = n$ mit Hilfe der DFT für $N = 2$ und $N = 4$. Da $\omega_0 = \frac{2\pi}{NT} = 1$, entspricht direkt $X_{\mathrm{DFT}}(n) = X(\omega = n)$.
b) Wir verdoppeln T. Wie lautet das Ergebnis jetzt? Argumentieren Sie, und berechnen Sie.
c) Wir halbieren T. Wie lautet das Ergebnis jetzt? Argumentieren Sie lediglich.

Lösung:
a) Für $N = 2$ lauten die Abtastwerte $x[0] = 1$, $x[1] = -1$. Daraus ergibt sich
$X(0) = 1 - 1 = 0$,
$X(1) = 1 \cdot e^{-j\pi \cdot 0} - 1 \cdot e^{-j\pi \cdot 1} = 2$.
Für $N = 4$ lauten die Abtastwerte $x[0] = 1$, $x[1] = 0$, $x[2] = -1$, $x[3] = 0$, und daraus
$X(0) = 1 + 0 - 1 + 0 = 0$,
$X(1) = 1 \cdot e^{-j(\pi/2) \cdot 0} - 1 \cdot e^{-j(\pi/2) \cdot 2} = 2$,
$X(2) = 1 \cdot e^{-j(\pi/2) \cdot 2 \cdot 0} - 1 \cdot e^{-j(\pi/2) \cdot 2 \cdot 2} = 0$,
$X(3) = 1 \cdot e^{-j(\pi/2) \cdot 3 \cdot 0} - 1 \cdot e^{-j(\pi/2) \cdot 3 \cdot 2} = 2$.
Im Fall $N = 4$ erhalten wir eine zusätzliche Spektrallinie bei $\omega = 3$. Dies ist ein Resultat der Periodizität der DFT im Frequenzbereich mit der Periode N. Da $x(t) = cos(t)$ eine Spektrallinie bei $\omega = -1$ besitzt, erhalten wir allgemein im Frequenzbereich immer eine zusätzliche Spektrallinie bei $n = N - 1$ (aliasing). Im Fall $N = 2$ fallen diese

beiden Linien zusammen, da $1 = n = N - n$ gilt. Wir sehen also nur eine Linie mit doppeltem Betrag 2. Für $N = 4, 8, 16, \ldots$ existieren diese Überlappungen nicht mehr, die Spektrallinien haben dann den Wert $X(n) = N/2$. Zu weiteren Betrachtungen siehe Übung 6.3.
b) Bei Verdoppelung von T wird die Funktion über 2 Perioden abgetastet. Wegen $\omega_0 = \frac{2\pi}{NT} = 1/2$ wären die Spektrallinien bei $n = 2$ und $n = -2$ zu erwarten, sowie bei Verschiebungen dieser beiden Linien um Vielfache von N. Für $N = 2$ betrachten wir die Frequenzwerte $[0, 1]$. In diesem Intervall erhalten wir keine Spektrallinien im Basisband. Aus Seitenbändern erhalten wir jedoch Werte durch Verschiebung (markiert durch $<>$) für $n = 2- < 2 >= 0$ und $n = -2+ < 2 >= 0$. Insgesamt existiert der Wert bei $n = 0$ also doppelt, andere Spektralanteile existieren nicht im Intervall.
Für $N = 4$ betrachten wir die Frequenzwerte $[0, 3]$. In diesem Intervall erhalten wir eine Spektrallinien im Basisband bei $n = 2$. Aus Seitenbändern erhalten wir Werte durch Verschiebung (markiert durch $<>$) für $n = -2+ < 4 >= 2$. Insgesamt existiert der Wert bei $n = 2$ also doppelt, andere Spektralanteile existieren nicht im Intervall. Für $N = 8, 16, \ldots$ existieren diese Überlappungen nicht mehr, die Spektrallinien haben dann den Wert $X(n) = N/2$.
Die Rechnung zeigt folgendes:
Für $N = 2$ lauten die Abtastwerte $x[0] = 1$, $x[1] = 1$, und daraus
$X(0) = 1 + 1 = 2$,
$X(1) = 1 \cdot e^{-j\pi \cdot 0} + 1 \cdot e^{-j\pi \cdot 1} = 0$.
Für $N = 4$ lauten die Abtastwerte $x[0] = 1$, $x[1] = -1$, $x[2] = 1$, $x[3] = -1$, und daraus
$X(0) = 1 - 1 + 1 - 1 = 0$,
$X(1) = 1 \cdot e^{-j(\pi/2) \cdot 0} - 1 \cdot e^{-j(\pi/2) \cdot 1} + 1 \cdot e^{-j(\pi/2) \cdot 2} - 1 \cdot e^{-j(\pi/2) \cdot 3} = 1 + j - 1 - j = 0$,
$X(2) = 1 \cdot e^{-j(\pi/2) \cdot 2 \cdot 0} - 1 \cdot e^{-j(\pi/2) \cdot 2 \cdot 1} + 1 \cdot e^{-j(\pi/2) \cdot 2 \cdot 2} - 1 \cdot e^{-j(\pi/2) \cdot 2 \cdot 3} = 1 + 1 + 1 + 1 = 4$,
$X(3) = 1 \cdot e^{-j(\pi/2) \cdot 3 \cdot 0} - 1 \cdot e^{-j(\pi/2) \cdot 3 \cdot 1} + 1 \cdot e^{-j(\pi/2) \cdot 3 \cdot 2} - 1 \cdot e^{-j(\pi/2) \cdot 3 \cdot 3} = 1 - j - 1 + j = 0$,
c) Für halbiertes T wird nur eine halbe Periode der Funktion $x(t) = cos(t)$ abgetastet. Da die DFT die Funktionen als periodisch (mit NT) fortgesetzt betrachtet, entspricht dies der analogen Fouriertransformierten der stückweise definierten Funktion $x(t) = cos(t - k \cdot \pi)$ für $t = [k\pi \ldots (k + 1)\pi)$. Diese Funktion hat bei allen Werten $k\pi$ einen Sprung der Höhe -2. Da die DFT an den Abtastwerten der analogen Fouriertransformierten dieser Funktion entspricht, erhält man durch diesen Leckeffekt ein völlig anderes Ergebnis, das sich auch durch aliasing nicht auf die DFT der Fälle a) und b) zurückführen lässt. Dasselbe Resultat gilt immer, wenn nicht mit ganzzahligen Vielfachen der Periode der ursprünglichen Funktion abgetastet wird.

□

6.2 Zusammenhang zwischen DFT und anderen Transformationen

Im vorigen Abschnitt wurde gezeigt, dass es sich bei der DFT um eine ein-eindeutige Transformation für endliche Folgen mit $L \leq N$ handelt. Daraus folgt, dass $X[n]$ in einer eindeutigen Beziehung zu anderen transformierten Größen stehen muss. Dies soll hier anhand der Z-Tansformation gezeigt werden, zur weiteren Vertiefung sei beispielsweise auf (Föllinger, 2000) verwiesen. Die Z-Transformation ist für eine Folge der Länge L wie folgt definiert:

$$\begin{aligned} X(z) &= \sum_{k=0}^{L-1} x[k] z^{-k} = \sum_{k=0}^{L-1} \left(IDFT\left\{X[n]\right\} \right) z^{-k} \\ &= \sum_{k=0}^{L-1} \left(\frac{1}{N} \sum_{n=0}^{N-1} X_{DFT}[n] e^{jn\frac{2\pi}{N}k} \right) z^{-k} \end{aligned} \tag{6.22}$$

Auf dem Einheitskreis der Z-Ebene, also für $z = e^{j\omega T} = e^{j\Phi}$ mit $\Phi = \omega T$ erhält man:

$$\begin{aligned} X(e^{j\Phi}) &= \frac{1}{N} \sum_{n=0}^{N-1} X_{DFT}[n] \sum_{k=0}^{L-1} e^{jn\frac{2\pi}{N}k} (e^{j\Phi})^{-k} = \sum_{n=0}^{N-1} \frac{X_{DFT}[n]}{N} \sum_{k=0}^{L-1} e^{j(n\frac{2\pi}{N}-\Phi)k} \\ &= \frac{1}{N} \sum_{n=0}^{N-1} X_{DFT}[n] \sum_{k=0}^{L-1} e^{j\varphi k} \quad \text{mit} \quad \varphi = n\frac{2\pi}{N} - \Phi \end{aligned} \tag{6.23}$$

Die innere Summe hat die Form einer geometrischen Reihe. Anstatt das Ergebnis einfach anzugeben, wollen wir es hier zu Demonstrationszwecken einmal explizit berechnen. Durch Multiplikation mit einem geeigneten Faktor und Indexverschiebung erhält man:

$$\begin{aligned} \sum_{n=0}^{L-1} e^{j\varphi k} &= \frac{1}{1-e^{j\varphi}} \left(1 - e^{\,j\varphi}\right) \sum_{n=0}^{L-1} e^{j\varphi k} = \frac{1}{1-e^{j\varphi}} \left(\sum_{n=0}^{L-1} e^{j\varphi k} - \sum_{n=1}^{L} e^{j\varphi k} \right) \\ &= \frac{1}{1-e^{\,j\varphi}} \left(\sum_{n=0}^{L-1} e^{j\varphi k} - \left[\sum_{n=0}^{L-1} e^{j\varphi k} - e^{j\varphi(0)} + e^{j\varphi L} \right] \right) \\ &= \frac{1}{1-e^{j\varphi}} \left(1 - e^{j\varphi L}\right) = \frac{1-e^{j\varphi L}}{1-e^{j\varphi}} \end{aligned} \tag{6.24}$$

Unter Verwendung der Beziehung

$$\sin(\alpha) = \frac{1}{2j} \left(e^{j\alpha} - e^{-j\alpha} \right) \tag{6.25}$$

erhält man durch weiteres Umformen

$$\sum_{n=0}^{L-1} e^{j\varphi k} = \frac{1-e^{j\varphi L}}{1-e^{j\varphi}} = \frac{e^{j(L\varphi/2)}}{e^{j(\varphi/2)}} \cdot \frac{e^{-j(L\varphi/2)} - e^{j(L\varphi/2)}}{e^{-j(\varphi/2)} - e^{j(\varphi/2)}}$$

$$= e^{j\frac{\varphi}{2}(L-1)} \frac{(e^{-j(L\varphi/2)} - e^{j(L\varphi/2)})/2j}{(e^{-j(\varphi/2)} - e^{j(\varphi/2)})/2j} = e^{j(L-1)\varphi/2} \frac{\sin(L\varphi/2)}{\sin(\varphi/2)}$$

$$= e^{j(L-1)\varphi/2} p(\varphi, L) \quad \text{mit} \quad p(\varphi, L) = \frac{\sin(L\varphi/2)}{\sin(\varphi/2)} \tag{6.26}$$

Die Funktion $p(\phi, L)$ wurde schon einmal im Zusammenhang mit der Rekonstrunktion endlicher Signale (Abschn. 4.4.4) betrachtet. Die Abb. 6.1 zeigt noch einmal ihr Aussehen.

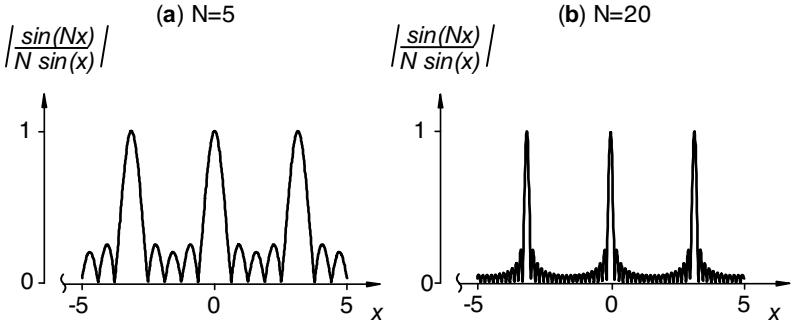


Abbildung 6.1. Verlauf der Funktion $|\sin(Nx)/N\sin(x)|$ mit $N = 5$ bzw. $N = 20$

Der Betrag der Funktion $p(\phi, L)$ nähert sich für große L immer mehr einer Summe von Delta-Funktionen für $\phi = 2k\pi$ an, denn hier hat die Nennerfunktion ihre Nullstellen. Nach der Regel von L'Hospital erhält man an diesen Stellen:

$$\lim_{\varphi \to 2\pi r} \frac{\sin(L\varphi/2)}{\sin(\varphi/2)} = \left[\frac{0}{0}\right] = \lim_{\varphi \to 2\pi r} \frac{L/2}{1/2} \frac{(\cos(L\varphi/2))}{\cos(\varphi/2)} = L\frac{\cos(\pi r L)}{\cos(\pi r)} = L\frac{(-1)^{rL}}{(-1)^{r}}$$

$$\lim_{\varphi \to 2\pi r} p(\varphi, L) = (-1)^{r(L-1)} L \tag{6.27}$$

Dies entspricht genau dem Verhalten in Abb. 6.1.

Setzt man das Ergebnis für die innere Summe (6.26) in die Gl. (6.23) ein, erhält man:

Theorem 6.2 (Interpolationseigenschaft der DFT).

$$X(e^{j\Phi}) = \frac{1}{N} \sum_{n=0}^{N-1} X_{DFT}[n] e^{\frac{j}{2}(n\frac{2\pi}{N} - \Phi)(L-1)} p(n\frac{2\pi}{N} - \Phi, L) \tag{6.28}$$

Für allgemeine Frequenzen $\omega = \Phi/T$ interpoliert *Gl.* (6.28) *also die Spektralwerte aus den Werten der DFT. Für die exakten Stützfrequenzen der DFT,*

$\omega = \Omega n/N$, liefert Gl. (6.28) die genauen Werte

$$X(\omega = \Omega\frac{n}{N}) = \frac{L}{N}X_{DFT}[n] \tag{6.29}$$

Um Gl. (6.29) zu zeigen, betrachten wir die Funktion $p(\phi, L)$ für alle Stellen $\phi = 2\pi r$. Hier gilt:

$$\begin{aligned}\varphi = n\frac{2\pi}{N} - \Phi \;\leftrightarrow\; 2\pi r = n\frac{2\pi}{N} - \omega T \;\leftrightarrow\\ \omega = \frac{2\pi}{T}\left(\frac{n}{N} - r\right) = \Omega\left(\frac{n}{N} - r\right)\end{aligned} \tag{6.30}$$

Damit ergibt sich aus Gl. (6.28) mit Hilfe von Gl. (6.27):

$$\begin{aligned}X\left(\omega = \Omega(\frac{n}{N} - r)\right) &= \frac{1}{N}X_{DFT}[n]e^{\frac{i}{2}(2\pi r)(L-1)}p(2\pi r, L)\\ &= \frac{1}{N}X_{DFT}[n](-1)^{\,r(L-1)}\left((-1)^{\,r(L-1)}L\right)\\ &= \frac{L}{N}X_{DFT}[n]\end{aligned} \tag{6.31}$$

Da das Signal bandbegrenzt sein soll, liegt ω innerhalb der Bandbreite Ω. Dies ist nur für $r = 0$ erfüllt. Es gilt also:

$$X(\omega = \Omega\frac{n}{N}) = \frac{L}{N}X_{DFT}[n] \tag{6.32}$$

Dies ist der Zusammenhang, der zu zeigen war. □

Die DFT stellt also nur einen eindeutigen Zusammenhang zwischen den Spektralwerten bei den diskretisierten Frequenzen und den Amplitudenwerten bei den diskretisierten Zeitpunkten dar. Bei anderen Frequenzwerten gilt die Interpolation nach Theorem 6.2.

Beispiel 6.2 – Interpolation mit DFT.

In Beispiel 6.1 hatten wir die Funktion $x(t) = cos(t)$ betrachtet. Wir hatten für die Zeitfenster $NT \in \{2\pi, 4\pi, \pi\}$ und für $N \in \{2, 4\}$ die DFT berechnet.
Wir untersuchen nun für $NT = 2\pi$ und $N = 2$, welche Form das aus den Spektralwerten der DFT rekonstruierte Spektrum von $x(t)$ annimmt.

Lösung:
Für $N = L = 2$ lauten die Abtastwerte $x[0] = 1$, $x[1] = -1$ und die Spektralwerte $X(0) = 0$, $X(1) = 2$. Mit Hilfes des Interpolationstheorems 6.2 erhalten wir

$$
\begin{aligned}
X(e^{j\Phi}) &= \frac{1}{2}\sum_{n=0}^{1} X_{DFT}[n] e^{\frac{j}{2}(n\pi-\Phi)} \frac{\sin(n\,2\pi - 2\Phi)}{\sin(n\pi - \Phi)} \\
&= e^{\frac{j}{2}(\pi-\Phi)} \frac{\sin(2\pi - 2\Phi)}{\sin(\pi - \Phi)} \\
&= -j\, e^{\frac{j}{2}(-\Phi)} \frac{\sin(2\Phi)}{\sin(\Phi)}
\end{aligned}
$$

und mit $\Phi = \omega T = \omega 2\pi/N = \omega\pi$:

$$
X(\omega) = -j\, e^{\frac{j}{2}(-\omega\pi)} \frac{\sin(2\omega\pi)}{\sin(\omega\pi)} \tag{6.33}
$$

Speziell für die interessierende Frequenz $\omega = 1$ gilt also

$$
X(\omega = 1) = -\lim_{\omega \to 1} \frac{\sin(2\omega\pi)}{\sin(\omega\pi)} = 2 \quad ,
$$

das heisst wir haben bestätigt, dass $X(\omega = 1) = X_{\mathrm{DFT}}(\omega = 1)$. Für andere Frequenzen gilt

$$
|X(\omega)| = \frac{\sin(2\omega\pi)}{\sin(\omega\pi)} \quad .
$$

Die ursprüngliche Spektrallinie $X(\omega) = \delta(1)$, die aus der analogen FT von $x(t) = cos(t)$ hervorging, wird also bei Interpolation aus den Werten der DFT „verschmiert" gemäss Gl. (6.33). Je grösser N ist, um so geringer ist diese Verschmierung, und das interpolierte Spektrum ist immer stärker konzentriert bei $\omega = 1$. Diese stärkere Konzentration kann am Betrag des interpolierten Spektrums betrachtet werden, der sich für wachsende N wie in Abb. 6.1 verhält. Zu weiteren Betrachtungen zur Interpolation bei einer einzelnen Spektrallinie siehe Übung 6.4. □

Damit aus den durch die DFT erhaltenen Spektralwerten die ursprüngliche Zeitfolge der Länge L rekonstruiert werden kann, müssen mindestens $N = L$ Abtastwerte vorhanden sein, wie im vorigen Abschnitt gezeigt wurde. Man kann die Anzahl N darüber hinaus aber beliebig erhöhen. Dadurch erhält man eine Überabtastung (engl. oversampling) und das Spektrum wird an mehr Stellen dargestellt. Im Grenzfall für $N \to \infty$ erhielte man wieder das kontinuierliche Spektrum selbst. Ein Informationsgewinn ergibt sich durch die Überabtastung allerdings nicht, zur Rekonstruktion des ursprünglichen Signals genügen $N = L$ Werte. Überabtastung wird aus Sicherheitsgründen benutzt, z.B. bei optischen Speichermedien (CD): werden einzelne Werte falsch abgetastet oder abgelesen oder gehen verloren, so ist die Qualitätseinbusse bei Überabtastung geringer.

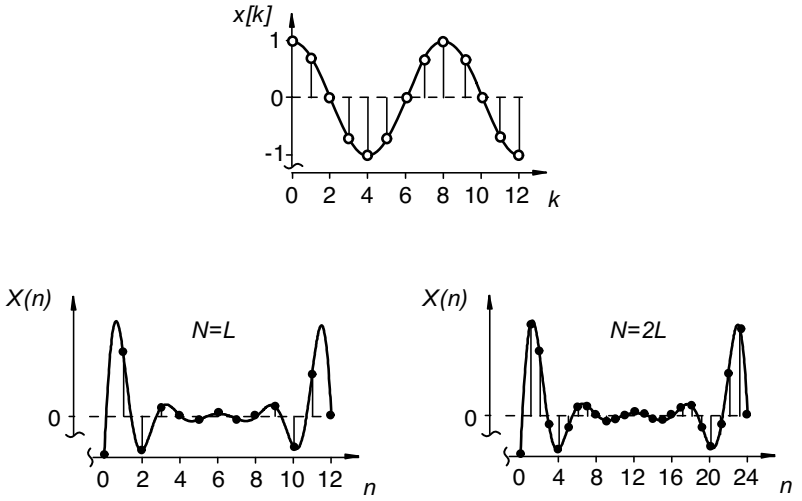


Abbildung 6.2. Abtastung einer zeitbegrenzten Signalfolge mit doppelt so vielen Abtastwerten. In beiden Fällen sind die Frequenzwerte der Abtastfolge exakt.

Beispiel 6.3 – Überabtastung.

Ein Beispiel für Überabtastung zeigt die Abb. 6.2. Eine zeitbegrenzte Signalfolge mit $L = 12$ Werten wird mit $N = 12$ und $N = 24$ abgetastet. In beiden Fällen sind die Frequenzwerte der Abtastfolge exakt, und die Folge ist auch exakt rekonstruierbar, da das Abtasttheorem erfüllt ist. Es ergibt sich kein Informationsgewinn durch Überabtastung. □

6.3 Eigenschaften der diskreten Fouriertransformation

Die Eigenschaften der DFT entsprechen im wesentlichen denen der analogen Fouriertransformation, es muss lediglich eine Anpassung an die diskreten Werte im Zeit- und Frequenzbereich vorgenommen werden. In Tabelle 6.1 werden die wichtigsten Eigenschaften kurz aufgelistet.

Eigenschaft	**Formel**
Definition	$\mathrm{DFT}\{x[k]\} = X[n]$ $\mathrm{IDFT}\{X[n]\} = x[k]$
Linearität	$\mathrm{DFT}\{\alpha x_1[k] + \beta x_2[k]\} = \alpha X_1[n] + \beta X_2[n]$
Zeitverschiebung	$\mathrm{DFT}\{x[k+i]\} = X[n]W_N^{-ni}$
Frequenzverschiebung	$\mathrm{IDFT}\{X[n+i]\} = x[k]W_N^{ki}$
Faltung	$\mathrm{DFT}\{x_1[k] * x_2[k]\} = X_1[n]X_2[n]$
Multiplikation	$\mathrm{DFT}\{x_1[n]x_2[n]\} = X_1[k] * X_2[k]$

Tabelle 6.1: Eigenschaften der Diskreten Fouriertransformation.

6.4 Fensterfolgen

Die diskrete Fouriertransformation kann nur auf endliche Folgen einer bestimmten Länge sinnvoll angewendet werden. Eine endliche Folge der Länge L gewinnt man durch die Multiplikation mit einer sogenannten Fensterfolge (engl. window). Im einfachsten Fall ist dies das Rechteckfenster:

$$w[k] = \begin{cases} 1 & \text{für } k = 0 \ldots L-1 \\ 0 & \text{sonst} \end{cases} . \tag{6.34}$$

Multipliziert man $w[n]$ mit der Folge $x[n]$, so erhält man deren Glieder von 0 bis $N-1$:

$$w[k]x[k] = x_N[k] = [x[0], x[1], \ldots x[L-1]] \tag{6.35}$$

Im Frequenzbereich entspricht die Multiplikation einer Faltung:

$$\text{DFT}\{w[k]x[k]\} = W[n] * X[n] \tag{6.36}$$

Die Fouriertransformierte der Rechteckfunktion ist eine sinc-Funktion (siehe Abb. 6.3a). Bei deren Faltung mit der Frequenzfolge des Signals ergibt sich ein ungünstiges Frequenzverhalten. Es entstehen Überschwinger, die sich auch mit beliebig vielen Stützstellen im Frequenzbereich nicht beseitigen lassen, das so genannte Gibbssche Phänomen.

Daher sollen nun einige gebräuchliche Fensterfolgen vorgestellt werden, die zu günstigeren spektralen Eigenschaften führen. Alle hier erwähnten Folgen lassen sich nach dem folgenden Cosinus-Schema definieren:

$$w[k] = \begin{cases} \alpha + \beta\cos(2\pi\frac{k}{N-1}) + \gamma\cos(4\pi\frac{k}{N-1}) & \text{für} \quad 0 \leqslant k \leqslant N-1 \\ 0 & \text{sonst} \end{cases} . \tag{6.37}$$

Nach ihren Parametern kann man nun vier Fensterfolgen unterscheiden. Das

Rechteck	$w^{\text{Re}}[n]$	1	0	0
Hanning	$w^{\text{Hn}}[n]$	0,5	0,5	0
Hamming	$w^{\text{Hm}}[n]$	0,54	0,46	0
Blackman	$w^{\text{Bl}}[n]$	0,42	0,5	0,08

Tabelle 6.2: Parameter der Fensterfolgen nach dem Cosinus-Schema

einfachste und bekannteste der aufgeführten Fenster ist das Hanningfenster[1]. Es führt im Vergleich zum Rechteckfenster zu einer wesentlich besseren Dämpfung im Sperrbereich (siehe Abb. 6.3b). Dieser Vorteil wird mit einem doppelt so breiten Durchlassbereich erkauft.

[1] In der Literatur ist ebenfalls die (richtige) Bezeichnung Hann-Fenster gebräuchlich.

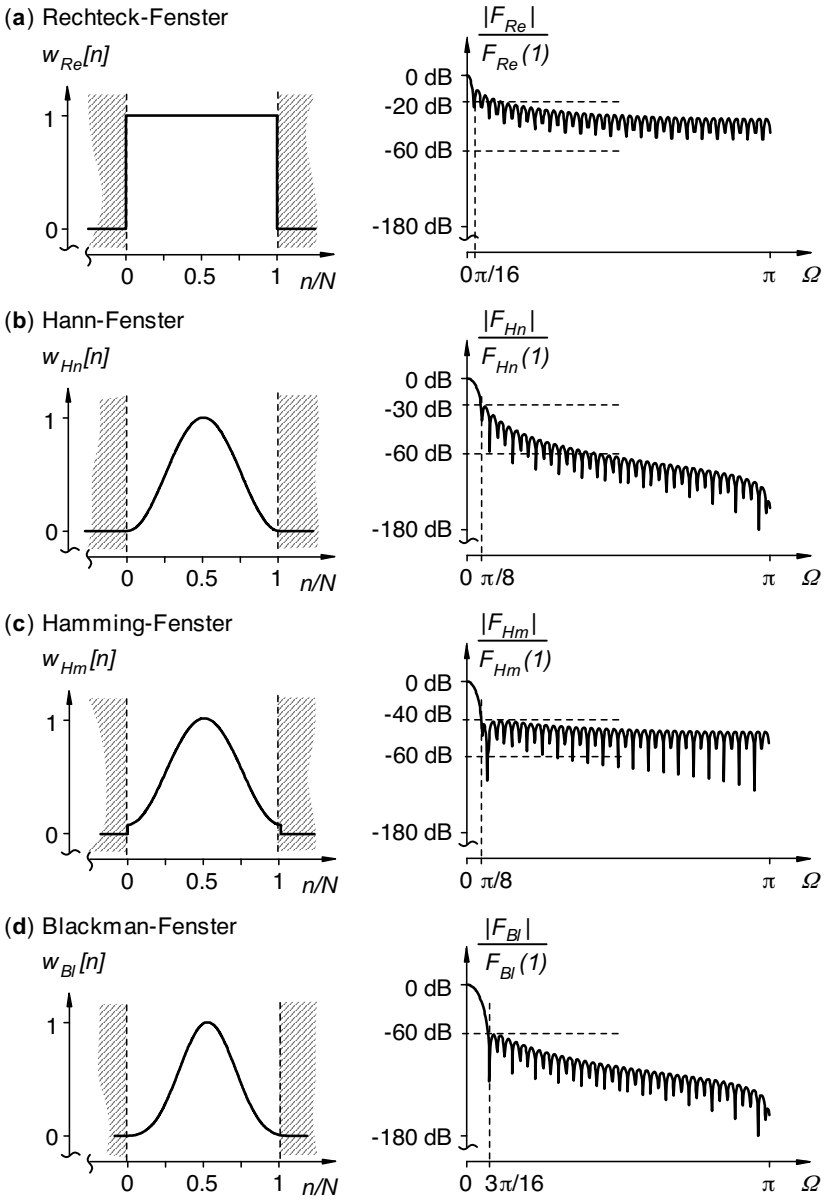


Abbildung 6.3. Verschiedene Fensterfolgen im Zeit- und Frequenzbereich. a) Rechteckfolge, b) Hanningfolge (auch Hannfolge genannt), c) Hammingfolge, d) Blackmanfolge

Das Hammingfenster stellt eine Optimierung des Hanningfensters dar. Wurden bei diesem die Parameter α und β eher intuitiv mit zweimal $0,5$ belegt, so werden sie bei der Hanningfolge so gewählt, dass das Hauptmaximum im Sperrbereich minimal wird. Obwohl der Unterschied in den Parametern nur gering wirkt, kann so die minimale Sperrdämpfung um $10dB$ erhöht werden(siehe Abb. 6.3c). Der Durchlassbereich entspricht dem der Hanningfolge.

Eine weitere Möglichkeit der Fensterung stellt die Blackmanfolge dar. Hier wird eine höhere Sperrdämpfung gegenüber den Hanning- und Hammingfenstern dadurch erreicht, dass ein noch breiterer Durchlassbereich in Kauf genommen wird (siehe Abb. 6.3d).

Die Tabelle 6.3 zeigt noch einmal zusammengefasst die Eigenschaften der einzelnen Fenster.

Name	**Fensterlänge** N	**Durchlassbereich** Ω_S	**min. Sperrdämpfung** a_{min}	**Abs. Maximalwert des Spektrums**
			(Tschebyscheff-Fenster)	
Rechteck	32	$\pi/16$	13 (21)	$32 = N(20)$
Hanning	32	$2\pi/16$	32	$15,6 = N/2$
Hamming	32	$2\pi/16$	42(47)	$16,8(16,6)$
Blackman	32	$3\pi/16$	58(74)	$13,0(14,3)$

Tabelle 6.3: Eigenschaften der verschiedenen Fensterfolgen.
In Klammern: Werte unter Verwendung des Tschebyscheff-Fensters

Im Vergleich zu den hier genannten Fensterfolgen kann die Sperrdämpfung durch die Verwendung der sogenannten Tschebyscheff-Fenster weiter erhöht werden, ohne dass dies zu einem breiteren Durchlassbereich führt. Tschebyscheff-Fenster erreichen die optimale Sperrdämpfung bei gleichmässiger Approximation im Sperrbereich. Da die theoretische Behandlung dieser Fensterfolgen recht aufwändig ist, sollen sie hier nur am Rande erwähnt sein. Die resultierenden Sperrdämpfungen und Maximalwerte sind in obiger Tabelle in Klammern so angegeben, dass sie zu denselben Fensterlängen und Durchlassbereichen korrespondieren wie die entsprechenden Fenster des Cosinus-Schemas. Man erkennt, dass eine $5 - 16dB$ höhere Dämpfung erreicht werden kann. Darüber hinaus sieht man, dass vor allem das Hamming-Fenster bereits eine sehr gute Annäherung an das korrespondierende Tschebyscheff-Fenster darstellt, was seine breite Verwendung erklärt. Näheres zum Tschebyscheff-Entwurf wird unter dem Thema „Digitale Filter“ , z.B. in (Kammeyer und Kroschel, 2002; von Grünigen, 2001) behandelt.

6.5 Die schnelle Fouriertransformation (FFT)

Dieser Abschnitt beschäftigt sich mit der Implementierung der DFT. Es soll gezeigt werden, wie mit geeigneten Algorithmen der erforderliche Rechenaufwand erheblich reduziert werden kann. Ausgangspunkt ist die am Anfang des Kapitels hergeleitete Definitionsgl. (6.20):

$$X_{DFT}[n] = \sum_{k=0}^{N-1} x[k] W_N^{kn}. \tag{6.38}$$

Für die rechnergestützte Bearbeitung des Problems bietet sich das Matrixschema an. Damit erhält man für den Zusammenhang zwischen der Zeitfolge $f[n]$ und der Frequenzfolge $F[n]$ folgende Schreibweise:

$$\begin{pmatrix} F[0] \\ F[1] \\ F[2] \\ \vdots \\ F[N-1] \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & W_N^1 & W_N^2 & \cdots & W_N^{N-1} \\ 1 & W_N^2 & W_N^4 & \cdots & W_N^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & W_N^{N-1} & W_N^{2(N-1)} & \cdots & W_N^{(N-1)^2} \end{pmatrix} \begin{pmatrix} f[0] \\ f[1] \\ f[2] \\ \vdots \\ f[N-1] \end{pmatrix}. \tag{6.39}$$

Zur Lösung dieser Matrixmultiplikation sind N^2 komplexe Multiplikationen und eben so viele komplexe Additionen nötig.

Eine häufig verwendete Möglichkeit zur Verringerung der nötigen Operationen ist die schnelle Fouriertransformation (engl. **F**ast **F**ourier **T**ransformation, FFT), siehe auch (Brigham, 1995). Diese lässt sich wie folgt aus der DFT herleiten. Zunächst wird vorausgesetzt, dass N eine Potenz von 2 ist. Ist dies nicht der Fall, kann dieser Zustand leicht durch das Anhängen von Nullen (*zero padding*) hergestellt werden. Nun teilt man den Vektor der Zeitfolge $f[n]$ in einen Vektor $\mathbf{f_g}$ der geraden Folgeglieder und einen Vektor $\mathbf{f_u}$ der ungeraden Folgeglieder. Dann multipliziert man beide mit einer Matrix $\mathbf{U}$ bzw. $\mathbf{V}$ so, dass die Summe beider Produkte gerade den Vektor $\mathbf{F_1}$ der ersten $N/2$ Folgeglieder der Frequenzfolge $F[n]$ ergibt.

Damit erhält man die folgende Gleichung:

$$\mathbf{F_1} = \mathbf{V f_g} + \mathbf{U f_u} = \mathbf{F_g} + \mathbf{F_u} \tag{6.40}$$

oder in ausgeschriebener Form

$$\begin{pmatrix} F[0] \\ F[1] \\ F[2] \\ \vdots \\ F[\frac{N}{2}-1] \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & W_N^2 & W_N^4 & \cdots & W_N^{N-2} \\ 1 & W_N^4 & W_N^8 & \cdots & W_N^{(N-2)2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & W_N^{2(N/2-1)} & W_N^{4(N/2-1)} & \cdots & W_N^{(N-2)(N/2-1)} \end{pmatrix} \begin{pmatrix} f[0] \\ f[2] \\ f[4] \\ \vdots \\ f[N-2] \end{pmatrix}$$
$$+ \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ W_N^1 & W_N^3 & W_N^5 & \cdots & W_N^{N-1} \\ W_N^2 & W_N^6 & W_N^{10} & \cdots & W_N^{(N-1)2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ W_N^{N/2-1} & W_N^{3(N/2-1)} & W_N^{5(N/2-1)} & \cdots & W_N^{(N-1)(N/2-1)} \end{pmatrix} \begin{pmatrix} f[1] \\ f[3] \\ f[5] \\ \vdots \\ f[N-1] \end{pmatrix} \tag{6.41}$$

Wenn man die beiden Matrizen $\mathbf{U}$ und $\mathbf{V}$ zeilenweise miteinander vergleicht, kann man erkennen, dass diese in einfacher Weise voneinander abhängen. Diese Abhängigkeit kann man mit Hilfe einer Diagonalmatrix $\mathbf{D}$ wie folgt beschreiben:

$$\mathbf{U} = \begin{pmatrix} W_N^0 & 0 & 0 & \cdots & 0 \\ 0 & W_N^1 & 0 & \cdots & 0 \\ 0 & 0 & W_N^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & W_N^{N/2-1} \end{pmatrix} \mathbf{V} = \mathbf{DV}. \tag{6.42}$$

Damit lassen sich die ersten $N/2$ Glieder der Frequenzfolge nach folgender Vorschrift berechnen:

$$\mathbf{F}_1 = \mathbf{V}\mathbf{f}_g + \mathbf{DV}\mathbf{f}_u = \mathbf{F}_g + \mathbf{F}_u. \tag{6.43}$$

Die Frage ist nun, ob sich auch die zweite Hälfte der Frequenzfolge mit diesen Ergebnissen bestimmen lässt. Dafür nutzt man die Tatsache, dass jedes Element des Vektors $\mathbf{F}_2$ der Frequenzfolgeglieder $N/2 \ldots N-1$ gegenüber dem entsprechenden Element des Vektors $\mathbf{F}_1$ um $N/2$ verschoben ist. Setzt man dies in die Definitionsgleichung ein, so erhält man:

$$\begin{aligned} X[N/2+m] &= \sum_{k=0}^{N-1} x[k] W_N^{k(N/2+m)} = \sum_{k=0}^{N-1} x[k] e^{-j\frac{2\pi}{N}k(N/2+m)} \\ &= \sum_{k=0}^{N-1} x[k] e^{-j\pi k} e^{-j\frac{2\pi}{N}km} = \sum_{k=0}^{N-1} x[k](-1)^k e^{-j\frac{2\pi}{N}km} \\ &= \sum_{k=0,\text{gerade}}^{N-1} x[k](-1)^k e^{-j\frac{2\pi}{N}km} + \sum_{k=1,\text{ungerade}}^{N-1} x[k](-1)^k e^{-j\frac{2\pi}{N}km} \\ &= \sum_{k=0,\text{gerade}}^{N-1} x[k] e^{-j\frac{2\pi}{N}km} - \sum_{k=1,\text{ungerade}}^{N-1} x[k] e^{-j\frac{2\pi}{N}km} \end{aligned} \tag{6.44}$$

Die Verschiebung um $N/2$ hat also zur Folge, dass der ungerade Anteil nicht addiert, sondern subtrahiert wird:

$$\mathbf{F}_2 = \mathbf{V}\mathbf{f}_g - \mathbf{D}\mathbf{V}\mathbf{f}_u = \mathbf{F}_g - \mathbf{F}_u. \tag{6.45}$$

Damit ist also ein Weg gefunden, die gesamte Frequenzfolge aus den Vektoren $\mathbf{f}_g$ und $\mathbf{f}_u$ sowie den Matrizen $\mathbf{D}$ und $\mathbf{V}$ zu bestimmen. Somit muss nicht mehr mit einer Matrix vom Format $(N \times N)$, sondern zweimal mit einer Matrix vom Format $(N/2 \times N/2)$ gerechnet werden. Dafür sind nur halb so viele Operationen notwendig. Zusätzlich muss allerdings die Matrix $\mathbf{U} = \mathbf{DV}$ berechnet werden. Da $\mathbf{D}$ eine Diagonalmatrix ist, ist der Aufwand hierfür gering. Wir fassen zusammen:

Theorem 6.3 (Schnelle Fourier-Transformation (FFT)).
Die diskrete Fourier-Transformation lässt sich durch Aufteilung in Berechnungsblöcke der halben Grösse als schnelle Fourier-Transformation (FFT) wie folgt durchführen:

$$\mathbf{F}_1 = \mathbf{V}\mathbf{f}_g + \mathbf{D}\mathbf{V}\mathbf{f}_u = \mathbf{F}_g + \mathbf{F}_u \tag{6.46}$$

$$\mathbf{F}_2 = \mathbf{V}\mathbf{f}_g - \mathbf{D}\mathbf{V}\mathbf{f}_u = \mathbf{F}_g - \mathbf{F}_u. \tag{6.47}$$

Der in Abb. 6.4 dargestellte Graph zeigt anschaulich den Ablauf der Berechnungen.

