LEHRBUCH

Rüdiger Seydel

Höhere Mathematik im Alltag

Vom Regenbogen bis zur digitalen Bildkompression

Springer Spektrum

Höhere Mathematik im Alltag

Rüdiger Seydel

Höhere Mathematik im Alltag

Vom Regenbogen bis zur digitalen Bildkompression

Springer Spektrum

Rüdiger Seydel
Mathematisches Institut
Universität zu Köln
Köln, Nordrhein-Westfalen, Deutschland

ISBN 978-3-662-64048-7 ISBN 978-3-662-64049-4 (eBook)
https://doi.org/10.1007/978-3-662-64049-4

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

Planung/Lektorat: Annika Denkert
Springer Spektrum ist ein Imprint der eingetragenen Gesellschaft Springer-Verlag GmbH, DE und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

Vorwort

Der Titel *Höhere Mathematik im Alltag* bedarf der Erklärung. Was meint Alltag, was bedeutet Höhere Mathematik? „Alltag" ist hier umfassend gemeint und nicht auf den privaten Alltag beschränkt. Das Spektrum der Themen reicht vom Naturereignis eines Regenbogens bis zur digitalen Methodik der Bildverarbeitung. Zum Alltag gehören auch Körperfunktionen wie die von Herz und Nerven, und klassische Maschinen, so etwa der Wankelmotor. Manche Beispiele sind eher einfache Konstruktionen, z. B. der Tonarm eines Plattenspielers, andere sind komplizierte Erfindungen wie das Farbfernsehen. Einiges ist abstrakt wie Frequenzmodulation, anderes anschaulich wie der Regenbogen oder die Kodierung von Farbbildern.

Ausgewählt wurden hier Themen des Alltags, deren Funktionsprinzipien sich gut mit Mathematik beschreiben lassen. Das Schwergewicht der Diskussion liegt auf den mathematischen Methoden, nicht auf den technischen oder biologischen Feinheiten der jeweiligen Fallstudie. Insofern ist dies kein Buch etwa über Bildkompression oder Herzschlag, sondern über ihre interessanten mathematischen Aspekte. Der Fokus auf den *Erkenntnisgewinn durch Mathematik* erlaubt es auch, einige Themen zu diskutieren, die heute vielleicht schon etwas veraltet sind. Der Kern ist jeweils die zugrunde liegende Mathematik, und die ist zeitlos.

Höhere Mathematik ist ein Terminus, der in der Universitätsausbildung von Ingenieuren klar definiert ist und Analysis, Lineare Algebra und gewöhnliche Differenzialgleichungen umfasst. Entsprechend spielen für dieses Buch Partielle Differenzialgleichungen und ausgefeilte Methoden der Numerik und der Stochastik keine Rolle. Die Beschränkung auf *Höhere Mathematik* bedingt eine Begrenzung der Fallstudien in Auswahl und Tiefe.

Die Thematik ist auch angeregt durch das Buch *R. Seydel, R. Bulirsch: Vom Regenbogen zum Farbfernsehen* aus dem Jahre 1986[1]. Schwerpunkt waren Themen aus Naturwissenschaft und Technik. Seit dieser Zeit hat sich die Mathematik neue Anwendungsbereiche erobert. Aktuellere Themen in dem vorliegenden Buch *Höhere Mathematik im Alltag* widmen sich zum Beispiel der Verarbeitung von Daten, einem wichtigen „Rohstoff" unserer Zeit. Da ist zum einen die Erfassung von Strukturen in großen Datenmengen, beispielsweise in

[1]Quelle: TUM-Skriptum des Autors von 1983.

Aktienkursen, und die JPEG-Komprimierung von Fotos. Ein anderes aktuelles Thema befasst sich mit der Schätzung und Filterung von Daten, die zum Beispiel aus Messungen stammen. Bei solchen Methoden und Modellen sind die Vorgaben aus Naturwissenschaft und Technik schwächer oder fehlen ganz, wie bei der Simulation von Wanderungsbewegungen in der Gesellschaft.

Das Buch stellt sich auch der historischen Entwicklung. So ist zum Beispiel das PAL-System beim Farbfernsehen, vor 50 Jahren eine geniale kulturelle Leistung, beim Digital-Fernsehen obsolet geworden. Trotzdem habe ich mich entschlossen, diese erfolgreiche Erfindung zu diskutieren, wegen des Erkenntnisgewinns durch Mathematik. Ähnliches gilt auch für den Wankelmotor. Solche Beispiele sind Technikgeschichte und behalten dennoch ihren Rang, insbesondere auch weil ihre Mathematik nicht veraltet.

Charakteristisch für eine Fallstudie ist, dass die jeweilige Anwendung und die benötigte Mathematik synchron diskutiert werden; die Mathematik wird nicht von der Anwendung getrennt. Der Index dieses Buches mit der Vielfalt der Begriffe mag illustrieren, wie verwoben die Mathematik mit verschiedensten Gebieten aus Natur und Technik ist. Mathematik und Anwendung inspirieren sich gegenseitig. Derart aufgebaute Fallstudien bilden das Gerüst des Buches. Die Essenz vieler Fallstudien ist in Übungsaufgaben niedergelegt, eine Tradition nicht nur an der Technischen Universität München.

Zur Beschreibung der jeweiligen Funktionsprinzipien werden Begriffe wie Phasensprung, Grenzzykel, Oberschwingung, Randmaximum oder Reihenentwicklung verwendet. Solche Phänomene können in der notwendigen Klarheit nur in der Sprache der Mathematik erläutert werden. Entsprechend sind die Ziele dieses Buches gesetzt: Durch das Studium mathematischer Fragestellungen soll das Verständnis wesentlicher Aspekte der behandelten Themen gefördert werden.

Durch die in die Fallstudien eingestreuten Aufgaben ist dieses Buch auch zum Selbststudium geeignet: Die Leserinnen und Leser sollten dabei mit Papier und Bleistift lesen, nach dem Studium jeder Aufgabenstellung mit der Lektüre innehalten und sich zunächst selbst an der Lösung versuchen. Auf diese Weise werden einerseits wichtige Gebiete angewandter Mathematik eingeübt, andererseits wird das Verständnis der zum Teil recht schwierigen Probleme erleichtert, zumal bei den Rechnungen nicht alle Zwischenschritte aufgeführt werden.

Die Reihenfolge der Fallstudien versucht mit einer Aufteilung nach Themenstellung, mathematischen Methoden und aufsteigender Schwierigkeit gleichzeitig verschiedenen Kriterien gerecht zu werden. Die jeweils zur Bearbeitung benötigten mathematischen Kenntnisse werden im Anhang in einer Tabelle aufgelistet. Diese Kenntnisse werden durch Vorlesungen oder Lehrbücher zur Höheren Mathematik vermittelt. Im Anhang finden sich einige Literaturhinweise dazu.

Mehr als 80 Abbildungen wurden mithilfe der Software *xfig* und *gnuplot* neu erstellt. Die Berechnungen habe ich unter *fortran* durchgeführt, bis auf die Singulärwertzerlegung, für die *matlab* eingesetzt wurde.

Köln
im Mai 2021

Rüdiger Seydel

Danksagung

Professor Roland Bulirsch hat im Rahmen von Lehrveranstaltungen zur Höheren Mathematik in den 1970er Jahren einige der Themen angeregt und die Tradition der Technischen Universität München weitergegeben. Mein langjähriger Freund und Kollege Klaus-Dieter Reinsch[2] hat das Manuskript aufmerksam studiert und mit seinem Wissen und seiner Erfahrung viele wertvolle Vorschläge beigesteuert. Ich bin Roland Bulirsch und Klaus-Dieter Reinsch zu großem Dank verpflichtet.

[2] 1951–2020.

Notationen

Die meisten Variablen, wie x, t, y, sind reelle Größen. Dabei ist t oft die Zeitvariable. Die reellen Zahlen werden mit dem Symbol $\mathbb{R}$ erfasst. Die Größen i, j, n, m, ν sind meist natürliche Zahlen, das geht jeweils aus dem Zusammenhang hervor. Komplexe Größen tauchen nur am Ende von Kap. 15 auf. Wenn x ein Spaltenvektor ist, dann ist x^{tr} der transponierte Vektor, also Zeilenvektor. Analog bezeichnet A^{tr} die transponierte Version einer Matrix A.

Dezimalzahlen werden in diesem Buch mit Dezimalkomma geschrieben, wie

$$\pi = 3{,}14159265$$
$$\sqrt{2} = 1{,}41421356$$

(gerundet).
So viele Dezimalstellen wie oben für π und $\sqrt{2}$ notiert, sind für die Ausführungen in diesem Buch nicht notwendig. Alle Dezimalbrüche werden auf eine übersichtliche Stellenzahl gerundet. Wenn wir beispielsweise schreiben

$$\gamma = 25{,}9,$$

dann heißt das im Allgemeinen *nicht* $\gamma = 25{,}90000000$! Fast alle angegebenen Dezimalzahlen sind Näherungen. Mit dieser generellen Verabredung vermeiden wir die ständige Wiederholung von „gerundet“. Die Bezeichnung $a \doteq b$ bedeutet Gleichheit bis zur letzten Stelle des für die Berechnung verwendeten Rechners.

Achsen der Figuren: Wenn in der Legende beispielsweise steht, dass $\varphi(t)$ aufgetragen ist, dann ist t auf der horizontalen Achse und φ auf der vertikalen Achse aufgetragen. Ansonsten sind die Achsen klar aus dem Zusammenhang, oder sie werden extra erklärt.

Qualifizierende Attribute wie beispielsweise „gering“ werden im Allgemeinen nicht quantifiziert, da dies der Diskussion nicht angemessen wäre. Auf Anführungsstriche verzichten wir hier.

Inhaltsverzeichnis

Regenbogen

1

Fallen die Strahlen der Sonne auf fein verteilte Wassertröpfchen, kann man, sofern man sich an an einem günstigen Standort befindet, ein prächtiges Naturereignis bewundern: den Regenbogen (Abb. 1.1). Dieses Schauspiel ist schon im Kleinen bei einem Wasserfall oder beim Rasensprengen sichtbar; die Wirkung ist natürlich weitaus großartiger bei Regen, wenn der Bogen – die Brücke am Himmel zur Erde hatten ihn die alten Völker genannt – in leuchtenden Farben am Himmel erscheint.

Die Höhe des Bogens am Himmel, manchmal sieht man dort sogar zwei Bögen, lässt sich berechnen, dabei wird deutlich, wie dieser farbige Bogen entsteht.

Brechungsgesetz

Zur Erklärung dieses Phänomens können wir uns sowohl auf geometrische Beziehungen als auch auf Extremalprinzipien stützen. Das Ineinandergreifen dieser beiden mathematischen Aspekte wollen wir zunächst bei der Lichtbrechung studieren. Das Brechungsgesetz (Snellius 1621) ist für die Entstehung des Regenbogens von grundlegender Bedeutung. Ohne Beschränkung der Allgemeinheit lässt sich das Brechungsgesetz an dem folgenden speziellen Fall erläutern (Abb. 1.2):

Aufgabe 1 *Ein Lichtstrahl werde auf dem Weg von* $P = (-1, 1)$ *nach* $Q = (1, -1)$ *an der* x*-Achse gebrochen. Aus der Forderung, dass er seinen Weg in minimaler Zeit zurücklegen soll, leite man das Brechungsgesetz der Optik ab. (Die Geschwindigkeit des Lichtstrahles sei* v_1 *für* $y > 0$ *und* v_2 *für* $y < 0$*, und es gelte* $0 < v_2 < v_1$*.)*

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_1

Abb. 1.1 Regenbogen mit Nebenregenbogen am Villehang Kendenich in Hürth bei Köln (Foto: Friederike Seydel 2007)

In dem Szenario von Abb. 1.2 liegt für $y < 0$ das optisch dichtere Medium, mit der geringeren Lichtgeschwindigkeit v_2. Der Lichtstrahl durchstößt die Trennfläche beider Medien im Punkt $D = (x, 0)$. Das Licht hat somit die beiden Strecken

$$\overline{\mathrm{PD}} = \sqrt{(1+x)^2 + 1} = \frac{1+x}{\sin\alpha_1}$$
$$\overline{\mathrm{DQ}} = \sqrt{(1-x)^2 + 1} = \frac{1-x}{\sin\alpha_2}$$

zurückzulegen. Die Brechungswinkel α_1, α_2 hängen von x ab: $\alpha_1(x), \alpha_2(x)$.

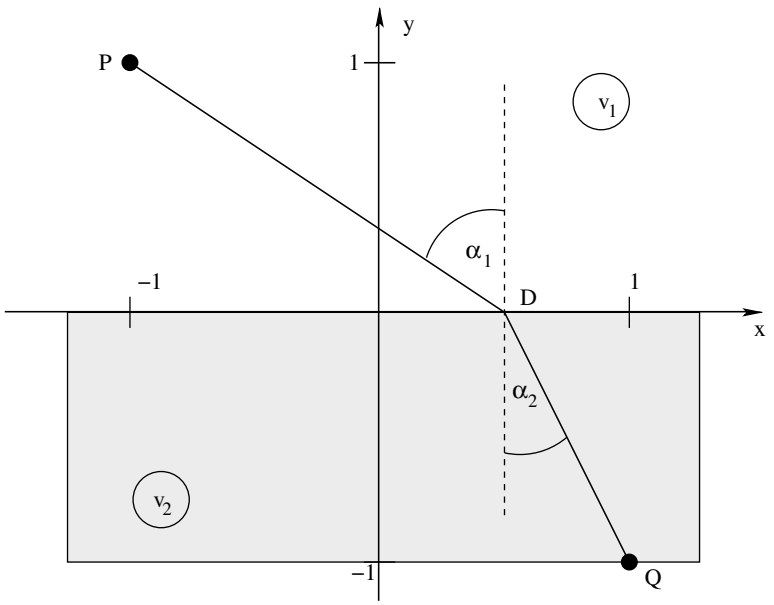

Abb. 1.2 Zum Brechungsgesetz von Snellius: Schnitt durch zwei Medien, deren Lichtgeschwindigkeiten v_1 und v_2 sind. Die x-Achse repräsentiert die (ebene) Grenzfläche zwischen den Medien. Ein Lichtstrahl führt von P nach Q; gestrichelt ist das Lot auf die Grenzfläche im Punkt D. Das Lot und der Punkt P definieren die gezeigte Schnittebene

Die Gesamtzeit T, die der Lichtstrahl von P nach Q benötigt, setzt sich aus den Quotienten von Weg und Geschwindigkeit zusammen, also

$$T(x) = \frac{\overline{\mathrm{PD}}}{v_1} + \frac{\overline{\mathrm{DQ}}}{v_2} = \frac{1}{v_1}\sqrt{(1+x)^2+1} + \frac{1}{v_2}\sqrt{(1-x)^2+1}.$$

Ein Extremum dieser Funktion ist bestimmt durch die Beziehung

$$\begin{aligned} 0 = \frac{\mathrm{d}T}{\mathrm{d}x} &= \frac{1}{v_1}\,\frac{x+1}{\sqrt{(1+x)^2+1}} + \frac{1}{v_2}\,\frac{x-1}{\sqrt{(1-x)^2+1}} \\ &= \frac{1}{v_1}\sin\,\alpha_1 - \frac{1}{v_2}\sin\,\alpha_2. \end{aligned}$$

Hieraus folgt das Brechungsgesetz

$$\frac{v_1}{v_2} = \frac{\sin\,\alpha_1}{\sin\,\alpha_2}.$$

Für große Werte von $|x|$ ist die Zeit $T(x)$ groß; T hat nur ein globales Minimum. Für den Quotienten $R := v_1/v_2$ gilt bei der getroffenen Anordnung $R > 1$.

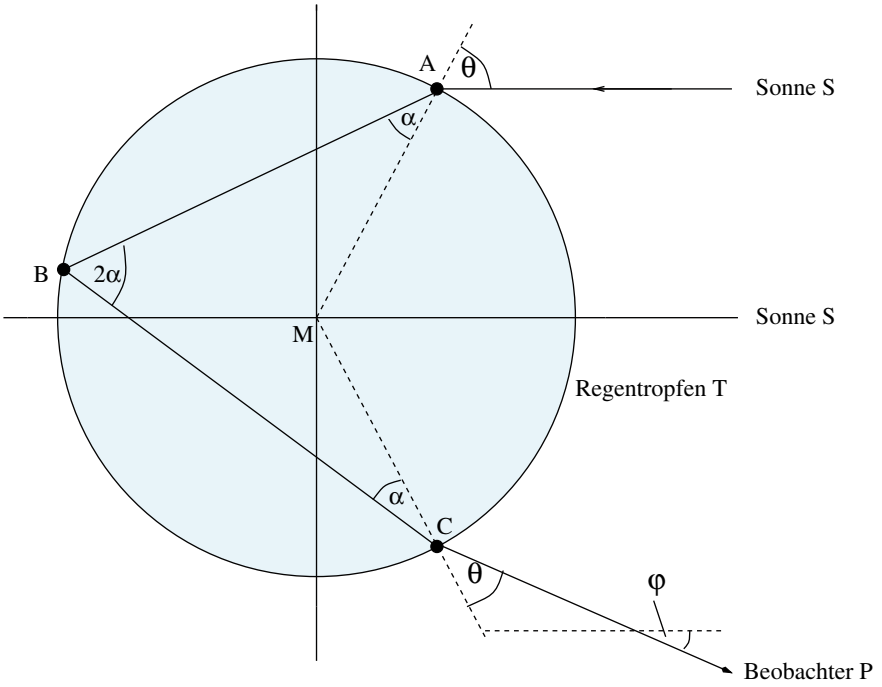

Abb. 1.3 Wassertropfen und Sonnenstrahl

Entstehung des Regenbogens

Zur Entstehung des Regenbogens vergleiche Abb. 1.3. Ein Lichtstrahl, der auf einen als kugelförmig angenommenen Regentropfen trifft, wird zum einen Teil reflektiert, zum anderen Teil ins Innere gebrochen. Der innere Teilstrahl wird, wenn er erneut die Oberfläche berührt, zum Teil ins Innere reflektiert, zum Teil verlässt das Licht den Wassertropfen. Diese Reflektion kann sich noch einige Male wiederholen, die Stärke des inneren Reststrahles nimmt dabei ab. Der Beobachter sieht insbesondere die Strahlen, die den Tropfen nach genau dreimaliger Berührung bzw. Durchdringung der Oberfläche (also eine Reflexion) verlassen; diese Strahlen bilden den Hauptregenbogen. Der schwächere Nebenregenbogen entsteht durch die Strahlen, welche die Oberfläche des Tropfens viermal berühren bzw. durchsetzen; dies entspricht einer zweimaligen Reflexion ins Innere.

Der betrachtete Regentropfen wird auf seiner ganzen Halbfläche von parallelen Sonnenstrahlen beschienen. Je nach dem, ob einfallende Sonnenstrahlen den Tropfen in der Nähe des Randes oder mehr in der Mitte treffen, unterscheiden sich die Richtungen *(Streuwinkel)* der Strahlen, die den Tropfen wieder verlassen. Diese Streuwinkel werden im Folgenden berechnet, insbesondere die Richtung ihrer maximalen Helligkeit.

Wir betrachten *eine* Regentropfen-Kugel mit Mittelpunkt M (vgl. Abb. 1.3). Die Sonne S (als unendlich ferner Punkt) und M definieren eine Gerade $\mathcal{G}$. Der Tropfen wird in voller Breite von zu $\mathcal{G}$ parallelen Sonnenstrahlen beschienen. Nehmen wir

einen Strahl heraus, der den Tropfen im Punkt A trifft. Die drei Punkte S, M, A (bzw. die Gerade $\mathcal{G}$ und der Punkt A) definieren eine Ebene. Wegen der Symmetrie der Anordnung wird die Fortsetzung des Lichtstrahls im Inneren der Ebene bleiben. Diese Ebene ist in der Abbildung dargestellt. Der Eintrittswinkel des Sonnenstrahls im Punkt A wird hier mit θ bezeichnet.

Der ankommende Strahl wird in A zum Teil nach außen reflektiert, zum Teil nach innen gebrochen. Bei dieser und bei folgenden Reflexionen (wie in B) verliert der Strahl jeweils Energie, weil ein Teil nach außen gebrochen wird. Im Folgenden diskutieren wir die Winkel derjenigen Strahlenfolge, welche die Tropfen-Oberfläche genau dreimal berührt (Abb. 1.3).

Nach dem Brechungsgesetz von Snellius genügen die Winkel θ und α der Beziehung

$$\frac{\sin\theta}{\sin\alpha} = R, \text{ z. B. } R = 1{,}331 \quad \text{für rotes Licht, blau: } R = 1{,}343.$$

Ein Lichtstrahl in Richtung des Beobachters P verlässt den Regentropfen unter dem Streuwinkel φ, der eine Funktion f des Eintrittswinkels ist,

$$\varphi = f(\theta), \ -\frac{\pi}{2} < \theta < \frac{\pi}{2}.$$

Der Streuwinkel φ ist gegen die Sonnenrichtung gemessen. Die Intensität dieser Strahlung ist maximal für θ^* mit $f'(\theta^*) = 0$. Der zugehörige Winkel $\varphi^* = f(\theta^*)$ maximaler Helligkeit gibt die „Höhe“ des Regenbogens an.