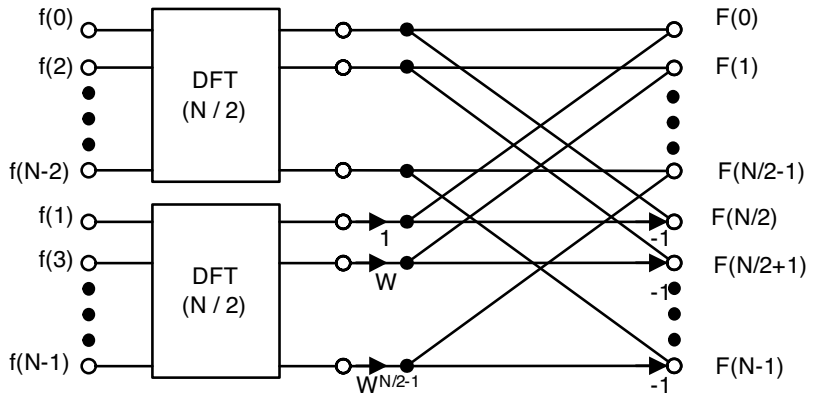


Abbildung 6.4. Erste Stufe der Radix-2-DFT bei Reduktion im Zeitbereich

Genauso wie man den Vektor $\mathbf{F}_N$ in die Vektoren $\mathbf{F}_1$ und $\mathbf{F}_2$ geteilt hat, kann man nun diese wiederum in zwei halb so große Vektoren aufteilen. Da N eine Potenz von zwei ist, kann man dies bis hin zu Vektoren mit nur einem Element fortsetzen. Durch geschicktes Ausnutzen der Periodizität des Drehoperators W_N^k erhält man zum Beispiel für $N = 8$ den Signalflussplan in Abb. 6.5.

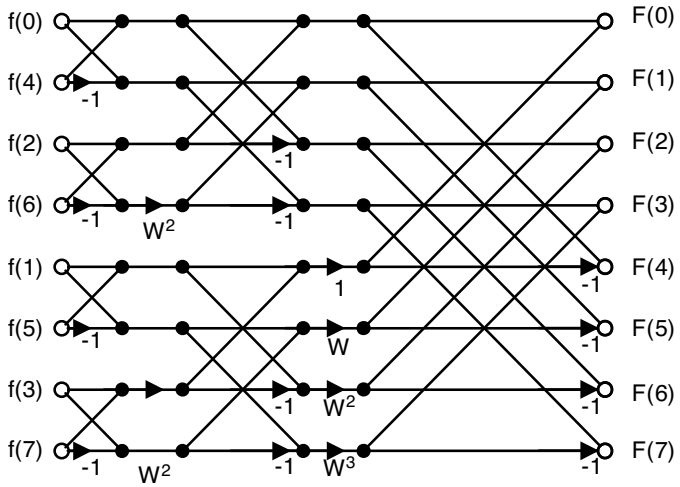


Abbildung 6.5. Vollständige Radix-2-DFT bei Reduktion im Zeitbereich für $N = 8$

Die eigentliche Definition der DFT wird zur Berechnung nach diesem Muster gar nicht mehr benötigt. An ihre Stelle tritt die sogenannte Schmetterlings-Operation („butterfly“) die durch den Schmetterlings-Graphen in Abb. 6.6 dargestellt werden kann:

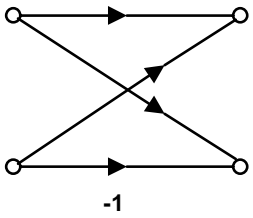


Abbildung 6.6. FFT für $N = 2$: „butterfly“ als Elementaroperation

Die regelmäßige Struktur der FFT wird dadurch erreicht, dass die Reihenfolge der Eingangsglieder aufgegeben wird. Die auf den ersten Blick etwas unübersichtliche Anordnung kann durch eine einfache Operation, die Bitumkehr, hergestellt werden. Dabei werden die höchstwertigsten Bits in die niederwertigsten gespiegelt. Die Tabelle 6.4 zeigt die Auswirkungen für das Beispiel $N = 8$. Sie kann aber auch für alle anderen Potenzen von 2 angewendet werden.

Durch die Anwendung des FFT-Algorithmus erreicht man, dass sich der Aufwand für die Berechnung der DFT auf $N \operatorname{ld}(N)$ Operationen beschränkt. Dies führt für große N zu enormen Einsparungen. Für $N = 1024$ lässt sich die

FFT beispielsweise mit nur 1% der Operationen lösen, die der einfache DFT-Algorithmus (mit N^2 Operationen) bräuchte. Selbst für $N = 513$ lohnt es sich somit, ein zero-padding von 511 Nullen durchzuführen, der Rechenaufwand mit FFT beträgt dann immer noch nur ca. 4% des Aufwandes mit DFT. Der

k	**Dualzahl**	**Bitumkehr**	k'
0	000	000	0
1	001	100	4
2	010	010	2
3	011	110	6
4	100	001	1
5	101	101	5
6	110	011	3
7	111	111	7

Tabelle 6.4: Bitumkehr-Indizierung für 3-stellige Bits.

hier beschriebene Algorithmus ist die Radix-2-FFT mit Reduktion im Zeitbereich. Alternativ dazu kann die Reduktion auch im Frequenzbereich erfolgen. Diesen Algorithmus erhält man, wenn im hier beschriebenen Verfahren die Zeitfolge mit der Frequenzfolge vertauscht wird. Eine weitere gebräuchliche Möglichkeit ist die Radix-4-FFT. Hier wird der Vektor nicht in zwei sondern vier Teilvektoren zerlegt. Durch den Radix-4-Algorithmus können weitere Rechenschritte eingespart werden, die Programmierung ist allerdings wesentlich aufwändiger als die der Radix-2-FFT.

Beispiel 6.4 – Berechnung der FFT.

Wir verfolgen das Beispiel 6.1 mit der Funktion $x(t) = cos(t)$ weiter. Sie werde an N Stellen $t = k \cdot 2\pi/N$ abgetastet, also $T = 2\pi/N$. Wir ermitteln die Spektralwerte bei den Frequenzen $\omega = n\frac{2\pi}{NT} = n$ mit Hilfe der FFT für $N = 2$ und $N = 4$.

Lösung:
Für $N = 2$ lauten die Abtastwerte $x[0] = 1$, $x[1] = -1$. Die FFT verlangt für $N = 2$ nur eine elementare butterfly-Operation. Bitumkehr ergibt keine Änderung. Das Ergebnis lautet:
$X(0) = x[0] + x[1] = 0,$
$X(1) = x[0] - x[1] = 2.$

Für $N = 4$ lauten die Abtastwerte $x[0] = 1$, $x[1] = 0$, $x[2] = -1$, $x[3] = 0$. Die Bitumkehr liefert als Eingangssequenz $(0, 2, 1, 3)$. Daraus berechnen wir den Ausgang der ersten Stufe einer Radix-2-FFT (siehe Abb. 6.5):
$X'(0) = x[0] + x[2] = 0,$
$X'(1) = x[0] - x[2] = 2,$
$X'(2) = x[1] + x[3] = 0,$

$X'(3) = x[1] - x[3] = 0.$
Daraus erhalten wir den Eingang der zweiten Stufe der FFT durch Multiplikation von $X'(3)$ mit W_N^2. Diese ergibt hier keine Änderung. Den Ausgang der zweiten Stufe erhalten wir gemäss Abb. 6.5 mit:
$X(0) = X'(0) + X'(2) = 0,$
$X(1) = X'(1) + X'(3) = 2,$
$X(2) = X'(0) - X'(2) = 0,$
$X(3) = X'(1) - X'(3) = 2.$
Dies entspricht genau dem Ergebnis aus Beispiel 6.1. □

6.6 Die inverse schnelle Fouriertransformation (IFFT)

Bei der Berechnung der Inversen DFT kann man folgendermaßen vorgehen. Man erkennt aus den Definitionsgleichungen, dass sich die Hin- und die Rücktransformation nur durch den Drehoperator und den Normierungsfaktor $1/N$ unterscheiden. Die IDFT kann also nach demselben Verfahren bestimmt werden wie die DFT. Es muss lediglich bei der Definition des Drehoperators das Vorzeichen des Exponenten geändert werden. Die anschließende Division durch N kann, da N eine Potenz von 2 ist, durch einfaches Verschieben der Bits erreicht werden.

Beispiel 6.5 – Inverse FFT.

Wir verfolgen das Beispiel 6.4 mit der Funktion $x(t) = cos(t)$ weiter. Ihre DFT wurde an den Stellen $X(n)$ für $\omega_0 = 1$ berechnet. Wir ermitteln die Werte der Funktion im Zeitbereich bei den Zeitpunkten $t = k \cdot 2\pi/N$ durch IFFT.

Lösung:
Für $N = 2$ lauten die Spektralwerte der DFT $X(0) = 0$, $X(1) = 2$. Die IFFT verlangt für $N = 2$ nur eine elementare butterfly-Operation. Bitumkehr ergibt keine Änderung. Das Ergebnis lautet:

$x[0] = (1/2)(X(0) + X(1)) = 1,$
$x[1] = (1/2)(X(0) - X(1)) = -1.$

Für $N = 4$ lauten die Spektralwerte der DFT $X(0) = 0$, $X(1) = 2$, $X(2) = 0$, $X(3) = 2$. Die Bitumkehr liefert als Eingangssequenz $(0, 2, 1, 3)$. Daraus berechnen wir den Ausgang der ersten Stufe einer Radix-2-IFFT (siehe Abb. 6.5):
$x'[0] = X(0) + X(2) = 0,$
$x'[1] = X(0) - X(2) = 0,$
$x'[2] = X(1) + X(3) = 4,$
$x'[3] = X(1) - X(3) = 0.$

Daraus erhalten wir den Eingang der zweiten Stufe der IFFT durch Multiplikation von $x'[3]$ mit W_N^{-2} - beachten Sie den negativen Exponenten! Diese Multiplikation ergibt hier keine Änderung. Den Ausgang der zweiten Stufe erhalten wir gemäss Abb. 6.5 mit:
$x[0] = (1/4)(x'[0] + x'[2]) = 1,$
$x[1] = (1/4)(x'[1] + x'[3]) = 0,$
$x[2] = (1/4)(x'[0] - x'[2]) = -1,$
$x[3] = (1/4)(x'[1] - x'[3]) = 0.$
Dies entspricht genau den ursprünglichen Abtastwerten der Funktion $x(t) = cos(t)$ bei $t = k \cdot \pi/2$.

□

Eine weitere Variante zur Berechnung der IDFT geht einen anderen Weg. Hier wird die Definitionsgl. (6.21) der inversen DFT umgeformt:

$$\begin{aligned}\text{IDFT}\{X[n]\} &= \frac{1}{N}\sum_{n=0}^{N-1} X[n]W_N^{-kn} = \frac{j}{N}\sum_{n=0}^{N-1}(-j)X[n]W_N^{-kn} \\ &= \frac{j}{N}\left[\sum_{n=0}^{N-1}(-j)^*(X[n])^*(W_N^{-kn})^*\right]^* = \frac{j}{N}\left[\sum_{n=0}^{N-1} jX^*[n]W_N^{kn}\right]^* \\ &= \frac{1}{N}\, j\left[\text{DFT}\{jX^*[n]\}\right]^* \end{aligned} \tag{6.48}$$

Dabei bedeutet der Zusammenhang

$$jZ^* = j\left(\text{Re}\{Z\} + j\text{Im}\{Z\}\right) = j\text{Re}\{Z\} - \text{Im}\{Z\} \tag{6.49}$$

eine Vertauschung von Real- und Imaginärteil, sowie Vorzeichenänderung des neuen Realteils.

Somit erhält man die IDFT, indem man zunächst Real- und Imaginärteil der Frequenzfolge vertauscht und das Vorzeichen des neuen Realteils ändert. Danach wendet man den FFT-Algorithmus an. Beim Ergebnis wird nun ein weiteres Mal der Real- mit dem Imaginärteil vertauscht und das Vorzeichen des neuen Realteils geändert. Multipliziert man nun noch mit dem Normierungsfaktor $1/N$, so erhält man das Ergebnis der IDFT. Der Vorteil dieser Variante ist, dass sowohl die Hin- als auch die Rücktransformation mit Hilfe desselben Algorithmus behandelt werden können. Man kann also z.B. dieselbe hardware benutzen und benötigt lediglich die zusätzlichen Vertauschungs- und Normierungsoperationen.

Übungen

Übung 6.1 – Fensterfolgen.

Gegeben sind zwei Sinussignale mit ähnlicher Frequenz aber unterschiedlicher Leistung

$$x(t) = U_1 \sin(2\pi f_1 t) + U_2 \sin(2\pi f_2 t)$$

mit $U_1 = 1\text{V}$, $U_2 = 10\text{mV}$, $f_1 = 990\text{Hz}$ und $f_2 = 1010\text{Hz}$. Das Signal wird mit einer Abtastfrequenz von $f_s = 4\text{KHz}$ mit $N = 1024$ Abtastwerten abgetastet. Untersuchen Sie qualitativ (hinsichtlich spektraler Auflösung und Leckeffekt) ob man die beiden Frequenzen des Signals bei Verwendung eines Rechteck- bzw. eines Hanning-Fensters erkennen kann. Welches Fenster ist demnach besser geeignet?

Übung 6.2 – Signalflußdiagramm der FFT.

Stellen Sie für $N = 4$ Abtastwerte, ausgehend von der Definition der diskreten Fouriertransformation das Signalflußdiagramm einer Radix-2-FFT auf!

Übung 6.3 – Periodizität der DFT für eine Winkelfunktion.

Betrachten Sie Beispiel 6.1 auf Seite 155. Weisen Sie zunächst nach, dass die Funktion $x(t) = cos(t)$ nur Spektrallinien bei $\omega = 1$ und $\omega = -1$ hat. Dazu siehe auch Gl. 4.4 auf Seite 90.

Weisen Sie nun durch direkte Berechnung der DFT für $NT = 2\pi$ nach, dass für $x(t) = cos(t)$ und für beliebige N gilt: $X(1) = X(N-1) \neq 0$.

Übung 6.4 – Interpolation.

Betrachten Sie die Funktion $x(t) = \exp(jt)$. Weisen Sie zunächst nach, dass diese Funktion nur eine Spektrallinie bei $\omega = 1$ hat. Berechnen Sie die DFT für $NT = 2\pi$ und weisen Sie nach, dass für beliebige N gilt: $X(1) = N$, $X(n) = 0$ (sonst).

Mit diesem Ergebnis berechnen Sie nun für beliebige N und $NT = 2\pi$ das Frequenzverhalten $X(\omega)$, das die DFT interpoliert. Verwenden Sie das Interpolationstheorem 6.2 auf Seite 158 und orientieren Sie sich am Beispiel 6.2 auf Seite 159. Zeigen Sie am Ergebnis, dass mit wachsendem N das interpolierte Spektrum sich immer mehr dem wahren annähert.

Übung 6.5 – FFT für 8 Abtastwerte.

Gegeben ist eine Folge $x[n]$ mit ($N = 8$ Abtastwerte)

$$x[n] = \begin{cases} +1 & 0 \leq n \leq 3 \\ -1 & 4 \leq n \leq 7 \end{cases}.$$

Bestimmen Sie die diskrete Fourier-Transformation der Folge mittels eines Signalflußdiagramms einer Radix-2-FFT!

Übung 6.6 – FFT bei unpassender Anzahl von Abtastwerten.

Wie kann man vorgehen, wenn die Zahl der Abtastwerte keine Zweierpotenz ist?

7

Stochastische Signalverarbeitung

7.1 Das Komplexitätsproblem

Im Rahmen der bisherigen Betrachtungen wurde davon ausgegangen, dass physikalische Systeme – die durch verschiedene mathematische Darstellungsformen beschrieben werden können – ein streng deterministisches Verhalten zeigen: Aus einer bekannten Systemfunktion, einem initialen Systemzustand und der Kenntnis des Eingangssignals soll es möglich sein, das Ausgangssignal fehlerfrei zu berechnen. Diese Feststellung besitzt zwar durchaus ihre allgemein gültige Richtigkeit, jedoch zeigt es sich, dass jedes beliebige reale physikalische System durch eine weit reichende Wechselwirkung mit seiner Umwelt derartig komplex ist, dass eine allumfassende exakte mathematische Beschreibung schlichtweg nicht möglich ist. Auf der anderen Seite gibt es physikalische Prozesse, die sich grundsätzlich einer deterministischen Betrachtung entziehen. Als Beispiele seinen hier die thermodynamischen Rauschprozesse und die generelle Unschärfe im Bereich der Quantenmechanik genannt.

7.2 Grenzen der deterministischen Betrachtungsweise

In der Praxis macht sich das Fehlen eines exakten Systemmodells bekanntermaßen in einer mehr oder minder starken Abweichung zwischen dem berechnetem und dem tatsächlichem Ausgangssignal bemerkbar. In Ermangelung eines genaueren Modells wird dieser Fehler als „zufällig“ (oder „statistisch“) bezeichnet. Eine genauere Erfassung von weiteren Einflussfaktoren auf ein System verringert zwar im Allgemeinen diesen Fehler; jedoch ist die daraus resultierende Komplexität der mathematischen Darstellung ab einem gewissen Grade nicht mehr vertretbar und die Systembeschreibung damit ggf. unbrauchbar. Ist man an dieser Stelle mit den Möglichkeiten der Signalverarbeitung am Ende?

Günstiger Weise stellt sich bei genauem Überdenken der Problemstellung in vielen Fällen gar nicht die Frage nach einer exakten Modellierung eines

physikalischen Systems. Dies soll am Beispiel der Spracherkennung erläutert werden.

7.3 Ein Beispiel aus der Sprachverarbeitung

Die Spracherkennung setzt sich zum Ziel, aus einem Schallsignal auf die ursprünglich gesprochene Wortfolge eines Sprechers zu schließen (Wendemuth, 2004). Bevor das Signal am Mikrofon angelangt, hat es jedoch eine enorme Zahl von Systemen mit einer entsprechend vielfältigen Zahl von unbekannten Parametern durchlaufen. Dies beginnt bei der Möglichkeit des Sprechers, dasselbe Wort akustisch unterschiedlich zu formen (lauter, höhere Tonlage, schneller, gereizt, ...). Wollte man Systemparameter finden, um diese akustische Charakteristik formal und kausal zu beschreiben, kämen z.B. Stimmung und Vorwissen des Sprechers sowie Grad der Störung durch Umgebungsgeräusche in Frage, sowie viele weitere mögliche Parameter. Hier wird bereits klar, dass eine formale Beschreibung dieses "Systems" der akustischen Signalbildung beim Sprecher extrem komplex ist und mit hoher Wahrscheinlichkeit formal nicht zufriedenstellend gelingt. Aber auch der weitere physikalische Weg des akustischen Signals vom Sprecher zum Mikrophon unterliegt grösstenteils unbekannten oder unvorhersehbaren Einflüssen wie etwa Störgeräuschen, der akustischen Umgebungscharakteristik, eventuellen Charakteristiken eines Übertragungskanals (Telefon) etc. Letztlich wird das Signal aufgenommen mit einem Mikrofon mit spezieller Frequenz- und Richtcharakteristik, variabler Ausrichtung im Raum, und es wird umgewandelt in ein elektrisches Signal, das zu allem Überfluss von thermischen und anderen Rauschsignalen überlagert wird.

In einem ersten Ansatz könnte man geneigt sein, jedes Teilsystem der gesamten Prozesskette zu analysieren und dessen Parameter durch Messungen zu bestimmen. Aus einer sehr großen Datenbank könnte man dann das beobachtete Sprachmuster heraussuchen und erhält das ursprünglich gesprochene Wort. Ganz offensichtlich stellt dies ein aussichtsloses und höchst unpraktikables Unterfangen dar – dass es auch anders gehen muss, dafür ist der Mensch selbst ein lebender Beweis.

7.4 Motivation der stochastischen Signalverarbeitung

Überdenken wir die Problemstellung im Falle der Spracherkennung noch einmal genauer, so stellen wir fest, dass es uns nicht interessiert, ob der Sprecher zehn oder zwölf Zentimeter vom Mikrofon entfernt ist, ob er laut oder leise, schnell oder langsam, gut artikuliert oder nachlässig spricht. All diese Parameter sind zwar Bestandteil des realen Systems und auf das konkret am Mikrofon anliegende Signal (als „**Einzelbeobachtung**“ oder auch „**Muster**“ bezeichnet) haben sie ohne Zweifel einen Einfluss. Das heißt aber nicht, dass all diese

Parameter für die *Unterscheidung* sämtlicher Worte aus dem Vokabular des Sprechers notwendig sind.

Man könnte sich vorstellen, dass die gesprochene Wortfolge ein deterministisches Signal sei, dem in vielfältigster Weise ein statistisches Rauschen additiv, multiplikativ oder durch Faltung überlagert wird. Der Gesamtprozess, der ein derart verrauschtes Signal erzeugt, soll im Folgenden als „**stochastischer Prozess**" bezeichnet werden.

Wir weisen hier darauf hin, dass der eben eingeführte Unterschied deterministisch – stochastisch in diesem Buch auf *Signale* beschränkt bleibt. Man kann darüber hinaus ebenso von stochastischen Systemen sprechen. Dies wären Systeme, deren Übertragungsverhalten eine Zufallskomponente aufweist. Etwa könnte die Übertragungsfunktion eines Telefonkanals eine Gesamtverstärkung oder eine Frequenzcharakteristik aufweisen, die durch Umgebungseinflüsse in gewissen Grenzen schwankt. Dies könnte z.B. beschrieben werden, indem die Parameter der Übertragungsfunktion des Kanals keine Konstanten sind, sondern modelliert werden als ein Zufallsprozess, der aus einer vorgegebenen Wahrscheinlichkeitsverteilung zum jeweiligen Zeitpunkt der Beschreibung den interessierenden aktuellen Parameter zufällig „zieht". Wir wollen solche stochastischen Systeme hier nicht beschreiben, sondern uns mit deterministischen Systemen und stochastischen Signalen beschäftigen.

Die Stochastische Signalverarbeitung setzt sich zum Ziel, die durch einen stochastischen Prozess erzeugten Signale sinnvoll weiter zu verarbeiten. Häufig besteht die Aufgabe – wie im oben genannten Beispiel – darin, das „Nutzsignal" aus dem verrauschten stochastischen Signal zuverlässig zu rekonstruieren. Warum dies möglich ist, und welche Methoden dazu verwendet werden können, ist Gegenstand dieses Kapitels.

7.5 Zeitdiskrete stochastische Prozesse

Dieses Kapitel bietet eine grundlegende Einführung in die Beschreibung der Eigenschaften und des Übertragungsverhaltens von deterministischen Systemen bei Anregung mit stochastischen Signalen. Dabei wird auf eine mathematisch erschöpfende Behandlung der wahrscheinlichkeitstheoretischen Begriffe und ihrer Anwendungen verzichtet. Hier sei auf die weitergehende Literatur (Behnen und Neuhaus, 1995; Beichelt, 1997; Böhme, 1998) verwiesen.

7.5.1 Grundlegende stochastische Begriffe

Im Gegensatz zu deterministischen Signalen interessiert bei stochastischen Signalen nicht nur ein bestimmter Signalverlauf, sondern vielmehr ein Modell, das die Menge aller möglichen zufälligen Signalverläufe hinreichend genau beschreibt. Solche Modelle bezeichnet man in der Wahrscheinlichkeitstheorie als stochastische Prozesse oder Zufallsprozesse.

Sei $X(k)$ eine Zufallsvariable über der Menge Ω aller Elementarereignisse ω und aller Teilmengen von Ω, die man als Ereignisalgebra $\mathcal{A}$ bezeichnet. $X(k)$ repräsentiere ein Modell für das zum Zeitpunkt k gemessene Signal. Wir bezeichnen dann die Gesamtheit dieser Zufallsvariablen für eine bestimmte Indexmenge I als stochastischen Prozess. Im Falle diskreter stochastischer Prozesse gilt $I \subset \mathbb{Z}$. Ein *diskreter stochastischer Prozess* ist also eine Abbildung

$$X: \quad \mathbb{Z} \times \Omega \to \mathbb{R}, \quad (k, \omega) \to X(k, \omega), \tag{7.1}$$

die jedem Zeitpunkt k und jedem Elementarereignis ω eine reelle Zahl zuordnet. Bei festem k erhält man eine Zufallsvariable $X(k)$. Für festes ω erhält man eine Zeitfunktion $x[k]$, die wir als *eine* mögliche Realisierung des stochastischen Prozesses auffassen können. Dies könnte z. B. ein zu bestimmten Zeitpunkten gemessener Spannungsverlauf an den Anschlüssen eines elektronischen Bauelements sein. Im Falle zeitdiskreter Prozesse spricht man auch von *Musterfolgen* $x[k]$. Bei endlicher Indexmenge handelt es sich bei diesen Musterfolgen um Vektoren $\mathbf{x}[k]$. Zur Veranschaulichung dieser Begriffe soll Bsp. 7.1 dienen.

Beispiel 7.1 – Würfeln als stochastischer Prozess.

Zwei Würfel werden gleichzeitig N mal geworfen, wobei jedesmal die Augensumme festgehalten wird.
Die Indexmenge I ist folglich die Menge $\{1, 2, \ldots, N\}$. Mögliche Elementarereignisse ω sind Paare von Augenzahlen wie z. B. $(2, 3)$, $(5, 5)$, $(1, 6)$ usw., so dass wir als Menge Ω aller Elementarereignisse das kartesische Produkt $\{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$ erhalten. Die Zufallsvariable X ist eine Abbildung, die jedem Elemntarereignis ω die Summe der Augenzahlen der beiden Würfel zuordnet. Es können also nur natürliche Zahlen von 2 bis 12 auftreten. Die N-fache Wiederholung dieses Zufallsexperiments stellt einen stochastischen Prozess dar. Eine mögliche Realisierung dieses Zufallsprozesses ist für $N = 5$ der Vektor

$$\mathbf{x}[k] = \begin{pmatrix} 3 & 7 & 4 & 2 & 12 \end{pmatrix}^T.$$

Man sieht, dass sogar für dieses verhältnismäßig einfache Beispiel die Zahl aller Realisierungen des stochastischen Prozesses 11^5 beträgt.

□

Eine Messung einer kompletten (zeitlichen) Folge von Stichproben an einem zeitdiskreten Prozess liefert *eine* Musterfolge $x[k]$. Eine Wiederholung des Experiments mit gleichen Anfangsbedingungen wird je nach Prozess mit hoher Wahrscheinlichkeit eine andere Musterfolge liefern. Soll nun der stochastische Prozess selbst beschrieben werden, reicht im Gegensatz zu den zeitdiskreten deterministischen Signalen nicht die Angabe *einer* Musterfolge, sondern vielmehr ist die Menge aller aufgrund der Wahrscheinlichkeitsvorgabe möglichen zufälligen Signalverläufe zu erfassen und in geeigneter Form zu beschreiben.

Beispiel 7.2 – Messung einer Musterfolge.

Es sei bekannt, dass ein System ein sinusförmiges Signal der Amplitude A und der Kreisfrequenz B erzeuge. Diesem Signal ist ein normalverteiltes Rauschsignal additiv überlagert:

$$x(t) = A \cdot \sin(B \cdot t) + N(\mu, \sigma)$$

Wir beobachten das Gesamtsignal $x(t)$ im Zeitraum 2 bis 8 und erhalten durch 60 Messungen die im folgenden Diagramm durch Punkte angedeutete Musterfolge $x[k]$. Gesucht ist ein Verfahren, mit dem sich aus den Werten dieser Musterfolge die Parameter A und B des oben genannten Systems ermitteln lassen. □

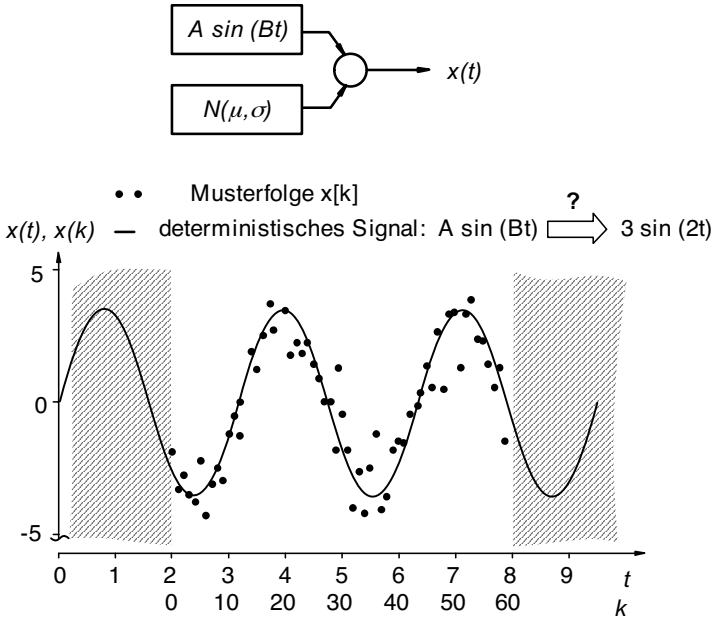


Abbildung 7.1. Messung einer Musterfolge.

Abb. 7.1 illustriert die statistische Sicht auf das Zustandekommen einer Musterfolge durch wiederholtes Anwenden eines Zufallsprozesses, hier $N(\mu, \sigma)$. Letzterer produziert fortlaufend reelle Werte gemäß einer u.U. zeitlich veränderlichen Wahrscheinlichkeitsdichtefunktion. Die Frage ist nun, mit welchen Parametern ein stochastischer Prozess beschrieben werden kann. Diese Parameter werden durch Mittelung aus dem gesamten Zufallsprozess ermittelt (nicht nur über eine seiner Musterfolgen). Man spricht daher auch von

Scharmittelwerten. In Abb. 7.2 ist der Unterschied zum *zeitlichen Mittelwert* dargestellt.

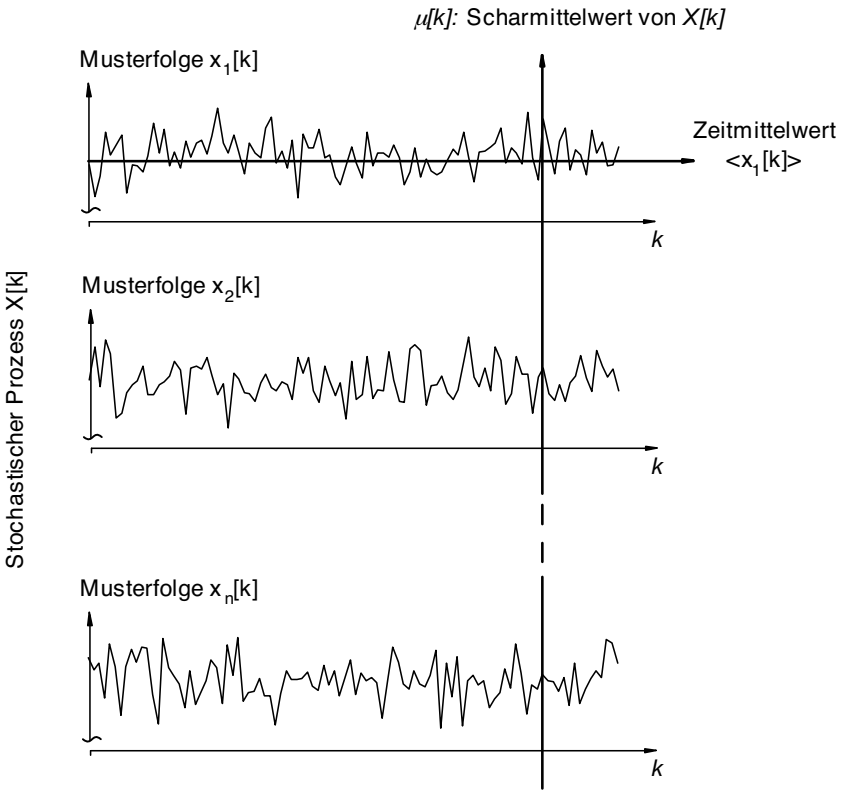


Abbildung 7.2. Scharmittelwert und Zeitmittelwert

Es muss deutlich darauf hingewiesen werden, dass Scharmittelwert und Zeitmittelwert im Allgemeinen *nicht identisch* sind!

In der Praxis ist es nicht von Belang, alle Realisierungen des stochastischen Prozesses zu erfassen. Stattdessen ist man eher bestrebt, die Beobachtungen des stochastischen Prozesses kompakter mit dem mathematischen Begriff der Wahrscheinlichkeit zu beschreiben. Dies geschieht in der Wahrscheinlichkeitstheorie auf zwei Arten.

- Die **Verteilungsfunktion** F_X gibt an, mit welcher Wahrscheinlichkeit P die Menge $\{X \leq x\}$ auftritt. Hierbei ist X eine bestimmte Zufallsvariable und x deren Realisierung.

$$x \rightarrow F_X(x) = P\{X \leq x\} \tag{7.2}$$

Bezogen auf Bsp. 7.1 bedeutet das, dass $F_X(x) = 1$ sein muss für $x \geq 12$, da die Zufallsvariable X lediglich natürliche Zahlen von 2 bis 12 produzieren kann. Hier ist folglich F_X eine treppenförmig steigende Funktion. Allgemein ist F_X eine monoton nicht fallende Funktion, die für $x \to \infty$ gegen eins strebt und für $x \to -\infty$ gegen null.

- Jede reelle Funktion $f_X(x) \geq 0$ mit

$$\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1 \tag{7.3}$$

heißt **Dichte** und definiert eindeutig die entsprechende Verteilungsfunktion

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\mathrm{d}y. \tag{7.4}$$

Das heißt insbesondere, dass $f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$, wenn F_X im Punkt x differenzierbar ist. Da die Verteilungsfunktion für den stochastischen Prozess aus Bsp. 7.1 eine treppenförmig steigende Funktion ist, ist seine Dichte eine Sprungfolge. Die Sprünge befinden sich an den natürlichen Zahlen von 2 bis 12. Die Höhe der Sprünge gibt an, wie wahrscheinlich es ist, dass die entsprechende Augensumme beobachtet wird. Daher gilt für den diskreten Fall $f_X(x) = P\{X = x\}$. Die Dichte gibt in diesem Fall unmittelbar die Wahrscheinlichkeiten der einzelnen Beobachtungen x wieder und die Integrale in (7.3) und (7.4) gehen somit in Summen über.

7.5.2 Eigenschaften stochastischer Prozesse

Ein stochastischer Prozess ist vollständig durch seine Verteilungsfunktion bzw. seine Dichte gekennzeichnet. Da sich diese Funktionen in der Praxis lediglich in Sonderfällen mit vertretbarem Aufwand schätzen lassen, werden die Eigenschaften eines Zufallsprozesses mit Hilfe seiner Erwartungswerte beschrieben. Diese Erwartungswerte stellen immer eine Mittelung über den gesamten Zufallsprozess dar, man spricht daher auch von *Scharmittelwerten*. Näheres zu Erwartungswerten ist beispielsweise in (Behnen und Neuhaus, 1995) zu finden. Hier sollen nur die wichtigsten Definitionen angegeben werden, wobei im Folgenden davon ausgegangen wird, dass es sich um diskrete Zufallsvariablen handelt, die aus einem beliebigen Zufallsprozess stammen. Im Allgemeinen muss folglich von einer Zeitabhängigkeit (Index k) der Erwartungswerte ausgegangen werden.

- **Erwartungswert** einer Funktion $g(X[k])$ einer Zufallsvariablen

$$E\{g(X[k])\} = \sum_{i=-\infty}^{\infty} g(x_i[k])P\{X_i[k] = x_i[k]\} \tag{7.5}$$

Es wird über alle Beobachtungen $x_i[k]$ summiert, die auftreten können. In Bsp. 7.1 ist dies die Summe über alle natürlichen Zahlen zwischen 2 und 12 einschließlich, wobei hier die Zufallsvariable nicht durch eine Funktion g abgebildet wird

- **Deterministischer Anteil** einer Zufallsvariablen („Mittelwert“)

$$\mu[k] = E\{X[k]\} \tag{7.6}$$

Diesen Wert kann man sich als eine Größe vorstellen, die in unabhängig voneinander wiederholten Messungen des selben Zufallsprozesses stets wiederkehrt.

- **Varianz** einer Zufallsvariablen

$$\sigma[k] = Var\{X[k]\} = E\left\{(X[k] - \mu[k])^2\right\} \tag{7.7}$$

Die Varianz ist ein Maß für die Streuung der einzelnen Beobachtungen um den deterministischen Anteil herum.