Aufgabe 2 *Man zeige, dass für den Streuwinkel gilt:*

$$\varphi = f(\theta) = 2\theta - 4\arcsin\left(\frac{\sin\theta}{R}\right).$$

Für die 4 Teilstrecken ergeben sich die folgenden Winkel:

$$\begin{aligned} &\text{SA}: \quad \text{Winkel} = \pi \\ &\text{AB}: \qquad\quad \pi + \theta - \alpha \\ &\text{BC}: \qquad\quad 2\pi + \theta - 3\alpha \\ &\text{CP}: \qquad\quad 2\pi + 2\theta - 4\alpha. \end{aligned}$$

Bis auf den Summanden 2π gilt also $\varphi = 2\theta - 4\alpha$. Wie bereits erwähnt, genügen die Winkel α und θ nach dem Brechungsgesetz der Beziehung

$$\sin\alpha = \frac{\sin\theta}{R},$$

also gilt für den Streuwinkel

$$\varphi = f(\theta) = 2\theta - 4\arcsin\left(\frac{\sin\theta}{R}\right).$$

Die unabhängige Variable θ gibt an, wo der von der Sonne kommende Lichtstrahl den Tropfen trifft, etwa am Rand $\left(|\theta| \approx \frac{\pi}{2}\right)$ oder eher in der Mitte ($\theta \approx 0$). Jedem Wert von θ entspricht ein Kleinkreis auf der Tropfen-Oberfläche. Da die Funktion $f(\theta)$ ungerade ist [$f(-\theta) = -f(\theta)$] und weil die Streuwinkel für negative Winkel θ vom Beobachter wegzeigen, genügt es, $f(\theta)$ für

$$0 \leq \theta < \frac{\pi}{2}$$

näher zu untersuchen.

Für die Beobachtung des Regenbogens ist wesentlich, in welchem Winkelbereich die Streuwinkel auftreten können. Um den Wertebereich

$$\varphi_{\min} \leq \varphi \leq \varphi_{\max}$$

bestimmen zu können, berechnen wir mit Hilfe der Ableitung

$$f'(\theta) = 2 - \frac{4}{\sqrt{1 - \frac{\sin^2\theta}{R^2}}} \cdot \frac{\cos\theta}{R}$$

mögliche Extremwerte von $f(\theta)$. Da für den betrachteten Winkelbereich $\cos\theta > 0$ gilt und arccos eine monoton fallende Funktion ist, gelten die folgenden Umformungen:

$$\begin{aligned}
f'(\theta) &\geq 0 \\
2R\sqrt{1 - \frac{\sin^2\theta}{R^2}} &\geq 4\cos\theta \\
\sqrt{R^2 - \sin^2\theta} &\geq 2\cos\theta \\
R^2 - \sin^2\theta &\geq 4\cos^2\theta \\
R^2 - 1 &\geq 3\cos^2\theta \\
\sqrt{\frac{R^2-1}{3}} &\geq \cos\theta \\
\arccos\sqrt{\frac{R^2-1}{3}} &\leq \theta.
\end{aligned}$$

Für rotes Licht ($R = 1{,}331$) ergibt sich für den Zahlenwert des extremalen Einfallswinkels

$$\theta^* := \arccos\sqrt{\frac{R^2-1}{3}} \approx 1{,}0389$$

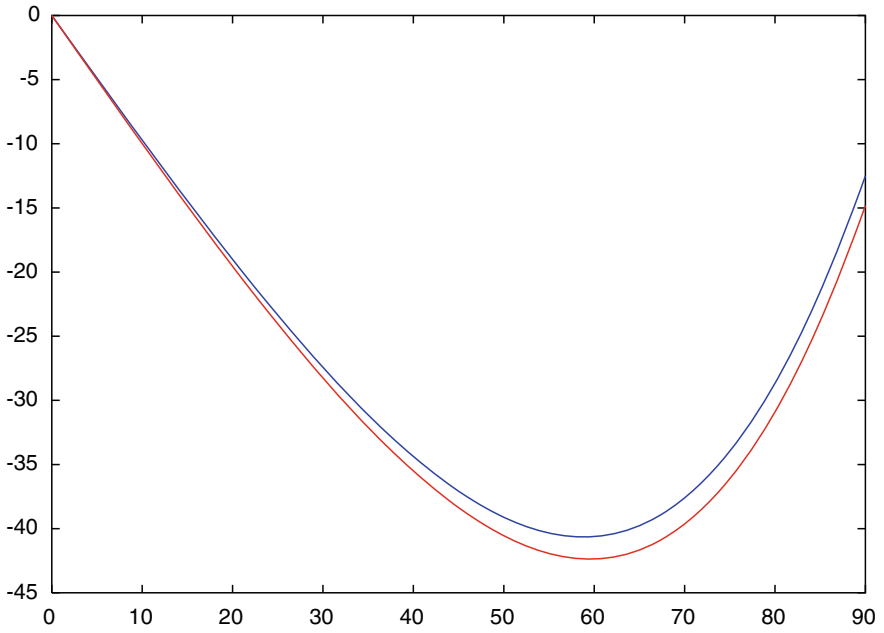

Abb. 1.4 Der Streuwinkel φ aufgetragen über dem Einfallswinkel θ, für rotes und für blaues Licht, jeweils im Gradmaß

Im Gradmaß ist das Winkel 59,5°. Wegen

$$\begin{aligned} f'(\theta^*) &= 0 \\ f'(\theta) &> 0 \text{ für } \theta^* < \theta \\ f'(\theta) &< 0 \text{ für } \theta^* > \theta \end{aligned}$$

handelt es sich bei $f(\theta^*)$ um ein Minimum, der Zahlenwert im Bogenmaß ist

$$\varphi^* = \varphi_{\min} = f(\theta^*) \approx -0{,}7395,$$

und im Gradmaß[1] ist das der Winkel $-42{,}4°$. Maxima von $f(\theta)$ liegen am Rand des Bereiches $0 \leq \theta < \frac{\pi}{2}$. Von den zugehörigen Werten $f(0)$ und $f\left(\frac{\pi}{2}\right)$ ist $f(0) = 0$ das absolute Maximum. Also gilt für den Bereich der Streuwinkel:

$$-42{,}4° \leq \varphi \leq 0 \text{ bzw. } |\varphi| \leq 42{,}4°.$$

In der Abb. 1.4 ist der Verlauf von $\varphi = f(\theta)$ aufgetragen, für rotes und für blaues Licht, jeweils in Grad.

[1] Im Folgenden, wegen der größeren Anschaulichkeit, wird auch häufig das Gradmaß verwendet.

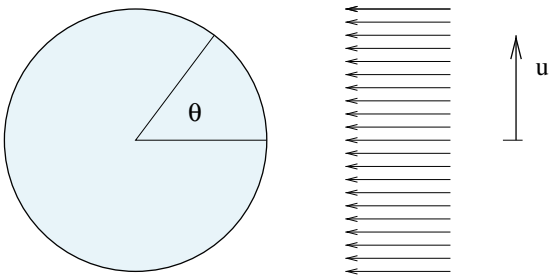

Abb. 1.5 Zur Intensität der Regenbogenstrahlung: Regentropfen mit einfallendem Sonnenlicht

Für die anderen Farben des Regenbogens ist der Bereich der Streuwinkel kleiner. So ergeben sich für Blau ($R = 1{,}343$) die Werte $\theta^* = 1{,}0268$ (oder $58{,}8°$) und $\varphi(\theta^*) = -0{,}7094$ (oder $-40{,}6°$).

In der Nachbarschaft eines Extremums ist die Änderung der Funktion gering. Das bedeutet in unserem Fall, dass für Winkel θ aus einem breiten Bereich um θ^* die zugehörigen Streuwinkel sich nur geringfügig unterscheiden (vgl. Abb. 1.4). Somit ist der Winkel $\varphi^* = f(\theta^*)$ durch eine große Anzahl nahezu paralleler Strahlen und deshalb durch hohe Lichtintensität ausgezeichnet. Diese Intensität reicht aus, um den Betrachter in Höhe φ^* (über der Verlängerung der Geraden Sonne-Beobachter) den vom Tropfen ausgehenden Lichtstrahl erkennen zu lassen.

Nun hat die Regenwolke viele Tropfen, und viele davon auf dem durch φ^* definierten Kegel. So sieht der Betrachter einen farbigen Kreisbogen, den Regenbogen.

Intensität

Die Intensität des Lichtstrahls, den der Beobachter sieht, ist eine nähere Untersuchung wert. Für einen vom Regentropfen ausgestrahlten schmalen Winkelbereich $\Delta\varphi$ stellt sich die Frage, wieviele Sonnenstrahlen eingehen. Tatsächlich ist nicht der Winkel θ ein Maß für die Menge der einfallenden Strahlen, sondern die Größe u senkrecht zu diesen Strahlen, $u = \sin\theta$ (Abb. 1.5). Die Intensität der Regenbogenstrahlung wird also durch die Funktion $f(\arcsin u)$ beschrieben, durch

$$g(u) := f(\arcsin u) = 2\arcsin u - 4\arcsin\frac{u}{R},\ 0 \le u < 1.$$

(Zur Vereinfachung betrachten wir einen „Einheitsregentropfen“ mit Radius 1, siehe Abb. 1.5.) Der Graph dieser Funktion g (in Abb. 1.6) ähnelt dem oben in Abb. 1.4 gezeichneten, er ist lediglich etwas verzerrt.