Da in der Regel nicht alle Musterfunktionen des Prozesses bekannt sind, und eine eventuell vorkommende Zeitabhängigkeit eine zusätzliche Erschwerung darstellt, können auch die Erwartungswerte nicht ohne weiteres bestimmt werden. Die folgenden zwei Eigenschaften von Zufallsprozessen erleichtern die Bestimmung von Erwartungswerten.

- **Stationarität:** Ein Zufallsprozess heißt stationär, wenn seine Eigenschaften (Verteilungsfunktion bzw. Dichte) nicht von der Wahl des Zeitursprungs abhängen. Dies ist in der Praxis schwierig zu überprüfen. Stattdessen ist meistens der mathematisch etwas schwächere Begriff der „Stationarität im weiteren Sinne“ relevant, der sich anhand der Erwartungswerte zeigen lässt. Im Folgenden ist diese Art von Stationarität gemeint, wenn von stationären Signalen oder Stationarität die Rede ist. Ein Merkmal für die Stationarität eines Zufallsprozesses ist, dass er zu jedem Zeitpunkt k dieselben Erwartungswerte hat:

$$E\{X[k]\} = \text{const} \quad \forall k \tag{7.8}$$

In diesem Fall spricht man von einem „Gleichanteil“. Das zweite Kriterium für Stationarität wird noch angegeben.

- **Ergodizität:** Ein Zufallsprozess heißt ergodisch, wenn die Stationarität erfüllt ist und darüber hinaus die zeitliche Mittelung über jede beliebige Musterfunktion mit der Zeit gegen den Gleichanteil des Prozesses strebt:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} x[k] = E\{X[k]\} = \text{const} \quad \forall k \tag{7.9}$$

Beispiel 7.3 – Stationäre Prozesse.

a) Sei ein Rauschprozess gegeben mit normalverteilter Wahrscheinlichkeitsdichte $N(\mu, \sigma)$. Mittelwert und Varianz des Rauschens hängen von der Zeit nicht ab. Andere Funktionen $g(\cdot)$ des Zufallsprozesses haben damit ebenfalls zeitlich konstante Erwartungswerte. Der Rauschprozess ist damit stationär.
b) Sei ein Prozess gegeben durch $X[k] = \sin(2\pi k/10) + N(\mu, \sigma)$. Dieser ist nicht stationär, da die sin-Funktion mit k nicht konstante Werte hat.
c) Sei ein Prozess gegeben durch

$$[k] = \sum_{k=1}^{10} \sin\left(\frac{2\pi}{10}(k+\kappa)\right) + N(\mu, \sigma)$$

Dieser Prozess ist stationär, da für alle k der Summen-Teil denselben Wert (Null) ergibt, und da bei der Mittelung über alle Musterfunktionen des Rauschprozesses zeitlich konstante Erwartungswerte auftreten. □

Beispiel 7.4 – Stationarität beim Würfeln.

Das Würfelexperiment aus Bsp. 7.1 erfüllt das Kriterium der Stationarität. Die zu erwartende Augensumme bleibt ja stets gleich unabhängig davon, zu welchem Zeitpunkt man würfelt. Der Grund ist, dass die Dichte gleich bleibt. Man würfelt stets mit den gleichen Würfeln. Der Prozess erfüllte dieses Kriterium nicht und wäre somit kein stationärer Prozess, wenn beispielsweise zu jedem Zeitpunkt k andere Würfel mit anderen Dichten verwendet würden. Dann wäre jedesmal eine andere Augensumme zu erwarten, und der Erwartungswert des Prozesses wäre zeitabhängig. □

Beispiel 7.5 – Ergodizität beim Würfeln.

Der Mittelwert der Augenzahl beim Würfeln (3,5) kann gemessen werden entweder aus den Ergebnissen beim *fortgesetzten* Werfen *eines* Würfels, oder den Ergebnissen beim *einmaligen* Werfen vieler zu dem ersten *identischer* Würfel. Dies gilt für das Gedankenexperiment. Es gilt auch noch mit realen Würfeln, wenn man von einer konstanten Qualität in der Produktion der Würfel ausgehen kann. Dies bedeutet für eine Spielbank, die qualitativ hochwertige Würfel verwendet, dass sie von Ergodizität ausgehen kann.
Im realen Fall setzt Ergodizität eine Konstanz in der Qualität der Produktion der Würfel voraus. Im Fall der *Kontrolle* der Produktion würde man gerade argumentieren, dass die Ergodizität nicht gegeben ist: Veränderungen in der Qualität der Produktion lassen sich aus dem Werfen *eines* Würfels eben nicht erkennen.

Es kann gezeigt werden, dass der Zufallsprozess aus Bsp. 7.1 auch ergodisch ist. Dazu ist zu zeigen, dass die oben angegebene zeitliche Mittelung mit „Wahrscheinlichkeit 1“ gegen den Gleichanteil strebt. Die Konvergenz mit Wahrscheinlichkeit 1 ist eine starke stochastische Konvergenzart. Näheres hierzu und zu anderen stochastischen Konvergenzarten findet man in (Böhme, 1998). □

Stationarität und Ergodizität stellen zwar eine starke Einschränkung der Allgemeinheit für einen Zufallsprozess dar, jedoch können sie für Rauschprozesse, mit denen sich die Signalverarbeitung in der Regel beschäftigt, als erfüllt angenommen werden. Der große Vorteil ergodischer Prozesse ist, dass hier aus den Eigenschaften einer einzigen (gemessenen) Musterfunktion $x[k]$ Aussagen für den gesamten Zufallsprozess $X[k]$ abgeleitet werden können. Die Ergodizität vorausgesetzt, gilt speziell für Mittelwert und Varianz eines zeitdiskreten Prozesses:

$$\text{Mittelwert:} \quad \mu = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{k=-N}^{N} x[k] \tag{7.10}$$

$$\text{Varianz:} \quad \sigma = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{k=-N}^{N} |x[k] - \mu|^{\,2} \tag{7.11}$$

7.6 Motivation zur Einführung der Korrelation

Von zentraler Bedeutung für die Beschreibung stochastischer Prozesse ist der Begriff der Korrelation. Worum es sich dabei handelt, wird in den folgenden Abschnitten beschrieben. Hier soll anhand eines einfachen Beispiels der Sinn und die Nützlichkeit der Korrelation gezeigt werden.

Von einem kausalen System sei bekannt, dass seine Impulsantwort $h[k]$ aus $(N+1)$ Termen besteht. Die Übertragungsfunktion erhält man durch die Transformation in den Z-Bereich:

$$H(z) = \sum_{k=0}^{N} h[k] z^{-k} \tag{7.12}$$

Speziell an der Stelle $z = 1$ ergibt sich

$$H(1) = \sum_{k=0}^{N} h[k] \tag{7.13}$$

Die Größe $H(1)$ soll nun experimentell bestimmt werden. Dazu legt man ein Signal $x[k]$ an den Eingang. Das Signal am Ausgang erhält man durch Faltung mit der Impulsantwort

$$y[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k] \tag{7.14}$$

Die Impulsantwort ist unbekannt! Wir wählen am Eingang die Sprungfolge ($x[k] = 1$ für $k = 0 \ldots N$) und erhalten:

$$y[n] = \sum_{k=0}^{N} 1 \cdot h[n-k] = \sum_{k=n-N}^{n} h[k] \tag{7.15}$$

wobei die Kausalität des Systems noch nicht berücksichtigt wurde. Speziell gilt für ein kausales System

$$y[n] = \sum_{k=0}^{n} h[k] \tag{7.16}$$

und

$$y(N) = H(1) \tag{7.17}$$

Praktisch läßt sich am Eingang die Sprungfolge nicht exakt realisieren. Wie jedes physikalische Signal ist auch die Eingangsfolge mit einem Fehler behaftet. Wir betrachten also die Musterfolge eines Zufallsprozesses mit folgender Gleichung:

$$x[k] = 1 + \varepsilon[k], \tag{7.18}$$

wobei $\varepsilon[k]$ Mittelwert 0 und Varianz σ haben möge. Für die Ausgangsfolge ergibt sich damit bei $n = N$:

$$\begin{aligned} y[n] &= \sum_{k=0}^{n} h[k]\,(1 + \varepsilon[k]) = \sum_{k=0}^{n} h[k] + \sum_{k=\,0}^{n} h[k]\varepsilon[k] \\ &= H(1) + \sum_{k=0}^{n} h[k]\varepsilon[k] \end{aligned} \tag{7.19}$$

Wir erhalten einen Störterm, den es zu beseitigen gilt. Eine erste Möglichkeit dazu ist von den physikalischen Messreihen her bekannt. Man kann die Messung mehrfach (M-mal) durchführen und dann den Mittelwert bilden:

$$\hat{y}[n] = \frac{1}{M} \sum_{m=1}^{M} y_m[n] = \frac{1}{M} \sum_{m=1}^{M} H(1) + \frac{1}{M} \sum_{m=1}^{M} \left(\sum_{k=0}^{n} h[k]\varepsilon_m[k] \right) \tag{7.20}$$

$$\hat{y}[n] = H(1) + \sum_{k=0}^{n} h[k] \left(\frac{1}{M} \sum_{m=1}^{M} \varepsilon_m[k] \right) \tag{7.21}$$

Da sich die Fehler ε_m zumindest gegenseitig aufheben, wird der Wert in der Klammer für große M beliebig klein. Durch eine große Zahl von Messungen

erhält man beliebig genaue Werte. Wie man allerdings auch von den physikalischen Messreihen her weiß, ist das eine recht ermüdende Angelegenheit. Außerdem können in der digitalen Signalverarbeitung viele Messungen nicht identisch wiederholt werden. Als Beispiel diene wieder die oben angeführte Sprachverarbeitung: eine bestimmte akustische Äußerung wird in dieser Form nur einmal gesagt, Wiederholungen sind oft nicht erwünscht oder nicht zumutbar und würden auch eine veränderte Akustik ergeben.

Es wäre daher wünschenswert, den Störterm bei nur einer bekannten Messreihe minimieren zu können. Dazu überlegt man sich, wie die Verringerung der Streuung durch Mehrfachmessung erreicht wurde. Bei einer *einzelnen* Messung geht nur der Fehler *eines* Eingangswertes in die Berechnung des Ausgangswerts ein. Bei mehrfacher Messung gehen mehrere Eingangswerte mit ihren Fehlern in die Berechnung eines Ausgangswerts ein. Diese heben sich teilweise gegenseitig auf und sorgen so für eine Verringerung der Streuung. Aus unserem Beispiel erkennt man aber, dass schon nach einmaliger Messung viele Eingangswerte vorliegen. Die Idee der Korrelation ist nun folgende: Statt die Eingangswerte an der selben Stelle aus den verschiedenen Messreihen zur Mittelung heranzuziehen, verwendet man die Eingangswerte der selben Messreihe an verschiedenen Stellen. Wir erzeugen eine Größe (die streng genommen von einem Parameter M abhängt) wie folgt:

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M} x[n+m]y[n] &= \frac{1}{M}\sum_{m=1}^{M} x[n+m]\left(\sum_{k=0}^{n} h[k]x[n-k]\right)\\ &= \sum_{k=0}^{n} h[k]\left(\frac{1}{M}\sum_{m=1}^{M} x[n+m]x[n-k]\right) \qquad (7.22)\end{aligned}$$

Wenn der Klammerausdruck gleich eins ist, erhält man also für $n = N$ das gesuchte $H(1)$. Durch Ausmultiplizieren ergibt sich:

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M} x\ [n+m]x[n-k] &= \frac{1}{M}\sum_{m=1}^{M}(1+\varepsilon[n+m])\,(1+\varepsilon[n-k])\\ &= \frac{1}{M}\sum_{m=1}^{M}(1+\varepsilon[n-k]+\varepsilon[n+m]+\varepsilon[n-k]\varepsilon[n+m])\\ &= 1+\frac{1}{M}\sum_{m=1}^{M}(\varepsilon[n-k]+\varepsilon[n+m]+\varepsilon[n-k]\varepsilon[n+m])\\ &= 1+\frac{\sigma}{M}\sum_{m=1}^{M}\delta[k+m] \qquad (7.23)\end{aligned}$$

Wie schon bei der Mehrfachmessung geht auch hier die Mittelung über die ε-Terme für großes M gegen null. Eine Ausnahme bilden die Produkte $\varepsilon[a]\varepsilon[b]$, welche für identische Indizes $a = b$ gleich der Varianz werden. Dies tritt in der

Summe über m einmal auf, falls k zwischen -1 und $-M$ liegt. Dargestellt ist dies über eine Folge von Dirac-Impulsen. Setzt man dieses Ergebnis wieder in die Formel ein, so erhält man:

$$\frac{1}{M}\sum_{m=1}^{M} x[n+m]y[n] = \sum_{k=0}^{n} h[k]\left(1+\frac{\sigma}{M}\sum_{m=1}^{M}\delta[k+m]\right) \tag{7.24}$$

Da $k \geq 0$, erscheint der letzte Varianzterm nicht in der Summe, und wir erhalten für $n = N$:

$$\lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M} x[N+m]y[N] = \sum_{k=0}^{N} h[k] = H(1) \tag{7.25}$$

Der hier benutzte Term vom Typ

$$\lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M} x[n+m]y[n] \tag{7.26}$$

entspricht nicht der exakten Definition der Korrelation, zeigt aber die Idee der Mittelung über Produkte aus *einer* Messung eines Prozesses. Die Verbindung zwischen der zeitlichen Mittelung durch (als durchführbar angenommene) Wiederholung des Experiments und dem hier hergeleiteten Ausdruck stellt die im Abschnitt 7.5.1 eingeführte Ergodizität her.

7.7 Die Autokorrelation

Nachdem im Abschnitt 7.6 die Nützlichkeit der Korrelation von Signalen gezeigt wurde, soll diese nun exakt definiert werden. Dieser Abschnitt beschäftigt sich mit der Autokorrelation, also der Korrelation einer Funktion mit sich selbst. Die Autokorrelation enthält eine Aussage darüber, wie schnell sich das Zufallssignal zeitlich ändern kann bzw. sagt aus, inwieweit der Signalwert zum Zeitpunkt k_1 den Signalwert zum Zeitpunkt k_2 beeinflusst. Sie ist für zeitdiskrete komplexe Signale wie folgt definiert:

$$\begin{aligned} s_{XX}(k_1,k_2) &= E\{X^*[k_1]X[k_2]\} \\ &= E\{X_1^* X_2\} \\ &= E\{(X_{1R}-jX_{1I})(X_{2R}+jX_{2I})\} \\ &= E\{X_{1R}X_{2R}-X_{1I}X_{2I}\}+jE\{X_{1R}X_{2I}-X_{1I}X_{2R}\} \end{aligned} \tag{7.27}$$

Für reelle Signale vereinfacht sich die Formel zu:

$$s_{XX}[k_1,k_2] = E\{X[k_1]X[k_2]\} = E\{X_1X_2\} \tag{7.28}$$

Offensichtlich ist die Autokorrelationsfolge abhängig von den beiden Variablen k_1 und k_2. Für einen stationären Prozess kann man ein Wertepaar k_1 und k_2

beliebig innerhalb des Definitionsbereichs verschieben, ohne dass sich eine Veränderung für s_{xx} ergibt. Die Autokorrelation ist in diesem Fall also nur von der Differenz $k_1 - k_2$ abhängig. Mit der Substitution $k_1 = k$, $k_2 = k + \kappa$ ergibt sich also:

$$s_{XX}[\kappa] = E\{X^*[k]X[k+\kappa]\} \tag{7.29}$$

Setzt man darüber hinaus Ergodizität voraus, kann man die Autokorrelation aus jeder beliebigen Musterfolge $x[k]$ bestimmen:

$$s_{XX}[\kappa] = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{k=-N}^{+N} x[k]x[k+\kappa] \tag{7.30}$$

Aus der Definition der Autokorrelation lassen sich einige interessante Eigenschaften ableiten. Beispielsweise ist die Autokorrelationsfolge konjugiert gerade:

$$\begin{aligned} s_{XX}[\kappa] &= E\{X^*[k]X[k+\kappa]\} = E\{X[k]X^*[k+\kappa]\}^* \\ &= E\{X[k-\kappa]X^*[k]\}^* = s^*_{XX}[-\kappa] \end{aligned} \tag{7.31}$$

Das bedeutet, dass die Autokorrelation eines reellen Prozesses gerade ist:

$$s_{XX}[\kappa] = s_{XX}[-\kappa] \tag{7.32}$$

Eine weitere wichtige Eigenschaft bezieht sich auf den Fall, dass die beiden betrachteten Zeitpunkte aufeinander fallen, also $k_1 = k_2$, $\kappa = 0$.

$$s_{XX}(0) = E\{X^*[k]X[k+0]\} = E\{|x[k]|^2\} = \sigma + |\mu|^2 \tag{7.33}$$

wobei die letztere Gleichheit, die den Zusammenhang zwischen linearem und quadratischem Mittelwert sowie der Varianz angibt, aus der Def. (7.7) folgt. Die Autokorrelation an der Stelle $\kappa = 0$ ist also gleich dem Mittelwert des Quadrats und somit ein Maß für die mittlere Leistung des Prozesses. Aus dieser Betrachtung wird auch deutlich, dass man die Autokorrelationsfolge als eine Art Verallgemeinerung des Mittelwertes des Quadrates auffassen kann, in dem Sinne, dass die Folge vor dem Quadrieren gegen sich selbst verschoben wird.

Zur formalen Berechnung einer Autokorrelationsfolge bietet sich das Matrixschema an: Es wird ein Zufallsprozess $X[k]$ der Länge n betrachtet. Man schreibt nun die Werte des Zufallsprozesses von $X(k)$ bis $X(k-n)$ in eine Spaltenmatrix $\mathbf{X}$ und die konjugiert komplexen Werte $X^*(k)$ bis $X^*(k-n)$ in eine Zeilenmatrix $\mathbf{X}^*$. Durch Multiplikation beider Matrizen erhält man die Autokorrelationsmatrix $\mathbf{S_{XX}}$:

$$\mathbf{S_{XX}} = E\{\mathbf{X}\mathbf{X}^*\} = E\left\{\begin{pmatrix} X(k) \\ X(k-1) \\ \vdots \\ X(k-n) \end{pmatrix} \left(X^*(k)\ X^*(k-1) \cdots\ X^*(k-n)\right)\right\}$$

$$= E\left\{\begin{pmatrix} X(k)X^*(k) & X(k)X^*(k-1) & \cdots & X(k)X^*(k-n) \\ X(k-1)X^*(k) & X(k-1)X^*(k-1) & & X(k-1)X^*(k-n) \\ \vdots & & \ddots & \vdots \\ X(k-n)X^*(k) & X(k-n)X^*(k-1) & \cdots & X(k-n)X^*(k-n) \end{pmatrix}\right\}$$

$$= \begin{pmatrix} s_{XX}(0) & s_{XX}(1) & \cdots & s_{XX}(n) \\ s^*_{XX}(1) & s_{XX}(0) & & s_{XX}(n-1) \\ \vdots & & \ddots & \vdots \\ s^*_{XX}(n) & s^*_{XX}(n-1) & \cdots & s_{XX}(0) \end{pmatrix} \tag{7.34}$$

Auch für die Autokorrelation ist es sinnvoll, eine vom Mittelwert befreite Kenngröße einzuführen. In Analogie zur Varianz definiert man daher für stationäre Prozesse die Autokovarianz wie folgt:

$$\begin{aligned} c_{XX}[\kappa] &= E\{(x^*[k] - \mu^*)(x[k+\kappa] - \mu)\} \\ &= E\{x^*[k]x[k+\kappa]\} - \mu E\{x^*[k]\} - \mu^* E\{x[k+\kappa]\} + \mu^*\mu \\ &= s_{XX}[\kappa] - |\mu|^2 = s_{XX}[\kappa] - \mu\mu^* \end{aligned} \tag{7.35}$$

Die Kovarianzmatrix ergibt sich daher wie folgt:

$$\mathbf{C_{XX}} = \mathbf{S_{XX}} - |\mu|^2 \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & & 1 \\ \vdots & & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \tag{7.36}$$

7.8 Kreuzkorrelation

Beschränkt sich die Betrachtung nicht auf *einen* Zufallsprozess, so genügt es nicht, jeden Prozess für sich zu beschreiben. Es muss darüber hinaus die gegenseitige Beeinflussung zwischen zwei Prozessen untersucht werden. Zu diesem Zweck definiert man die Kreuzkorrelation zwischen zwei stationären Prozessen:

$$s_{XY}[\kappa] = E\{X^*[k]Y[k+\kappa]\} \tag{7.37}$$

Genau wie die Autokorrelationsfolge ist auch die Kreuzkorrelationsfolge konjugiert symmetrisch (man beachte die Vertauschung der Indices X,Y):

$$\begin{aligned} s_{XY}[\kappa] &= E\{X^*[k]Y[k+\kappa]\} = E\{X[k]Y^*[k+\kappa]\}^* \\ &= E\{X[k-\kappa]Y^*[k]\}^* = s^*_{YX}[-\kappa] \end{aligned} \tag{7.38}$$

Auch für die Kreuzkorrelation wird analog zur Autokovarianz eine mittelwertfreie Kenngröße definiert, die Kreuzkovarianzfolge :

$$\begin{aligned} c_{XY}[\kappa] &= E\left\{[X^*[k] - \mu_x^*][Y[k+\kappa] - \mu_Y]\right\} \\ &= s_{XY}[\kappa] - \mu_X^* \mu_Y \end{aligned} \tag{7.39}$$

Mit Hilfe der Kreuzkorrelation lassen sich zwei Spezialfälle für die Kopplung zweier Prozesse angeben.

- **Unkorreliertheit** : Zwei Prozesse sind *unkorreliert*, wenn sich ihre Kreuzkorrelation aus dem Produkt der Erwartungswerte der beiden Einzelprozesse ergibt.

  $$s_{XY}(\kappa) = E\{X^*[k]Y[k+\kappa]\} = E\{X^*[k]\}E\{Y[k+\kappa]\} \tag{7.40}$$

  Die Kreuzkovarianz ist in diesem Falle gleich null.

  $$\begin{aligned} c_{XY}[\kappa] &= E\{(X^*[k] - \mu_X^*)(y[k+\kappa] - \mu_Y)\} \\ &= E\{X^*[k]\}E\{Y[k+\kappa]\} - \mu_Y E\{X^*[k]\} - \mu_X^* E\{Y[k+\kappa]\} + \mu_X^* \mu_Y \\ &= \mu_X^* \mu_Y - \mu_X^* \mu_Y - \mu_X^* \mu_Y + \mu_X^* \mu_Y \\ &= 0 \end{aligned} \tag{7.41}$$

- **Orthogonalität**: Zwei Prozesse heißen *orthogonal*, wenn ihre Kreuzkorrelationsfolge für alle κ den Wert null annimmt.

  $$s_{XY}[\kappa] = E\{X^*[k]Y[k+\kappa]\} = 0 \tag{7.42}$$

  Zwei unkorrelierte Prozesse sind auch orthogonal, wenn mindestens einer der Prozesse mittelwertfrei ist.

Zur Illustration der Korrelation und der Orthogonalität bringen wir nun ein Beispiel, in dem deutlich wird, dass die beiden Eigenschaften unabhängig voneinander auftreten können.

Beispiel 7.6 – Korrelation und Orthogonalität.

a) unkorrelierte, orthogonale Prozesse: Zu jedem Zeitpunkt k werden zwei Münzen X, Y geworfen, wobei „Kopf“ mit 1 und „Zahl“ mit -1 bewertet werden:

$$E\{X^*[k] \cdot Y[k+\kappa]\} = 0 \quad ; \quad E\{X^*[k]\} \cdot E\{Y[k+\kappa]\} = 0$$

b) unkorrelierte, nicht orthogonale Prozesse: Zu jedem Zeitpunkt k werden zwei Münzen X, Y geworfen, wobei „Kopf“ mit 1 und „Zahl“ mit 0 bewertet werden:

$$E\{X^*[k] \cdot Y[k+\kappa]\} = 1/4 \quad ; \quad E\{X^*[k]\} \cdot E\{Y[k+\kappa]\} = 1/4$$

c) korrelierte, nicht orthogonale Prozesse; Zu jedem Zeitpunkt k wird *eine* Münze A geworfen, wobei „Kopf“ mit $\{X[k] = 6, Y[k] = 1\}$ und „Zahl“ mit $\{X[k] = -3, Y[k] = 2\}$ bewertet wird. Dann ist

$$E\{X^*[k] \cdot Y[k]\} = 0 \quad ; \quad E\{X^*[k]\} \cdot E\{Y[k]\} = 9/4$$
$$E\{X^*[k] \cdot Y[k+\kappa]\} = 9/4 \quad ; \quad E\{X^*[k]\} \cdot E\{Y[k+\kappa]\} = 9/4 \; ; \; \kappa \neq 0$$

Für $\kappa = 0$ ist die Unkorreliertheitsbedingung und für $\kappa \neq 0$ die Orthogonalitätsbedingung verletzt, insgesamt sind die Prozesse damit korreliert und nicht orthogonal.

Man überzeuge sich, dass a),b),c) ergodische Prozesse sind.

d) korrelierte, orthogonale Prozesse:
Zum Zeitpunkt $T_0 = -\infty$ wurde eine Münze A geworfen, wobei wiederum „Kopf“ mit $\{X[T_0] = 6, Y[T_0] = 1\}$ und „Zahl“ mit $\{X[T_0] = -3, Y[T_0] = 2\}$ bewertet wurden.
Für *alle* Zeitpunkte wird nun gesetzt $X[k] = X[T_0]$ und $Y[k] = Y[T_0]$. Dann ist

$$E\{X^*[k] \cdot Y[k+\kappa]\} = 0 \quad ; \quad E\{X^*[k]\} \cdot E\{Y[k+\kappa]\} = 9/4 \quad ; \quad \forall\kappa$$

und damit sind die Prozesse korreliert und orthogonal. Der Fall d) ist nicht ergodisch. Die Prozesse wurden vor unendlich langer Zeit *präpariert*[1], so dass wir bei Anwendung des Zeitmittels $<\ldots>$ über einen beliebig langen Zeitraum (auch unendlich lang, falls $k = -\infty$ nicht enthalten ist, also z.B. $k = 0 \ldots \infty$) entweder erhalten, falls die Münze A „Kopf“ zeigte:

$$<X^*[k] \cdot Y[k+\kappa]> = 6 \quad ; \quad <X^*[k]> \cdot <Y[k+\kappa]> = 6 \quad ; \quad \forall\kappa$$

oder, falls die Münze A „Zahl“ zeigte:

$$<X^*[k] \cdot Y[k+\kappa]> = -6 \quad ; \quad <X^*[k]> \cdot <Y[k+\kappa]> = -6 \quad ; \quad \forall\kappa$$

Damit ändert sich bei Anwendung des Zeitmittels auch das Ergebnis: der Prozess erscheint uns für jeden Ausgang des Münzwurfes von A jeweils unkorreliert und nicht orthogonal, also in Bezug auf Korreliertheit und Orthogonalität genau das Gegenteil des Ergebnisses nach Anwendung des Scharmittels. □

[1] Die Formulierung „Präparierung eines Prozesses“ wurde der Quantenmechanik entlehnt.

7.9 Spektraldarstellung stochastischer Prozesse

Wie schon bei den deterministischen Signalen lassen sich auch viele Eigenschaften der stochastischen Signale besser im Frequenzbereich beschreiben. Die spektrale Darstellung zeitdiskreter Signale erhält man mit Hilfe der zeitdiskreten Fouriertransformation:

$$X(e^{j\Phi}) = \sum_{k=-\infty}^{\infty} x[k]e^{-j\omega Tk} = \sum_{k=-\infty}^{\infty} x[k]e^{-j\Phi k} \quad \text{mit} \quad \Phi = \omega T \tag{7.43}$$

Wir betrachten die Fouriertransformation der Autokorrelationsfolge:

$$S_{XX}(e^{j\Phi}) = \sum_{\kappa=-\infty}^{\infty} s_{XX}[\kappa]e^{-j\Phi\kappa} \tag{7.44}$$

Es stellt sich die Frage, welche anschauliche Bedeutung diese Funktion hat. Dazu stellen wir zunächst fest, dass man durch die Fourier-Rücktransformation für $\kappa = 0$ aus ihr eine Aussage über die mittlere Leistung des Signals gewinnen kann. Wir wiederholen diese Rücktransformation hier nochmals zur Übung:

$$\begin{aligned}
\frac{1}{2\pi}\int_{-\pi}^{\pi} S_{XX}(e^{j\Phi})\mathrm{d}\Phi &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\sum_{\kappa=-\infty}^{\infty} s_{XX}[\kappa]e^{-j\Phi\kappa}\right]\mathrm{d}\Phi \\
&= \sum_{\kappa=-\infty}^{\infty}\frac{1}{2\pi}s_{XX}[\kappa]\int_{-\pi}^{\pi} e^{-j\Phi\kappa}\mathrm{d}\Phi \\
&= \sum_{\kappa=-\infty}^{\infty}\frac{1}{2\pi}s_{XX}[\kappa](2\pi\delta[\kappa]) \\
&= s_{XX}(0) = E\{|x[k]|^2\}
\end{aligned} \tag{7.45}$$

Der letzte Term stellt die mittlere Leistung des Signals dar. Diese lässt sich also durch Integration aus der Fouriertransformierten der Autokorrelationsfolge gewinnen. Folglich handelt es sich bei der Fouriertransformierten der Autokorrelationsfolge um eine *(Auto-)Leistungsdichte*. Wir erhalten damit das folgende

Theorem 7.1 (Wiener-Khintschine Theorem).
Die spektrale Autoleistungsdichte $S_{\mathrm{XX}}(e^{j\,\omega\,T})$ ergibt sich durch zeitdiskrete Fouriertransformation aus der Autokorrelationsfolge $s_{\mathrm{XX}}[\kappa]$.

Es ist also möglich, die mittlere Leistung eines Prozesses sowohl im Zeitbereich durch die Autokorrelationsfolge zu bestimmen, als auch im Frequenzbereich durch Integration über die spektrale Leistungsdichte. Da die Autokorrelationsfolge konjugiert gerade ist, ergibt sich:

$$S_{XX}(e^{j\Phi}) = \sum_{\kappa=-\infty}^{-1} s_{XX}[\kappa]e^{-j\Phi\kappa} + s_{XX}[0]e^{0} + \sum_{\kappa=1}^{\infty} s_{XX}[\kappa]e^{-j\Phi\kappa}$$

$$= s_{XX}(0) + \sum_{\kappa=1}^{\infty} \left(s^{*}_{XX}[\kappa]e^{j\Phi\kappa} + s_{XX}[\kappa]e^{-j\Phi\kappa}\right)$$

$$= s_{XX}(0) + 2\mathrm{Re}\{\sum_{\kappa=1}^{\infty} s_{XX}[\kappa]e^{-j\Phi\kappa}\} \tag{7.46}$$

Die Autoleistungsdichte ist also eine rein reelle Funktion. Zur Veranschaulichung soll nun die Autoleistungsdichte für zwei spezielle Zufallsprozesse angegeben werden, den mittelwertfreien, unkorrelierten Prozess und den gleichmäßig korrelierten Prozess.

Beispiel 7.7 – Mittelwertfreier, unkorrelierter (weißer) Prozess.

Für den *mittelwertfreien, unkorrelierten Prozess* ergibt sich der Wert der Autokorrelation an der Stelle $\kappa = 0$ nach der bereits früher hergeleiteten Beziehung (4.53)

$$s_{XX}[0] = \sigma + |\mu|^2$$

Auf Grund der Mittelwertfreiheit des Prozesses gilt $\mu = 0$ und es muss daher nicht zwischen Autokorrelation und Autokovarianz unterschieden werden. An allen Stellen $\kappa \neq 0$ ist die Autokorrelationsfolge auf Grund der Unkorreliertheit des Prozesses gleich null.

$$s_{XX}[\kappa] = \begin{cases} \sigma & \text{für} \quad \kappa = 0 \\ 0 & \text{sonst} \end{cases} \tag{7.47}$$

Durch die Fouriertransformation ergibt sich damit:

$$S_{XX}(e^{j\Phi}) = \sum_{\kappa=-\infty}^{\infty} s_{XX}[\kappa]\ e^{-j\Phi\kappa} = s_{XX}[0]e^{-j\Phi(0)} = \sigma \tag{7.48}$$

Die Autoleistungsdichte ist für alle Frequenzen gleich der Varianz des Signals. Da alle Frequenzen also den selben Beitrag zur Leistungsdichte bringen, spricht man auch von einem *weißen Prozess.* □

Beispiel 7.8 – Gleichmässig korrelierter Prozess.

Der zweite Spezialfall ist der *gleichmäßig korrelierte Prozess*. Die Autokorrelationsfolge hat hier einen konstanten Wert für alle κ:

$$s_{XX}[\kappa] = s_{XX}[0] \quad \forall\kappa \tag{7.49}$$

Für die Autoleistungsdichte gilt also:

$$
\begin{aligned}
S_{XX}(e^{j\Phi}) &= \sum_{\kappa=-\infty}^{\infty} s_{XX}[\kappa] e^{-j\Phi\kappa} \\
&= s_{XX}[0] \sum_{\kappa=-\infty}^{\infty} e^{-j\Phi\kappa} \\
&= s_{XX}[0] \frac{2\pi}{T} \sum_{\kappa=-\infty}^{\infty} \delta[\omega - \kappa\Omega] \quad \text{mit} \quad \Omega = \frac{2\pi}{T} \\
&= s_{XX}[0]\, 2\pi \sum_{\kappa=-\infty}^{\infty} \delta[\Phi - 2\pi\kappa] \qquad (7.50)
\end{aligned}
$$

wobei Theorem 4.1 auf Seite 92 benutzt wurde.

Die Autoleistungsdichte ist also ein *Impulszug*: Sie hat bei allen ganzzahligen Vielfachen der Grundfrequenz einen Delta-Impuls, der mit $2\pi s_{XX}[0]$ gewichtet ist, für alle anderen Stellen verschwindet sie. □

7.10 Transformation durch lineare Systeme

Bisher haben wir die Eigenschaften von Zufallsprozessen beschrieben und Kenngrössen dafür angegeben. Wir interessieren uns nun dafür, wie sich der Ausgang eines LTI-Systems verhält, wenn an seinem Eingang ein Zufallsprozess angelegt wird. Natürlich wird dies auch am Ausgang zu einem stochastischen Verhalten führen. Konkret interessieren wir uns also dafür, wie die Kenngrössen des Prozesses am Ausgang sich als Funktion der Kenngrössen des Prozesses am Eingang sowie der Kenngrössen des LTI-Systems beschreiben lassen (Orfanidis, 1988).

7.10.1 Übertragung von Autokorrelationsfolge und Autoleistungsdichte

In diesem Abschnitt soll das Verhalten von Autokorrelation und spektraler Leistungsdichte eines Signals bei der Übertragung durch ein LTI-System betrachtet werden. Der Eingangsprozess und die Impulsantwort $h[k]$ des Systems werden als bekannt angenommen. Wir interessieren uns nun für die Signaleigenschaften am Ausgang.

Bei Kenntnis einer Musterfolge $x[k]$ am Eingang läßt sich die zugehörige Musterfolge am Ausgang durch die Faltung mit der Impulsantwort bestimmen:

$$
y[k] = \sum_{i=-\infty}^{\infty} h[i] x[k-i] \qquad (7.51)
$$

Damit ergibt sich für die Autokorrelationsfolge am Ausgang folgendes:

$$\begin{aligned} s_{YY}[\kappa] &= E\{Y^*[k]Y[k+\kappa]\} \\ &= E\left\{\sum_{i=-\infty}^{\infty} h^*[i]X^*[k-i]\sum_{j=-\infty}^{\infty} h[j]X[k+\kappa-j]\right\} \\ &= \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} h^*[i]h[j]E\{X^*[k-i]X[k+\kappa-j]\} \\ &= \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} h^*[i]h[j]s_{XX}[i-j+\kappa] \\ &= \sum_{m=-\infty}^{\infty}\left[\sum_{i=-\infty}^{\infty} h^*[i]h[i+m]\right] s_{XX}[\kappa-m] \end{aligned} \tag{7.52}$$

Der eingeklammerte Teil hat die Form einer Faltungssumme.

$$\sum_{i=-\infty}^{\infty} h^*[i]h[i+m] = h[m] * h^*[-m] =: s_{hh}[m] \tag{7.53}$$

Die so definierte Größe $s_{hh}[k]$ bezeichnet man als *Systemkorrelationsfolge*. Damit erhält man weiter:

$$\begin{aligned} s_{YY}[\kappa] &= \sum_{m=-\infty}^{\infty} s_{hh}[m]s_{XX}[\kappa-m] \\ &= s_{hh}[\kappa] * s_{XX}[\kappa] = h[\kappa] * h^*[-\kappa] * s_{XX}[\kappa] \end{aligned} \tag{7.54}$$

Die Ausgangskorrelation läßt sich also durch die Faltung von Eingangs- und Systemkorrelation bestimmen. Durch die Übertragung in den Frequenzbereich läßt sich die schwierige Operation der Faltung durch die einfache Operation der Multiplikation wie folgt ersetzen:

$$\begin{aligned} S_{YY}(e^{j\Phi}) &= FT\{s_{YY}[\kappa]\} \\ &= FT\{h[\kappa] * h^*[-\kappa] * s_{XX}[\kappa]\} \\ &= FT\{h[\kappa]\}FT\{h^*[-\kappa]\}FT\{s_{XX}[\kappa]\} \\ &= H(e^{j\Phi})H^*(e^{j\Phi})S_{XX}(e^{j\Phi}) \\ &= \left|H(e^{j\Phi})\right|^2 S_{XX}(e^{j\Phi}) \end{aligned} \tag{7.55}$$

Wir geben weiter unten dazu das Beispiel 7.9.

7.10.2 Kreuzkorrelation zwischen Eingangs- und Ausgangsprozess

Zur Beschreibung eines Systems ist es günstig, die Korrelation zwischen Eingang und Ausgang zu bestimmen. Dazu bildet man die Kreuzkorrelationsfolge

$$\begin{aligned} s_{XY}[\kappa] &= E\{X^*[k]Y[k+\kappa]\} \\ &= E\left\{X^*[k]\sum_{i=-\infty}^{\infty} h[i]X[k+\kappa-i]\right\} \\ &= \sum_{i=-\infty}^{\infty} h[i]E\left\{X^*[k]X[k+\kappa-i]\right\} \\ &= \sum_{i=-\infty}^{\infty} h[i]s_{XX}[\kappa-i] \\ &= h[\kappa] * s_{XX}[\kappa] \end{aligned} \tag{7.56}$$

Durch die Fourier-Transformation ergibt sich daraus

$$S_{XY}(e^{j\Phi}) = H(e^{j\Phi})S_{XX}(e^{j\Phi}) \tag{7.57}$$

Beispiel 7.9 – Übertragung eines weißen, mittelwertfreien Prozesses.

Wählt man als Eingangssignal einen weißen, mittelwertfreien Prozess mit

$$\begin{aligned} s_{XX}[\kappa] &= \delta[\kappa]\sigma_X \\ S_{XX}(e^{j\Phi}) &= \sigma_X \end{aligned}$$

so erhält man am Ausgang die Autoleistungsdichte:

$$\begin{aligned} s_{YY}[\kappa] &= s_{hh}[\kappa]s_{XX}[\kappa] = s_{hh}[\kappa] * \delta[\kappa]\,\sigma_X \\ &= \sigma_X s_{hh}[\kappa] = \sigma_X h[\kappa]h^*[-\kappa] \end{aligned}$$

und

$$S_{YY}(e^{j\Phi}) = \left|H(e^{j\Phi})\right|^2 S_{XX}(e^{j\Phi}) = \sigma_X \left|H(e^{j\Phi})\right|^2$$

Die Ausgangs-Autoleistungsdichte eines weißen Prozesses bei Durchgang durch ein System ist also eine rein reelle Funktion, deren Frequenzverhalten ausschliesslich durch den Betrag der Systemübertragungsfunktion bestimmt ist.

Die Kreuzkorrelation zwischen Eingang und Ausgang bestimmen wir zu

$$s_{XY}[\kappa] = h[\kappa] * \delta[\kappa]\sigma_X = \sigma_X h[\kappa] \tag{7.58}$$

$$S_{XY}(e^{j\Phi}) = \sigma_X H(e^{j\Phi}) \tag{7.59}$$

Impulsantwort und Übertragungsfunktion eines Systems kann man also bestimmen, indem man es mit einem weißen Prozess speist und die Kreuzkorrelation zwischen Eingangs- und Ausgangsprozess misst.

□

7.11 Schätzung der Autokorrelationsfolge

Dieser Abschnitt beschäftigt sich mit der Bestimmung der Autokorrelationsfolge (AKF) eines Prozesses. Diese kann nicht direkt gemessen werden. Um trotzdem Aussagen über die Eigenschaften des Prozesses machen zu können, wurden verschiedene Schätzverfahren entwickelt.

7.11.1 Erwartungstreue und konsistente AKF-Schätzung

Wie im vorigen Kapitel besprochen wurde, ist die Autokorrelationsfolge wie folgt definiert:

$$s_{XX}[k_1, k_2] = E\{X[k_1]X[k_2]\} \tag{7.60}$$

Um die Autokorrelationsfolge bestimmen zu können, muss der Prozess über einen gewissen Zeitraum beobachtet werden. Daraus ergibt sich schon die erste Einschränkung. Damit das Ergebnis sinnvoll ist, muss der Prozess wenigstens über diesen Zeitraum stationär sein. Daher:

$$s_{XX}[\kappa] = E\{X[k]X[k+\kappa]\} \tag{7.61}$$

Da nie der gesamte Zufallsprozess $X[k]$ bekannt ist, sondern immer nur eine Musterfolge gemessen werden kann, muss darüber hinaus Ergodizität vorausgesetzt werden. Nur in diesem Fall lassen sich die Eigenschaften des Prozesses aus einer beliebigen Musterfolge ableiten:

$$s_{XX}[\kappa] = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{k=-N}^{N} x[k]x[k+\kappa] \tag{7.62}$$

Praktisch kann natürlich nur eine Musterfolge der endlichen Länge N betrachtet werden. Um zu zeigen, dass nun nur noch eine Schätzgröße vorliegt, kennzeichnen wir das Ergebnis mit einem „$\hat{}$“:

$$\hat{s}_{XX}[\kappa] = \frac{1}{N} \sum_{k=0}^{N-1} x[k]x[k+\kappa] \tag{7.63}$$

Der Zusammenhang des positiven und negativen Teils der Autokorrelationsfolge kann über die Substitution $k' = k - \kappa$ wie folgt hergeleitet werden:

$$\begin{aligned} \hat{s}_{XX}[-\kappa] &= \frac{1}{N} \sum_{k=0}^{N-1} x[k]x[k-\kappa] \\ &\overset{[k'=k-\kappa]}{=} \frac{1}{N} \sum_{k'=-\kappa}^{N-1-\kappa} x[k']x[k'+\kappa] \\ &\overset{\text{[Stationarität]}}{=} \frac{1}{N} \sum_{k'=0}^{N-1} x[k']x[k'+\kappa] = \hat{s}_{XX}[\kappa] \end{aligned} \tag{7.64}$$

Wie die Autokorrelationsfolge für reelle Signale selbst ist auch ihr Schätzwert gerade. Da positiver und negativer Teil identisch sind, wird im folgenden nur noch der positive Teil berücksichtigt.

Aus der zeitlichen Begrenzung ergibt sich ein weiteres Problem: Um das Produkt unter der Summe zu bilden, muss die Folge um κ gegen sich selbst verschoben werden. Dies führt zu Folgegliedern, die außerhalb des beobachteten Intervalls $0 \ldots N-1$ liegen und folglich nicht bekannt sind. Damit liegen für die Stelle κ nicht N Produkte zur Summierung, wie in (7.62), sondern nur $N - |\kappa|$ vor. Um diese Tatsache in die Formel (7.63) aufnehmen zu können, muss das Vorzeichen von κ berücksichtigt werden.

$$\begin{aligned}\hat{s}_{XX}[|\kappa|] &= \frac{1}{N} \sum_{k=0}^{N-1-|\kappa|} x[k]x[k+|\kappa|] \\ &= \frac{1}{N} \sum_{k=|\kappa|}^{N-1} x[k]x[k-|\kappa|]\end{aligned}$$

Neben der Anpassung der Summationsgrenzen stellt sich auch die Frage, inwieweit der Normierungsfaktor noch sinnvoll ist, wenn nur noch über $N - |\kappa|$ Elemente summiert wird. Dies wird für die weiteren Betrachtungen von Ausschlag gebender Bedeutung sein. Wir setzen daher zunächst einen unbestimmten Normierungsfaktor $1/M$ an und erhalten die folgende allgemeine Schätzfunktion für die Autokorrelationsfolge:

$$\begin{aligned}\hat{s}_{XX}[|\kappa|] = \hat{s}_{XX}[-|\kappa|] &= \frac{1}{M} \sum_{k=0}^{N-1-|\kappa|} x[k]x[k+|\kappa|] \\ &= \frac{1}{M} \sum_{k=|\kappa|}^{N-1} x[k]x[k-|\kappa|] \qquad (7.65)\end{aligned}$$

Für die Wahl des Normierungsfaktor bieten sich zwei plausible Möglichkeiten an. Zum einen kann man sich konsequent an die Definition der Autokorrelationsfolge halten und $M = N$ wählen. Andererseits kann man berücksichtigen, dass mit steigendem $|\kappa|$ immer weniger Summationselemente zur Verfügung stehen und $M = N - |\kappa|$ definieren.

Zur Beurteilung eines Schätzwertes können verschiedene Kriterien definiert werden. Zur Beurteilung der Güte einer AKF-Schätzung sind zwei Kriterien besonders aussagekräftig: die Erwartungstreue und die Konsistenz. Diese sollen nun für die beiden verschiedenen Varianten der Normierung bestimmt werden.

Das erste Kriterium, die *Erwartungstreue*, ist wie folgt definiert:

$$E\{\hat{s}_{XX}[|\kappa|]\} = s_{XX}[|\kappa|] \qquad (7.66)$$

Ein Schätzwert ist also erwartungstreu, wenn sein Erwartungswert mit dem wahren Wert übereinstimmt. Für unsere Schätzfolge ergibt sich:

$$
\begin{aligned}
E\{\hat{s}_{XX}[|\kappa|]\} &= E\left\{\frac{1}{M}\sum_{k=0}^{N-1-|\kappa|} X[k]X[k+\kappa]\right\} \\
&= \frac{1}{M}\sum_{k=0}^{N-1-|\kappa|} E\{X[k]X[k+\kappa]\} \\
&= \frac{1}{M}\sum_{k=0}^{N-1-|\kappa|} s_{XX}[|\kappa|] \\
&= \frac{N-|\kappa|}{M} s_{XX}[|\kappa|] \qquad (7.67)
\end{aligned}
$$

Man erkennt sofort, dass für $M = N - |\kappa|$ die Schätzung erwartungstreu ist. Für $M = N$ dagegen wird die wahre Autokorrelationsfolge mit dem Faktor $(N - |\kappa|)/N$ bewertet. Dies kann man sich anschaulich wie die Multiplikation mit einer dreieckigen Fensterfolge, dem Bartlett-Fenster, vorstellen (siehe Abb. 7.3).

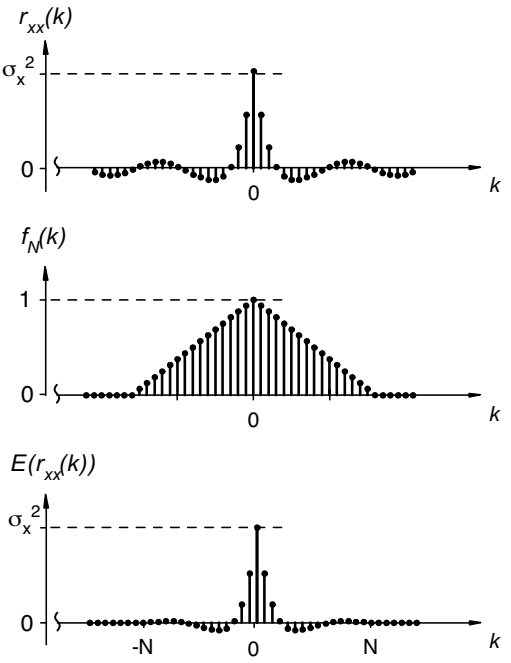


Abbildung 7.3. Modifikation durch Bartlettfenster (Mitte).

Das Bartlettfenster ist dabei wie folgt definiert:

$$f_N^B[|\kappa|] = \begin{cases} \frac{N-|\kappa|}{N} & \text{für} \quad |\kappa| \leq N-1 \\ 0 & \text{sonst} \end{cases} \tag{7.68}$$

Die AKF-Schätzung kann in diesem Fall nur für $N \gg |\kappa|$ als annähernd erwartungstreu angesehen werden.

Das zweite Kriterium ist die *Konsistenz*. Man spricht von einem konsistenten Schätzwert, wenn die Varianz des Schätzers für $N \to \infty$ gegen null geht.

$$\lim_{N\to\infty} Var\left\{\hat{s}_{XX}[|\kappa|]\right\} = 0 \tag{7.69}$$

Um die Formel für die Varianz mit erträglichem Aufwand herleiten zu können, wird im Folgenden der Spezialfall eines mittelwertfreien, unkorrelierten Prozesses behandelt. Die Mittelwertfreiheit ist wie folgt definiert:

$$E\left\{X[k]\right\} = 0 \tag{7.70}$$

Ist der Prozess darüber hinaus unkorreliert, so ist die Autokorrelationsfolge an allen Stellen $\kappa \neq 0$ gleich null:

$$s_{XX}[|\kappa|] = \sigma\,\delta[\kappa] = \begin{cases} \sigma & \text{für} \quad \kappa = 0 \\ 0 & \text{sonst} \end{cases} \tag{7.71}$$

Die Varianz lässt sich wie folgt aus dem quadratischen und linearen Mittelwert berechnen.

$$Var\left\{\hat{s}_{XX}[|\kappa|]\right\} = E\left\{\hat{s}_{XX}^2[|\kappa|]\right\} - \left|E\left\{\hat{s}_{XX}[|\kappa|]\right\}\right|^2 \tag{7.72}$$

Der lineare Mittelwert wurde schon im Zusammenhang mit der Erwartungstreue bestimmt. Daher gilt:

$$\begin{aligned} \left|E\left\{\hat{s}_{XX}[|\kappa|]\right\}\right|^2 &= \left(\frac{N-|\kappa|}{M} s_{XX}[|\kappa|]\right)^2 \\ &= \delta[\kappa]\left(\frac{N-|\kappa|}{M}\sigma\right)^2 \end{aligned} \tag{7.73}$$

Nun berechnen wir den Mittelwert des Quadrates:

$$\begin{aligned} E\left\{\hat{s}_{XX}^2[|\kappa|]\right\} &= E\left\{\frac{1}{M^2}\sum_{i=0}^{N-1-|\kappa|} X[i]X[i+|\kappa|] \sum_{j=0}^{N-1-|\kappa|} X[j]X[j+|\kappa|]\right\} \\ &= \frac{1}{M^2}\sum_{i=0}^{N-1-|\kappa|}\sum_{j=0}^{N-1-|\kappa|} E\left\{X[i]X[i+|\kappa|]X[j]X[j+|\kappa|]\right\} \end{aligned} \tag{7.74}$$

Wir erhalten einen Erwartungswert über das Produkt der Zufallsvariablen des Prozesses an vier Stellen. Da wir den Prozess als unkorreliert angenommen

haben, kann man den Erwartungswert des Produktes mit dem Produkt der Erwartungswerte der einzelnen Faktoren gleichsetzen, sofern sich diese nicht auf dieselbe Stelle beziehen:

$$\begin{aligned} E\{X[i]X[i+\kappa]X[j]X[j+\kappa]\} &= E\{X[i]\}E\{X[i+\kappa]\}E\{X[j]\}E\{X[j+\kappa]\} \\ &= 0 \quad \text{für} \quad i \neq i+\kappa \neq j \neq j+\kappa \end{aligned} \tag{7.75}$$

Tritt ein Index mehrfach auf, wird also das Produkt von Folgegliedern an der selben Stelle betrachtet, so kommt die Unkorreliertheit nicht zum Tragen, und es muss der Erwartungswert über die potenzierte Größe gebildet werden. Damit ist der Erwartungswert im Fall $\kappa = 0$ und $i \neq j$ ungleich null, wenn er die folgende Form annimmt:

$$\begin{aligned} E\{X[i]X[i+\kappa]X[j]X[j+\kappa]\} &\overset{[\kappa=0]}{=} E\left\{X^2[i]X^2[j]\right\} \\ &\overset{[i\neq j]}{=} E\left\{X^2[i]\right\}E\left\{X^2[j]\right\} \\ &= (s_{XX}[0])^2 = \sigma^2 \end{aligned} \tag{7.76}$$

Im Fall $\kappa = 0$ und $i = j$ ergibt sich ein einzelner Wert:

$$E\{X[i]X[i+\kappa]X[j]X[j+\kappa]\} \overset{[\kappa=0\ ,\ i=j]}{=} E\left\{X^4[i]\right\} \tag{7.77}$$

Für den Mittelwert des Quadrates ergibt sich daraus für den Fall $\kappa = 0$:

$$\begin{aligned} E\left\{\hat{s}^2_{XX}[0]\right\} &= \frac{1}{M^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} E\{X[i]X[i]X[j]X[j]\} \\ &= \frac{1}{M^2}\sum_{i=0}^{N-1}\sum_{j\neq i,\ j=0}^{N-1}\sigma^2 + \frac{1}{M^2}\sum_{i=0}^{N-1} E\left\{X^4[i]\right\} \\ &= \frac{N^2-N}{M^2}\sigma^2 + \frac{N\,E\left\{X^4\right\}}{M^2} \end{aligned} \tag{7.78}$$

Für den Fall $\kappa \neq 0$ ist der Erwartungswert nur dann von null verschieden, wenn gilt $i = j$:

$$\begin{aligned} E\left\{\hat{s}^2_{XX}[|\kappa|]\right\} &= \frac{1}{M^2}\sum_{i=0}^{N-1-|\kappa|}\sum_{j=0}^{N-1-|\kappa|} E\{X[i]X[i+\kappa]X[j]X[j+\kappa]\} \\ &= \frac{1}{M^2}\sum_{i=0}^{N-1-|\kappa|} E\left\{X^2[i]X^2[i+\kappa]\right\} = \frac{1}{M^2}\sum_{i=0}^{N-1-|\kappa|}\sigma^2 \\ &= \frac{N-|\kappa|}{M^2}\sigma^2 \end{aligned} \tag{7.79}$$

Damit erhält man für die Varianz:

$$Var\left\{\hat{s}_{XX}[|\kappa|]\right\}\ |_{\kappa=0} = \frac{(N^2-N)\sigma^2 + NE\left\{X^4\right\}}{M^2} - \frac{N^2}{M^2}\sigma^2$$

$$= \frac{N(E\left\{X^4\right\} - \sigma^2)}{M^2}$$

$$\overset{[M=N-|\kappa|=N]}{=} \frac{E\left\{X^4\right\} - \sigma^2}{N} = \mathcal{O}(1/N) \tag{7.80}$$

$$Var\left\{\hat{s}_{XX}[|\kappa|]\right\}\ |_{\kappa\neq 0} = \frac{N-|\kappa|}{M^2}\sigma^2 \tag{7.81}$$

Aus Gl. (7.80) erkennt man zunächst, dass bei beschränktem $E\left\{X^4\right\}$ die Varianz für $\kappa = 0$ mit der Ordnung $\mathcal{O}(1/N)$ klein wird. Aus Gl. (7.81) liest man ab, welchen Einfluss der Normierungsfaktor auf die Konsistenz der Schätzung für $\kappa \neq 0$ hat (siehe auch Abb. 7.4 unten). Für die erwartungstreue Schätzung mit $M = N - |\kappa|$ erhält man den Ausdruck

$$Var\left\{\hat{s}_{XX}[|\kappa|]\right\} = \frac{N-|\kappa|}{(N-|\kappa|)^2}\sigma^2 = \frac{1}{N-|\kappa|}\sigma^2 \tag{7.82}$$

Dieser Ausdruck geht nur dann gegen null, wenn N groß gegenüber κ ist. Für $|\kappa| = N-1$ dagegen geht die Varianz der AKF-Schätzung gegen σ^2. Die Zunahme der Varianz mit steigendem κ war auch zu erwarten, denn es stehen ja immer weniger Werte zur Mittelung zur Verfügung. Fazit: *Die erwartungstreue Schätzung ist nicht konsistent.*

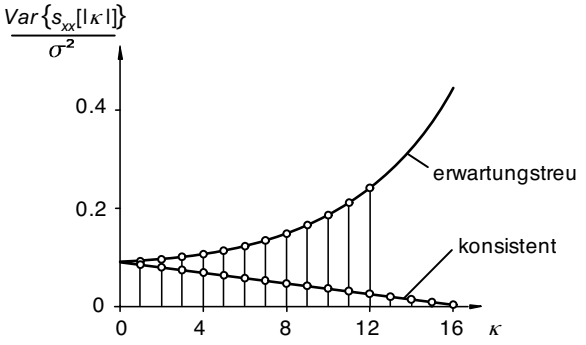


Abbildung 7.4. Varianz des erwartungstreuen (obere Linie) und des konsistenten (untere Linie) Schätzers als Funktion von κ.

Für die Normierung mit $M = N$ (konsistente Schätzung) ergibt sich:

$$Var\left\{\hat{s}_{xx}[|\kappa|]\right\} = \frac{N-|\kappa|}{N^2}\sigma^2 \tag{7.83}$$

Für steigendes N geht die Varianz für alle $|\kappa|$ gegen null. Der Schätzwert ist also konsistent. Mit steigendem κ nimmt die Varianz sogar ab. Diese positive

Eigenschaft wird allerdings dadurch relativiert, dass im gleichen Zuge die Erwartungstreue abnimmt. *Die konsistente Schätzung ist nicht erwartungstreu.*

Wir fassen unsere Ergebnisse zusammen in folgendem

Theorem 7.2 (Konsistente und erwartungstreue Schätzer).
Die AKF eines mittelwertfreien, unkorrelierten Prozesses kann wie folgt geschätzt werden.

$$\textit{Konsistent}: \quad \hat{s}_{XX}[|\kappa|] = \frac{1}{N} \sum_{k=0}^{N-1-|\kappa|} x[k]x[k+|\kappa|] \tag{7.84}$$

$$\textit{Erwartungstreu}: \quad \hat{s}_{XX}[|\kappa|] = \frac{1}{N-|\kappa|} \sum_{k=0}^{N-1-|\kappa|} x[k]x[k+|\kappa|] \tag{7.85}$$

Die Varianz beider Schätzer geht für $\kappa = 0$ *mit* $\mathcal{O}(1/N)$. *Für* $\kappa \neq 0$ *gilt:*

$$\textit{Konsistent}: \quad Var\{\hat{s}_{XX}[|\kappa|]\} = \frac{N-|\kappa|}{N^2}\sigma^2 \overset{[\forall\kappa]}{=} \mathcal{O}(1/N) \tag{7.86}$$

$$\textit{Erwartungstreu}: \quad Var\{\hat{s}_{XX}[|\kappa|]\} = \frac{1}{N-|\kappa|}\sigma^2 \overset{[\exists\kappa]}{\neq} \mathcal{O}(1/N) \tag{7.87}$$

Die konsistente Schätzung der AKF ist nicht erwartungstreu, und die erwartungstreue Schätzung ist nicht konsistent. Das Verhalten der Varianz als Funktion von $|\kappa|$ *wird in Abb. 7.4 deutlich.*

7.11.2 Schätzung mit Hilfe der FFT

Dieser folgende Abschnitt beschäftigt sich mit der praktischen Berechnung der AKF-Schätzung. Dabei beschränken wir uns auf die konsistente Schätzung. Wegen der Symmetrie der Schätzer, Gl. (7.65), setzen wir ohne Beschränkung der Allgemeinheit κ als nichtnegativ voraus:

$$\hat{s}_{XX}[\kappa] = \frac{1}{N} \sum_{k=0}^{N-1-\kappa} x[k]x[k+\kappa] \quad \text{mit} \quad 0 \leq \kappa \leq M \tag{7.88}$$

Die Idee zur Lösung des Problems ist folgende: Die AKF-Schätzung hat eine ähnliche Struktur wie die diskrete Faltung. Es ist bekannt, dass sich die Bestimmung der Faltungssumme im Frequenzbereich erheblich vereinfacht. Außerdem haben wir mit der schnellen Fouriertransformation (FFT) einen Algorithmus kennen gelernt, mit dem man die Transformation in den Frequenzbereich und wieder zurück relativ günstig durchführen kann. Es bietet sich also an, die Berechnung in den Frequenzbereich zu verlegen.

Dazu muss das Problem zunächst an die diskrete Fouriertransformation angepasst werden. Die DFT setzt sowohl im Zeitbereich als auch im Frequenzbereich periodische Folgen voraus. Die AKF-Schätzung wird jedoch aus einer

endlichen Folge $x[k]$ der Länge N berechnet. Um daraus eine periodische Folge zu machen, muss $x[k]$ zyklisch wiederholt werden. Dabei muss berücksichtigt werden, dass die Folge $x[k]$ zur Berechnung der AKF-Schätzung gegen sich selbst verschoben werden wird. Bei periodisch wiederholten Folgen wird dabei der Wert, der auf der linken Seite herausfällt, auf der rechten Seite wieder hineingeschoben (siehe Abb. 7.5).

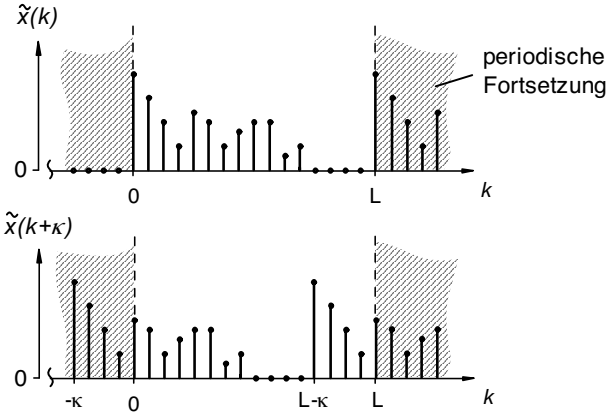


Abbildung 7.5. Verschiebung periodischer Folgen.

Dieses Problem ist von der periodischen bzw. aperiodischen Faltung her bekannt. Um eine Überschneidung der verschiedenen Perioden und damit eine Verfälschung der AKF-Schätzung an der Stelle κ zu verhindern, muss die Periode also mit mindestens κ Nullstellen aufgefüllt werden. Bezeichnet man die größte, für die AKF-Schätzung noch interessante Stelle mit M, so muss für die Periodenlänge L gelten:

$$L \geqslant N + M \tag{7.89}$$

Im Allgemeinen wird man L als eine Potenz von 2 wählen, um das Problem mittels FFT behandeln zu können.

Das periodische Signal, das aus $x[k]$ hervorgeht, kann damit folgendermaßen definiert werden:

$$\tilde{x}[k] = \begin{cases} x[k] & 0 \leq k \leq N-1 \\ 0 & N \leq k \leq L-1 \end{cases} \tag{7.90}$$

$$\tilde{x}[k] = \tilde{x}[k + iL] \quad \text{mit} \quad i \in \mathbb{Z} \tag{7.91}$$

Daraus erhält man die periodische AKF-Schätzung

$$\tilde{s}_{XX}[\kappa] = \frac{1}{N} \sum_{k=0}^{L-1-\kappa} \tilde{x}[k]\tilde{x}[k+\kappa] \tag{7.92}$$

Aus den oben gemachten Betrachtungen folgt, dass periodische und aperiodische Schätzung im Bereich $0 \ldots M$ übereinstimmen:

$$\hat{s}_{XX}[\kappa] = \tilde{s}_{xx}[\kappa] \quad \text{mit} \quad 0 \leq \kappa \leq M \tag{7.93}$$

Nun bringt man die periodische Schätzung durch die Substitution $k = k + \kappa$ in die Form einer Faltungssumme:

$$\begin{aligned}
\tilde{s}_{XX}[\kappa] &= \frac{1}{N} \sum_{k=0}^{L-1-\kappa} \tilde{x}[k]\tilde{x}[k+\kappa] \\
&= \frac{1}{N} \sum_{k=\kappa}^{L-1} \tilde{x}[k]\tilde{x}[k-\kappa] = \frac{1}{N} \sum_{k=\kappa}^{L-1} \tilde{x}[k]\tilde{x}[-(\kappa-k)] \\
&= \frac{1}{N} \tilde{x}[\kappa] * \tilde{x}[-\kappa]
\end{aligned} \tag{7.94}$$

Durch die diskrete Fouriertransformation erhält man nun

$$\begin{aligned}
\text{DFT}\,\{\tilde{s}_{XX}(\kappa)\} &= \frac{1}{N} \text{DFT}\,\{\tilde{x}[\kappa]\}\,\text{DFT}\,\{\tilde{x}[-\kappa]\} \\
&= \frac{1}{N} X[n]X[n] \qquad (7.95) \\
\tilde{s}_{XX}[\kappa] &= \frac{1}{N} \text{IDFT}\left\{X^2[n]\right\} \qquad (7.96)
\end{aligned}$$

Damit läßt sich eine einfache Schrittfolge zur Bestimmung der AKF-Schätzung angeben:

1. Wählen einer geeigneten Periodenlänge L ($L > N + M$, $L = 2^n$), $x[k]$ durch Anhängen von Nullstellen auf die Länge L bringen
2. FFT-Algorithmus liefert $X[n]$
3. Quadrieren von $X[n]$
4. FFT-Algorithmus liefert Rücktransformation
5. Normierung mit $1/N$

7.12 Zusammenfassung und Ausblick

Wir haben nunmehr alle Werkzeuge zur Hand, stochastische Größen zu beschreiben und ihre Transformation durch LTI-Systeme mittels Korrelationen anzugeben. Auch die Systemeigenschaften lassen sich durch stochastische Größen ermitteln.

Allerdings setzt die Schätzung der notwendigen Korrelationsfunktionen eine große Anzahl von Messwerten voraus, die unter den Voraussetzungen der Stationarität und Ergodizität erhoben werden müssen. Und selbst wenn dies gegeben ist, ist die Schätzung nicht gleichzeitig erwartungstreu und konsistent.

Da Stationarität oft nicht über die zum Erheben langer Messreihen notwendigen Zeitspannen vorausgesetzt werden kann, wollen wir uns im nächsten Kapitel einer völlig anderen, *modellgestützten* Methode der Beschreibung von Zufallsprozessen zuwenden. Diese ist modellgestützt und vermeidet damit den Nachteil des Benutzens langer, als stationär vorauszusetzender Zeitreihen.

Übungen

Übung 7.1 – Erwartungswert und Varianz.

Betrachten Sie den Zufallsprozess aus Beispiel 7.1. Berechnen Sie den Erwartungswert und die Varianz dieses Zufallsprozesses.

Übung 7.2 – Fragen zu stochastischen Signalen: Wahr oder falsch?.

Prüfen Sie die folgenden Aussagen auf ihre Richtigkeit.

a. Ein stationäres Signal kann einen von der Zeit unabhängigen Mittelwert, jedoch eine zeitabhängige Varianz haben.
b. Ergodische Signale haben die Eigenschaft, dass sich der Erwartungswert sowie die Autokorrelationsfolge aus einer einzigen Realisierung $x(t)$ bestimmen lassen.
c. Aus dem Leistungsdichtespektrum lässt sich ohne Zusatzinformation die AKF ermitteln.
d. Das Leistungsdichtespektrum lässt sich nicht für periodische Signale berechnen.

Übung 7.3 – Leistung eines stochastischen Stromes.

Ein Widerstand von $R/\Omega = 5$ werde von einem zufälligen Strom mit der Autokorrelationsfolge $s_{II}[\tau]/A^2 = 2e^{-|\tau|}$ durchflossen.

a. Wie groß ist die verbrauchte Leistung P?
b. Wie lautet die spektrale Leistungsdichte $S_{II}(j\omega)$?

Übung 7.4 – Filtern von Rauschen.

Gegeben ist eine Rauschquelle, die ein weißes mittelwertfreies Rauschen mit der Varianz σ_X erzeugt. Außerdem ist ein Übertragungssystem gegeben, das durch die Differenzengleichung $y[n] = 1/2x[n] + 1/2x[n-1]$ vollständig charakterisiert sei.

a. Geben Sie die Autokorrelationsfolge und die Autoleistungsdichte für den Rauschprozess an (Skizze).
b. Bestimmen Sie die Impulsantwort $h[k]$ und die Systemkorrelation $s_{hh}[\kappa]$ für das Übertragungssystem.
c. Das Übertragungssystem wird nun von der Rauschquelle gespeist. Bestimmen Sie die Autokorrelationsfolge $s_{YY}[\kappa]$ des Ausgangssignals. Was können Sie über die Varianz σ_Y am Ausgang aussagen?

8

Modellsysteme

Die im vorigen Kapitel behandelten klassischen Verfahren haben einige entscheidende Nachteile. Zur Bestimmung der Autokorrelationsfolge können nur endlich viele Werte aufgenommen werden. Die zeitliche Fensterung, die sich daraus für die AKF ergibt, führt zu ungünstigen Spektraleigenschaften, wie wir im Kapitel über die diskrete Fouriertransformation gesehen haben. Insbesondere schmalbandige Prozesse können so nur unzureichend dargestellt werden. Darüber hinaus ist es problematisch, Schätzwerte zu bestimmen, die sowohl erwartungstreu als auch konsistent sind. Dazu ist es nötig, dass eine große Anzahl von Folgegliedern für die Berechnung zur Verfügung steht, was wiederum zu einem hohen Rechenaufwand führt. Außerdem sind viele Prozesse nicht streng stationär. Bei solchen Prozessen kann Stationarität nur für kurze Zeitspannen angenommen werden, z.B. bei der Produktion akustischer Laute, wodurch sich die Forderung nach einer *Kurzzeitschätzung der Autokorrelationsfolge* ergibt. Dies ist mit den traditionellen Verfahren nicht befriedigend zu realisieren.

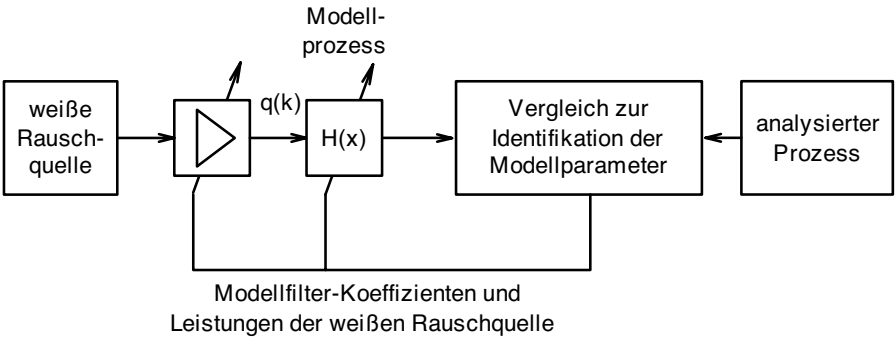


Abbildung 8.1. Modellsystem für einen Rauschprozess

Einen völlig anderen Ansatz bieten die sogenannten modellgestützten Verfahren, deren Funktionsprinzip in Abb. 8.1 dargestellt ist. Man versucht hier die Entstehung des Prozesses nachzubilden. Dazu nimmt man ein lineares System an, dessen *Ordnung*, also die Anzahl der Parameter, vorgegeben wird. Die Werte der Parameter sind zunächst unbekannt. Das lineare System wird mit weißem Rauschen der Leistung 1 gespeist. Dieses Rauschen wird auf eine zunächst unbekannte Leistung $q[k]$ verstärkt. Die Leistung der resultierenden Rauschquelle $q[k]$ sowie die Parameter des Systems werden zusammen als *Modellparameter* bezeichnet. Sie werden nun so gewählt, dass das Ausgangssignal $x[k]$ (der *Modellprozess*) das Spektrum des zu analysierenden Prozesses möglichst gut annähert. Diese *Adaption* der Modellparameter wird in Abb. 8.1 durch einen diagonalen Pfeil durch das entsprechende Schaltungselement dargestellt.

Da die Anregung durch einen Rauschprozess $q[k]$ geschieht, können zur Bestimmung der Modellparameter nicht unmittelbar die (stochastischen) Eingangs- und Ausgangsgrößen verwendet werden, sondern es müssen wieder Korrelationsfolgen benutzt werden. Man beachte aber folgenden Unterschied: in der bisherigen Vorgehensweise wurden Korrelationsfolgen als direkt determinierende Größen des Prozesses $x[k]$ gemessen, über die Entstehung von $x[k]$ wurden keine Annahmen getroffen. Im Gegensatz dazu wird bei dem modellgestützten Verfahren $x[k]$ aus einem linearen Prozess mit Rauschanregung $q[k]$ erzeugt. Korrelationsfolgen werden jetzt nur als Mittel zur Bestimmung der Modellparameter benötigt; dies ermöglicht es, auch Kurzzeitkorrelationen zu verwenden.

Falls der wahre Prozess durch die Struktur des Modells korrekt abgebildet wird, ist die Parameterschätzung exakt. Diese Kenntnis kann benutzt werden, um a-priori Wissen über die Prozessstruktur einzubauen, z.B. bekannte Ordnung des zu analysierenden Prozesses. Es wird also nicht mehr die (Langzeit-)AKF selbst, sondern es werden die Parameter eines Übertragungssystems geschätzt. Dadurch verringert sich die Zahl der zu schätzenden Größen in der Regel erheblich. Um diese zu bestimmen, genügt ein kurzer Beobachtungszeitraum, was zu befriedigenden Ergebnissen auch für kurzzeitstationäre Prozesse führt.

Absichtlich wird hier nicht die übliche Notation Eingang: $x[k]$, Ausgang: $y[k]$ gewählt. Stattdessen schreiben wir $q[k]$ als Anregung und $x[k]$ für den eigentlichen „Ausgang“ des Modellprozesses. Dies hat zwei Gründe: 1.) Durch die Bezeichnung $q[k]$ für die Anregung wird deutlich, dass es sich um ein Rauschsignal handelt. 2.) Der Modellprozess ist oft unbekannt, der Beobachter sieht also erst $x[k]$. Er kann nun $x[k]$ als Eingangssignal eines Analysesystems verwenden, welches den Prozess identifiziert. Wir werden solche Verfahren im Laufe dieses Kapitels kennenlernen. Der Ausgang des Analysesystems ist dann der „eigentliche“ Systemausgang. Durch die Wahl der Notation können also Modellsystem und Analysesystem unterschieden werden.

8.1 Einfaches Modellsystem: Markov-Prozess

Die einfachste Realisierung einer Parameterschätzung wird bei einem Prozess mit nur einem Parameter vorgenommen. Einen solchen Prozess bezeichnet man als Markov-Prozess. Der Markov-Prozess ist dadurch definiert, dass er nur auf den unmittelbar vorhergehenden Zustand zurückgreift. Er gehört zu den autoregressiven Modellen, auf die wir im nächsten Abschnitt noch näher eingehen wollen. Der Markov-Prozess ist kausal und hat die folgende Differenzengleichung:

$$x[k] = q[k] - a_1 x[k-1]. \tag{8.1}$$

Dabei steht $x[k]$ hier für den Ausgang des Prozesses. Diese Notation ist so gewählt, weil $x[k]$ als Eingangsgröße für das nachfolgende Analysesystem dient, welches die Modellparameter schätzt. Der Prozess wird angeregt mit weißem, mittelwertfreien Rauschen $q[k]$, für das gilt:

$$s_{QQ}[\kappa] = \delta[\kappa]\sigma_Q \tag{8.2}$$

und

$$S_{QQ}(e^{j\Phi}) = \sigma_Q. \tag{8.3}$$

Durch die Transformation in den Z-Bereich erhält man die Übertragungsfunktion:

$$\begin{aligned} X(z) &= Q(z) - a_1 z^{-1} X(z) \\ X(z)(1 + a_1 z^{-1}) &= Q(z) \\ H(z) = \frac{X(z)}{Q(z)} &= \frac{1}{1 \; + \; a_1 z^{-1}} \end{aligned} \tag{8.4}$$

Das Verhalten des gesamten Systems wird also nur durch einen Parameter a_1 bestimmt. Wir wollen nun zeigen, in welcher Weise die Autokorrelationsfolge und die Autoleistungsdichte des Ausgangsprozesses $X(n)$ von diesem Parameter a_1 abhängen. Die dazu benötigten Gleichungen wurden bereits im Abschnitt 7.10 des Kapitels *Stochastische Signalverarbeitung* hergeleitet. Wir stellen sie hier noch einmal in Kurzform zusammen und wenden sie auf den Markov-Prozess an.