Die Intensität kann auf einfache Weise quantifiziert werden: Mit Hilfe der Linearisierung von $g(u)$ erhält man die Beziehung

$$\Delta g := g(u_1) - g(u_2) \approx g'(u_2)(u_1 - u_2) =: g'(u_2)\Delta u.$$

Hieraus folgt für die Lichtintensität

$$\Delta u \approx \frac{1}{g'(u)}\,\Delta g,$$

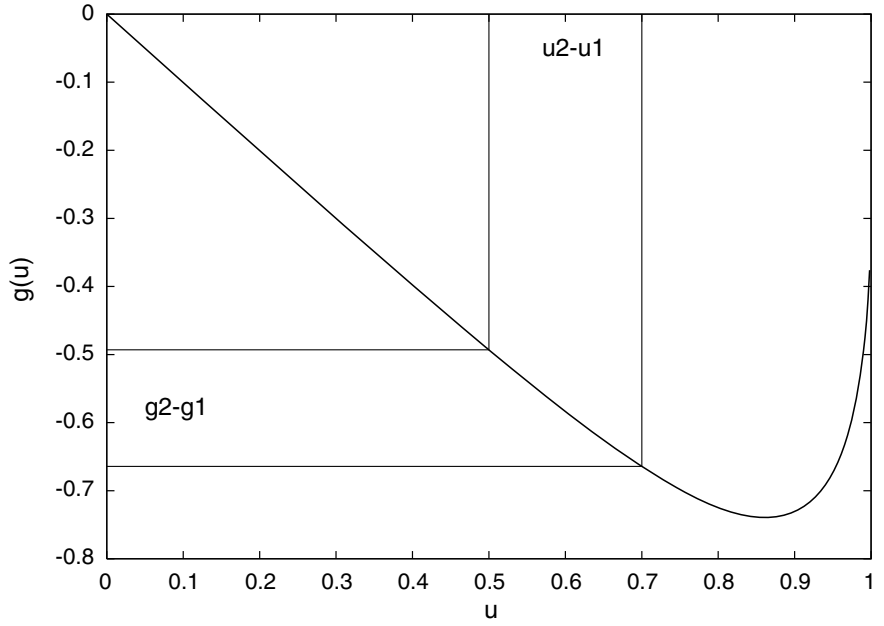

Abb. 1.6 Intensität g des Regenbogenstrahls; u ist der normierte Abstand zur zentralen Achse, vergleiche Abb. 1.5. Die Beziehung zwischen einem Abstand $u_2 - u_1$ und den entsprechenden g-Werten wird kritisch am Minimum des Graphen. Dort tragen sehr viele eingehende Strahlen zur Intensität des Streu-Strahls bei

d. h. für den Wert u^* mit $g'(u^*) = 0$ wird die Intensität des Streuwinkels maximal. Man kann es auch andersherum sagen: Für alle anderen Winkel ist die Intensität zu schwach, um vom Beobachter bemerkt zu werden. Ein breiter Bereich von einfallenden Sonnenstrahlen (Δu groß) konzentriert sich auf eine kleine Umgebung der Streuwinkel (Δg bzw. $\Delta\varphi$ klein). Der sichtbare Winkelbereich von φ besteht im Wesentlichen nur aus $\varphi^* = f(\theta^*)$. Die Strahlung des restlichen Bereiches ($f(\theta^*) < \varphi \leq 0$) interferiert mit den Strahlen anderer Wellenlängen zu weißem Licht; aus diesem Grunde erscheint das Himmelslicht „innerhalb“ des Regenbogens heller.

Nachdem wir die Geometrie der Brechung in der Ebene diskutiert haben, sei festgehalten, dass für jede Ebene durch $\mathcal{G}$ der gleiche Winkel φ^* resultiert. Dies definiert einen Kreiskegel von Lichtstrahlen der gleichen Farbe, mit Öffnungswinkel $2\varphi^*$.

Nun tritt die beobachtende Person P auf. Parallel zur Geraden $\mathcal{G}$ definiert sie die Gerade $\mathcal{G}_\mathrm{P}$. Diese Gerade $\mathcal{G}_\mathrm{P}$ definiert ein Büschel von Ebenen, die $\mathcal{G}_\mathrm{P}$ enthalten. In der Abb. 1.7 sind zwei Ebenen des Büschels skizziert. Wenn die Regenwolke groß und dicht genug ist, gibt es eine Schar solcher Ebenen, sodass auf jeder dieser Ebenen sich mindestens ein Tropfen T befindet, dessen mit φ^* emittierter Strahl genau P trifft. Deswegen ist die Höhe des Regenbogens, gemessen als Winkel über $\mathcal{G}_\mathrm{P}$, genau φ^*! Für jeden Tropfen gibt es eine solche Ebene, wie sie in Abb. 1.3 skizziert ist.

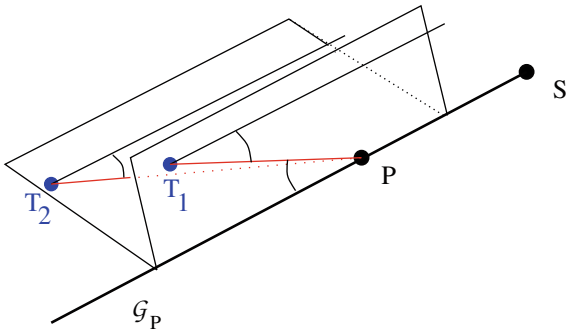

Abb. 1.7 Die beobachtende Person P sieht unter anderem Strahlen von Tropfen T_1 und T_2. Der angedeutete Winkel ist φ^*, und die Strahlen bilden einen Kegel

Diese Situation, illustriert in der Abb. 1.7, erklärt warum der Regenbogen ein Teil des Kreisbogens ist, und an welcher Stelle am Himmel er sichtbar ist.

Nebenregenbogen

Der interessierte Leser möge jetzt noch die entsprechenden Untersuchungen für den Nebenregenbogen durchführen, vergleiche die Illustration der Abb. 1.8.

Die Lösung sei kurz zusammengefasst: Für den Streuwinkel ψ des Nebenregenbogens gilt

$$\psi(\theta) = -\pi - 2\theta + 6 \arcsin\left(\frac{\sin\theta}{R}\right)$$

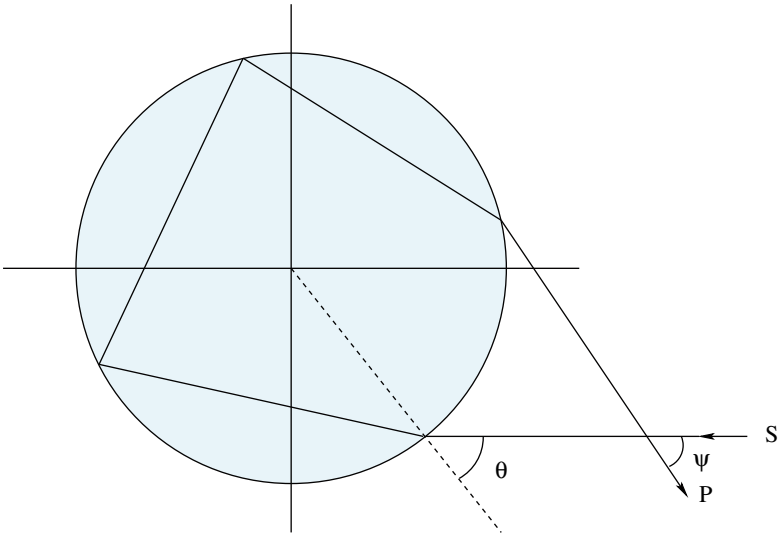

Abb. 1.8 Illustration zum Nebenregenbogen, Aufbau wie in Abb. 1.3

mit Extremum bei

$$\theta^* = \arccos\sqrt{\frac{R^2-1}{8}}.$$

Für rotes Licht ergibt sich ein Maximum bei $\theta^* = 1{,}255$ oder $71{,}9°$, mit $\psi(\theta^*) = -0{,}879$ oder $-50{,}4°$, und für blaues Licht $\theta^* = 1{,}248$ oder $71{,}5°$ mit $\psi(\theta^*) = -0{,}933$ oder $-53{,}5°$.

Die Farbreihenfolge ist hier umgekehrt gegenüber dem Hauptregenbogen, entsprechend ist das Extremum $\psi(\theta^*)$ ein Maximum, vergleiche die Abb. 1.8. Die Überlegungen bezüglich der Intensität gelten analog. Da der Himmel oberhalb des Nebenregenbogens wiederum heller erscheint, wirkt nur der Bereich zwischen den

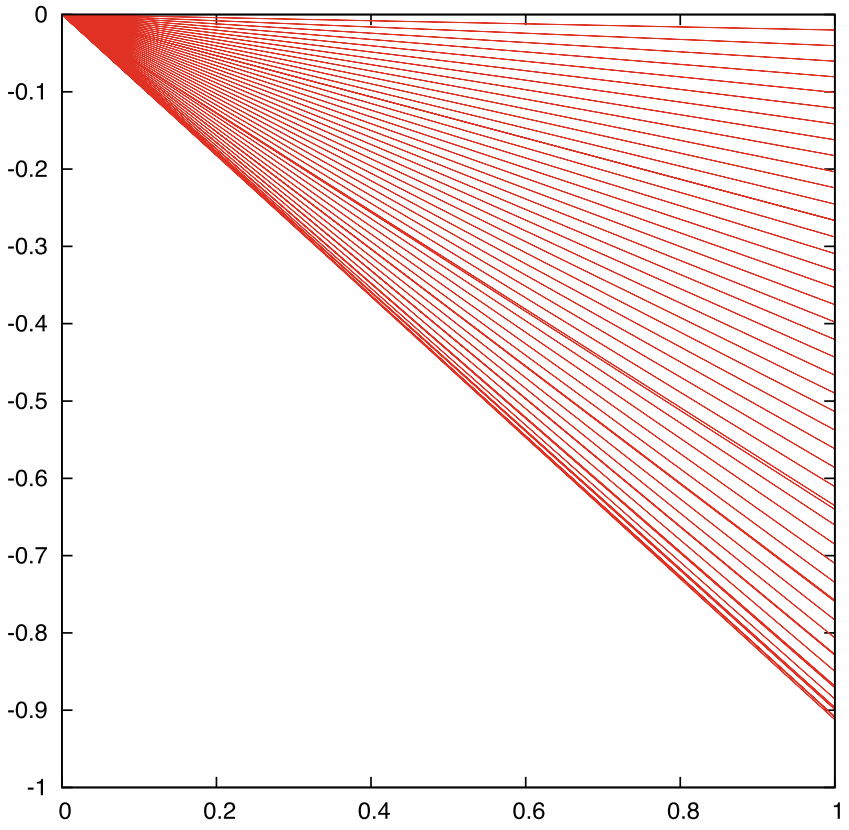

Abb. 1.9 (Geometrische Anordnung wie in Abb. 1.3) Ein fiktives Einheitsquadrat am Himmel definiert eine Ebene. Ein Regentropfen ist im Punkt (0, 0) angenommen. Man denke sich 50 gleichabständige Sonnenstrahlen in dieser Ebene, die waagerecht von rechts auf den Regentropfen fallen. Der Regentropfen ist hier punktförmig, und entsprechend sind die einfallenden Sonnenstrahlen nicht gezeigt, sie fallen mit der Oberkante des Quadrats zusammen. Im Bild dargestellt sind aber die jeweiligen ausgehenden roten Streu-Strahlen des Hauptregenbogens. Die Häufung der Strahlen an $\varphi^* = -42{,}4°$ deutet sich an. Die resultierende höhere Helligkeit dort kann im Druck nicht wiedergegeben werden

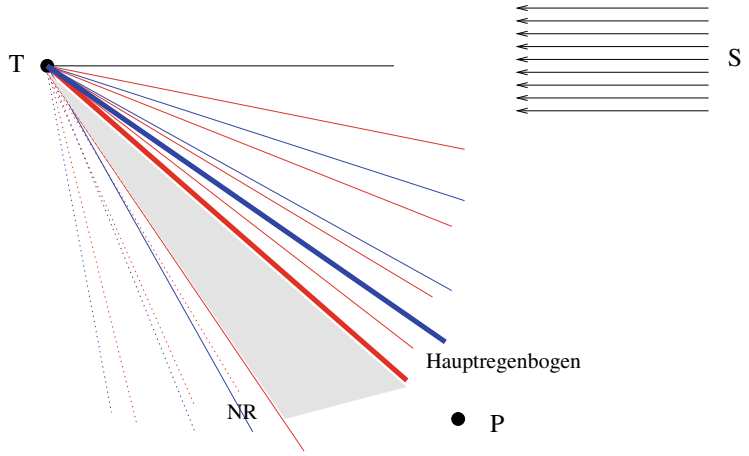

Abb. 1.10 Kaustiken von Hauptregenbogen (Strahlen durchgezogen, analog Abb. 1.9) und Nebenregenbogen (NR, Strahlen gepunktet), jeweils für rotes und blaues Licht, schematisch

beiden Bögen

$$42{,}4^\circ \leq \text{Höhe} \leq 50{,}4^\circ$$

als dunkleres Band. Es sei nochmals darauf hingewiesen, dass diese Höhenwinkel über der Verlängerung der Achse $\mathcal{G}_P$ Sonne-Beobachter gemessen sind. Diese Überlegungen erklären auch, warum man Regenbögen im Allgemeinen nur bei niedrigeren Sonnenständen beobachten kann.

Kaustik

Der Regenbogen kann auch als Kaustik erklärt werden. In der Abb. 1.9 ist ein einzelner Regentropfen mit dem von ihm ausgehenden Halb-Bündel von Strahlen illustriert. Die eingezeichneten Strahlen trennen jeweils 2 % der Lichtmenge des Halbbündels. Wie in der Abb. 1.9 sichtbar, wird der Abstand der Strahlen enger nahe an der Einhüllenden (Winkel $\varphi^* = -42{,}4^\circ$). Diese Konzentration von Helligkeit ist eine Kaustik. „Unterhalb“ der Kaustik treten keine Strahlen auf, dieses entspricht dem dunklen Band oberhalb (!) des Hauptregenbogens. Auch die Strahlen des Nebenregenbogens bilden eine Kaustik. In der Abb. 1.10 sind die Kaustiken beider Regenbögen angedeutet, mit dem dunklen Band dazwischen.

Literatur

Grundlage dieser Diskussion war der Artikel

Nussenzveig, H. M.: The Theory of the Rainbow. Sci. Am. **236**(4), 116–127 (1977)

Kontur des Kreiskolbenmotors

2

In der abendländischen Geschichte der Wissenschaft haben *Zykloide* eine prominente Rolle gespielt, waren doch Epizykeln die Grundlage der Astronomie, angefangen bei Hipparchos und Ptolemäus bis einschließlich Kopernikus.

Zykloiden spielten auch außerhalb der Astronomie eine Rolle. Die berühmte Aufgabe von Johann Bernoulli *„Man finde die Kurve kürzester Fallzeit, welche zwei gegebene Punkte verbindet“,* zog 1696/1697 das Interesse der scharfsinnigsten Mathematiker auf sich. Die Lösungskurve ist eine Zykloide. (Lösungen der Aufgabe wurden unter anderem von Leibniz und Newton eingesandt.) Eine andere klassische Anwendung von Zykloiden ist das Zykloidenpendel, bei ihm hat die Schwingungszeit für jede Amplitude denselben Wert (wichtig für Präzisionspendeluhren, Huygens 1673). Auch heute noch faszinieren Zykloiden, man denke nur, wie hübsch sich bei Nacht die Bewegung von seitlich beleuchteten, zwischen Fahrradspeichen geklemmten Reflektoren ausmacht.

Eine technische Anwendung von Zykloiden ist der Kreiskolbenmotor, populär geworden als „Wankelmotor“ (nach F. Wankel; weitere Erfindungen hierzu von H.D. Paschke). In der Kraftfahrzeug-Großserienfertigung hat sich der Wankelmotor bisher nicht durchgesetzt, aber wegen seiner kompakten Ausmaße ist er immer noch von Interesse.

Der Drehkolben des Wankelmotors hat einen Querschnitt, der einem gleichseitigen Dreieck ähnelt. Die Bewegung dieses Läufers setzt sich zusammen aus einer Drehung um die eigene Achse und einer Kreisbewegung der Achse; über einen Exzenter wird die Bewegung des Drehkolbens auf die Antriebsachse übertragen (schematisch in Abb. 2.2 und 2.3). Eine derartige Bewegung führt zu der Frage, wie die Abmessungen der Konstruktion zu wählen sind, damit die drei Ecken des Drehkolbens ständig die Gehäusewand berühren und auf diese Weise drei Kammern trennen. Die Kurve dieser Gehäusewand ist eine Zykloide (genauer: *Epitrochoide*).

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_2

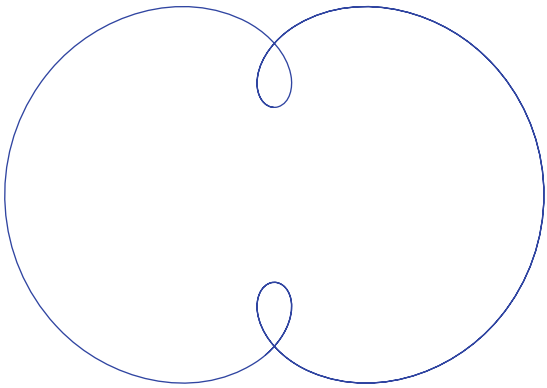

Abb. 2.1 Eine spezielle Zykloide (Definition später im Text)

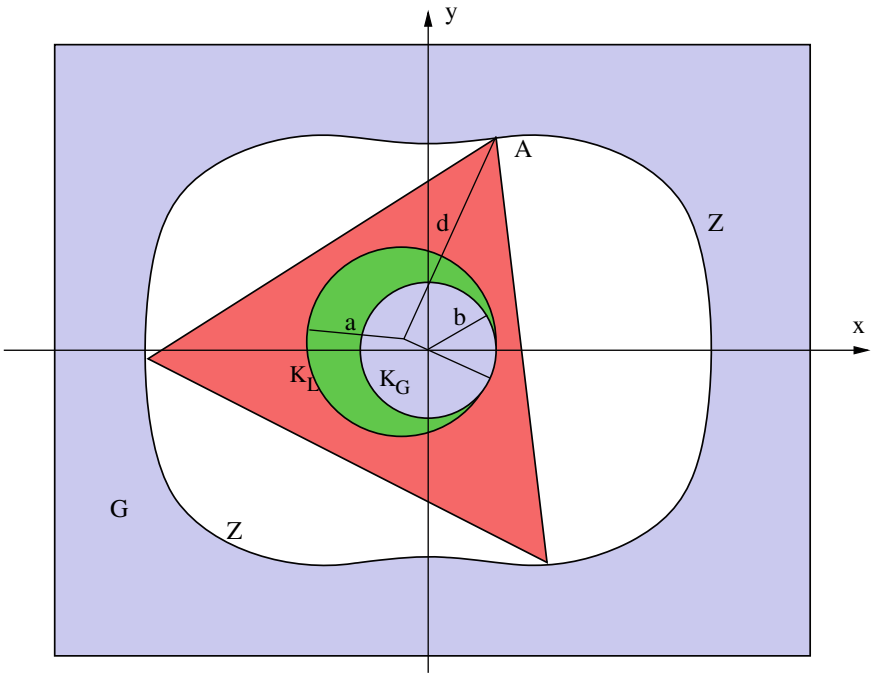

Abb. 2.2 Schematischer Aufbau eines Wankelmotors: Querschnitt 1. Der Rotationskolben, hier vereinfacht als (rotes) Dreieck dargestellt, sitzt auf einem Exzenter (grün), und wird von drei Verbrennungskammern (weiß) umgeben. G ist das Gehäuse. Der Kreis K_L mit Radius a rollt auf dem feststehenden Kreis K_G mit Radius b ab, dabei werden der exzentrische Läufer und der Rotationskolben in Drehung versetzt. Die Kurve Z, die der Punkt A beschreibt, ist eine Zykloide

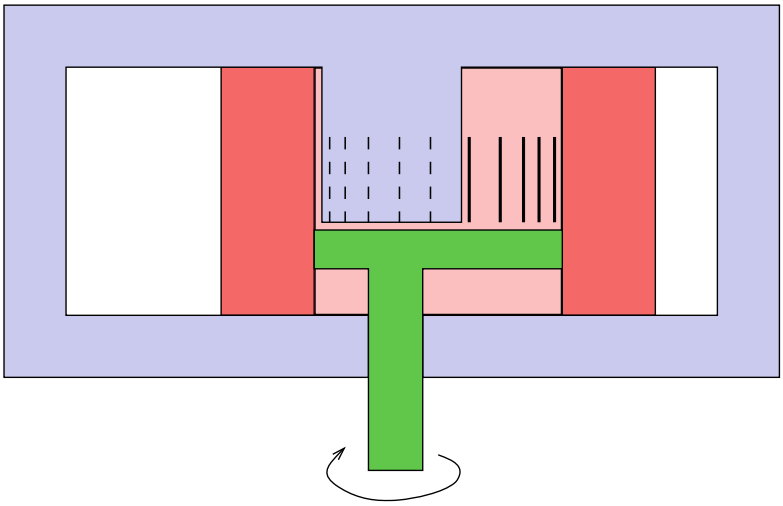

Abb. 2.3 Schematischer Aufbau eines Wankelmotors: Querschnitt 2, orthogonal zu Querschnitt 1 von Abb. 2.2; Farben analog. Der Zahnkranz ist angedeutet, der ein Durchrutschen des Zylinders verhindert

Die mathematische Fragestellung besteht in der Aufstellung und Diskussion einer Parameterdarstellung sowie in der Berechnung der Länge dieser Zykloide.