Für die Autoleistungsdichte gilt

$$\begin{aligned} S_{XX}(e^{j\Phi}) = \sigma_Q \left|H(e^{j\Phi})\right|^2 &= \frac{\sigma_Q}{\left|1 + a_1 e^{-j\Phi}\right|^2} \\ &= \frac{\sigma_Q}{1 + 2Re\left\{a_1 e^{-j\Phi}\right\} + \left|a_1\right|^2}. \end{aligned} \tag{8.5}$$

Für reellwertige Koeffizienten a_1 ergibt sich speziell

$$S_{XX}(e^{j\Phi}) = \frac{\sigma_Q}{1 + 2a_1 \cos(\Phi) + a_1^2} \tag{8.6}$$

Die Autokorrelationsfolge am Ausgang erhält man aufgrund der Beziehung:

$$s_{XX}[\kappa] = \sigma_Q s_{hh}[\kappa] = \sigma_Q h[\kappa] * h[-\kappa] \tag{8.7}$$

Dabei berechnet man die Impulsantwort aus der inversen Z-Transformation der kausalen Übertragungsfunktion

$$h[k] = Z^{-1}\{H(z)\} = Z^{-1}\left\{\frac{1}{1+a_1 z^{-1}}\right\} = \begin{cases}(-a_1)^k & \text{für } k \geq 0\\ 0 & \text{sonst}\end{cases}. \tag{8.8}$$

Damit ergibt sich die Systemkorrelation

$$\begin{aligned} s_{hh}[|\kappa|] &= \sum_{k=0}^{\infty}(-a_1)^k(-a_1)^{k+|\kappa|} \\ &= (-a_1)^{|\kappa|}\sum_{k=0}^{\infty}(-a_1)^{2k} \\ &= \frac{(-a_1)^{|\kappa|}}{1-(-a_1)^2}, \end{aligned} \tag{8.9}$$

wobei von der Beziehung $\sum_{k=0}^{\infty} b^k = \frac{1}{1-b}$ mit $b = (-a_1)^2$ für geometrische Reihen Gebrauch gemacht wurde. Für die Autokorrelationsfolge gilt dann

$$s_{XX}[|\kappa|] = \frac{(-a_1)^{|\kappa|}}{1-(-a_1)^2}\sigma_Q \tag{8.10}$$

oder in rekursiver Darstellung

$$\begin{aligned} s_{XX}(|\kappa|+1) &= \frac{(-a_1)^{|\kappa|+1}}{1-(-a_1)^2}\sigma_Q \\ &= (-a_1)\left(\frac{(-a_1)^{|\kappa|}}{1-(-a_1)^2}\sigma_Q\right) \\ &= (-a_1)s_{XX}[|\kappa|]. \end{aligned} \tag{8.11}$$

Man erkennt, dass der Markov-Prozess (und autoregressive Modelle allgemein) nicht von endlichen Autokorrelationsfolgen ausgeht wie die traditionellen Verfahren, sondern von einer unendlichen AKF, bei der jedes Folgeglied aus seinen Vorgängern bestimmt werden kann. Die Abb. 8.2 zeigt, dass trotz der sehr einfachen Struktur des Modells Spektren verschiedener Charakteristik erzeugt werden können. Für negative a_1 erhält man ein Tiefpass-Prozess, für positive a_1 einen Hochpass-Prozess und mit komplexem a_1 kann schließlich ein Bandpassprozess realisiert werden. Durch die Veränderung eines einzigen Parameters bei einem sehr einfachen Prozess erster Ordnung können also bereits ganz verschiedene Verhalten des Systems hervorgerufen werden.

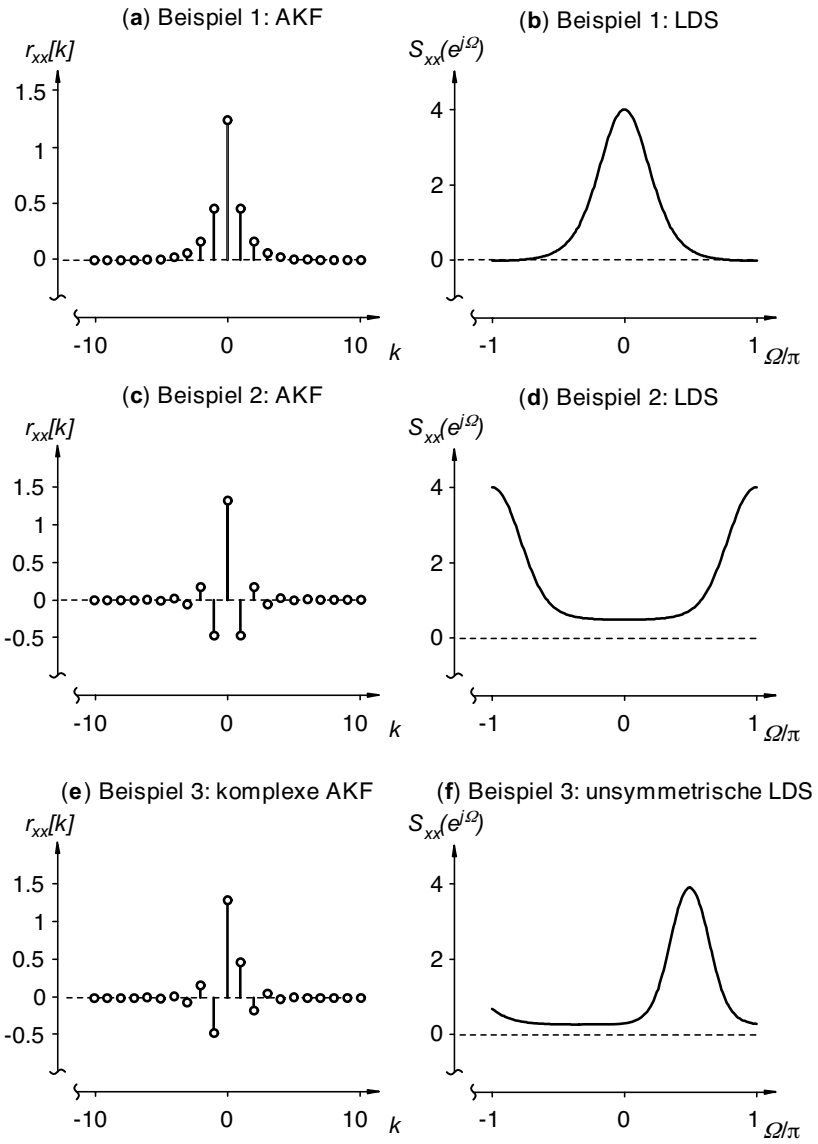


Abbildung 8.2. AKF und LDS eines Markov-Prozesses für verschiedene Parameter

8.2 AR, MA, ARMA-Modelle

Nachdem im vorigen Abschnitt schon der Markov-Prozess als ein einfaches Beispiel für einen autoregressives Modell vorgestellt wurde, sollen nun noch einmal allgemein die verschiedenen Modellformen vorgestellt werden. Nach der Struktur der Übertragungsfunktion werden drei verschiedene Typen unterschieden. Die Signalflusspläne dazu sind in Abb. 8.3 dargestellt.

- Das *autoregressive Modell* (AR-Modell)

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{\nu=1}^{n} a_\nu z^{-\nu}} \tag{8.12}$$

  Autoregressiv bedeutet, dass nur (auto) der Wert am Ausgang zur Rückkopplung (regressiv, feedback) benutzt wird. Man spricht auch von einem *all-pole-System*, da die Übertragungsfunktion ausschließlich Pole enthält.
- Das *moving-average Modell* (MA-Modell)

$$H(z) = B(z) = 1 + \sum_{\mu=1}^{m} a_\mu z^{-\mu} \tag{8.13}$$

  Der Name Moving-average (gleitender Mittelwert) leitet sich daraus her, dass gleitende Mittelwerte ebenso durch Gewichtung des Eingangs und seiner Vorgängerwerte (forward-loop) berechnet werden. Es handelt sich um ein sogenanntes *all-zero-System*, da die Übertragungsfunktion ausschliesslich Nullstellen enthält.
- Das *autoregressive moving-average Modell* (ARMA-Modell)

$$H(z) = \frac{B(z)}{A(z)} = \frac{1 + \sum_{\mu=1}^{m} b_\mu z^{-\mu}}{1 + \sum_{\nu=1}^{n} a_\nu z^{-\nu}} \tag{8.14}$$

  Dieses enthält beide Funktionalitäten (AR und MA). Die ARMA-Systeme entsprechen der allgemeinen Differenzengleichung, siehe Theorem 3.11 auf Seite 64. Nach diesem Theorem kann auch eine zu Abb. 8.3(c) äquivalente kanonische Schaltung gewählt werden, die mit $\max\{N, M\}$ vielen Verzögerern auskommt.

Von den vorgestellten Systemen ist das *autoregressive Modell* das gebräuchlichste. Dies liegt zum einen daran, dass es theoretisch einfach zu behandeln ist: Die Schätzung der Parameter des AR-Modells führt auf ein lineares Gleichungssystem, bei MA- und ARMA-Modellen dagegen auf nichtlineare Gleichungssysteme. Zum anderen liegen die Stärken des AR-Modells gerade da, wo die traditionellen Verfahren ihre Schwächen haben, bei der Darstellung schmalbandiger Prozesse, siehe Abb. 8.4. Da nur ein endlicher Ausschnitt (Fenster) aus der Reihe der Folgenwerte bekannt ist, führt die zeitliche Fensterung bei den traditionellen Verfahren zu einer starken Verbreiterung des

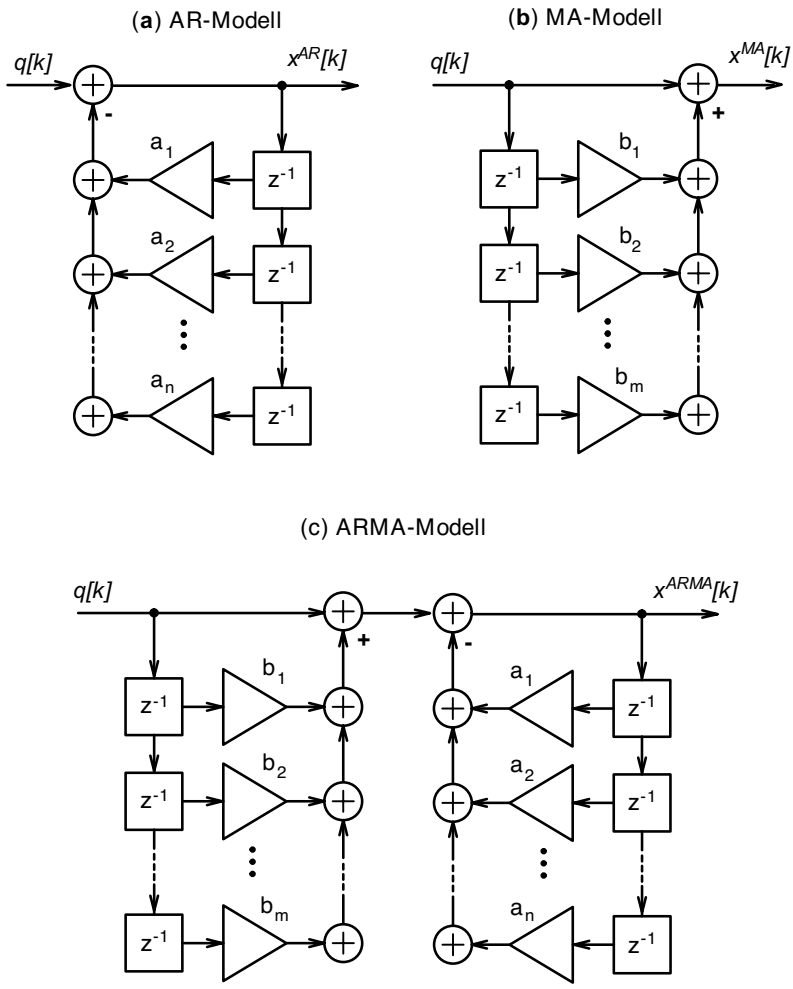


Abbildung 8.3. Signalflusspläne zu AR-, MA-, und ARMA-Modellsystemen

Impulses im Frequenzbereich. Dies ist bereits von der Fouriertransformation des Rechteckfensters bekannt, welches dem AKF-Fenster äquivalent ist. In der Abb. 8.4 wird dies im mittleren Bild gut sichtbar. Das AR-Modell dagegen kann den Deltaanteil gut durch eine Polstelle annähern, wie man ebenfalls in der Abbildung sieht. Bei der Darstellung breitbandiger Prozesse ist das AR-Modell allerdings weniger geeignet.

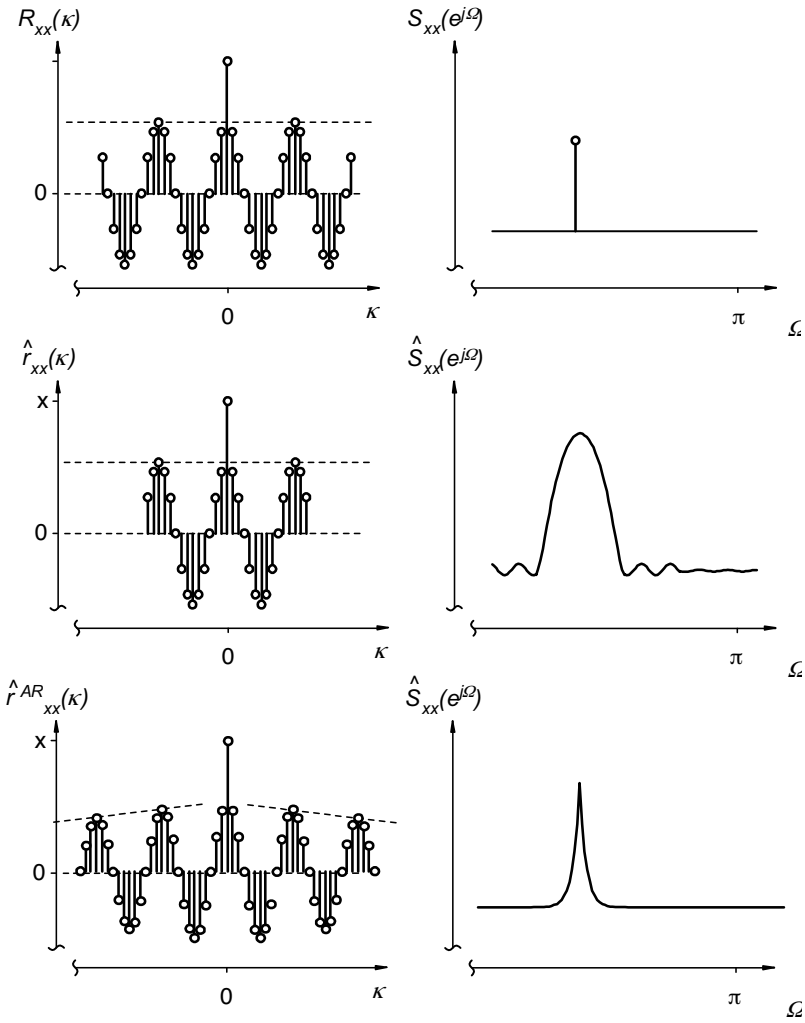


Abbildung 8.4. Vergleich von traditioneller und autoregressiver Schätzung einer Sinusschwingung. Links: Autokorrelation, rechts: Autoleistungsdichte. a) tatsächlicher Verlauf b) AKF-Schätzung, traditionell c) Modellsystem (AR)

8.3 Yule-Walker Gleichung

Um das Spektrum eines Zufallsprozesses mit Hilfe eines Modellsystems schätzen zu können, muss eine Beziehung zwischen den Modellparametern $(q, \{a_i\})$ und der Autokorrelationsfolge $s_{XX}[k]$ gefunden werden. Diese Beziehung soll hier für das autoregressive Modell hergeleitet werden. Wir verlangen dazu im Prinzip, dass alle Werte $x[k]$ gerade durch einen Modellprozess n-ter Ordnung erzeugt werden. Da es sich um einen stochastischen Prozess handelt,

mitteln wir diese Erzeugungsbedingung über viele Realisationen des Prozesses, so dass wir die Erzeugungsbedingung als Funktion der (Kurzzeit-)AKF darstellen können. Diese Bedingung ergibt dann Gleichungen für die n Parameter $a_1 ... a_n$.

Wir betrachten wie schon im vorigen Kapitel einen reellen Prozess. Zunächst setzt man die Differenzengleichung des Systems

$$x[k] = q[k] - \sum_{\nu=1}^{n} a_\nu x[k-\nu] \tag{8.15}$$

in die Definitionsgleichung der Autokorrelationsfolge ein. Dadurch erhalten wir weitere, schon bekannte Korrelationen:

$$\begin{aligned} s_{XX}[\kappa] &= E\{X(k)X(k+\kappa)\} \\ s_{XX}[\kappa] &= E\left\{X(k)\left[Q(k+\kappa) - \sum_{\nu=1}^{n} a_\nu X(k-\nu+\kappa)\right)\right\} \\ s_{XX}[\kappa] &= E\left\{X(k)Q(k+\kappa)\right\} - \sum_{\nu=1}^{n} a_\nu E\left\{X(k)X(k-\nu+\kappa)\right\}. \end{aligned} \tag{8.16}$$

Man beachte, dass wir hier nichts über die Erwartungswerte ausgesagt haben. Sie werden laut Definition als Mittelung über verschiedene Realisationen des Prozesses gebildet. Nehmen wir Ergodizität an, so können sie innerhalb einer Realisierung des Prozesses gebildet werden. Über die Länge des dabei verwendeten Zeitfensters ist hier noch nichts ausgesagt.

Der erste Erwartungswert in $s_{XX}[k]$ ist (nach Definition) die Kreuzkorrelation zwischen Eingang Q und Ausgang X:

$$E\left\{X(k)Q(k+\kappa)\right\} = s_{XQ}[\kappa]. \tag{8.17}$$

Für die Kreuzkorrelationsfolge haben wir bereits hergeleitet:

$$s_{XQ}[\kappa] = s_{XX}[\kappa] * h[\kappa] = \sigma_Q \delta[\kappa] * h[\kappa] = \sigma_Q h[\kappa] \tag{8.18}$$

Da es sich am Eingang, und damit auch am Ausgang, um einen weißen Prozess handelt, ist es plausibel anzunehmen, dass auch Eingangs- und Ausgangsprozess unkorreliert sind. Daher braucht nur die Stelle $\kappa = 0$ der Kreuzkorrelationsfolge betrachtet zu werden. Setzen wir Kausalität voraus, können wir aus der Differenzengleichung des Systems entnehmen, dass

$$x[0] = q[0]. \tag{8.19}$$

gilt. Damit ist auch die Impulsantwort an der Stelle $\kappa = 0$ gegeben, denn wegen der Kausalität gilt für den ersten Zeitschritt:

$$h[0] = x[0]/q[0] = 1. \tag{8.20}$$

Für die Kreuzkorrelation erhalten wir damit

$$s_{XQ}[\kappa] = \begin{cases} \sigma_Q & \text{für } \kappa = 0 \\ 0 & \text{sonst} \end{cases}. \tag{8.21}$$

Wir fahren nun fort mit der Betrachtung von $s_{XX}[\kappa]$. In diesem Ausdruck ist der zweite Erwartungswert unter der Summe gleich der verschobenen Autokorrelation am Ausgang. Damit ergibt sich:

$$\sum_{\nu=1}^{n} a_\nu E\left\{X(k)X(k-\nu+\kappa)\right\} = \sum_{\nu=1}^{n} a_\nu s_{XX}[\kappa-\nu] \tag{8.22}$$

Mit den Ergebnissen aus Gl. (8.22) erhalten wir folgende Formel für die AKF am Ausgang:

$$s_{XX}[\kappa] = \begin{cases} \sigma_Q - \sum_{\nu=1}^{n} a_\nu s_{XX}[\nu] & \text{für } \kappa = 0 \\ -\sum_{\nu=1}^{n} a_\nu s_{XX}[\kappa-\nu] & \text{für } \kappa > 0 \end{cases}. \tag{8.23}$$

Für $\kappa > 0$ ergeben sich n Gleichungen mit denen die n Koeffizienten a_i bestimmt werden können. Die Rauschleistung des Eingangsprozesses kann anschließend aus s_{XX} mit $\kappa = 0$ berechnet werden.

Bringt man das Gleichungssystem in eine Matrixform, so erhält man:

$$\mathbf{S_{XX}a} = -\mathbf{s_{XX}}. \tag{8.24}$$

Ausgeschrieben sieht die Matrixgleichung folgendermaßen aus:

$$\begin{pmatrix} s_{XX}[0] & s_{XX}[-1] & \cdots & s_{XX}[-(n-1)] \\ s_{XX}[1] & s_{XX}[0] & \cdots & s_{XX}[-(n-2)] \\ \vdots & \vdots & \ddots & \vdots \\ s_{XX}[n-1] & s_{XX}[n-2] & \cdots & s_{XX}[0] \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} -s_{XX}[1] \\ -s_{XX}[2] \\ \vdots \\ -s_{XX}[n] \end{pmatrix} \tag{8.25}$$

Für reellwertige Signale ist die AKF symmetrisch, so dass $s_{XX}[k] = s_{XX}[-k]$ gilt. In diesem Fall enthält die Matrixgleichung also nur n zu messende Größen $s_{XX}[k]$.

Den Koeffizientenvektor $\mathbf{a}$ erhält man nun durch die linksseitige Multiplikation mit der inversen Matrix S_{XX}^{-1}. Es folgt

Theorem 8.1 (Yule-Walker-Gleichung).

$$\mathbf{a} = -\mathbf{S_{XX}}^{-1}\mathbf{s_{XX}} \tag{8.26}$$

Diese Beziehung bezeichnet man als ***Yule-Walker-Gleichung****. Zur vollständigen Beschreibung des Modellprozesses benötigen wir noch die Rauschleistung σ_Q. Diese erhält man aus der Gl.* (8.23) *mit $\kappa = 0$:*

$$s_{XX}[0] = \sigma_Q - \sum_{\nu=1}^{n} a_\nu s_{XX}[\nu] = \sigma_Q - \mathbf{s_{XX}}^T\mathbf{a} \tag{8.27}$$

$$\sigma_Q = s_{XX}[0] + \mathbf{s_{XX}}^T\mathbf{a} = s_{XX}[0] - \mathbf{s_{XX}}^T\mathbf{S_{XX}}^{-1}\mathbf{s_{XX}} \tag{8.28}$$

Mit Hilfe der Yule-Walker-Gleichung können wir (bei reellwertigen Signalen) aus den Gliedern $s_{XX}[0]...s_{XX}[n]$ der Autokorrelationsfolge die Koeffizienten des Prozesses berechnen.

Beispiel 8.1 – Yule-Walker-Gleichung.

Gegeben sei ein System 2. Ordnung mit den Modellkoeffizienten $a_1 = 1/4, \quad a_2 = -1/2$. Die Systemgleichung lautet damit

$$x[n] = q[n] - \frac{1}{4}x[n-1] + \frac{1}{2}x[n-2] \quad [q(n) = \text{Anregung}].$$

Vor dem Zeitpunkt t=0 sind alle Grössen im System Null. Das System wird dann angeregt mit einer delta-Folge $q[n] = \delta[n]$.
Wir bestimmen die Autokorrelationsfolge $s_{XX}(\kappa)$. Dies könnten wir gemäss Gl. (7.54) durch Faltung der AKF der anregenden Delta-Folge mit der Systemkorrelationsfolge des übertragenden Systems tun. (Siehe dazu auch Übung 8.3.) Hier können wir aber auch die Yule-Walker-Gleichung zu diesem Zweck benutzen, und zwar am geeignetsten in der Form der Gln. (8.23). Da die anregenden Folge eine Delta-Folge ist, erhalten wir aus Gl. (8.23) das folgende allgemeine Gleichungssystem für Systeme 2. Ordnung, diesmal umgeschrieben zur Bestimmung der AKF-Werte:

$$\begin{pmatrix} 1 & a_1 & a_2 \\ a_1 & 1+a_2 & 0 \\ a_2 & a_1 & 1 \end{pmatrix} \begin{pmatrix} s_{XX}[0] \\ s_{XX}[1] \\ s_{XX}[2] \end{pmatrix} = \begin{pmatrix} \sigma_Q \\ 0 \\ 0 \end{pmatrix} \tag{8.29}$$

(Zum Gültigkeitsbereich dieser Gleichung siehe die Übung 8.4.) Wir setzen die Koeffizienten und die Rauschleistung der Delta-Folge (= 1) ein und erhalten

$$\begin{pmatrix} 1 & 1/4 & -1/2 \\ 1/4 & 1/2 & 0 \\ -1/2 & 1/4 & 1 \end{pmatrix} \begin{pmatrix} s_{XX}[0] \\ s_{XX}[1] \\ s_{XX}[2] \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

Dieses Gleichungssystem wird gelöst durch

$$s_{XX}[0] = \frac{16}{9} = 1.777, \; s_{XX}[1] = -\frac{8}{9} = -0.888, \; s_{XX}[2] = \frac{10}{9} = 1.111 \tag{8.30}$$

Weitere Werte erhalten wir durch Anwendung der Rekursion aus Gl. (8.23), also

$$s_{XX}[\kappa] = -a_1 s_{XX}[\kappa - 1] - a_2 s_{XX}[\kappa - 2] \quad \text{für } \kappa > 2$$

und damit z.B.

$$s_{XX}[3] = -a_1 s_{XX}[2] - a_2 s_{XX}[1] = -13/18 = -0.7222 \tag{8.31}$$

Man beachte, dass wir auf diese Weise bei einem System der Ordnung N immer zunächst ein Gleichungssystem der Ordnung $N+1$ lösen müssen und weitere Werte durch Rekursion erhalten. Dies macht vor allem dann Sinn, wenn wir nur an wenigen Werten der AKF interessiert sind. Die Lösung desselben Problems durch Faltung der AKF der anregenden Delta-Folge mit der Systemkorrelationsfolge des übertragenden Systems bzw. durch Fouriertransformation dieser Folgen (Gl. 7.55) ist zunächst mit mehr Aufwand verbunden, liefert dann aber einfacher die höheren Werte der AKF.
Ein weiterer interessanter Aspekt dieser Rechnung ist, dass die AKF bei genauer Kenntnis des Eingangsprozesses exakt berechnet werden kann. Im Gegensatz dazu stehen Schätzprobleme der AKF bei *Messung* mit endlicher Folgenlänge, wie wir gleich sehen werden. □

8.4 Lösung der Yule-Walker-Gleichung für endliche Merkmalsfolgen

Es sei noch einmal, wie schon oben bei der Betrachtung der Erwartungswerte, wiederholt, dass wir mit den Yule-Walker-Gleichungen noch nichts darüber ausgesagt haben, wie wir eigentlich zu den darin benötigten Gliedern der Autokorrelationsfolge gelangen. Zunächst gilt allgemein, dass bei gegebenen n Gliedern der Autokorrelationsfolge, die Koeffizienten eines linearen autoregressiven Modells mittels der Yule-Walker-Gleichung bestimmt werden.

Wir unterscheiden jetzt die Fälle der *Langzeitstationarität* und der *Kurzzeitstationarität*. Im Falle der Langzeitstationarität können die Glieder $s_{XX}[k]$ aus einer langen Merkmalsfolge, typischerweise mehrere hundert Glieder, mit hoher Genauigkeit gewonnen werden. Die AKF sagt in diesem Fall ausreichend viel über die Eigenschaften des Signals und damit des Prozesses aus. Da viele, genau bestimmte Glieder der AKF zur Verfügung stehen, kann auch eine Yule-Walker-Gleichung für einen Prozess hoher Ordnung bestimmt werden. Dabei wird möglicherweise über die Langzeit-AKF hinaus zusätzliche Information über den Prozess gewonnen: falls die Koeffizienten a_k des Modellprozesses ab einem bestimmten k_{max} Null oder mindestens sehr klein werden, kann davon ausgegangen werden, dass der erzeugende Prozess tatsächlich nur von der Ordnung k_{max} war. Diese Information hätte alleine aus der AKF nicht gewonnen werden können. Falls solch ein Abbruch der Koeffizienten nicht beobachtet wird, ist der Modellprozess in voller Ordnung n anzunehmen. In diesem Fall sind die beiden Darstellungen, mit n Gliedern der AKF oder n Koeffizienten des Modellprozesses, gleichwertig und auch gleich komplex in ihrer Darstellungsweise. Der letztere Fall wird jedoch, wenn wir bei Langzeitstationarität von mehreren hundert verfügbaren Gliedern der Merkmalsfolge ausgehen, praktisch nicht auftreten, da ein erzeugender AR-Prozess von einer derartig hohen tatsächlichen Komplexität (Ordnung > 100) praktisch nie beobachtet wird.

Daher können wir festhalten, dass in jedem Fall eine starke Reduktion der Anzahl der zur Prozessbeschreibung notwendigen Parameter erreicht werden kann. Sind die (wenigen) Koeffizienten a_k des Modellprozesses berechnet, so können alle Glieder der AKF über die Yule-Walker-Gleichung in der oben angegebenen Form

$$s_{XX}[\kappa] = -\sum_{\nu=1}^{n} a_\nu s_{XX}[\kappa - \nu] \quad \text{für} \quad \kappa > 0 \tag{8.32}$$

rekursiv ermittelt werden. Dies gilt auch über die n bekannten Glieder hinaus. Es kann also die *gesamte* AKF aus dem berechneten Modellprozess bestimmt werden. Wir betrachten jetzt den Fall der Kurzzeitstationarität. Sei der Prozess stationär über s Folgenglieder, etwa mehrere zehn viele. Die Glieder der AKF können nur aus maximal s Signalwerten bestimmt werden. Das bedeutet insbesondere, dass auch nur die Glieder $s_{XX}[0]..s_{XX}[s]$ bestimmt werden können. Wie wir wissen, entstehen dabei beträchtliche Fehler, die vor allem für hohe Indexwerte k (nahe s) auftreten.

Gehen wir aber davon aus, dass der Prozess von „normaler" Ordnung (< 10) ist, so sind gemäß der Yule-Walker-Gleichung auch nur soviele Glieder der AKF notwendig. Für diese kann noch eine ausreichende Genauigkeit angenommen werden. Darüber hinaus enthält die Yule-Walker-Gleichung den Fehler (in linearer Kombination mit Gewichten a_k) auf beiden Seiten der Gleichung, was man in folgender Form gut sehen kann:

$$\mathbf{S_{XX}a} = -\mathbf{s_{XX}} \tag{8.33}$$

Dadurch wird erreicht, dass sich ein eventuell vorhandener Fehler in der Bestimmung der Glieder der AKF gleichmäßig auf beide Seiten der Gleichung aufteilt, was den resultierenden Fehler in der Bestimmung der Koeffizienten a_k deutlich vermindert. Wir erkennen, dass damit selbst im Fall der Kurzzeitstationarität ein Modellprozess kleiner Ordnung (z.B. < 10) noch ausreichend sicher geschätzt werden kann. Damit bleiben alle oben für die Langzeitstationarität genannten Vorteile erhalten: Verringerung der Anzahl der notwendigen Parameter, Ermittlung der tatsächlichen Ordnung des erzeugenden Prozesses sowie die Möglichkeit der Berechnung beliebig vieler Glieder der AKF über Rekursion. Insbesondere diese Berechenbarkeit entsteht erst durch die Modellannahme, bei direkter Ausrechnung der AKF wäre sie bis zur Ordnung s fehlerbehaftet und über s hinaus unmöglich.

Nach diesen allgemeinen Betrachtungen kann man fragen, ob es nicht auf systematische Weise gelingt, im Falle endlicher Merkmalsfolgen geeignete Schätzgleichungen für die Modellparameter anzugeben. Wir werden diese Frage im Abschnitt *Burg-Algorithmus* wieder aufnehmen.

Wir geben nun ein Beispiel für die Berechnung der Modellkoeffizienten durch die Yule-Walker-Gleichung, und dabei auftretende Probleme.

Beispiel 8.2 – Yule-Walker-Gleichung für endliche Merkmalsfolgen.

Wie in Beispiel 8.1 sei ein System 2. Ordnung mit den Modellkoeffizienten $a_1 = 1/4, \quad a_2 = -1/2$ gegeben. Die Systemgleichung lautet

$$x[n] = q[n] - \frac{1}{4}x[n-1] + \frac{1}{2}x[n-2] \quad [\text{q(n)} = \text{Anregung}].$$

Vor dem Zeitpunkt $t = 0$ sind alle Grössen im System Null. Das System wird dann angeregt mit einer delta-Folge $q[n] = \delta[n]$. Es ergeben sich die ersten Folgenglieder auf 4 Nachkommastellen genau:
$x[0] = 1, x[1] = -0.25, x[2] = 0.5625, x[3] = -0.2656$
Hieraus schätzen wir jetzt die zugehörige konsistente Autokorrelationsfunktion mit N=4. Natürlich ergibt sich dabei durch das endliche N ein Fehler, wie oben diskutiert. Es gilt

$$\hat{s}_{XX}[|\kappa|] = \frac{1}{N} \sum_{k=0}^{N-1-|\kappa|} x[k]x[k+\kappa]$$

und damit auf 4 Nachkommastellen genau:

$$\begin{aligned}
\hat{s}_{XX}[0] &= \frac{1}{4}\,[1 + 0.125 + 0.3164 + 0.0705] = 0.3780 \\
\hat{s}_{XX}[\pm 1] &= \frac{1}{4}\,[-0.25 - 0.1406 - 0.1494] = -0.1350 \\
\hat{s}_{XX}[\pm 2] &= \frac{1}{4}\,[0.5625 + 0.0664] = 0.1572 \\
\hat{s}_{XX}[\pm 3] &= \frac{1}{4}\,[-0.2656] = -0.0664
\end{aligned} \tag{8.34}$$

Wir vergleichen mit den exakten Werten der AKF, die wir in Beispiel 8.1 in den Gln. (8.30) und (8.31) erhalten haben. Diese müssen wir wegen der endlichen Anzahl (N=4) der Messwerte hier mit $1/N$ normieren, um mit der konsistenten Schätzung vergleichen zu können. Wir haben also als exakte Werte
$s_{XX}[0] = 0.4444, \; s_{XX}[1] = -0.2222,$
$s_{XX}[2] = 0.2777, \; s_{XX}[3] = -0.1805$
Wir erkennen eine relativ gute Übereinstimmung zu den konsistenten AKF-Werten $\hat{s}_{XX}[\kappa]$, die mit zunehmendem κ wie erwartet schlechter wird. Aus den Werten $\hat{s}_{XX}[\kappa]$ schätzen wir nun die Koeffizienten für ein autoregressives Modell 2-ter Ordnung mit Hilfe der Yule-Walker-Gleichung:

$$\begin{pmatrix} 0.3780 & -0.1350 \\ -0.1350 & 0.3780 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 0.1350 \\ -0.1572 \end{pmatrix}$$

Dieses Gleichungssystem wird durch $a_1 = 0.2391, \; a_2 = -0.3305$ gelöst. Wir erkennen eine sehr gute Schätzung für a_1 und eine weniger

gute (aber mit richtigem Vorzeichen) für a_2. Die Abweichungen resultieren aus der geringen Anzahl von Messwerten, sie sind am höchsten für hohe Indices der Modellkoeffizienten. Mit steigender Anzahl von Messwerten werden die Abweichungen geringer. □

8.5 Lineare Prädiktion und Wiener-Hopf-Gleichung

Wir hatten bereits einleitend bemerkt, dass $x[k]$ als Eingang eines Analysefilters benutzt wird. Wenn dieses Analysefilter exakt den Modellprozess invertiert, erhalten wir am Ausgang des Analysefilters wieder das Rauschen $q[k]$. Im allgemeinen Fall, d.h. falls der Modellprozess nicht exakt invertiert wird, erhalten wir also ein stochastisches Fehlersignal $e[k]$.

Dieses Analysefilter mit der Gesamtübertragungsfunktion $P_e(z)$ wird wie folgt implementiert: Die Vorschaltung einer Verzögerung z^{-1} vor ein *Prädiktionsfilter* $P(z)$ hat den Grund, dass die abgeleiteten Gleichungen für die Koeffizienten von $P(z)$ besonders einfach werden. Die Abb. 8.5 zeigt den prinzipiellen Aufbau eines Prädiktionsfehlerfilters.