Aufgabe *Die Skizze von Abb. 2.2 zeigt einen Profilschnitt des Wankelmotors. Mit dem Gehäuse G ist der Kreis K_G mit Radius b starr verbunden, auf diesem Kreis rollt der Läuferkreis K_L mit Radius a ab. Ein Punkt A, der an K_L befestigt ist, beschreibt dabei die Kontur Z (Epitrochoide). Der Winkel t ist der Abrollwinkel des großen Kreises, α der des feststehenden kleinen Kreises (vergleiche auch die Abb. 2.4 und 2.5).*

a) *Man gebe α als Funktion von t an (Startposition $\alpha = t = 0$).*
b) *Man gebe eine Parameterdarstellung $x(t)$, $y(t)$ von Z an.*
c) *Für einen festen Punkt auf K_L berechne man den Umfang seiner Bahn bei einer vollen Umdrehung des Läufers (Zurückführung auf Bogenlänge einer Ellipse).*

(Zur Illustration sei $a = 3$; $b = 2$; $d = 7$ angenommen.)

Parameterdarstellung
Als Startposition nehmen wir $\alpha = t = 0$ an, diese Normierung wird zu einer achsensymmetrischen Lage der Zykloide führen. Beim Abrollen der beiden Kreise aufeinander sei vorausgesetzt, dass die Kreise nicht gleiten.[1] Dann sind $b\alpha$ und at

[1] kann durch Zahnkränze sichergestellt werden.

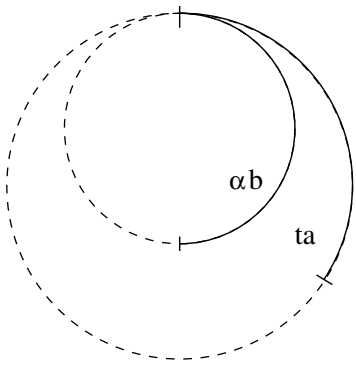

Abb. 2.4 Bogenlängen auf den Kreisen K_G und K_L

gleichlange Kreisbogenstücke, es gilt also (illustriert in Abb. 2.4)

$$\alpha = \frac{a}{b}t.$$

Um eine Parameterdarstellung der Zykloide zu konstruieren, ist die Abb. 2.5 hilfreich.

Betrachten wir das in Abb. 2.5 angezeigte (gelbe) Dreieck mit Hypotenusen-Länge $a - b$, dann ist die Summe der drei Innenwinkel π, also

$$\pi = (t - \beta) + \frac{\pi}{2} + (\pi - \alpha).$$

Deswegen gilt für die Winkel

$$\beta = \frac{\pi}{2} - (\alpha - t),$$

und nach den Gesetzen der Trigonometrie folgt

$$\cos\beta = \sin(\alpha - t), \quad \sin\beta = \cos(\alpha - t).$$

Die Koordinaten des Läufer-Eckpunktes A im $(\overline{x}, \overline{y})$-Koordinatensystem sind

$$(\overline{x},\ \overline{y}) = (d\cos\beta,\ d\sin\beta) = (d\sin(\alpha - t),\ d\cos(\alpha - t)).$$

Die Verschiebung der beiden Koordinatensysteme ist

$$\begin{aligned} x - \overline{x} &= -(a - b)\sin\alpha, \\ y - \overline{y} &= -(a - b)\cos\alpha. \end{aligned}$$

Zusammen haben wir

$$\begin{aligned} x &= d\sin(\alpha - t) - (a - b)\sin\alpha, \\ y &= d\cos(\alpha - t) - (a - b)\cos\alpha. \end{aligned}$$

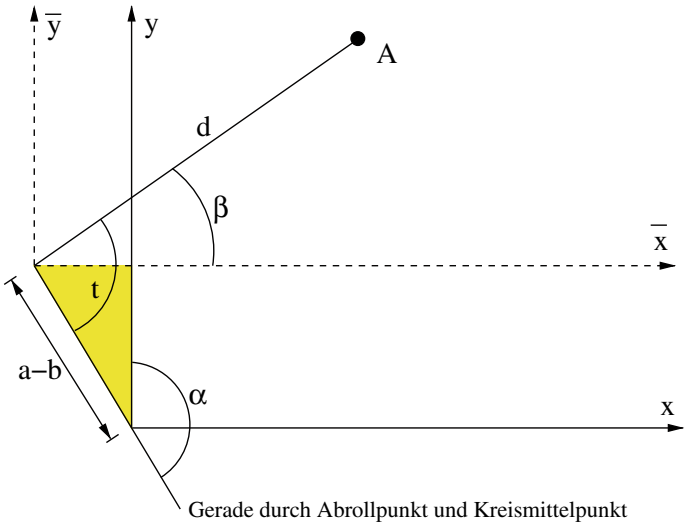

Abb. 2.5 Bezeichnungen für die Analyse, insbesondere der Winkel des gelb eingefärbten Dreiecks

Nach Einsetzen der Beziehung $\alpha = ta/b$ ergibt sich die Parameterdarstellung der Bewegung des Punktes A,

$$x(t) = d \sin\left(t\left(\frac{a}{b} - 1\right)\right) - (a - b)\sin\frac{a}{b}t,$$
$$y(t) = d \cos\left(t\left(\frac{a}{b} - 1\right)\right) - (a - b)\cos\frac{a}{b}t.$$

Diese Beziehung lässt sich deuten als Überlagerung der Kreisbewegung

$$x_1(t) = d \sin\left(\frac{a}{b} - 1\right)t\,, \qquad y_1(t) = d \cos\left(\frac{a}{b} - 1\right)t$$

mit

$$x_2(t) = (a - b)\sin\frac{a}{b}t\,, \qquad y_2(t) = (a - b)\cos\frac{a}{b}t.$$

Der große Kreis $x_1(t)$, $y_1(t)$ hat als Periode den Wert

$$p_1 = 2\pi\frac{1}{\frac{a}{b} - 1} = 2\pi\frac{b}{a}\frac{a}{a - b},$$

der kleine Kreis $x_2(t)$, $y_2(t)$ hat die Periode

$$p_2 = 2\pi\frac{b}{a}.$$

Diese Überlagerung der beiden Kreisbewegungen hat man sich wie in Abb. 2.6 gezeichnet vorzustellen. Der kleine (x_2, y_2)-Kreis mit Radius $(a - b)$ bewegt sich

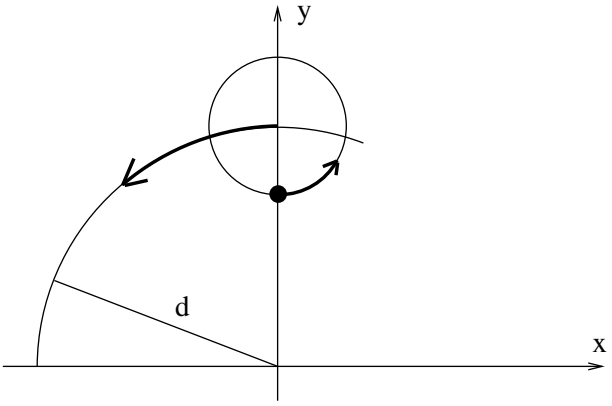

Abb. 2.6 Überlagerung von zwei Kreisbewegungen

planetenartig mit seinem Mittelpunkt auf dem großen (x_1, y_1)-Kreis mit Radius d. Der Punkt A dreht sich dabei auf dem kleinen Kreis. Auf diese Weise lässt sich die Zykloide konstruieren.

Noch frei sind die Werte der Radien a und b. Aus technischen Gründen ist hier nicht jede Wahl möglich. Der Wankelmotor soll drei Kammern haben. Dies bedeutet, dass die Zykloide nicht nur durch den Punkt A erzeugt wird, sondern gleichzeitig die beiden anderen Eckpunkte des Läufers sich entlang der Zykloide bewegen müssen. In dem Bild der überlagerten Kreisbewegung von Abb. 2.6 müssen sich demnach zwei zusätzliche kleine Kreise (Mittelpunkte um $\frac{2\pi}{3}$ bzw. $\frac{4\pi}{3}$ verschoben) auf dem großen Kreis drehen. Damit die Bewegung des Läufer-Dreiecks hineinpasst, müssen die Perioden durch die Beziehung

$$p_1 = 3p_2$$

verknüpft sein. In diesem Fall tritt nach Verschiebung der drei kleinen Kreise entlang des großen Kreises um $\frac{2\pi}{3}$ wieder die gleiche Situation wie zu Beginn ein. Die Forderung nach 3 Kammern führt also auf

$$\frac{a}{a-b} = 3. \tag{2.1}$$

Hieraus folgt, dass das Verhältnis zwischen den Radien a und b

$$\frac{a}{b} = \frac{3}{2}$$

sein muss.

Eine volle Umdrehung eines Läufer-Eckpunktes erfolgt für

$$0 \leq t \leq 4\pi, \quad 0 \leq \alpha \leq 6\pi;$$

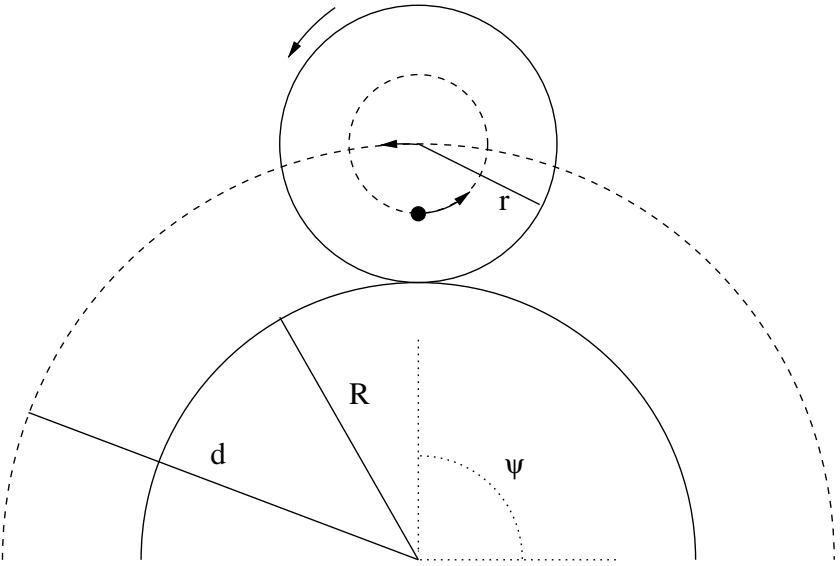

Abb. 2.7 Zykloide als Abrollvorgang (durchgezogene Kreise): Das Verhältnis der Abrollradien ist hier 2:1, der kleine Kreis rollt bei einem Umlauf genau zweimal ab; er führt dabei drei Drehungen aus. Die Bewegung von Abb. 2.6 ist hier gestrichelt gezeichnet

die Parameterdarstellung der für den Wankelmotor notwendigen Zykloide ist somit

$$\begin{aligned} x(t) &= d \sin \tfrac{t}{2} - (a-b) \sin \tfrac{3t}{2} \\ y(t) &= d \cos \tfrac{t}{2} - (a-b) \cos \tfrac{3t}{2} \end{aligned} \tag{2.2}$$

mit $a - b = \frac{a}{3} = \frac{b}{2}$.