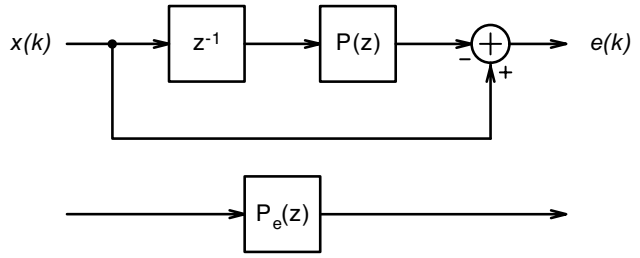


Abbildung 8.5. Signalflussplan des Prädiktionsfehlerfilters

Das Eingangssignal $x[k]$ läuft durch zwei parallele Zweige. Der erste Zweig enthält ein Verzögerungsglied und das Prädiktionsfilter $P(z)$. Das Prädiktionsfilter liefert aus den vergangenen Eingangswerten einen Schätzwert für den neuen Eingangswert. Da es ein autoregressives Modell invertieren soll, wird es als nichtrekursives, lineares System der Ordnung m realisiert. Es hat daher eine Übertragungsfunktion der folgenden Form (vgl. MA-Modell, Gl. 8.13):

$$P(z) = \sum_{\mu=1}^{m} p_\mu z^{-\mu+1}. \tag{8.35}$$

Der neue Eingangswert läuft ohne Verzögerung über den zweiten Zweig und wird mit seinem Schätzwert verglichen. Am Ausgang erhält man daraus den *Prädiktionsfehler* $e[k]$. Damit ergibt sich für das Gesamtsystem die Differenzengleichung

$$e[k] = x[k] - \sum_{\mu=1}^{m} p_\mu x[k-\mu] \tag{8.36}$$

sowie die Übertragungsfunktion

$$P_e(z) = 1 - z^{-1}P(z) = 1 - \sum_{\mu=1}^{m} p_\mu z^{-\mu} \tag{8.37}$$

Das Prädiktionsfilter soll nun dahingehend optimiert werden, dass die Leistung des Prädiktionsfehlers $e[k]$ minimal wird. Der Prädiktionsfehler $e[k]$ ist als Musterfolge eines Zufallsprozesses aufzufassen. Für den Zufallsprozesses soll hier auch ein kleiner Buchstabe verwendet werden, um Verwechslungen mit dem Erwartungswert E zu vermeiden. Das Optimierungskriterium lautet daher:

$$E\left\{e^2(k)\right\} \rightarrow \min \tag{8.38}$$

Setzen wir voraus, dass die Eingangsfolge $x[k]$ eine Musterfolge aus einem autoregressiven Prozess ist, und daher durch das AR-Modell exakt beschrieben werden kann, so ist das obige Kriterium genau dann erfüllt, wenn am Ausgang wiederum ein weißer Prozess entsteht. Man nennt daher das Prädiktionsfehlerfilter in diesem Fall auch *pre-whitening-Filter*. Der weiße Ausgangsprozess wird erreicht, indem das *pre-whitening-Filter* die Auswirkungen des autoregressiven Filters gerade kompensiert, also eine inverse Übertragungsfunktion hat. Daraus können die Koeffizienten p_i des Prädiktionsfilters berechnet werden:

$$\begin{aligned} P_e(z) &= \frac{1}{A(z)} \\ 1 - \sum_{\mu=1}^{m} p_\mu z^{-\mu} &= 1 + \sum_{\nu=1}^{n} a_\nu z^{-\nu}. \end{aligned} \tag{8.39}$$

Man erkennt, dass das Prädiktionsfilter und das AR-Modell von der selben Ordnung n sein müssen. Außerdem ist durch Koeffizientenvergleich ersichtlich, dass gilt:

$$p_\nu = -a_\nu. \tag{8.40}$$

Verwendet man die Matrixschreibweise und die Yule-Walker-Gleichung (Theorem 8.1 auf Seite 214), so erhält man:

Theorem 8.2 (Wiener-Hopf-Gleichung).

$$\begin{aligned} \mathbf{p} &= -\mathbf{a} \\ \mathbf{p} &= \mathbf{S_{XX}^{-1}} \mathbf{s_{XX}}. \end{aligned} \tag{8.41}$$

Diesen Zusammenhang bezeichnet man als ***Wiener-Hopf-Gleichung****.*

Ist der Eingangsprozess nicht autoregressiv, wie wir bisher angenommen haben, so erhält man mit den **Wiener-Hopf-Koeffizienten** keinen weißen Prozess am Ausgang. Man kann jedoch zeigen, dass die Leistung des Ausgangsprozesses trotzdem minimal bleibt. Damit ist das Optimierungskriterium (minimaler Fehler des Ausganges) auch dann erfüllt, wenn der Eingangsprozess nicht a-priori als autoregressiv angenommen wird.

Zusammenfassend gilt folgendes: Wir haben mit der Wiener-Hopf-Gleichung in jedem Fall eine optimale Berechnungsvorschrift für die Koeffizienten eines nichtrekursiven Systems der Ordnung m gefunden. Der Fehler am Ausgang dieses Systems hat minimale Leistung bei der Modellierung eines beliebigen Prozesses. Im Spezialfall, dass der Eingangsprozess gerade ein AR-Prozess der Ordnung m ist, können die Parameter des Eingangsprozesses exakt rekonstruiert werden.

Beispiel 8.3 – Wiener-Hopf-Gleichung.

Wegen der grossen formalen Ähnlichkeit zur Yule-Walker-Gleichung greifen wir nochmals zurück auf das Besipiel 8.2. Gegeben sei ein AR-System 2. Ordnung mit den Modellkoeffizienten $a_1 = 1/4$, $a_2 = -1/2$, und es werden nach Anregung durch eine Delta-Folge die ersten 4 Folgenglieder gemessen.
Wegen der Gl. (8.41)

$$\mathbf{p} = -\mathbf{a}$$

gilt für ein Prädiktorsystem 2. Ordnung $p_1 = -0.2391$, $p_2 = 0.3305$.
Es gelten dieselben Argumente hinsichtlich der Abweichungen von den theoretisch richtigen Werten, die schon in Besipiel 8.2 genannt wurden. Wir widmen uns der Frage, was eine falsche Prädiktorordnung für Auswirkungen hat. Wird ein Prädiktionssystem der Ordnung 1 angesetzt, erhält man mit der Wiener-Hopf-Gleichung

$$0.3780\, p_1 = -0.1350$$

und damit $p_1 = -0.3571$. Dies ist immer noch ähnlich dem theoretischen Wert von -0.25, aber natürlich fehlt der Einfluss des 2. Verzögerungsgliedes. □

8.6 Orthogonalität des Prädiktionsfehlerfilters

Aus der Minimierung des Fehlerprozesses $e(k)$ ergibt sich eine weitere, sehr nützliche Eigenschaft:

Theorem 8.3 (Orthogonalität des Prädiktionsfehlerfilters).
Der Ausgangsprozess eines Prädiktionsfehlerfilters ist orthogonal zu allen Vergangenheitswerten des Eingangsprozesses $X(k+\kappa)$. Prädiktionsfehlerfilter werden daher auch als „novelty detector" bezeichnet.

Der Begriff der Orthogonalität wurde bereits im Abschnitt 7.8 eingeführt. Das Filter ist nützlich, da es die „neuen" Komponenten des Eingangsprozesses identifiziert. Mit „neu" ist dabei „orthogonal zu allen vorherigen Werten" gemeint. Zum Nachweis der Orthogonalität geht man von der Definitionsgleichung der Kreuzkorrelation aus:

$$s_{eX}[\kappa] = E\{e(k)X(k-\kappa)\}. \tag{8.42}$$

Durch Einsetzen der Differenzengleichung des Prädiktionsfehlerfilters erhält man:

$$\begin{aligned} s_{eX}[\kappa] &= E\left\{\left(X(k) - \sum_{\mu=1}^{m} p_\mu X(k-\mu)\right) X(k-\kappa)\right\} \\ &= E\{X(k)X(k-\kappa)\} - \sum_{\mu=1}^{m} p_\mu E\{X(k-\mu)X(k-\kappa)\} \\ &= s_{XX}[\kappa] - \sum_{\mu=1}^{m} p_\mu s_{XX}[\kappa-\mu] \end{aligned} \tag{8.43}$$

Wie wir bei der Herleitung der Yule-Walker-Gleichung gezeigt haben, gilt für die Summe:

$$\sum_{\mu=1}^{m} p_\mu s_{XX}[\kappa-\mu] = -\sum_{\mu=1}^{m} a_\mu s_{XX}[\kappa-\mu] = s_{XX}[\kappa] \quad \text{für} \quad \kappa > 0 \tag{8.44}$$

Damit ergibt sich für die Kreuzkorrelation:

$$s_{eX}[\kappa] = E\{e(k)X(k-\kappa)\} = s_{XX}[\kappa] - s_{XX}[\kappa] = 0 \quad \text{für} \quad \kappa > 0 \tag{8.45}$$

Offensichtlich haben wir damit die Orthogonalität gezeigt. □

Beispiel 8.4 – Orthogonalität bei Prädiktionssystemen.

Wir führen das Beispiel 8.2 weiter. Gegeben sei ein Prädiktor 2. Ordnung mit den Modellkoeffizienten $p_1 = -1/4, \quad p_2 = 1/2$. Für die Eingangswerte aus dem AR-System des Beispiels 8.2 galt:
$x[n<0] = 0, x[0] = 1, x[1] = -0.25, x[2] = 0.5625, x[3] = -0.2656$
Für den Prädiktionsfehler gilt nach Gl. (8.36)

$$e[k] = x[k] - \sum_{\mu=1}^{2} p_\mu x[k-\mu] = x[k] + \frac{1}{4}x[k-1] - \frac{1}{2}x[k-2]$$

Daraus erhalten wir die Fehlerwerte:
$e[n<0] = 0, \quad e[0] = 1,$
$e[1] = -0.25 + 1/4 = 0,$
$e[2] = 0.5625 + (-0.25)/4 - 1/2 = 0,$
$e[3] = -0.2656 + 0.5625/4 - (-0.25)/2 = 0.$

Das Fehlersignal entspricht also genau dem ursprünglichen anregenden Signal des Eingangs-AR-Systems (Delta-Folge).
Wir untersuchen die Orthogonalität zwischen diesem Fehlersignal und dem Eingangssignal $x[k]$. Nach Gl. (8.42) ist für $\kappa > 0$ zu untersuchen:

$$s_{eX}[\kappa] = E\{e(k)X(k-\kappa)\}.$$

Wir erhalten in der Tat

$$s_{eX}[\kappa] = E\{e(k)X(k-\kappa)\} = 0 \quad (\kappa > 0),$$

was die Orthogonalität des Ausgangsfehlersignals zu allen vorherigen Werten des Eingangssignals zeigt. □

Orthogonalität gilt im Allgemeinen bei einem Prädiktorsystem der Ordnung n nur solange, wie nicht mehr als n Werte in der Eingangsfolge vorliegen. Diese Einschränkung ist einleuchtend, da ein lineares System n-ter Ordnung nur n Orthogonalitätsbedingungen exakt erfüllen kann. Im Falle einer längeren Eingangsfolge können wir dieses Ergebnis für die Kreuzkorrelation zur so genannten *Lückenfunktion* oder *gap-function* verallgemeinern:

$$\begin{aligned} g_n[\kappa] &= E\{e(k)X(k-\kappa)\} \\ g_n[\kappa] &= 0 \quad 1 \le \kappa \le n \end{aligned} \tag{8.46}$$

Dabei entspricht n der Ordnung des Prädiktionssystems. Außerhalb des Intervalls $1...n$ wird die Lückenfunktion im Allgemeinen von Null verschiedene Werte annehmen, wie Abb. 8.6 zeigt. Die Lückenfunktion ist von zentraler Be-

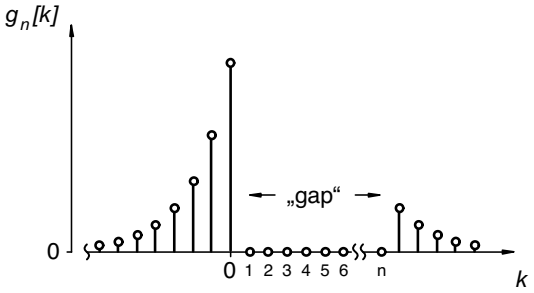


Abbildung 8.6. Lückenfunktion ("gap-function")

deutung bei der Herleitung des **Levinson-Durbin-Algorithmus** im nächsten Abschnitt. Dort wird, ausgehend von einer Lückenfunktion für ein System der Ordnung n, eine Lückenfunktion für ein System der Ordnung $n+1$ konstruiert: das System mit dieser um 1 höheren Ordnung kann offensichtlich eine weitere Bedingung erfüllen, da nun $n+1$ lineare Gleichungen für $n+1$ Unbekannte zur Verfügung stehen.

8.7 Levinson-Durbin-Rekursion

In den vorangegangenen Abschnitten haben wir gezeigt, dass man einen Zufallsprozess mit Hilfe von Modellsystemen schätzen kann. Wir haben den Zusammenhang zwischen der Autokorrelationsfolge des Prozesses und den Parametern des Modells hergeleitet, für das AR-Modell also die Yule-Walker-Gleichung und für die lineare Prädiktion die Wiener-Hopf-Gleichung. Beide Gleichungssysteme haben eine fast identische Struktur. In beiden Fällen muss zur Lösung eine Matrixinversion einer Autokorrelationsmatrix vorgenommen werden, was mit einem hohen Rechenaufwand verbunden ist. Zudem haben die resultierenden Modellparameter von Systemen niedrigerer Ordnung nichts mit Parametern von Systemen höherer Ordnung zu tun, wie das Beispiel 8.3 zur Wiener-Hopf-Gleichung zeigte: für jede Ordnung des Modellsystems müssen alle Parameter immer wieder unabhängig voneinander neu berechnet werden.

Einen besseren Ansatz zur Berechnung der Modellparameter bietet der **Levinson-Durbin-Algorithmus** (oder Levinson-Durbin-Rekursion). Dieser geht nicht den Weg über die Matrixinversion, sondern berechnet das Modell $(r+1)$-ter Ordnung rekursiv aus dem Modell r-ter Ordnung. Dies führt zu einem wesentlich effizienteren Berechnungsverfahren.

Natürlich beschreiben beide Verfahren denselben Prozess und berechnen somit auch dieselben Modellparameter. Das schwierige Problem der Matrixinversion wird zudem von dem in der Anwendung relativ einfachen Levinson-Durbin-Algorithmus gelöst. Dass es sich dabei nicht um ein Mysterium handelt, welches das Problem „vereinfacht“ , hat einen mathematischen Grund:

Wir betrachten noch einmal die zu invertierende Matrix $\mathbf{S}$:

$$\mathbf{S} = \begin{pmatrix} s_{XX}[0] & s_{XX}[-1] & \cdots & s_{XX}[-(n-1)] \\ s_{XX}[1] & s_{XX}[0] & \cdots & {}_{XX}[-(n-2)] \\ \vdots & \vdots & \ddots & \vdots \\ s_{XX}[n-1] & s_{XX}[n-2] & \cdots & s_{XX}[0] \end{pmatrix} \tag{8.47}$$

Für reellwertige Signale ist darüber hinaus bekanntlich die AKF symmetrisch, so dass $s_{XX}[k] = s_{XX}[-k]$ gilt. In diesem Fall enthält die Matrix nicht, wie im allgemeinen Fall n^2, sondern nur n unterschiedliche Größen $s_{XX}[k]$. Diese (in n lineare) Anzahl von Unbekannten ermöglicht allgemeine Invertierungsverfahren, die mit n einfachen Schritten auskommen. Solche Invertierungsverfahren müssen speziell für den jeweiligen Matrixtyp formuliert werden. In unserem Fall handelt es sich um eine Matrix von so genannter *Toeplitz-Struktur*. Sie ist für reellwertige Signale[1] derart aufgebaut, dass jede Zeile aus der vorherigen durch Verschieben aller Werte um eine Position

[1] Beschränken wir uns nicht auf reellwertige Signale, so können wir benutzen, dass für ergodische, stationäre Prozesse $s^*_{XX}[k] = s_{XX}[-k]$ gilt. Wegen der konjugierten Komplexheit ist damit die Korrelationsmatrix allgemein von *hermitescher Toeplitzform*. Die Aussagen über Toeplitzmatrizen gelten dann äquivalent.

nach rechts entsteht. Der ganz rechts „herausfallende“ Wert wird links wieder „hineingeschoben“ , so dass eine zyklische Verschiebung der Zeilenwerte entsteht.

Der Levinson-Durbin-Algorithmus ist also gerade ein besonderes Verfahren, welches Toeplitz-Matrizen invertiert. Und da Toeplitz-Matrizen eine spezielle, einfache zyklische Struktur mit insgesamt nur n Unbekannten haben, lässt sich die Inversion mit geringem Aufwand durchführen.

An dieser Stelle können wir auch auf die einschlägige mathematische Literatur verweisen. Wir erhalten aber einen sehr guten Einblick in die Struktur der Modellsysteme, wenn wir den Levinson-Durbin-Algorithmus herleiten und dabei sehen, wie das Modell $(r+1)$-ter Ordnung rekursiv aus dem Modell r-ter Ordnung berechnet wird.

Bei der Herleitung der Levinson-Durbin-Rekursion gehen wir von den Gleichungen aus, die wir im vorigen Abschnitt für die lineare Prädiktion gefunden haben. Wir wiederholen zunächst die Übertragungsfunktion, die Differenzengleichung und die Lückenfunktion. Für ein System r-ter Ordnung lauten diese wie folgt:

Übertragungsfunktion:

$$A_r(z) = 1 + \sum_{\nu=1}^{r} a_\nu^r z^{-\nu} = \sum_{\nu=0}^{r} a_\nu^r z^{-\nu} \quad \text{mit} \quad a_0^r = 1 \tag{8.48}$$

Differenzengleichung:

$$e[k] = \sum_{\nu=0}^{r} a_\nu^r x[k-\nu] \quad \text{mit} \quad a_0^r = 1 \tag{8.49}$$

Lückenfunktion:

$$\begin{aligned} g_r[\kappa] &= E\left\{e_r(k)X(k-\kappa)\right\} \\ g_r[\kappa] &= 0 \quad 1 \le \kappa \le r \end{aligned} \tag{8.50}$$

Das hochgestellte r bei den Modellparametern ${a_\nu}^r$ steht dabei nicht für eine Potenz, sondern es soll auf die Zugehörigkeit des Parameters zum Modell r-ter Ordnung hinweisen.

Aus dem Modell r-ter Ordnung soll nun das Modell $(r+1)$-ter Ordnung bestimmt werden. Dazu betrachten wir zunächst die Lückenfunktion. Für ein System r-ter Ordnung verschwindet die Lückenfunktion g_r im Bereich von 1 bis r. Die Lückenfunktion g_{r+1} muss zusätzlich an der Stelle $r+1$ gleich null sein. Um dies zu erreichen wird g_r zunächst an der y-Achse gespiegelt und dann um $r+1$ nach rechts verschoben, wie Abb. 8.7 zeigt.

Die so erhaltene Funktion

$$g_r^*[\kappa] = g_r[r+1-\kappa] \tag{8.51}$$

ist ebenfalls im Bereich von 1 bis r gleich null. Für die Stelle $r+1$ der ursprünglichen Funktion gilt jedoch:

$$g_r^*[r+1] = g_r[0]. \tag{8.52}$$

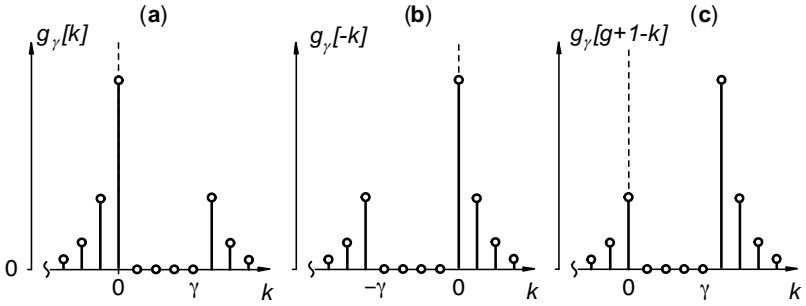


Abbildung 8.7. Modifizierte Lückenfunktion

Man kann nun beliebige Vielfache von g_r und g_r^* miteinander addieren, ohne dass sich eine Veränderung im Bereich $1 \ldots r$ ergibt. Wählt man den Faktor für die Folge g_r^* so, dass sich beide Folgen an der Stelle $r+1$ zu null addieren, so erhält man die Lückenfunktion der Ordnung $r+1$:

$$\begin{aligned} g_{r+1}[\kappa] &= g_r[\kappa] - \gamma_{r+1} g_r^*[\kappa] \\ &= g_r[\kappa] - \gamma_{r+1} g_r[r+1-\kappa]. \end{aligned} \tag{8.53}$$

Den Faktor γ_{r+1} bezeichnet man als **PARCOR-Koeffizienten**, wobei PARCOR abkürzend für *PARtial CORrelation* steht. Der Sinn dieser Bezeichnung wird gleich deutlich werden. Der PARCOR-Koeffizient γ_{r+1} wird so gewählt, dass gilt:

$$g_{r+1}[r+1] = g_r[r+1] - \gamma_{r+1} g_r[0] = 0. \tag{8.54}$$

Daher ist

$$\gamma_{r+1} = \frac{g_r[r+1]}{g_r[0]}. \tag{8.55}$$

Setzt man die Definitionsgleichung der Lückenfunktion ein, so erhält man

$$\gamma_{r+1} = \frac{E\{e_r(k) X(k-(r+1))\}}{g_r[0]} \tag{8.56}$$

Damit erschließt sich der Sinn der Bezeichnung *Partial Correlation*: γ_{r+1} ist die auf $g_r[0]$ normierte partielle Korrelation zwischen dem Prädiktorfehler und der $(r+1)$-fach verzögerten Prädiktor-Eingangsgröße.

Für unsere Zwecke benötigen wir nun die Abhängigkeit des PARCOR-Koeffizienten von den Parametern des Systems r-ter Ordnung. Dazu setzen wir die Differenzengleichung des Systems in die Definitionsgleichung der Lückenfunktion ein:

$$\begin{aligned} g_r[\kappa] &= E\left\{e_r(k)X(k-\kappa)\right\} = E\left\{\left(\sum_{\nu=0}^{r} a_\nu^r X(k-\nu)\right) X(k-\kappa)\right\} \\ &= \sum_{\nu=0}^{r} a_\nu^r E\left\{\left(X(k-\nu)\right) X(k-\kappa)\right\} \\ &= \sum_{\nu=0}^{r} a_\nu^r s_{XX}[\kappa-\nu]. \end{aligned} \tag{8.57}$$

Diese Gleichung setzen wir wiederum in Gl. (8.55) für den PARCOR-Koeffizienten ein:

$$\begin{aligned} \gamma_{r+1} = \frac{g_r[r+1]}{g_r[0]} &= \frac{\sum_{\nu=0}^{r} a_\nu^r s_{XX}[r+1-\nu]}{\sum_{\nu=0}^{r} a_\nu^r s_{XX}[-\nu]} \\ &= \frac{\sum_{\nu=0}^{r} a_\nu^r s_{XX}[r+1-\nu]}{\sum_{\nu=0}^{r} a_\nu^r s_{XX}[\nu]} \end{aligned} \tag{8.58}$$

Wir haben also folgendes erreicht: Wir können aus den Parametern des Modells r-ter Ordnung den PARCOR-Koeffizienten γ_{r+1} und damit auch die Lückenfunktion g_{r+1} bestimmen. Um nun die Modellparameter des Systems $(r+1)$-ter Ordnung zu finden, gehen wir von der rekursiven Formel für g_{r+1} aus und nutzen wie schon beim PARCOR-Koeffizienten die Abhängigkeit zwischen Lückenfunktion und Modellparametern:

$$\begin{aligned} g_{r+1}[\kappa] &= g_r[\kappa] - \gamma_{r+1} g_r[r+1-\kappa] \\ \sum_{\nu=0}^{r+1} a_\nu^{r+1} s_{XX}[\kappa-\nu] &= \sum_{\nu=0}^{r} a_\nu^r s_{XX}[\kappa-\nu] \\ &\quad - \gamma_{r+1} \sum_{\nu=0}^{r} a_\nu^r s_{XX}[r+1-\kappa-\nu]. \end{aligned} \tag{8.59}$$

Wir wollen nun die rechte Summe so umschreiben, dass alle Terme zum Zeitpunkt $\kappa-\nu$ erscheinen - dann können wir alle Summen termweise vergleichen. Die gewünschte Umschreibung erreichen wir mit den Substitutionen in der rechten Summe: $\nu \rightarrow \mu = r+1-\nu$ (Gl. 8.60), und dann der erneuten Notation $\mu \rightarrow \nu = \mu$. Es folgt die Ausnutzung von $s[n] = s[-n]$ für reelle Signale und die korrekte Angabe der Reihenfolge der Summationsgrenzen (Gl. 8.61):

$$\sum_{\nu=0}^{r+1} a_\nu^{r+1} s_{XX}[\kappa-\nu] = \sum_{\nu=0}^{r} a_\nu^r s_{XX}[\kappa-\nu] - \gamma_{r+1} \sum_{\mu=r+1}^{1} a_{r+1-\mu}^r s_{XX}[\mu-\kappa] \tag{8.60}$$

$$\sum_{\nu=0}^{r+1} a_\nu^{r+1} s_{XX}[\kappa-\nu] = \sum_{\nu=0}^{r} a_\nu^r s_{XX}[\kappa-\nu] - \gamma_{r+1} \sum_{\nu=1}^{r+1} a_{r+1-\nu}^r s_{XX}[\kappa-\nu]. \tag{8.61}$$

Durch den Vergleich der Koeffizienten der AKF-Glieder an der selben Position erhält man:

$$a_0^{r+1} = a_0^r = 1 \quad \text{für} \quad \nu = 0 \tag{8.62}$$

$$a_\nu^{r+1} = a_\nu^r - \gamma_{r+1} a_{r+1-\nu}^r \quad \text{für} \quad \mu = \nu, 1 \le \mu, \nu \le r \tag{8.63}$$

$$a_{r+1}^{r+1} = \gamma_{r+1} a_0^r = \gamma_{r+1} \quad \text{für} \quad \mu = \nu = r+1. \tag{8.64}$$

Diese Ergebnisse lassen sich übersichtlich in Vektorform darstellen:

$$\begin{pmatrix} a_0^{r+1} \\ a_1^{r+1} \\ a_2^{r+1} \\ \vdots \\ a_r^{r+1} \\ a_{r+1}^{r+1} \end{pmatrix} = \begin{pmatrix} 1 \\ a_1^r \\ a_2^r \\ \vdots \\ a_r^r \\ 0 \end{pmatrix} - \gamma_{r+1} \begin{pmatrix} 0 \\ a_r^r \\ a_{r-1}^r \\ \vdots \\ a_1^r \\ 1 \end{pmatrix} \tag{8.65}$$

Die Ausgangsleistung nach der r-ter Stufe des Prädiktionssystems können wir durch Anwendung der Wiener-Hopf-Gleichung zu

$$\sigma_E^r = \sigma + \sum_{\nu=1}^{r} a_\nu^r s_{XX}[\nu] = s_{XX}[0] + \sum_{\nu=1}^{r} a_\nu^r s_{XX}[\nu] = \sum_{\nu=0}^{r} a_\nu^r s_{XX}[\nu] \tag{8.66}$$

angeben. Genau dieselbe Berechnungsvorschrift haben wir auch für $g_r[0]$ erhalten. Daher:

$$g_r(0) = \sigma_E^r \tag{8.67}$$

Die Levinson-Durbin-Rekursion liefert also ohne weiteren Aufwand auch die Leistung am Ausgang jeder Stufe.

Somit kann man für die Levinson-Durbin-Rekursion die folgende Schrittfolge angeben:

Theorem 8.4 (Levinson-Durbin-Rekursion).

- *Initialisierung* $(r = 0)$*:*

$$a_0^0 = 1 \tag{8.68}$$

$$\sigma_E^0 = s_{XX}[0] \tag{8.69}$$

- *1. Iterationsschritt* $(r = 1)$*:*

$$\gamma_1 = \frac{s_{XX}[1]}{s_{XX}[0]} \tag{8.70}$$

$$\begin{pmatrix} a_0^1 \\ a_1^1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \gamma_1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{8.71}$$

$$\sigma_E^1 = s_{XX}[0] + a_1^1 s_{XX}[1] = (1 - |\gamma_1|^2)\sigma_E^0 \tag{8.72}$$

- *r-ter Iterationsschritt* $(r = 2...n)$*:*

$$\gamma_r = \frac{\sum_{\nu=0}^{r-1} a_\nu^{r-1} s_{XX}[r-\nu]}{\sum_{\nu=0}^{r-1} a_\nu^{r-1} s_{XX}[\nu]} \tag{8.73}$$

$$\begin{pmatrix} a_0^r \\ a_1^r \\ \vdots \\ a_{r-1}^r \\ a_r^r \end{pmatrix} = \begin{pmatrix} 1 \\ a_1^{r-1} \\ \vdots \\ a_{r-1}^{r-1} \\ 0 \end{pmatrix} - \gamma_r \begin{pmatrix} 0 \\ a_{r-1}^{r-1} \\ \vdots \\ a_1^{r-1} \\ 1 \end{pmatrix} \tag{8.74}$$

$$\sigma_E^r = \sum_{\nu=0}^{r} a_\nu^r s_{XX}[\nu] = (1 - |\gamma_r|^2)\sigma_E^{r-1} \tag{8.75}$$

Da $1 - |\gamma|^2 < 1$*, sinkt die Fehlerleistung mit jeder weiteren Stufe!*

- *Resultat: Prädiktor bzw. AR-Modell n-ter Ordnung*

$$A_n(z) = \sum_{\nu=0}^{n} a_\nu^n z^{-\nu} \tag{8.76}$$

Die Levinson-Durbin-Rekursion bietet uns eine Möglichkeit, die Parameter eines AR-Modells n-ter Ordnung rekursiv zu bestimmen. Das erfordert einen wesentlich geringeren Aufwand als die Berechnung mittels Matrixinversion.

Beispiel 8.5 – Levinson-Durbin-Rekursion.

Gegeben sei wie in Beispiel 8.1 ein System 2. Ordnung mit den Modellkoeffizienten $a_1 = 1/4, \quad a_2 = -1/2$. Die Systemgleichung lautet damit

$$x[n] = q[n] - \frac{1}{4}x[n-1] + \frac{1}{2}x[n-2] \quad \text{[q(n) = Anregung]}.$$

Vor dem Zeitpunkt t=0 sind alle Grössen im System Null. Das System wird dann angeregt mit einer delta-Folge $q[n] = \delta[n]$.
Wir hatten in Beispiel 8.1 die ersten 4 exakten Werte der Autokorrelationsfolge $s_{XX}(\kappa)$ bestimmt, sie lauteten gemäss Gln. (8.30) und (8.31):

$$s_{XX}[0] = 16/9 = 1.777, \quad s_{XX}[1] = -8/9 = -0.888,$$
$$s_{XX}[2] = 10/9 = 1.111, \quad s_{XX}[3] = -13/18 = -0.7222$$

Diese Werte benutzen wir nun zum Durchführen einer Levinson-Durbin-Rekursion:

- Initialisierung ($r = 0$):

$$a_0^0 = 1$$
$$\sigma_E^0 = s_{XX}[0] = 16/9$$

- 1. Iterationsschritt ($r = 1$):

$$\gamma_1 = \frac{s_{XX}[1]}{s_{XX}[0]} = -1/2$$
$$\begin{pmatrix} a_0^1 \\ a_1^1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \gamma_1 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \end{pmatrix}$$
$$\sigma_E^1 = s_{XX}[0] + a_1^1 s_{XX}[1]$$
$$= 16/9 + 1/2 \cdot (-8/9) = 4/3$$
$$\text{bzw. } \sigma_E^1 = (1 - |\gamma_1|^2)\sigma_E^0 = (1 - 1/4) \cdot 16/9 = 4/3$$

- 2. Iterationsschritt:

$$\gamma_2 = \frac{\sum_{\nu=0}^{1} a_\nu^1 s_{XX}[r-\nu]}{\sum_{\nu=0}^{1} a_\nu^1 s_{XX}[\nu]}$$
$$= \frac{a_0^1 s_{XX}[2] + a_1^1 s_{XX}[1]}{a_0^1 s_{XX}[0] + a_1^1 s_{XX}[1]} = \frac{1 \cdot 10/9 + 1/2 \cdot (-8/9)}{1 \cdot 16/9 + 1/2 \cdot (-8/9)} = 1/2$$
$$\begin{pmatrix} a_0^2 \\ a_1^2 \\ a_2^2 \end{pmatrix} = \begin{pmatrix} 1 \\ a_1^1 \\ 0 \end{pmatrix} - \gamma_2 \begin{pmatrix} 0 \\ a_1^1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/2 \\ 0 \end{pmatrix} - 1/2 \begin{pmatrix} 0 \\ 1/2 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1/4 \\ -1/2 \end{pmatrix}$$
$$\sigma_E^2 = s_{XX}[0] + a_1^2 s_{XX}[1] + a_1^2 s_{XX}[2]$$
$$= 16/9 + 1/4 \cdot (-8/9) - 1/2(10/9) = 1$$
$$\text{bzw. } \sigma_E^2 = (1 - |\gamma_2|^2)\sigma_E^1 = (1 - 1/4) \cdot (4/3) = 1$$

Das Resultat zeigt, dass bei Verwenden der exakten Werte der Autokorrelationsfunktion sowohl die Modellparameter wie auch die Leistung des Eingangssignals mit Hilfe der Levinson-Durbin-Rekursion korrekt berechnet werden.
Zum Berechnen einer Levinson-Durbin-Rekursion mit gemessenen (fehlerhaften) Werten der AKF siehe Übung 8.5. □

Nachteilig bei der Levinson-Durbin-Rekursion ist allerdings, dass für jede Ordnung des Modells alle Parameter neu berechnet werden müssen. Es wäre wünschenswert, wenn die Parameter aus Modellen geringerer Ordnung beibehalten werden könnten und in jedem Iterationsschritt nur der jeweils neue Parameter berechnet werden müsste. Mit dieser Problematik befasst sich der folgende Abschnitt zu den Lattice-Strukturen.

8.8 Modellsysteme in Lattice-Struktur

Im vorigen Kapitel haben wir uns mit der Modellierung von Rauschsignalen beschäftigt. Wir haben den Zusammenhang zwischen Autokorrelationsfolge und den Modellparametern hergeleitet, die Yule-Walker-Gleichung für autoregressive Modelle und die Wiener-Hopf-Gleichung für lineare Prädiktionssysteme. Schließlich haben wir mit der Levinson-Durbin-Rekursion einen Algorithmus gefunden, der die Modellparameter mit Hilfe der PARCOR-Koeffizienten rekursiv ermittelt. Allerdings hat der Levinson-Durbin-Algorithmus den Nachteil, dass für jede Ordnung r alle Modellparameter a_i^r neu berechnet werden müssen.

Diesen Aufwand würde man gerne sparen, zumal wir gesehen haben, dass ein Modellsystem r-ter Ordnung durch seine r PARCOR-Koeffizienten bereits eindeutig bestimmt ist. Gesucht wird daher eine Gleichung, die das Ausgangssignal nicht in Abhängigkeit von den Modellparametern, sondern von den PARCOR-Koeffizienten darstellt. Dieses Verfahren soll zudem im „Baukastenprinzip“ funktionieren: jedes Verfahren r-ter Ordnung wird nicht modifiziert. Für ein Verfahren der Ordnung $(r+1)$ wird vielmehr eine Stufe mit identischer *Struktur* an die vorhergehenden Stufen angehängt. Da jede Stufe - wie wir sehen werden - zwei parallele, miteinander verschaltete Zweige enthält, sieht die Aneinanderreihung solcher Stufen aus wie ein Gitter: daher der Name *Lattice-Struktur* für diese Anordnung.

Mit jeder zusätzlichen Stufe soll sich der Prädiktorfehler verringern. Wir wollen jetzt die Struktur der einzelnen Stufen, und die optimalen Parameter solch einer Lattice-Struktur herleiten.

8.8.1 Ableitung der Analysegleichungen

Für die Herleitung der Lattice-Struktur und seiner Parameter betrachten wir zunächst wieder einen reellen Prozess, der durch ein lineares Prädiktionssystem übertragen wird. Wir werden zur Strukturbestimmung zunächst zeigen, dass sich ein Prädiktionsfilter interpretieren lässt als ein System, welches in einem Zweig einen Systemwert zeitlich vorwärts und im parallelen Zweig einen anderen Systemwert zeitlich rückwärts transportiert. Beide Systemwerte werden wir mathematisch identifizieren und benennen. Dabei werden wir ihre Verknüpfungen aufzeigen und daraus die Parameter einer jeden Lattice-Stufe angeben können.