Abrollbewegung

An dieser Stelle soll diskutiert werden, welche Art von Zykloide vorliegt. Bisher haben wir die Bewegung als Planetenbewegung gedeutet (Abb. 2.6) mit den Radien d und $(a - b)$. Mit anderen Radien lässt sich die Bewegung auch als Abrollvorgang ansehen, bei der ein Kreis vom Radius r auf einem Kreis vom Radius R abrollt. Ein Punkt auf der r-Kreisscheibe im Abstand λr zum r-Mittelpunkt bewegt sich dann entlang der Zykloide (Abb. 2.7).

Für diese Abrollbewegung bezeichne ψ den Winkel zum Abrollpunkt und damit zum Mittelpunkt des r-Kreises. Wie oft dreht sich der r-Kreis, bis er einmal um den R-Kreis herum ist? Zunächst R/r mal – das Verhältnis der abrollenden Bogenlängen. Dazu kommt 1 für die Umrundung des R-Kreises.[2] Also lässt sich die Abrollbewe-

[2] An einem Beispiel, etwa mit $r = 1$ und $R = 2$, visualisiere man sich diesen Sachverhalt.

gung darstellen durch

$$x = (R+r)\sin\psi - \lambda r \sin\frac{R+r}{r}\psi\,,$$
$$y = (R+r)\cos\psi - \lambda r \cos\frac{R+r}{r}\psi.$$

Durch Vergleich mit (2.2) findet man folgende Beziehung zur Planetenbewegung:

$$\psi = \frac{t}{2}\,,\ R = \frac{2}{3}d\,,\ r = \frac{d}{3}\,,\ \lambda = \frac{a}{d}\,.$$

Den zur Illustration angenommenen Zahlenwerten $d = 7, a = 3, b = 2$ entsprechen die Werte

$$R = \frac{14}{3}\,,\ r = \frac{7}{3}\,,\ \lambda = \frac{3}{7}\,.$$

Diese Art von Abrollbewegung eines Kreises *auf* einem anderen Kreis ist eine Epitrochoide, lediglich der Spezialfall $\lambda = 1$, also $d = a$, führt zu einer Epizykloide. In der Abb. 2.7 sind die Kreise der Planetenbewegung gestrichelt, die der Abrollbewegung ausgezogen gezeichnet.

Die Abb. 2.8 zeigt für $a = 3$, $b = 2$ und drei Werte von d ($d = 3$, $d = 5$, $d = 7$) die zugehörigen Zykloiden. Offenbar hat die Kurve von $d = 3$ eine Art von Rückkehrpunkt, und es stellt sich die Frage, wie die Kurve für $d < 3$ aussehen mag. Die Auflösung zeigt Abb. 2.1: Dort ist die Kurve für $d = 2$ gezeigt, und man sieht ein Schleife ähnlich wie sie auch bei Planetenbewegungen aufzutreten scheint.

Die Epitrochoide des Wankelmotors ist symmetrisch zu den x/y-Achsen. Der Abstand der Trochoide zum 0-Punkt hat zwei Maxima und zwei Minima. Bei einem Umlauf nimmt das Volumen einer jeden Kammer folglich zwei maximale und zwei minimale Werte an. Entsprechend ist der Wankelmotor ein Viertaktmotor: Die vier Phasen Ansaugen, Verdichten, Arbeiten und Ausschieben kann man den Maxima und Minima zuordnen. Drei der vier Takte finden gleichzeitig statt.

Bahn-Umfang

Es bleibt noch die Berechnung des Bahn-Umfangs für einen festen Punkt auf dem Läufer. Diese Bogenlänge L berechnet sich aus

$$L = \int_0^{4\pi} \sqrt{\dot{x}^2 + \dot{y}^2}\,\mathrm{d}t.$$

Ausgehend von (2.2) erhält man nach Einsetzen von

$$\dot{x} = \frac{d}{2}\cos\frac{t}{2} - \frac{3}{2}(a-b)\cos\frac{3t}{2}$$
$$\dot{y} = -\frac{d}{2}\sin\frac{t}{2} + \frac{3}{2}(a-b)\sin\frac{3t}{2}$$

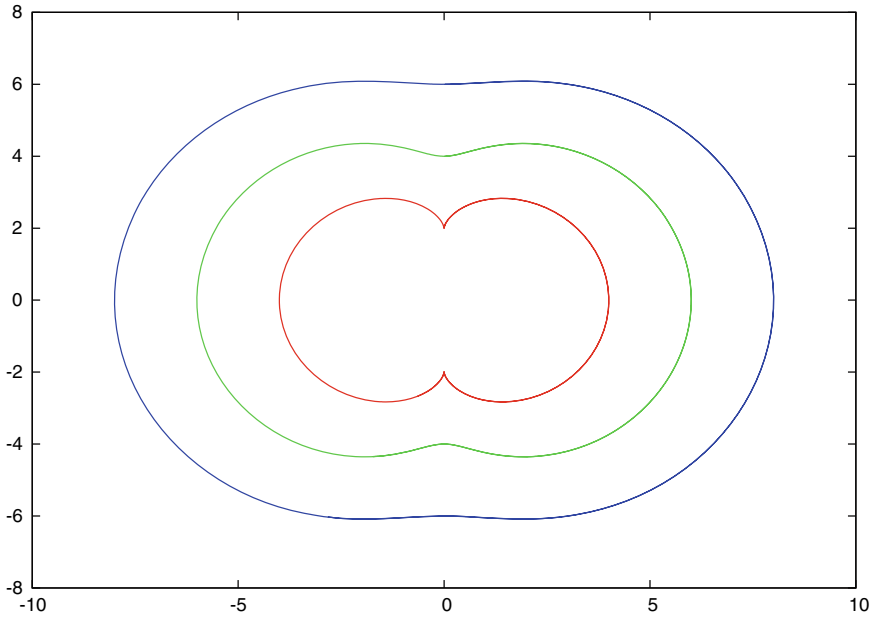

Abb. 2.8 (x, y)-Ebene. Zykloiden (2.2) für $a = 3$, $b = 2$ und drei Werte von d ($d = 3$ in rot, $d = 5$ in grün, $d = 7$ in blau)

das Integral

$$L = \frac{1}{2}\int_0^{4\pi} \left(d^2 + 9(a-b)^2 - 6d(a-b)\left(\cos\frac{t}{2}\cos\frac{3t}{2} + \sin\frac{t}{2}\sin\frac{3t}{2}\right)\right)^{1/2} \mathrm{d}t.$$

Mit Hilfe der Additionstheoreme der trigonometrischen Funktionen formen wir um:

$$\begin{aligned}\cos\frac{t}{2}\cos\frac{3t}{2} + \sin\frac{t}{2}\sin\frac{3t}{2} &= \cos\left(\frac{3t}{2} - \frac{t}{2}\right)\\ &= \cos t = 2\cos^2\frac{t}{2} - 1\end{aligned}$$

also

$$\begin{aligned}L &= \frac{1}{2}\int_0^{4\pi} \left(d^2 + 9(a-b)^2 + 6d(a-b) - 12d(a-b)\cos^2\frac{t}{2}\right)^{1/2} \mathrm{d}t\\ &= \frac{1}{2}\int_0^{4\pi} \left((d + 3(a-b))^2 - 12d(a-b)\cos^2\frac{t}{2}\right)^{1/2} \mathrm{d}t.\end{aligned}$$

Die Substitution $t = 2\varphi$ ergibt

$$L = \int_0^{2\pi} \sqrt{(d + 3(a-b))^2 - 12d(a-b)\cos^2\varphi}\,\mathrm{d}\varphi$$
$$= (d + 3(a-b)) \int_0^{2\pi} \sqrt{1 - \frac{12d(a-b)}{(d+3(a-b))^2}\cos^2\varphi}\,\mathrm{d}\varphi.$$

Mit den Abkürzungen

$$B := d + 3(a-b) = a + d$$
$$k^2 := \frac{12d(a-b)}{(d+3(a-b))^2} = \frac{4ad}{(a+d)^2}$$

haben wir unter Berücksichtigung von (2.1) die Bogenlänge durch ein elliptisches Integral 2. Gattung ausgedrückt:

$$L = B \int_0^{2\pi} \sqrt{1 - k^2\cos^2\varphi}\,\mathrm{d}\varphi.$$

Derartige Integrale sind im Allgemeinen nicht in geschlossener Form lösbar. Lediglich im Sonderfall $k^2 = 1$ (hier $d = 3$, $a = 3$, $b = 2$, also Epizykloide) gilt

$$L = B \int_0^{2\pi} |\sin\varphi|\,\mathrm{d}\varphi = 2B \int_0^{\pi} \sin\varphi\,\mathrm{d}\varphi = 4B.$$

Ein Punkt auf dem Läuferkreis $\mathrm{K_L}$ führt bei einer vollen Umdrehung demnach eine Bewegung mit Weglänge $L = 24$ aus.

Bei unseren Zahlenwerten ($d = 7$) ergibt sich für den Weg eines Läufer-Eckpunktes

$$k^2 = 0{,}84,\ B = 10.$$

Hier ist man auf eine Tabelle oder auf numerische Näherungsverfahren angewiesen. Zunächst bringen wir das Integral auf eine „Standardform“ (Legendre-Form). Es gilt wegen der Eigenschaften von $\cos^2\varphi$ und $\sin^2\varphi$

$$\int_0^{2\pi} \sqrt{1 - k^2\cos^2\varphi}\,\mathrm{d}\varphi = 4\int_0^{\pi/2} \sqrt{1 - k^2\cos^2\varphi}\,\mathrm{d}\varphi = 4\int_0^{\pi/2} \sqrt{1 - k^2\sin^2\varphi}\,\mathrm{d}\varphi.$$

Das letztere der Integrale wird mit $E(k)$ bezeichnet, es ist das vollständige elliptische Integral 2. Art. Damit ist die Bogenlänge der Epitrochoide gegeben durch

$$L = 4B\, E(k).$$

Für $k^2 = 0{,}84$ ergibt die Tabelle (oder das Computerprogramm für $E(k)$) den Wert $E(k) = 1{,}15065...$; also ist die Bogenlänge bei unseren Zahlenwerten

$$L \approx 46.$$

Exkurs zur numerischen Berechnung
Die Berechnung des elliptischen Integrals $E(k)$ soll hier nur kurz angedeutet werden. Das Integral, umgeformt zu

$$E(k) = \int_0^{\pi/2} \sqrt{\cos^2\varphi + (1-k^2)\sin^2\varphi}\, \mathrm{d}\varphi,$$

kann mit Hilfe der Bartky-Transformationen in eine Folge von Integralen

$$E(k) = \int_0^{\pi/2} G_\nu(R_\nu)\, \mathrm{d}\varphi,\ \ \nu = 0, 1, 2, \ldots$$

umgeformt werden, mit Funktionen[3] G_ν und

$$\begin{aligned} R_\nu &= \sqrt{m_\nu^2 \cos^2\varphi + n_\nu^2 \sin^2\varphi}\,, \quad \text{mit} \\ m_0 &= 1,\ n_0 = \sqrt{1-k^2}, \\ m_\nu &= \frac{1}{2}\,(m_{\nu-1} + n_{\nu-1})\,,\ n_\nu = \sqrt{n_{\nu-1} m_{\nu-1}} \quad (\nu \geq 1). \end{aligned}$$

Die Folge der m_ν, n_ν strebt sehr rasch gegen den gemeinsamen Grenzwert des Gaußschen arithmetisch-geometrischen Mittels; nach wenigen (etwa l) Schritten gilt $m_l \doteq n_l$ und also

$$R_l = \sqrt{m_l^2 \cos^2\varphi + n_l^2 \sin^2\varphi} \doteq m_l.$$

[3] Die Bartky-Transformation und die Funktionen G_ν sind bei der Thematik dieses Kapitels ohne Interesse; wir verweisen auf Spezialliteratur.