Wir gehen von einem Prädiktionssystem der Ordnung $(r+1)$ aus. Sein Prädiktionsfehler genügt folgender Differenzengleichung:

$$e_{r+1}[k] = x[k] + \sum_{\nu=1}^{r+1} a_\nu^{r+1} x[k-\nu] \tag{8.77}$$

Wir erhalten damit den Prädiktionsfehler der Ordnung $(r+1)$ aus den Parametern des Systems $(r+1)$-ter Ordnung. Wir ersetzen nun die Parameter ${a_\nu}^{r+1}$ durch ihre rekursive Berechnungsvorschrift entsprechend des Levinson-Durbin-Algorithmus und erhalten so eine Abhängigkeit von den Systemparametern der Ordnung r:

$$\begin{aligned} e_{r+1}[k] &= x[k] + \sum_{\nu=1}^{r} \left(\left(a_\nu^r - \gamma_{r+1} a_{r+1-\nu}^r \right) x[k-\nu] \right) + \gamma_{r+1} x[k-(r+1)] \\ &= \left(x[k] + \sum_{\nu=1}^{r} a_\nu^r x[k-\nu] \right) \\ &\quad - \gamma_{r+1} \left(x[k-(r+1)] + \sum_{\nu=1}^{r} a_{r+1-\nu}^r x[k-\nu] \right) \end{aligned} \tag{8.78}$$

Wie man unschwer erkennen kann handelt es sich beim ersten Klammerausdruck um den Prädiktionsfehler r-ter Ordnung:

$$e_r[k] = x[k] + \sum_{\nu=1}^{r} a_\nu^r x[k-\nu]. \tag{8.79}$$

Der zweite Klammerausdruck ist dem ersten in seiner Struktur sehr ähnlich. Es erweist sich als sinnvoll, eine neue Größe zu seiner Beschreibung einzuführen. Wir definieren dazu den sogenannten *Rückwärts-Prädiktionsfehler* wie folgt:

$$b_r[k] = x[k-r] + \sum_{\mu=1}^{r} a_\mu^r x[k-r+\mu]. \tag{8.80}$$

Mit Hilfe der Substitution $\nu \to \mu = r+1-\nu$ können wir nun den zweiten Klammerausdruck in Gl. (8.78) in die Form eines Rückwärts-Prädiktionsfehler bringen:

$$\begin{aligned} x[k-(r+1)] + \sum_{\nu=1}^{r} a_{r+1-\nu}^r x[k-\nu] &= x[(k-1)-r] \\ &\quad + \sum_{\mu=1}^{r} a_\mu^r x[(k-1)-r+\mu] \\ &= b_r[k-1]. \end{aligned} \tag{8.81}$$

Für den Prädiktionsfehler der Ordnung $r+1$, im Folgenden zur besseren Unterscheidung *Vorwärts-Prädiktionsfehler* genannt, gilt somit:

$$e_{r+1}[k] = e_r[k] - \gamma_{r+1} b_r[k-1] \tag{8.82}$$

Wir können somit den Vorwärts-Prädiktionsfehler rekursiv aus Vorwärts- und Rückwärts-Prädiktionsfehler der nächst geringeren Ordnung bestimmen. Das hilft uns allerdings nur dann weiter, wenn wir auch den Rückwärts-Prädiktionsfehler in gleicher Weise rekursiv berechnen können. Wie bereits bei der Vorwärtsrekursion, Gl. (8.82), gehen wir auch hier von der Definitionsgleichung aus und ersetzen die Modellparameter mit Hilfe der Levinson-Durbin-Rekursion:

$$\begin{aligned} b_{r+1}[k] &= x[k-(r+1)] \\ &\quad + \sum_{\mu=1}^{r} \left(\left(a_\mu^r - \gamma_{r+1} a_{r+1-\mu}^r \right) x[k-(r+1)+\mu] \right) + \gamma_{r+1} x[k] \\ &= x[k-(r+1)] \\ &\quad + \sum_{\nu=1}^{r} a_\mu^r x[k-(r+1)+\mu] - \gamma_{r+1} x[k] \\ &\quad + \sum_{\mu=1}^{r} a_{r+1-\mu}^r x[k-(r+1)+\mu] \end{aligned} \tag{8.83}$$

Der erste Ausdruck beschreibt den Rückwärts-Prädiktionsfehler r-ter Ordnung. Für den zweiten erhält man nach der Substitution $\mu \rightarrow \nu = r+1-\mu$ den Vorwärts-Prädiktionsfehler r-ter Ordnung. Damit ergibt sich:

$$b_{r+1}[k] = b_r[k-1] - \gamma_{r+1} e_r[k]. \tag{8.84}$$

Somit haben wir eine Möglichkeit gefunden Vorwärts- und Rückwärts-Prädiktionsfehler der Ordnung $r+1$ mit Hilfe des PARCOR-Koeffizienten γ_{r+1} aus den Prädiktionsfehlern r-ter Ordnung zu ermitteln. Die entsprechende Struktur bezeichnet man als Lattice-Struktur. Für sie fassen wir noch einmal die Lattice-Gleichungen zusammen:

Theorem 8.5 (Lattice-Gleichungen).

$$\begin{aligned} e_{r+1}[k] &= e_r[k] - \gamma_{r+1} b_r[k-1] \\ b_{r+1}[k] &= b_r[k-1] - \gamma_{r+1} e_r[k] \end{aligned} \tag{8.85}$$

Abb. 8.8 zeigt eine einfache Stufe dieser Lattice-Struktur für den Übergang von der Ordnung r zur Ordnung $r+1$.

Schaltet man nun n Stufen dieser Art hintereinander (siehe Abb. 8.9) so erhält man am Ausgang den Prädiktionsfehler n-ter Ordnung. Die Übertragungsfunktion bezüglich des Vorwärts-Prädiktionszweigs ist also identisch mit

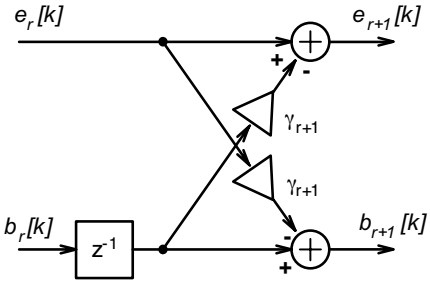


Abbildung 8.8. Einfache Lattice-Stufe

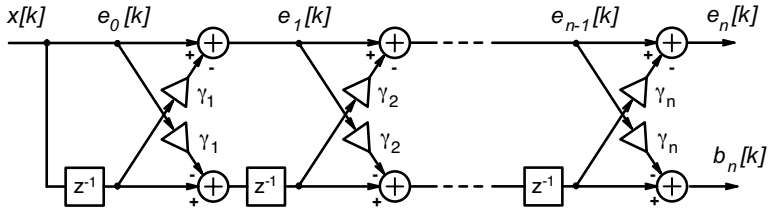


Abbildung 8.9. Prädiktionsfehlerfilter n-ter Ordnung in Lattice-Struktur

der Übertragungsfunktion des linearen Prädiktionssystems n-ter Ordnung. Den zusätzlich bestimmten Rückwärts-Prädiktionsfehler betrachten wir zunächst nur als Hilfsgröße.

Der wesentliche Vorteil, den man durch die Verwendung der Lattice-Struktur erhält, ist die Möglichkeit, das System zu erweitern ohne die bestehenden Teile verändern zu müssen.

Beispiel 8.6 – Lattice-Struktur.

Gegeben sei wie in Beispiel 8.1 ein AR-System 2. Ordnung mit den Modellkoeffizienten $a_1 = 1/4, \quad a_2 = -1/2$. Ein entsprechender Prädiktor hat also 2 Stufen, er sei in Lattice-Struktur realisiert. Der Ausgang des Prädiktors $e_2[n]$ entspricht dem oberen Zweig in Abb. 8.9 bei der gestrichelten Linie.
Der Prädiktor hat nach der Wiener-Hopf-Gleichung die Koeffizienten $p_1 = -1/4, \quad p_2 = 1/2$, und die Fehlerfunktion lautet damit

$$e_2[n] = x[n] + \frac{1}{4}x[n-1] - \frac{1}{2}x[n-2] \tag{8.86}$$

Vor dem Zeitpunkt t=0 sind alle Grössen im System Null. Der Prediktor wird nun angeregt mit den Ausgangswerten des AR-Systems $x[0] = 1, x[1] = -0.25, x[2] = 0.5625, x[3] = -0.2656$

(siehe Beispiel 8.2). Die PARCOR-Koeffizienten des Prädiktors hatten wir mit Hilfe der Levinson-Durbin-Rekursion in Beispiel 8.5 bestimmt zu:

$$\gamma_1 = -1/2$$
$$\gamma_2 = 1/2$$

Wir überzeugen uns anhand von Abb. 8.9, dass mit diesen PARCOR-Koeffizienten gerade die Prädiktor-Gleichung realisiert wird. Dazu verfolgen wir die möglichen Wege des Eingangsignals und seiner Zeitverzögerungen.
Der Ausgang des Prädiktors $e_2[n]$ entsteht ohne Zeitverzögerung aus $x[n]$, mit einfacher Zeitverzögerung aus $-\gamma_1 x[n-1]$ (erste Stufe) und aus $(-\gamma_1)(-\gamma_2)x[n-1]$ (zweite Stufe), sowie mit zweifacher Zeitverzögerung aus $-\gamma_2 x[n-2]$. In Summe ergibt sich also

$$e_2[n] = x[n] + \frac{1}{2}x[n-1] - \frac{1}{4}x[n-1] - \frac{1}{2}x[n-2]$$

Dies entspricht genau der verlangten Prädiktorgleichung (8.86).
Die Werte der Vorwärts– und Rückwärtsprädiktionsfehler werden in Übung 8.6 berechnet. □

8.8.2 Inverses Filter

Im vorigen Abschnitt haben wir eine Lattice-Struktur hergeleitet, die einem linearen Prädiktionsfehlerfilter entspricht. Wir fragen uns nun, ob man auch ein autoregressives Modell durch eine Lattice-Struktur ersetzen kann. Das AR-Modell hat gerade die inverse Übertragungsfunktion des Prädiktionssystems, siehe Abb. 8.10. Anders ausgedrückt, ein autoregressiver Prozess n-ter

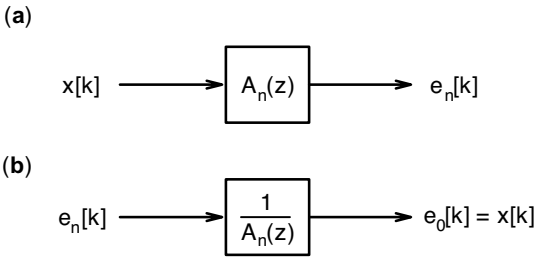


Abbildung 8.10. a) Prädiktionssystem (Analysefilter) b) AR-Modell (Synthese-Filter)

Ordnung kann mit einem entsprechenden Prädiktionsfehlerfilter n-ter Ordnung exakt vorhergesagt werden. Diese Eigenschaft können wir in der Lattice-

Struktur dadurch nachbilden, dass die Signalflussrichtung für den Vorwärts-Prädiktionsfehler umgekehrt wird. Dazu ist es lediglich notwendig, die Gleichung für den Vorwärts-Prädiktionsfehler (Theorem 8.5) nach $e_r[k]$ umzustellen. Aus dem Prädiktionsfehler n-ter Ordnung wird der Prädiktionsfehler $e_0[k]$ rekonstruiert, der dem vorhergesagten Prozess entspricht. Damit ergeben sich die folgenden Gleichungen für die Lattice-Struktur eines AR-Prozesses:

Theorem 8.6 (Inverse Lattice-Struktur).

$$e_r[k] = e_{r+1}[k] + \gamma_{r+1} b_r[k-1] \tag{8.87}$$

$$b_{r+1}[k] = b_r[k-1] - \gamma_{r+1} e_r[k] \tag{8.88}$$

Die Signalflussrichtung für den Vorwärts-Prädiktionsfehler wird im Gegensatz zur Lattice-Struktur eines Prädiktionssystems umgedreht. Die so erhaltene Struktur zeigen die Abb. 8.11 und 8.12.

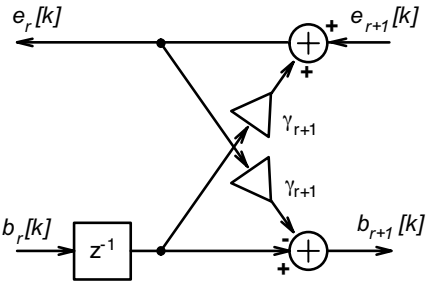


Abbildung 8.11. Inverse Lattice-Stufe

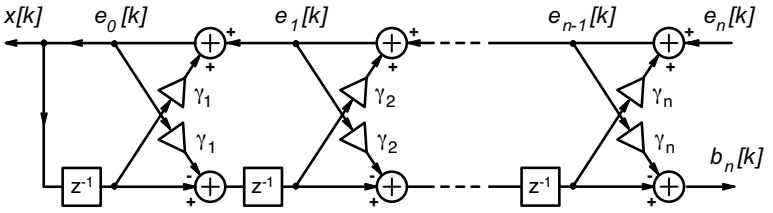


Abbildung 8.12. AR-Modell n-ter Ordnung in Lattice-Struktur

Beispiel 8.7 – Inverses Filter in Lattice-Struktur.

Wir erzeugen ein inverses Filter zu dem Prädiktionssystem aus Beispiel 8.6. Dieses inverse Filter hat also 2 Stufen, sein Ausgang $x[n]$

entspricht dem oberen Zweig in Abb. 8.12 nach Anregung durch eine Delta-Folge $e_2[n] = \delta[n]$ (von rechts).
Wir überzeugen uns anhand dieser Abbildung, dass wieder ein AR-System erzeugt wird (siehe Beispiel 8.2). Dazu verfolgen wir die möglichen Wege des Eingangsignals und seiner Zeitverzögerungen.
Die PARCOR-Koeffizienten des Prädiktors hatten wir mit Hilfe der Levinson-Durbin-Rekursion in Beispiel 8.5 bestimmt zu:

$$\gamma_1 = -1/2$$
$$\gamma_2 = 1/2$$

Der Ausgang des inversen Filters $x[n]$ entsteht ohne Zeitverzögerung aus $e_2[n]$, mit einfacher Zeitverzögerung aus $\gamma_1 x[n-1]$ (linke Stufe) und aus $(-\gamma_1)\gamma_2 x[n-1]$ (rechte Stufe), sowie mit zweifacher Zeitverzögerung aus $\gamma_2 x[n-2]$. In Summe ergibt sich also

$$x[n] = e_2[n] - \frac{1}{2}x[n-1] + \frac{1}{4}x[n-1] + \frac{1}{2}x[n-2]$$

Regen wir nun mit $e_2[n] = q[n]$ an, ergibt sich

$$x[n] = q[n] - \frac{1}{4}x[n-1] + \frac{1}{2}x[n-2]$$

Dies entspricht genau der verlangten Gleichung des AR-Systems aus Beispiel 8.2.
Die Werte der Vorwärts– und Rückwärtsprädiktionsfehler werden in Übung 8.7 berechnet. □

8.9 Orthogonalität des Rückwärts-Prädiktionsfehlers

Wir betrachten ein Lattice mit R Stufen. In den beiden vorigen Abschnitten haben wir den Rückwärts-Prädiktionsfehler $b_r[k]$ als eine Hilfsgröße zur Realisierung der Lattice-Struktur eingeführt. Er hat darüber hinaus aber eine weitere sehr interessante Eigenschaft, die wir in folgendem Theorem formulieren:

Theorem 8.7 (Orthogonalität der Rückwärts-Prädiktionsfehler).
Alle Rückwärts-Prädiktionsfehler der verschiedenen Stufen sind zum selben Zeitpunkt paarweise orthogonal:

$$E\{b_r[k]b_q[k]\} = \sigma_r \delta_{rq} \quad r, q \leq R. \tag{8.89}$$

Die Gültigkeit dieses Theorems für ein Beispiel überprüfen wir in Übung 8.7.

Die Idee zum Nachweis von Theorem 8.7 ist es, den obigen Erwartungswert auf die Lückenfunktion zurückzuführen. Diese hatten wir im vorigen Abschnitt wie folgt definiert:

$$g_r[\kappa] = E\{e_r(k)X(k-\kappa)\}$$
$$g_r[\kappa] = 0 \quad 1 \le \kappa \le r \tag{8.90}$$

Im Orthogonalitäts-Kriterium haben wir das Produkt aus zwei Rückwärts-Prädiktionsfehlern unter dem Erwarungswert. Einen ersetzt man nun durch seine Definitionsgleichung

$$b_r[k] = x[k-r] + \sum_{\mu=1}^{r} a_\mu^r x[k-r+\mu])$$
$$= \sum_{\mu=0}^{r} a_\mu^r x[k-r+\mu], \quad \text{mit} \quad a_0^r = 1, \tag{8.91}$$

den zweiten durch die Summe von zwei Vorwärts-Prädiktionsfehlern. Wir müssen dabei beachten, dass wir nicht Rekursionsgleichungen für Vorwärts-Prädiktionsfehler benutzen, die auf Stufen zurückgreifen, die noch gar nicht existieren. Diese Gefahr besteht, wenn wir die Rekursionsgleichung für den Vorwärts-Prädiktionsfehler

$$e_q[k] = e_{q-1}[k] - \gamma_q b_{q-1}[k-1] \tag{8.92}$$

einfach nach dem Rückwärts-Prädiktionsfehler umformen. Dies ergibt

$$b_q[k-1] = \frac{1}{\gamma_{q+1}}\left(e_q[k] - e_{q+1}[k]\right) \tag{8.93}$$

und damit

$$b_q[k] = \frac{1}{\gamma_{q+1}}\left(e_q[k+1] - e_{q+1}[k+1]\right) \tag{8.94}$$

In der letzten Zeile wird aber der Rückwärts-Prädiktionsfehler der Stufe q auf einen Vorwärts-Prädiktionsfehler der Stufe $q+1$ bezogen. Dies führt in der letzten Stufe R des Lattices dazu, dass wir für $q = R$ einen Vorwärts-Prädiktionsfehler der Stufe $R+1$ benutzen: diesen haben wir aber gar nicht zur Verfügung!

Wir müssen daher zunächst für jede Stufe den Rückwärts-Prädiktionsfehler auf Vorwärts-Prädiktionsfehler beziehen, die nicht höheren Stufen entsprechen. Dies gelingt, wenn wir die beiden Definitionsgleichungen der Vorwärts- und Rückwärts-Prädiktionsfehler benutzen und ineinander einsetzen. Die Definitionsgleichungen lauten:

$$e_q[k] = e_{q-1}[k] - \gamma_q b_{q-1}[k-1]$$
$$b_q[k] = b_{q-1}[k-1] - \gamma_q b_{q-1}[k] \tag{8.95}$$

Wir lösen die Gleichung (8.95) nach $b_{q-1}[k-1]$ auf und setzen den gefundenen Ausdruck in Gl. (8.95) ein:

$$\gamma_q b_q[k] = -e_q[k] + (1-\gamma_q^2)e_{q-1}[k] = -e_q[k] + \frac{\sigma_r}{\sigma_{r-1}}e_{q-1}[k] \tag{8.96}$$

Damit haben wir das Ziel erreicht, den Rückwärts-Prädiktionsfehler (Stufe q) auf Vorwärts-Prädiktionsfehler beziehen, die nicht höheren Stufen entsprechen (Stufen q, $q-1$).

Wir können nun die o.a. Idee der Rückführung auf gapped-Funktionen verfolgen. Dazu nehmen wir ohne Einschränkung der Allgemeinheit an, dass $q \geq r$ gilt. Setzen wir nun die Gleichungen (8.96) und (8.91) in die Gl. (8.89) ein, so erhalten wir:

$$\begin{aligned} E\{b_r[k]b_q[k]\} &= E\left\{\left(\sum_{\mu=0}^{r} a_\mu^r x[k-r+\mu]\right)\left(\frac{1}{\gamma_{q+1}}\left(e_q[k+1]-e_{q+1}[k+1]\right)\right)\right\} \\ &= \frac{1}{\gamma_{q+1}}\sum_{\mu=0}^{r} a_\mu^r E\{e_q[k+1]x[k-r+\mu]\} \\ &- a_\mu^r E\{e_{q+1}[k+1]x[k-r+\mu]\} \\ &= \frac{1}{\gamma_{q+1}}\sum_{\mu=0}^{r} a_\mu^r \left(g_q[r+1-\mu]-g_{q+1}[r+1-\mu]\right) \end{aligned} \tag{8.97}$$

Die Argumente der beiden Lückenfunktionen sind abhängig vom Summationsindex μ. An den Summationsgrenzen erkennen wir, dass das Argument κ von $g[\kappa]$ alle Werte mit $1 \leq \kappa \leq r+1$ durchläuft. Da wir angenommen haben, dass $q \geq r$ und aufgrund der Eigenschaften der Lückenfunktion, liefert die Funktion $g_{q+1}[\kappa]$ für alle Summanden den Wert null. Für den Fall $q > r$ gilt dasselbe für die andere Lückenfunkion $g_q[\kappa]$. Nur wenn $q = r$ ergibt sich an der Stelle $\kappa = r+1$ ein von null verschiedener Wert:

$$E\{b_r[k]b_r[k]\} = \frac{1}{\gamma_{r+1}} a_0^r g_r[r+1] = \frac{1}{\gamma_{r+1}} g_r[r+1] \tag{8.98}$$

Mit den Beziehungen (8.56) und (8.67) ergibt sich:

$$E\{b_r[k]b_q[k]\} = g_r[0]\delta_{q,r} = \sigma_r \delta_{q,r} \tag{8.99}$$

Damit ist die Orthogonalität der Rückwärts-Prädiktionsfehler gezeigt. □

Große praktische Bedeutung hat die Orthogonalitäts-Eigenschaft für die Realisierung von adaptiven Systemen. Als adaptiv bezeichnet man solche Systeme, bei denen die Parameter bei laufendem Betrieb durch geeignete Algorithmen eingestellt werden. Als Beispiel soll hier der adaptive Echoentzerrer angeführt werden. Abb. 8.13 a) zeigt dazu einen konventionellen Entzerrer in Transversalform, Abb. 8.13 b) eine Realisierung mit Hilfe der Lattice-Struktur.

In beiden Fällen wird der Ausgang des Entzerrers $y[k]$ mit einem Referenzsignal $d[k]$ verglichen. Das Differenzsignal $\epsilon[k]$ speist einen adaptiven Algorithmus. Durch Einstellen der Parameter g_i bzw. h_i soll nun das Differenzsignal $\epsilon[k]$ minimiert werden. Die Konvergenzgeschwindigkeit des adaptiven

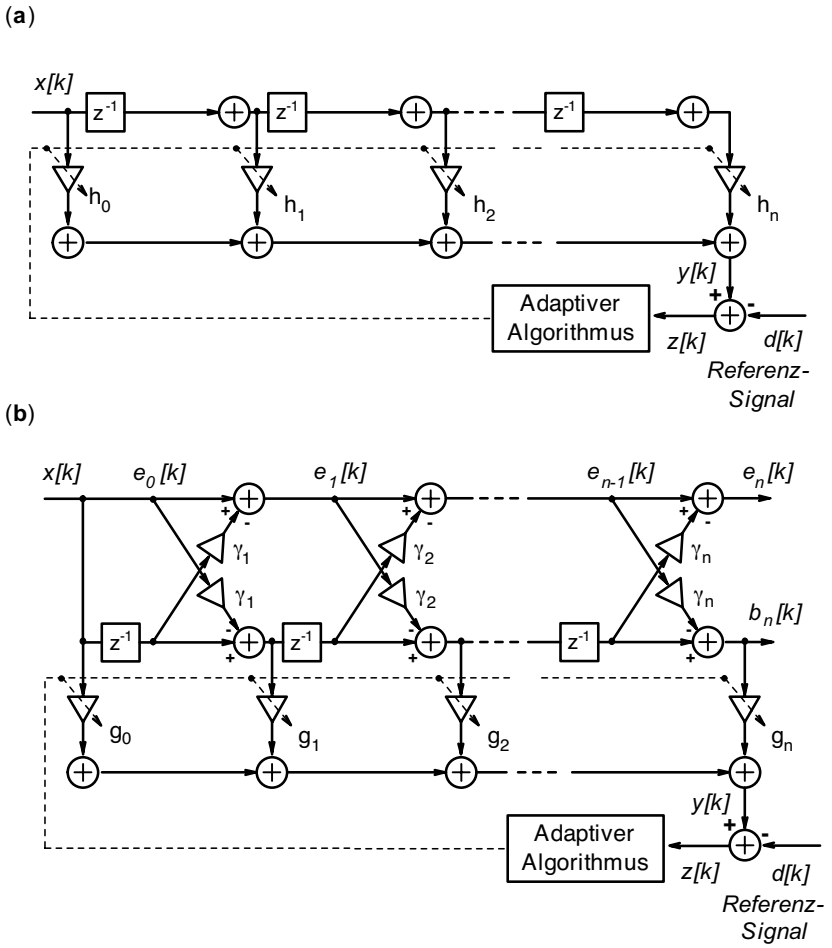


Abbildung 8.13. Adaptiver Entzerrer n-ter Ordnung: a) Transversalform, b) Lattice-Struktur

Algorithmus hängt maßgeblich davon ab, in wie weit die Eingänge des Summierers miteinander korreliert sind. Verwendet man die Lattice-Struktur, so liegen an den Summierer-Eingängen die orthogonalen und damit unkorrelierten Rückwärts-Prädiktionsfehler. Somit erreicht man durch die Verwendung der Lattice-Struktur eine erhebliche Steigerung der Konvergenzgeschwindigkeit. Außerdem ist diese Realisierung weitgehend robust gegen Anfangsbedingungen, das Einschwingverhalten ist optimal.

8.10 Gram-Schmidt-Orthogonalisierung

Mathematisch lässt sich die stufenweise Orthogonalität der Rückwärts-Prädiktionsfehler als *Gram-Schmidt-Orthogonalisierung* der eingehenden Signale durch das r-stufige Filter interpretieren. Dabei werden schrittweise zueinander orthogonale Funktionen (Vektoren) aufgebaut, wobei mit solchen Vektoren begonnen wird, die die stärkste Dekorrelation (Ent-Korrelierung) bewirken. Der Rückwärts-Prädiktionsfehler erfährt also von Stufe zu Stufe eine Dekorrelation, wobei die größtmögliche Dekorrelation zuerst durchgeführt wird. Obwohl sie in vielfältigen Textbüchern behandelt wird, wollen wir die Gram-Schmidt-Orthogonalisierung allgemein einführen, mit Beispielen illustrieren, und dann untersuchen, inwiefern solch eine Orthogonalisierung von Lattice-Strukturen durchgeführt wird.

8.10.1 Lineare Räume, Basen, innere Produkte

Eine Gram-Schmidt-Orthogonalisierung ist die schrittweise Entwicklung einer *Basis* aus *Beispieldaten.* Wir erinnern dazu zunächst an einige wichtige Eigenschaften einer Basis und eines linearen Raumes.

Eine Basis ist eine minimale (linear unabhängige) Darstellung eines linearen Raumes. Bezeichnen wir die Elemente der Basis mit $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$, so kann jedes Element $\mathbf{x}$ des linearen Raumes durch eine lineare Überlagerung der Elemente der Basis dargestellt werden. Definieren wir also Koeffizienten $(a_1, a_2, \ldots, a_n)$, so gilt für alle $\mathbf{x}$:

$$\mathbf{x} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + ... + a_n\mathbf{e}_n \quad \forall \mathbf{x} \tag{8.100}$$

Gleichung (8.100) wird auch als *Entwicklung* eines Elements $\mathbf{x}$ nach der Basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ bezeichnet. Um die Orthogonalität zu definieren, benötigen wir ein *inneres Produkt* aus zwei Elementen $\mathbf{x}, \mathbf{y}$ des linearen Raumes. Dieses wird in der Literatur mit einem geklammerten Paar $(\mathbf{xy})$ oder auch einem Punkt $\mathbf{x}\bullet\mathbf{y}$ (engl. „dot-product“) bezeichnet. Es gilt nun:

Zwei Elemente $\mathbf{x}, \mathbf{y}$ des linearen Raumes sind *orthogonal*, wenn ihr inneres Produkt verschwindet, also $(\mathbf{xy}) = 0$.

Insbesondere sind natürlich die Elemente der Basis auch Elemente des linearen Raumes, und wir können die Orthogonalitätseigenschaft daher auch auf die Basis anwenden. Es gilt:

Eine *orthogonale Basis* hat die Eigenschaft, dass *jedes Paar* von Basiselementen orthogonal ist, also

$$(\mathbf{e}_i\mathbf{e}_j) = 0 \quad \forall i \neq j \tag{8.101}$$

Zuletzt führen wir mit Hilfe des inneren Produktes eine Normierung ein: Ein Element $\mathbf{x}$ des Raumes ist „auf Eins normiert“ oder einfach „normiert“ , wenn gilt $(\mathbf{xx}) = 1$. Ein Kunstwort für „orthogonal und normiert“ ist „orthonormiert“ . Eine orthonormierte Basis ist also eine orthogonale Basis, für

die zusätzlich gilt $(\mathbf{e}_i\mathbf{e}_i) = 1$ für alle $\mathbf{e}_i$. Orthonormale Basen haben bezüglich der o.a. Entwicklung der Elemente sehr angenehme Eigenschaften. Um einen der Koeffizienten a_k zu bestimmen, bilden wir das innere Produkt aus der Entwicklungsgleichung mit dem zugehörigen Basiselement $\mathbf{e}_k$. Es gilt:

$$\begin{aligned}(\mathbf{e}_k\mathbf{x}) &= (\mathbf{e}_k a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + ... + a_n\mathbf{e}_n) \\ &= a_1(\mathbf{e}_k\mathbf{e}_1) + a_2(\mathbf{e}_k\mathbf{e}_2) + ... + a_n(\mathbf{e}_k\mathbf{e}_n) \\ &= a_k \end{aligned} \tag{8.102}$$

Dabei wurde sowohl die Linearitätseigenschaft des inneren Produktes als auch die Orthonormiertheit der Basis ausgenutzt, die alle inneren Produkte der letzten Summe verschwinden lässt bis auf eines, nämlich $(\mathbf{e}_k\mathbf{e}_k) = 1$. Somit ergeben sich bei orthonormalen Basen die Koeffizienten eines Elementes $\mathbf{x}$ einfach als inneres Produkt von x mit den entsprechenden Basiselementen.

Aus gutem Grund haben wir uns bisher nicht festgelegt, wie die Elemente unserer linearen Räume und wie das innere Produkt aussehen sollen. Dies wird sich jetzt als hilfreich erweisen, denn wir werden an zwei Beispielen sehen, dass man das Konzept der linearen Räume auf ganz unterschiedliche Typen von Elementen anwenden kann.

Beispiel 8.8 – Vektorraum.

Als erstes Beispiel mögen uns konventionelle *Vektoren* dienen. Die Elemente $\mathbf{x}$ sind dann N-dimensionale Vektoren und das innere Produkt ist das Skalarprodukt zweier Vektoren, gegeben durch

$$(\mathbf{xy}) = \sum_{k=1}^{N} x_k y_k \tag{8.103}$$

Eine orthonormale Basis wird z.B. gebildet aus den Einheitsvektoren eines rechtwinkligen Koordinatensystems. Im Beispiel zweier Dimensionen ist also

$$\mathbf{e}_1 = (1,0), \quad \mathbf{e}_2 = (0,1) \tag{8.104}$$

eine orthonormale Basis. Aber auch

$$\mathbf{f}_1 = \frac{1}{\sqrt{2}}(1,1), \quad \mathbf{f}_2 = \frac{1}{\sqrt{2}}(-1,1) \tag{8.105}$$

ist eine orthonormale Basis. Allgemein ist also die Basisdarstellung *nicht* identisch mit der Koordinatendarstellung! Sie ist sogar völlig unabhängig davon, wie wir jetzt illustrieren:
Betrachten wir z.B. den Vektor $\mathbf{x} = 2\mathbf{e}_1$, so lässt sich dieser auch darstellen als

$$\mathbf{x} = 2\mathbf{e}_1 = \sqrt{2}(\mathbf{f}_1 - \mathbf{f}_2) \tag{8.106}$$

Man beachte, dass wir bisher noch gar keine Koordinatendarstellung angegeben haben! Diese ist völlig unabhängig von der Basis: Wählen wir als Koordinatenachsen die $\mathbf{e}$-Basisvektoren, so gilt $\mathbf{x} = (2,0)$.

Wählen wir die **f**-Basisvektoren, so gilt $\mathbf{x} = (\sqrt{2}, -\sqrt{2})$. Wählen wir eine noch andere orthonormale Basis $\{\mathbf{g}_1, \mathbf{g}_2\}$, so gilt nach dem Entwicklungssatz $\mathbf{x} = a_1\mathbf{g}_1 + a_2\mathbf{g}_2$ und damit in **g**-Koordinaten $\mathbf{x} = (a_1, a_2)$ mit

$$a_1 = (\mathbf{g}_1\mathbf{x}) = 2(\mathbf{g}_1\mathbf{e}_1) = \sqrt{2}[(\mathbf{g}_1\mathbf{f}_1) - (\mathbf{g}_1\mathbf{f}_2)] \tag{8.107}$$

$$a_2 = (\mathbf{g}_2\mathbf{x}) = 2(\mathbf{g}_2\mathbf{e}_1) = \sqrt{2}[(\mathbf{g}_2\mathbf{f}_1) - (\mathbf{g}_2\mathbf{f}_2)] \tag{8.108}$$

Das heisst, die Koordinaten (a_1, a_2) in der Darstellung der **g**-Basis lassen sich aus beiden anderen Darstellungen (und auch aus jeder anderen Basis) berechnen. □

Beispiel 8.9 – Fourier-Reihe.

Als zweites Beispiel betrachten wir die Fourier-Reihe. Dazu definiert man N diskrete Frequenzen innerhalb der Bandbreite Ω des Signals (Abtastzeit T):

$$\omega = n\omega_0 = n\frac{\Omega}{N} = n\frac{2\pi}{NT}; \quad \Omega = \frac{2\pi}{T} \tag{8.109}$$

Für die zeitdiskrete Fouriertransformation (Fourier-Reihe) erhält man:

$$X(n) = \sum_{k=-\infty}^{\infty} x[k]e^{-j\omega kT} = \sum_{k=-\infty}^{\infty} x[k]e^{-jn(\frac{2\pi}{N})k} \tag{8.110}$$

bzw.

$$X(n) = \sum_{k=-\infty}^{\infty} x[k]W_N^{nk} \tag{8.111}$$

Dabei ist $W_N = e^{-j\frac{2\pi}{N}}$ der komplexe Drehoperator.
Inwiefern entspricht dies der Darstellung eines linearen Raumes? Offensichtlich gilt $x[k] = a_k$, es handelt sich dabei also um Komponenten in einer Basis. Wenn wir diese Analogie fortsetzen, so muss die Basis gegeben sein durch $W_N^{nk} = e_k$. Die Komponente $x[k]$ erhalten wir dann analog durch Bilden des inneren Produktes, also $x[k] = (e_k X(n)) = (W_N^{nk} X(n))$, und nach Einsetzen von $X(n)$:

$$\begin{aligned} x[k] = (W_N^{nk} X(n)) &= (W_N^{nk} \sum_{m=-\infty}^{\infty} x[m]W_N^{nm}) \\ &= \sum_{m=-\infty}^{\infty} x[m](W_N^{nk} W_N^{nm}) \end{aligned} \tag{8.112}$$

Dies ist offensichtlich genau dann richtig, wenn die $\{W_N^{nk}\}$ eine orthonormale Basis bilden, da dann in der Summe nur ein Term übrig

bleibt. Wir können nun fragen, wie ein inneres Produkt aussehen muss, das diese Eigenschaft hat. Die Antwort sieht wie folgt aus:

$$(W_N^{nk} W_N^{nm}) = \frac{1}{N} \sum_{r=1}^{N} (W_N^{nk})^r (W_N^{nm})^{(-r)} = \delta_{km} \tag{8.113}$$

Dass dies in der Tat gilt, erkennt man durch Einsetzen des komplexen Drehoperators in die Gleichung:

$$(W_N^{nk} W_N^{nm}) = \frac{1}{N} \sum_{r=1}^{N} e^{-j\frac{2\pi}{N} nr(k-m)} = \delta_{km} \tag{8.114}$$

Denn wenn $k = m$, sind die Exponentialterme identisch 1. Wenn $k \neq m$, so dreht der komplexe Drehoperator gerade $(k-m)n$ mal über den Einheitskreis, und die Summe aller Terme wird damit Null.
Wie wir sehen, ist dieses innere Produkt sehr ähnlich dem oben angegebenen inneren Produkt für Vektoren (Skalarprodukt):

$$(\mathbf{xy}) = \sum_{r=1}^{N} x_r y_r \tag{8.115}$$

In beiden Fällen wird also über N „Komponenten“ summiert. □

8.10.2 Prinzip der Gram-Schmidt-Orthogonalisierung

Eine orthogonale Basis soll nun aus Beispieldaten $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ schrittweise aufgebaut werden. Dies geschieht nach folgendem *Gram-Schmidt-Schema* für die Basisvektoren $\{\mathbf{e}_1, \mathbf{e}_2, ..., e_n\}$:

$$\mathbf{e}_1 = \mathbf{x}_1 \tag{8.116}$$

$$\mathbf{e}_2 = \mathbf{x}_2 - b_1 \mathbf{e}_1, \tag{8.117}$$

wobei wegen der Basiseigenschaft $(e_1 e_2) = 0$ gelten soll. Mithin

$$0 = (\mathbf{e}_1 \mathbf{e}_2) = (\mathbf{e}_1 \mathbf{x}_2) - b_1 (\mathbf{e}_1 \mathbf{e}_1), \quad \text{also} \quad b_1 = \frac{\mathbf{e}_1 x_2}{\mathbf{e}_1 \mathbf{e}_1}. \tag{8.118}$$

Für das k-te Element soll gelten

$$\mathbf{e}_k = \mathbf{x}_k - \sum_{i=1}^{k-1} b_i^k \mathbf{e}_i \tag{8.119}$$

das heisst das neue Basiselement $\mathbf{e}_k$ soll senkrecht (orthogonal) zu dem bisher aufgespannten Raum, mithin senkrecht zu allen bisher bekannten Beispieldaten stehen. Dies wird erreicht mit Hilfe des neuen Beispiels $\mathbf{x}_k$, wobei davon ausgegangen wird, dass dieses tatsächlich eine Komponente senkrecht zu

dem bisher aufgespannten Raum enthält. Das neue Basiselement wird daher auch *Innovation* genannt, und die Entwicklung eines Elementes x des linearen Raumes in die *Gram-Schmidt-Basis* heisst auch *Entwicklung nach Innovationen.*

Die Koeffizienten erhalten einen hochgestellten Index k, um deutlich zu machen, dass für jede (k-te) Stufe wieder andere Koeffizienten berechnet werden. Wegen der Orthogonalität gilt

$$(\mathbf{e}_j\mathbf{e}_k) = (\mathbf{e}_j\mathbf{x}_k) - \sum_{i=1}^{k-1} b_i^k(\mathbf{e}_j\mathbf{e}_i) = (\mathbf{e}_j\mathbf{x}_k) - b_j^k(\mathbf{e}_j\mathbf{e}_j) = 0 \quad (\forall j \neq k) \tag{8.120}$$

also

$$b_j^k = \frac{(\mathbf{e}_j\mathbf{x}_k)}{(\mathbf{e}_j\mathbf{e}_j)} \tag{8.121}$$

und damit

$$\mathbf{e}_k = \mathbf{x}_k - \sum_{i=1}^{k-1} \frac{(\mathbf{e}_i\mathbf{x}_k)}{(\mathbf{e}_i\mathbf{e}_i)}\mathbf{e}_i \tag{8.122}$$

Offensichtlich macht dies nur solange Sinn, wie $k \leq N$, also kleiner als die Dimension des linearen Raumes ist.

Wir haben schrittweise eine Basis aus einer Serie von Beispieldaten aufgebaut. Dies ist z.B. extrem nützlich, wenn diese Beispieldaten nicht alle gleichzeitig vorliegen, sondern als ein „Strom“ von Daten zeitlich nacheinander eintreffen. Die Basis kann dann schritthaltend aufgebaut werden.

Das dargestellte Verfahren hat die Eigenschaft, dass es *rekursiv* arbeitet: für die Bestimmung des Basiselements $\mathbf{e}_k$ müssen alle vorherigen Basiselemente bekannt sein.

Wir wählen nun eine übersichtliche Darstellung, die diesen Prozess im ganzen wiedergibt. Dies kann in Matrixschreibweise wie folgt geschehen:

$$\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \vdots \\ \mathbf{x}_k \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ b_1^1 & 1 & 0 & \cdots & 0 \\ b_1^2 & b_2^2 & 1 & \cdots & 0 \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ b_1^k & \cdots & \cdots & b_k^k & 1 \end{pmatrix} \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \\ \vdots \\ \mathbf{e}_k \end{pmatrix} \tag{8.123}$$

Man beachte, dass es sich bei den $\mathbf{x}$ wie auch bei den $\mathbf{e}$ um Elemente des linearen Raumes handelt, die (wie an obigen Beispielen gesehen) z.B. Vektoren oder funktionale Ausdrücke sein können.

8.10.3 Geschlossene Lösung für die Gram-Schmidt-Orthogonalisierung

Wir überlegen uns nun, wie wir zu einer geschlossenen (nicht rekursiven) Lösung für die Basiselemente $\mathbf{e}_k$ kommen können. Offensichtlich ist dazu die

Invertierung der Matrix mit den b_k-Einträgen erforderlich. Wir wollen diese Matrix $\mathbf{B}$ nennen, ihre Inverse $\mathbf{A}$. Dann gilt

$$\begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \\ \vdots \\ \mathbf{e}_k \end{pmatrix} = \mathbf{B}^{-1} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \vdots \\ \mathbf{x}_k \end{pmatrix} = \mathbf{A} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \vdots \\ \mathbf{x}_k \end{pmatrix} \tag{8.124}$$

Die Invertierung wollen wir jedoch wegen des hohen Rechenaufwands nicht durchführen. Wir können aber eine Aussage über die *Struktur* von $\mathbf{A}$ machen. Dazu hilft uns folgende Überlegung: Wir schreiben symbolisch

$$\mathbf{B} = \mathbf{U}_B + \mathbf{L}_B \tag{8.125}$$

Dies bedeutet, dass $\mathbf{B}$ aus einer Einheitsmatrix (**U**nit Matrix) $\mathbf{U}$ und einer unteren (oder linken) Dreiecksmatrix (**L**eft Triangular Matrix) $\mathbf{L}$ besteht. Für die inverse Matrix $\mathbf{A}$ schreiben wir allgemein in ähnlicher Darstellung

$$\mathbf{A} = \mathbf{D}_A + \mathbf{L}_A + \mathbf{R}_A \tag{8.126}$$

wobei $\mathbf{D}$ für eine Diagonalmatrix und $\mathbf{R}$ für eine obere (oder rechte) Dreiecksmatrix steht. Wir können nun folgende Aussagen über die *Struktur* der Produkte von Matrizen der angegebenen Form machen (diese lassen sich leicht durch komponentenweises Aufschreiben verifizieren):

$$\mathbf{DX} = \mathbf{X} \quad (\mathbf{X} = \text{Matrix beliebiger Struktur}) \tag{8.127}$$
$$\mathbf{XD} = \mathbf{X} \tag{8.128}$$
$$\mathbf{LL} = \mathbf{L} \tag{8.129}$$
$$\mathbf{RR} = \mathbf{R} \tag{8.130}$$
$$\mathbf{LR} = \mathbf{L} + \mathbf{D} + \mathbf{R} \tag{8.131}$$
$$\mathbf{RL} = \mathbf{L} + \mathbf{D} + \mathbf{R} \tag{8.132}$$

Mit diesen Zusammenhängen schreiben wir jetzt für die aus dem Produkt von $\mathbf{B}$ und seiner Inversen entstehenden Einheitsmatrix $\mathbf{U} = \mathbf{BA}$:

$$\begin{aligned} \mathbf{U} &= \mathbf{BA} \\ &= (\mathbf{U}_B + \mathbf{L}_B)(\mathbf{D}_A + \mathbf{L}_A + \mathbf{R}_A) \\ &= \mathbf{U}_B\mathbf{D}_A + \mathbf{L} + (\mathbf{U}_B + \mathbf{L}_B)\mathbf{R}_A \end{aligned} \tag{8.133}$$

Der letzte Term sind dabei alles Multiplikationen mit $\mathbf{R}_A$. Alle anderen Multiplikationen ergeben eine Diagonalmatrix $\mathbf{U}_B\mathbf{D}_A$ sowie eine linke Dreiecksmatrix.

Wir erkennen daraus folgendes:

1. Im Resultat wird der Beitrag zur rechten Dreiecksmatrix ausschliesslich durch $(\mathbf{U_B}+\mathbf{L_B})\mathbf{R_A}$ gebildet. Dieser Beitrag muss, da die Einheitsmatrix entstehen soll, Null werden. Damit ist $\mathbf{R_A} = \mathbf{0}$.
2. Mit $\mathbf{R}_A = \mathbf{0}$ entsteht im Resultat die diagonale Komponente ausschliesslich durch $\mathbf{U}_B\mathbf{D}_A$. Da dies die Einheitsmatrix ergeben soll, gilt offensichtlich $\mathbf{D}_A = \mathbf{U}$.
3. Die verbleibende linke Dreiecksmatrix $\mathbf{L}$ muss im Resultat ebenfalls Null werden. Zur Berechnung steht nunmehr noch der Anteil $\mathbf{L}_A$ der Matrix $\mathbf{A}$ zur Verfügung, es gilt

$$\begin{aligned}\mathbf{0} = \mathbf{L} &= \mathbf{U}_B\mathbf{L}_A + \mathbf{L}_B\mathbf{D}_A + \mathbf{L}_B\mathbf{L}_A \\ &= \mathbf{L}_A + \mathbf{L}_B + \mathbf{L}_B\mathbf{L}_A \\ &= (\mathbf{U} + \mathbf{L}_B)\mathbf{L}_A + \mathbf{L}_B \end{aligned} \tag{8.134}$$

   mit

$$\mathbf{L}_A = -(\mathbf{U} + \mathbf{L}_B)^{-1}\mathbf{L}_B \tag{8.135}$$

   Dabei wurden die bisherigen Resultate in die Gleichung eingesetzt. Offensichtlich ist das Resultat mit den gegebenen Matrizen auch tatsächlich erzielbar, denn die rechte Seite der letzten Gleichung ergibt wieder eine linke Dreiecksmatrix.

Zusammenfassend stellen wir fest, dass die Inverse einer **UL**-Matrix wieder eine **UL**-Matrix ist. Damit können wir die Inversion von $\mathbf{B}$ der Struktur nach durchführen und eine geschlossene Lösung für die Basiselemente $\mathbf{e}_k$ angeben:

$$\begin{pmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{e}_3\\ \vdots\\ \mathbf{e}_k\end{pmatrix} = \begin{pmatrix}1 & 0 & 0 & \cdots & 0\\ a_1^1 & 1 & 0 & \cdots & 0\\ a_1^2 & a_2^2 & 1 & \cdots & 0\\ \vdots & \cdots & \cdots & \cdots & \vdots\\ a_1^k & \cdots & \cdots & a_k^k & 1\end{pmatrix}\begin{pmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \mathbf{x}_3\\ \vdots\\ \mathbf{x}_k\end{pmatrix} \tag{8.136}$$

Wir halten noch einmal fest, dass diese Darstellung *notwendig* für eine Gram-Schmidt-Orthogonalisierung ist. Wir können sie also im folgenden zur Identifikation verwenden. Die Orthogonalität der Basen ist damit noch nicht garantiert, für sie müssen die entsprechenden Orthogonalitätsgleichungen verifiziert werden.

8.10.4 Berechnung des Prädiktionsfehlers: eine Gram-Schmidt-Orthogonalisierung?

Die Definition für den Rückwärts-Prädiktionsfehler lautete für jede Stufe r:

$$b^r[k] = x[k-r] + \sum_{\mu=1}^{r} a_\mu^r x[k-r+\mu] \tag{8.137}$$

Wir schreiben nun auch diese Definition für alle Stufen r übersichtlich in Matrixform:

$$\begin{pmatrix} b^0[k] \\ b^1[k] \\ b^2[k] \\ \vdots \\ b^r[k] \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_1^1 & 1 & 0 & \cdots & 0 \\ a_2^2 & a_1^2 & 1 & \cdots & 0 \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ a_r^r & \cdots & \cdots & a_1^r & 1 \end{pmatrix} \begin{pmatrix} x[k] \\ x[k-1] \\ x[k-2] \\ \vdots \\ x[k-r] \end{pmatrix} \tag{8.138}$$

Durch Vergleich mit der Matrixdarstellung für die geschlossene Lösung der Gram-Schmidt-Orthogonalisierung erkennt man, dass es sich (gegeben die noch zu überprüfende Orthogonalität der entstehenden Elemente) in der Tat um dasselbe Verfahren handelt: Die Lattice-Struktur realisiert eine stufenweise Berechnung des Rückwärts-Prädiktionsfehlers, die genau einer Gram-Schmidt-Orthogonalisierung entspricht. Dabei wird als erster (Beispiel)Wert $x[k]$ zur Verfügung gestellt, als zweiter $x[k-1]$ etc. Dies entspricht genau dem zeitlichen Rückwärts-Auftreten dieser Werte. Die Koeffizienten der Matrix sind entsprechend geordnet.

Für die Rückwärts-Prädiktionsfehler haben wir die Orthogonalität bereits explizit nachgewiesen, indem wir uns die *gapped-Funktion* zunutze gemacht haben. Die Lattice-Struktur produziert also mit Hilfe einer Gram-Schmidt-Orthogonalisierung eine Basis aus Rückwärts-Prädiktionsfehlern, die alle *zum selben Zeitpunkt* (k) orthogonal zueinander sind. Aus diesem Grunde werden die Rückwärts-Prädiktionsfehler gerne als diagonalisierte Information über die Daten benutzt, um z.B. adaptive Filter zu realisieren.

Wie verhält es sich nun mit den Vorwärts-Prädiktionsfehlern? Dazu benutzen wir wieder die Definition

$$e^r[k] = x[k] + \sum_{\mu=1}^{r} a_\mu^r x[k-\mu] \tag{8.139}$$

und schreiben auch diese übersichtlich in Matrixdarstellung, wobei wir besonders auf die Zeitpunkte achten müssen:

$$\begin{pmatrix} e^0[k-r] \\ e^1[k-r+1] \\ e^2[k-r+2] \\ \vdots \\ e^r[k] \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_1^1 & 1 & 0 & \cdots & 0 \\ a_2^2 & a_1^2 & 1 & \cdots & 0 \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ a_r^r & \cdots & \cdots & a_1^r & 1 \end{pmatrix} \begin{pmatrix} x[k-r] \\ x[k-r+1] \\ x[k-r+2] \\ \vdots \\ x[k] \end{pmatrix} \tag{8.140}$$

Durch Vergleich mit der Matrixdarstellung für die geschlossene Lösung der Gram-Schmidt-Orthogonalisierung erkennt man, dass auch hier (gegeben die noch zu überprüfende Orthogonalität der entstehenden Elemente) mit Hilfe der r Werte x eine Gram-Schmidt-Orthogonalisierung vorgenommen wird. Die zeitliche Ordnung der x-Werte entspricht dieser Vorwärtsrichtung. Allerdings kann dabei wegen der linken Seite der Matrixgleichung nur eine orthogonale

Basis aus Vorwärts-Prädiktionsfehlern zu *verschiedenen* Zeitpunkten gebildet werden. Aus diesem Grunde können die Vorwärts-Prädiktionsfehler nur dann als gleichzeitige diagonalisierte Information über die Daten benutzt werden, wenn weitere Zeitverzögerungsglieder zwischengeschaltet sind. Wegen der dabei auftretenden Verzögerung werden sie daher typischerweise nicht benutzt.

Wir haben die Orthogonalität der Basis aus Prädiktionsfehlern daraus geschlossen, dass die Definitionsgleichungen genau der Gram-Schmidt-Struktur entsprechen und dass die Orthogonalität der Elemente nachgewiesen ist. Für die Rückwärts-Prädiktionsfehler hatten wir dies weiter oben auch explizit nachgewiesen, indem wir uns die *gapped-Funktion* zunutze gemacht haben. Dies ist in gleicher Weise mit den Vorwärts-Prädiktionsfehlern (wie angegeben, zu verschiedenen Zeitpunkten) möglich. Dem Leser sei dies als Übung 8.8 empfohlen.

8.11 Ausblick: Der Burg-Algorithmus

Wir haben bereits im Abschnitt 8.4 die Auswirkungen von Kurz- und Langzeitstationarität auf die Schätzung der Autokorrelationsfolge und damit auf die Lösung der Yule-Walker-Gleichung diskutiert. Gleiches gilt -wegen identischer Struktur - für die Wiener-Hopf-Gleichung. Die Schätzproblematik entstand aus dem Ansatz, dass wir eine korrekte Identifikation der Parameter des erzeugenden Modells (Yule-Walker) bzw. einen minimalen Fehler der Prädiktion (Wiener-Hopf) fordern und dabei die Werte der Autokorrelationsfolge als *gegeben* annehmen. *Anschließend* wendeten wir uns der Frage zu, wie genau diese Folgenwerte berechnet werden können.

Wir wollen uns nun fragen, ob es nicht auf systematische Weise gelingt, im Falle endlicher Merkmalsfolgen geeignete (neue) Schätzgleichungen für die Modellparameter anzugeben. Dazu sollte ein Verfahren angegeben werden, das die Probleme a) der fehlerhaften Berechnung der AKF-Werte bei endlichen Merkmalsfolgen (welche sich nicht beseitigen lässt) und b) der Identifikation der Modellparameter gleichzeitig angeht.

Betrachten wir zunächst das simple *Abschneiden* der Folgenglieder der AKF. Hier entsteht bekanntlich der Effekt, dass wir entweder eine erwartungstreue und nicht konsistente, oder aber eine nicht erwartungstreue aber konsistente Schätzung der AKF vornehmen können. Dies führt zu Verfälschungen der Schätzwerte der AKF und damit der Modellparameter.

Ein anderer Ansatz ist es, die durch das Abschneiden bedingten Einschwingphasen und Ausschwingphasen in der Berechnung der AKF herauszunehmen. D.h. die AKF wird bei einem Prädiktor n-ten Grades und Vorliegen von N Meßwerten nur für diejenigen Werte $s[k]$ berechnet, wo $n<k<N-n$ ist, die anderen Werte von $s[k]$ werden als unbestimmt betrachtet. Dieses Verfahren heisst *Kovarianzmethode*. Es zeigt sich, siehe (Kammeyer und Kroschel, 2002), dass damit zwar eine erwartungstreue Schätzung realisiert wird, dass aber das entstehende Prädiktorfilter nicht zwingend stabil ist.

Man versucht nun, die Vorteile der Kovarianzmethode beizubehalten, das Verfahren aber so zu modifizieren, dass das entstehende Filter minimalphasig, mithin stabil ist. Dies gelingt, indem wir uns die Methodologie der Lattice-Strukturen zunutze machen: daher ist dieser Abschnitt an dieser relativ fortgeschrittenen Stelle des Buches angesiedelt. Wir wählen als Optimierungskriterium nicht wie bisher die Minimierung des Vorwärts-Prädiktionsfehlers, sondern minimieren die *Summe aus Vorwärts- und Rückwärts-Prädiktionsfehler*. Dabei nutzen wir aus, dass die Leistungen beider Fehler gleich sind. Die PARCOR-Koeffizienten $\hat{\gamma}_r$ dieses Verfahrens erhalten wir dann aus dem Verschwinden der partiellen Ableitungen

$$\frac{\delta}{\delta\hat{\gamma}_r}\sum_{k=r}^{N-1}(|e^r[k]|^2+|b^r[k]|^2)=0 \quad r=1\ldots n \tag{8.141}$$

Dieses Verfahren trägt den Namen **Burg-Algorithmus** und wird auch als „Methode der Maximalen Entropie“ bezeichnet, (Burg, 1967). Die Stabilität des entstehenden Filters ist garantiert. Da die PARCOR-Koeffizienten direkt gewonnen werden, kann ein Lattice-Filter direkt angegeben werden. Die Koeffizienten der dazugehörigen Analyse- bzw. Synthesefilter ergeben sich aus den PARCOR-Koeffizienten mit Hilfe der Gleichungen der Levinson-Durbin-Rekursion.

Mit dem Burg-Algorithmus ist ein Verfahren beschrieben, welches stabile Filter aus endlichen Merkmalsfolgen schätzt. Der Algorithmus ist damit geeignet für die Berechnung von Filtern aus kurzzeitstationären Daten. Auf mathematische Details soll hier nicht eingegangen werden, näheres siehe z.B. (Kammeyer und Kroschel, 2002).

8.12 Beispiel Sprachverarbeitung

Wir haben uns in den vorhergehenden Abschnitten ausführlich mit den Modellsystemen zur Spektralschätzung beschäftigt. Wir haben uns dabei auf die autoregressiven Modelle beschränkt, da diese in der Praxis die bei weitem größte Bedeutung haben. Der wesentliche Vorteil der AR-Modelle gegenüber den traditionellen Verfahren ist, dass auf Grund der wenigen zu schätzenden Größen auch mit kurzen Messreihen eine zufriedenstellende Spektralschätzung erreicht werden kann.

Diese Eigenschaft ist besonders wichtig in der automatischen Sprachverarbeitung. Wir beschränken uns hier auf einen kleinen Ausschnitt der Signalverarbeitung innerhalb der automatischen Sprachverarbeitung, für einen grösseren Überblick siehe z.B. das Lehrbuch (Wendemuth, 2004).

Da sich das Sprachsignal ständig ändert, kann nur für sehr kurze Zeitspannen (wenige Millisekunden) Stationarität angenommen werden. Traditionelle Verfahren sind unter diesen Bedingungen völlig ungeeignet. Ein weiterer Vorteil ist, dass sich der Vorgang der Spracherzeugung gut mit Hilfe

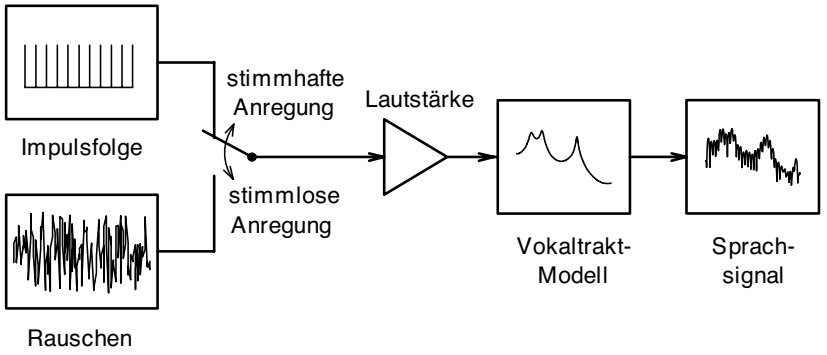


Abbildung 8.14. Quelle-Filter-Modell der Spracherzeugung

des AR-Modells beschreiben läßt. Abb. 8.14 zeigt dazu ein einfaches Quelle-Filter-Modell der Spracherzeugung von G. Fant (Fant, 1960). Die Stimmbänder erzeugen eine Erregerfrequenz, im Fall stimmhafter Laute ein Impulsfolge („pitch") bzw. weißes Rauschen für einen stimmlosen Laut. Hier besteht ein Unterschied zum autoregressiven Modell, bei dem wir grundsätzlich von einer Anregung durch weißes Rauschen ausgegangen sind. Aus der Erregerfrequenz wird nun im *Vokaltrakt* (Mund, Nase und Rachen) ein bestimmter Klang geformt. Dies kann gut durch rekursive Filter nachgebildet werden. Charakteristisch für einen Laut ist also nicht die Anregung, sondern die Struktur des Vokaltraktes. Diesen gilt es zu modellieren (Synthesefilter = AR-Prozess, für bekannte Laute) und zu analysieren (Analysefilter, für unbekannte Laute).

In Abb. 8.15 a) wird das originale Spektrum eines Ausschnitts des Vokals „a" mit einer Schätzung durch das AR-Modell verglichen. Man erkennt den charakteristischen Verlauf der Schätzung mit den ersten drei Resonanzstellen bei $1200Hz$, $2600Hz$ und $3400Hz$. In der Sprachverarbeitung bezeichnet man diese Resonanzen als *Formanten*. Der gewählte Ausschnitt hat eine zeitliche Länge von ca. $30ms$, in der das Sprachsignal als hinreichend stationär angesehen werden kann. Während das tatsächliche Spektrum aus 512 FFT-Werten besteht, werden für die AR-Schätzung lediglich 16 Koeffizienten für das geglättete Spektrum benötigt. Darüber hinaus zeigt Abb. 8.15 b) den zeitlichen Verlauf des Prädiktionsfehlers des geschätzten Prozesses. Bei einem autoregressiven Prozess würde man ein weißes Rauschen erwarten, das Diagramm dagegen zeigt eine verrauschte Impulsfolge. Ursache hierfür ist die Tatsache, dass bei der Erzeugung stimmhafter Laute das erregende Signal eine Impulsfolge ist, siehe Abb. 8.14. Wir sehen also, dass trotz dieses Unterschieds zum AR-Modell eine zufriedenstellende Schätzung des Spachsignals möglich ist.

Abb. 8.16 a) zeigt den charakteristischen Verlauf der Spektren der Vokale „e" und „i" mit einer Prädiktorordnung $n = 16$. Typisch für den Vokal „i" ist eine etwa gleichbleibende Ausprägung der zweiten und dritten Formantspitzen. Beim Vokal „e" dagegen fallen die Spitzen zu höheren Frequenzen hin ab.

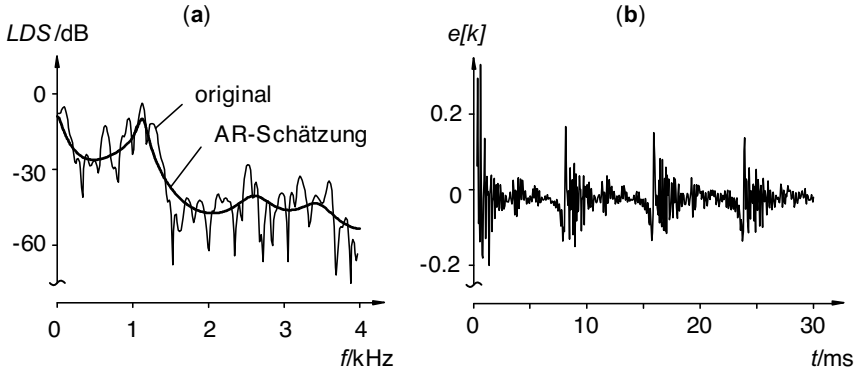


Abbildung 8.15. Spektralanalyse des Vokals "a": a) Vergleich des tatsächlichen Spektrums mit der AR-Schätzung, b) Prädiktionsfehler des Prozesses

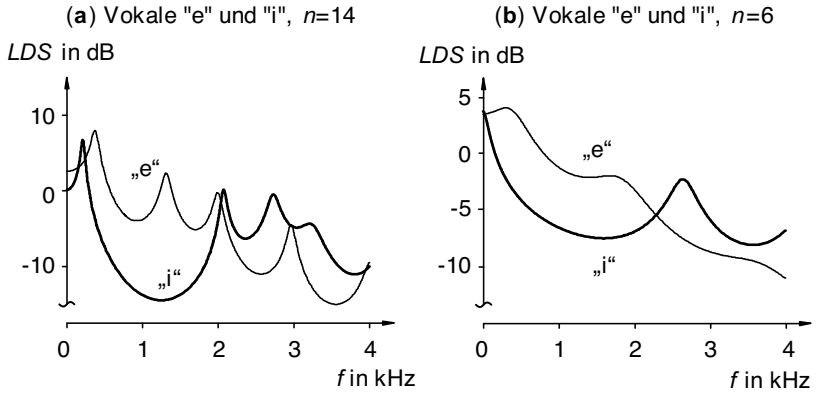


Abbildung 8.16. Spektralanalyse der Vokale "e" und "i": a) Prädiktionssystem der Ordnung n=16, b) Prädiktionssystem der Ordnung n=6

Zum Vergleich wird in Abb. 8.16 b) die Verwendung eines Systems mit einer zu geringen Ordnung demonstriert. Die Formanten können nicht mehr richtig dargestellt werden, ihre Zahl verringert sich durch die Glättung drastisch, so dass eine genaue Bestimmung der beiden Vokale erheblich erschwert wird.

Übungen

Übung 8.1 – Autoregressiver Prozess.

Gegeben sei eine Folge $x[n]$ bei der mit Hilfe eines Fensters die folgenden Funktionswerte ausgeschnitten wurden: $x[0] = 1$, $x[1] = -2$, $x[2] = 5$, $x[3] = 3$.

Außerhalb des Fensters sind alle Werte gleich Null.

1. Schätzen Sie die zugehörige konsistente AKF $\hat{s}_{XX}[k]$ mit $N = 4$.
2. Berechnen Sie aus $\hat{s}_{XX}[k]$ die Prädiktorkoeffizienten für ein AR-Modell zweiter Ordnung.
3. Wie lautet die Übertragungsfunktion?

Übung 8.2 – AKF mit Systemkorrelationsfolge.

Betrachten Sie ein AR-System 2. Ordnung mit den Modellkoeffizienten
$a_1 = 1/4, \quad a_2 = -1/2$ wie in Beispiel 8.1. Vor dem Zeitpunkt t=0 sind alle Grössen im System Null. Das System wird dann angeregt mit einer delta-Folge $q[n] = \delta[n]$.

a) Bestimmen Sie die die ersten 4 Terme der wahren Autokorrelationsfolge $s_{XX}(\kappa)$ mit Hilfe der Gleichung (7.54) durch Faltung der AKF der anregenden Delta-Folge mit der Systemkorrelationsfolge des übertragenden Systems, und anschliessender Rücktransformation.
b) Vergleichen Sie mit den Ergebnissen aus Beispiel 8.1.

Übung 8.3 – Yule-Walker-Gleichung.

Betrachten Sie ein AR-System 2. Ordnung mit den Modellkoeffizienten
$a_1 = 1/8, \quad a_2 = -1/2$.

Vor dem Zeitpunkt t=0 sind alle Grössen im System Null. Das System wird dann angeregt mit einer delta-Folge $q[n] = \delta[n]$.

Bestimmen Sie die die ersten 4 Terme der wahren Autokorrelationsfolge $s_{XX}(\kappa)$
a) mit Hilfe der Gleichung (7.54) durch Faltung der AKF der anregenden Delta-Folge mit der Systemkorrelationsfolge des übertragenden Systems, und anschliessender Rücktransformation.
b) mit Hilfe des modifizierten Yule-Walker-Gleichungssysstems (8.29) (siehe Beispiel 8.1).

Übung 8.4 – Versagen der Yule-Walker-Gleichung?.

Betrachten Sie erneut ein AR-System 2. Ordnung mit den Modellkoeffizienten
$a_1 = 1/2, \quad a_2 = -1/2$.

Vor dem Zeitpunkt t=0 sind alle Grössen im System Null. Das System wird dann angeregt mit einer delta-Folge $q[n] = \delta[n]$.

Man versucht, die ersten 3 Terme der Autokorrelationsfolge $s_{XX}(\kappa)$ mit Hilfe des modifizierten Yule-Walker-Gleichungssystems (8.29) aus Beispiel 8.1 zu bestimmen. Zeigen Sie, dass das nicht möglich ist, da die Matrix des Gleichungssystems (8.29) singulär wird. Um dies zu zeigen, berechnen Sie die Determinante der Matrix.

Durch die Singularitätseigenschaft kann das Gleichungssystem nicht gelöst werden. Warum? Hat die Yule-Walker-Gleichung versagt?

Wenn Sie dazu keine passende Antwort haben, schauen Sie nochmals in den Abschnitt 3.3.1 auf Seite 67.

Übung 8.5 – Levinson-Durbin-Rekursion.

Betrachten Sie ein AR-System 2. Ordnung mit den Modellkoeffizienten
$a_1 = 1/4, \quad a_2 = -1/2$ wie in Beispiel 8.2. Vor dem Zeitpunkt t=0 sind alle Grössen im System Null. Das System wird dann angeregt mit einer delta-Folge $q[n] = \delta[n]$.

Die Werte der zugehörigen Schätzung einer konsistenten Autokorrelationsfunktion mit N=4 wurden in Beispiel 8.2 berechnet, die Ergebnisse finden sich in Gln. (8.34). Benutzen Sie diese Ergebnisse, um mit Hilfe einer Levinson-Durbin-Rekursion Modellparameter 1. und 2. Ordnung zu bestimmen. Vergleichen Sie die Ergebnisse für beide Ordnungen mit den Resultaten $a_1^2 = 0.2391, \quad a_2^2 = -0.3305$ aus Beispiel 8.2 bzw. $a_1^1 = -p_1^1 = 0.3571$ aus Beispiel 8.3.

Übung 8.6 – Lattice-Struktur.

Betrachten Sie das 2-stufige Prädiktionsfilter in Lattice-Struktur aus Beispiel 8.6. Regen Sie dieses Filter an mit den Werten aus Beispiel 8.2,
$x[0] = 1, x[1] = -0.25, x[2] = 0.5625, x[3] = -0.2656$.

Berechnen Sie die Werte des Vorwärts–Prädiktionsfehlers $e_0[n]$, $e_1[n]$, $e_2[n]$ und des Rückwärts–Prädiktionsfehlers $b_0[n]$, $b_1[n]$ zu den Zeitpunkten $n = 0, 1, 2, 3$. Vergleichen Sie die Ergebnisse für $e_2[n]$ mit denen aus Beispiel 8.4.

Übung 8.7 – Inverse Lattice-Struktur.

Betrachten Sie das 2-stufige inverse Filter in Lattice-Struktur aus Beispiel 8.7. Berechnen Sie bei Anregung mit $e_2[n] = \delta[n]$ die Werte des Vorwärts–Prädiktionsfehlers $e_0[n]$, $e_1[n]$, $e_2[n]$ und des Rückwärts–Prädiktionsfehlers $b_0[n]$, $b_1[n]$, $b_2[n]$ zu den Zeitpunkten $n = 0, 1, 2, 3$. Vergleichen Sie die Ergebnisse für $x[n] = e_0[n]$ mit denen aus Beispiel 8.2,
$x[0] = 1, x[1] = -0.25, x[2] = 0.5625, x[3] = -0.2656$.

Überprüfen Sie die Gültigkeit von Theorem 8.7 auf Seite 237, indem Sie für die Rückwärts–Prädiktionsfehler Ergodizität annehmen und alle paarweisen Erwartungswerte über die Zeitpunkte $n = \{0, 1, 2, 3\}$ bilden.

Übung 8.8 – Orthogonalität des Vorwärts-Prädiktionsfehlers.

Gegeben sei eine Lattice-Struktur. Für den Vorwärts-Prädiktionsfehlers gilt Gl. (8.140). Aus dieser Gleichung erkennt man, dass es sich bei der Berechnung um eine Gram-Schmidt-Orthogonalisierung handeln kann. Beachten Sie die verschiedenen Zeitpunkte!

Zeigen Sie analog zum Beweis von Theorem 8.7 auf Seite 237, dass die Vorwärts-Prädiktionsfehler zu den angegebenen Zeitpunkten paarweise orthogonal sind. Begründen Sie, warum es sich deshalb, zusammen mit der genannten Gl. (8.140), bei der Berechnung der Vorwärts-Prädiktionsfehler zu den angegebenen Zeitpunkten um eine Gram-Schmidt-Orthogonalisierung handelt.

Abbildungsverzeichnis

Verzeichnis der Beispiele

Verzeichnis der Übungen

Literatur

[Argand 1813] ARGAND, Jean-Robert: Essai sur une manière de représenter les quantités imaginaires dans les constructions géométriques. In: *Annales de mathématique pures et appliquées* Bd. V 4 a, Joseph Gergonne (Hrsg.), 1813, S. 133–147

[Argand 1815] ARGAND, Jean-Robert: Réflexions sur la nouvelle théorie d'analyse. In: *Annales de mathématique pures et appliquées* Bd. T. V a, Joseph Gergonne (Hrsg.), 1815, S. 197–210. – vorausgegangene Briefwechsel erschienen bei Gergonne 1813-1814

[Bacon 1605] BACON, Francis: *The tvvoo bookes of Francis Bacon, of the proficience and aduancement of learning, diuine and humane. To the King. At London. By Francis Bacon, Viscount Saint Alban, Baron of Verulam.* London : "Printed for Henrie Tomes, and are to be sould at his shop at Graies Inne Gate in Holborne.", 1605

[Bacon 1640] BACON, Francis: *Of the advancement and proficience of learning: or, The partitions of sciences? Translated by Gilbert Wats.* Oxford University : Leon. Lichfield, for Rob. Young, I& Ed. Forrest, 1640

[Behnen und Neuhaus 1995] BEHNEN, Konrad ; NEUHAUS, Georg: *Grundkurs Stochastik.* Stuttgart : Teubner, 1995

[Beichelt 1997] BEICHELT, Frank: *Stochastische Prozesse für Ingenieure.* Stuttgart : Teubner, 1997

[Böhme 1998] BÖHME, Johann F.: *Stochastische Signale.* Stuttgart : Teubner, 1998

[Brigham 1995] BRIGHAM, Elbert O.: *FFT, Schnelle Fourier-Transformation.* Oldenbourg, 1995. – ISBN 3486231774

[Bronstein u. a. 1999] BRONSTEIN, Ilja N. ; SEMENDJAJEW, Konstantin A. ; MUSIOL, Gerhard: *Taschenbuch der Mathematik.* Frankfurt am Main : Harri Deutsch, 1999

[Burg 1967] BURG, J. B.: Maximum Entropy Spectral Analysis. In: *In Proc. 37 th Meeting of Society of Exploration Geophysicists* Bd. 1, 1967

[Cover und Thomas 1991] COVER, Thomas M. ; THOMAS, Joy. A.: *Elements of Information Theory.* New York : Wiley, 1991. – ISBN 0-471-06259-6

[Fant 1960] FANT, Gunnar: *The Acoustic Theory of Speech Production.* Mouton & Co., 1960

[Feigenbaum 1980] FEIGENBAUM, M.: Universal behavior in non-linear systems. In: *Los Alamos Science* 1 (1980), S. 4–27

[Fliege 1991] FLIEGE, Norbert: *Systemtheorie.* Teubner, 1991

[Föllinger 2000] FÖLLINGER, Otto: *Laplace-, Fourier- und z-Transformation.* Hüthig Verlag, 2000. – ISBN 3778527061

[von Grünigen 2001] GRÜNIGEN, Daniel von: *Digitale Signalverarbeitung.* Fachbuchverlag Leipzig, 2001. – ISBN 3-446-21445-3

[Kammeyer und Kroschel 2002] KAMMEYER, Karl D. ; KROSCHEL, Kristian: *Digitale Signalverarbeitung - Filterung und Spektralanalyse, Neuauflage.* Teubner, 2002. – Ideal und kompakt für hier verwendete Signalverarbeitungsverfahren. – ISBN 3-519-46122-6

[Kay 1993] KAY, Steven M.: *Fundamentals of Statistical Signal Processing: Estimation Theory.* Upper Saddle River : Prentice-Hall, 1993. – Bias-Variance-Verfahren, Schätzer, Bayes-Theorie. – ISBN 0-13-345711-7

[Kay 1998] KAY, Steven M.: *Fundamentals of Statistical Signal Processing: Detection Theory.* Upper Saddle River : Prentice-Hall, 1998. – Zufällige Signale und Rauschmodelle. – ISBN 0-13-504135-X

[Nyquist 1924] NYQUIST, H.: Certain factors affecting telegraph speed. In: *Bell System Technical Journal* Bd. 3, 1924, S. 324–346. – Erste Formulierung der notwendigen Bandbreite für Informationsübertragung

[Nyquist 1928] NYQUIST, H.: Certain topics in telegraph transmission theory. In: *Trans. AIEE* Bd. 47, 1928, S. 617–644. – Erste Formulierung des Abtasttheorems, noch kein Beweis

[Oppenheim und Schafer 1975] OPPENHEIM, Alan V. ; SCHAFER, Ronald W.: *Digital Signal Processing.* Prentice Hall, 1975

[Orfanidis 1988] ORFANIDIS, Sophocles J.: *Optimum Signal Processing.* 2. Aufl. New York : Macmillan, 1988. – Sehr ausführliches Buch über stochastische Signale und Prozesse. – ISBN 0-02-389380-X

[Petrova 1973] PETROVA, S. S.: From the history of the analytic proofs of the fundamental theorem of algebra (Russian). In: *History and methodology of the natural sciences* Bd. XIV: Mathematics, mechanics. Moskau, 1973, S. 167–172. – Zum Fundamentalsatz der Algebra

[Proakis und Minolakis 1996] PROAKIS, John G. ; MINOLAKIS, Dimitris G.: *Digital Signal Processing.* Prentice Hall, 1996

[Remmert 2001] REMMERT, Reinhold: *Funktionentheorie Bd. 1.* 5. Aufl. Heidelberg : Springer, 2001. – Klassiker der Funktionentheorie, mit historischen Anmerkungen. – ISBN 3540418555

[Schüßler 1994] SCHÜSSLER, Hans-Wilhelm: *Digitale Signalverarbeitung 1.* 4. Aufl. Berlin : Springer, 1994. – ISBN 3-540-57428-X

[Shannon 1949] SHANNON, C.E.: Communication in the presence of noise. In: *Proc. IRE* Bd. 1, 1949, S. 10–21. – Der Klassiker zum Abtasttheorem, mit Beweis. Gilt als Grundlage der Informationstheorie.

[Wendemuth 2004] WENDEMUTH, Andreas: *Grundlagen der Stochastischen Sprachverarbeitung.* München : Oldenbourg Wissenschaftsverlag, 2004. – ISBN 3486274651

Index