(Im Beispiel war $k^2 = 0{,}84$; dann ist $l = 5$ und $|m_5 - n_5| \leq 10^{-12}$.) Das Integral mit dem Integranden $G_l(m_l)$ lässt sich dann elementar integrieren.

Literatur

Baier, O.: Die Kinematik der Drehkolben- und Kreiskolbenmaschinen und ihre Fertigungsmöglichkeiten. VDI-Berichte Nr. **45**, 31–37 (1960)

über Zykloiden in der Geschichte

Hildebrandt, S., Tromba, A.: Mathematics and Optimal Form. Scientific American Library, New York (1985)

zur Berechnung der elliptischen Integrale

Bulirsch, R., Stoer, J.: Darstellung von Funktionen in Rechenautomaten. In: Sauer, R., Szabó, I. (Hrsg.), Mathematische Hilfsmittel des Ingenieurs. Bd. III. Springer, Berlin (1968)
Bulirsch, R.: An extension of the Bartky-transformation to incomplete elliptic integrals of the third kind. Numer. Math. **13**, 266–284 (1969)

Die Aufgabenstellung zu Beginn des Kapitels greift auf eine ältere Aufgabe zurück, die an der TU München schon vor 1975 entwickelt wurde.
Eine gute Quelle zu technischen Aspekten ist der Eintrag „Wankelmotor“ bei Wikipedia

Lateraler Abtastfehler bei Schallplatten

3

Das Schneiden einer herkömmlichen Schallplatte erfolgt tangential; daher liegen in den Rillen die zusammengehörigen Impulse auf den beiden Flanken genau „gegenüber“. Nach Möglichkeit sollten die auf den beiden Rillenflanken sich entsprechenden Signale auch gleichzeitig abgetastet werden, andernfalls entstehen bei der Wiedergabe Verzerrungen. Ein fehlerfreies Abtasten von Schallplatten ist mit einem Tangentialtonarm (angedeutet in Abb. 3.1) zwar möglich, aber aufwendig. Konstruktiv einfacher sind die Tonarme, die in einem Punkt drehbar gelagert sind (Abb. 3.2). Bei diesen üblichen Tonarmen muss bei Abtastnadeln mit elliptischem Querschnitt jedoch ein tangentialer Spurfehlerwinkel γ in Kauf genommen werden.

Der Winkel γ könnte beliebig klein sein, wenn man nur die „Einbautiefe“ a und die Länge l des Tonarms sehr groß wählen könnte. Diese beiden Größen unterliegen jedoch Beschränkungen: Es steht nur ein begrenztes Raumangebot zur Verfügung, außerdem sind beim Tonarm die mit der Länge zunehmende Masse und die Fähigkeit zur Eigenschwingung zu berücksichtigen. Überdies können die Größen a und l nicht unabhängig voneinander gewählt werden; ihr Definitionsbereich muss garantieren, dass *alle* Rillen abgetastet werden. Das bedeutet, dass der Radius r der aktuellen Nadelposition N alle Werte zwischen dem Innenradius R_i und dem Außenradius R_a annehmen muss:

$$R_\mathrm{i} \leq r \leq R_\mathrm{a}.$$

R_i entspricht der innersten Rille einer Schallplatte, R_a der äußersten.

Eine einfache geometrische Überlegung hilft hier weiter (Abb. 3.3): Zeichnet man zur Schallplatte die beiden extremen Nadel- und Auflagepositionen N,A bzw. N',A' ein, so liest man ab, dass $a - R_\mathrm{i} < l < a + R_\mathrm{i}$ gelten muss. Dies ist äquivalent zur Abspiel-Bedingung

$$|l - a| < R_\mathrm{i}.$$

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_3

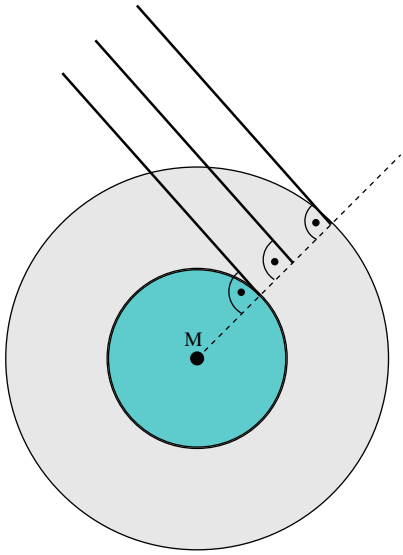

Abb. 3.1 Schallplatte mit Mittelpunkt M und Rillenbereich (grau). Angedeutet ist tangentiales Abspielen

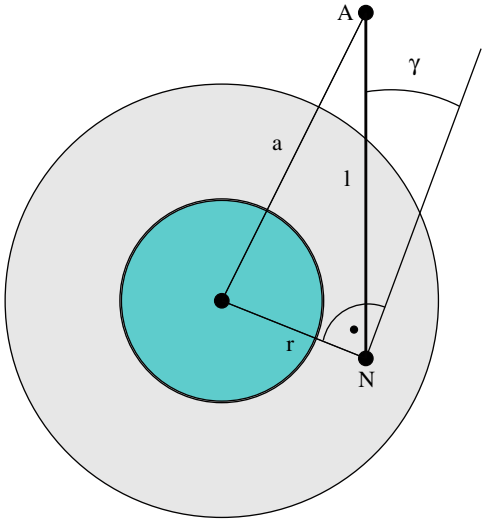

Abb. 3.2 Im Punkt A drehbar gelagerter gerader Tonarm mit Einbautiefe a, Tonarmlänge l und Spurfehlerwinkel γ. r ist der Abstand zwischen Nadel N und Mittelpunkt M

Ist diese Ungleichung erfüllt, so nimmt die Variable r sowohl den Wert R_a als auch R_i an.

Der Spurfehlerwinkel γ ist deutlich kleiner bei den üblichen gekröpften Tonarmen (Abb. 3.4). Dann hängt γ nicht nur von a und l ab, sondern auch vom Kröpfungswinkel β. Durch Variieren von a, l und β soll versucht werden, den Spurfehlerwinkel und damit die Verzerrung klein zu halten (Abb. 3.4). Die Forderung $|l - a| < R_\mathrm{i}$ hindert a und l nicht, unrealistisch große Werte anzunehmen. Wir werden daher im Folgenden eine der drei Größen a, l, β als fest vorzugeben annehmen. Der Optimie-

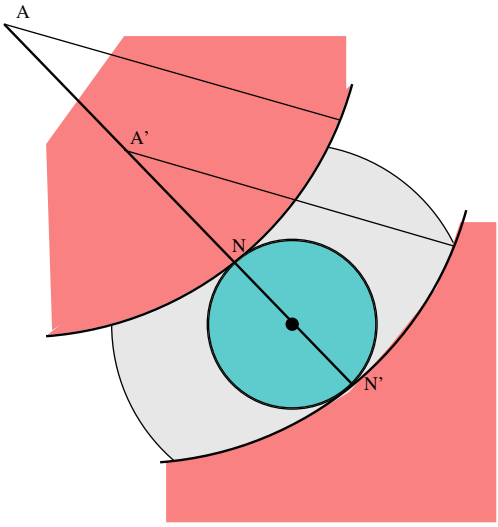

Abb. 3.3 Extreme Auflagepositionen mit Abspielradien: Der rote Bereich ist nicht zulässig

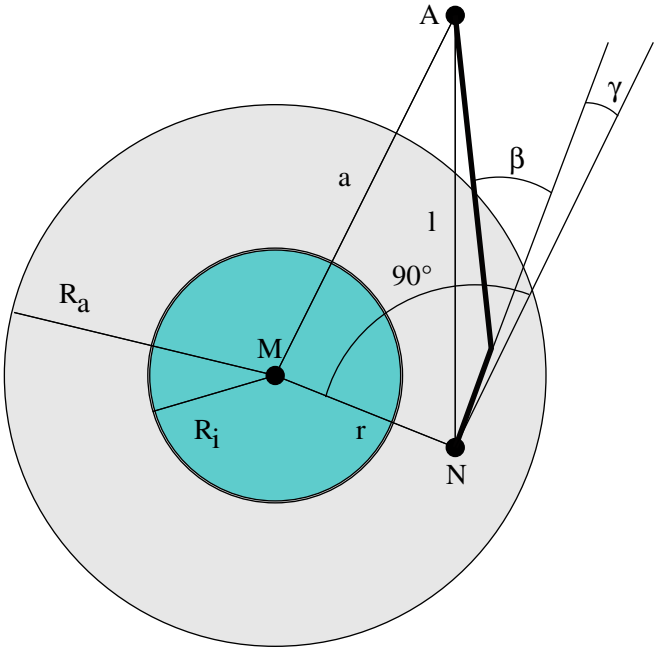

Abb. 3.4 Schallplatte mit Innenradius R_i, Außenradius R_a, gekröpftem Tonarm mit effektiver Tonarmlänge l, Kröpfungswinkel β, und Spurfehlerwinkel γ

rungsprozess soll dann die beiden freien Parameter bestimmen. Die Abspielbedingung $|l - a| < R_\mathrm{i}$ wird dann von selbst erfüllt sein.

Aufgabe *Der Tonarm eines Plattenspielers ist im Punkt A drehbar gelagert. Beim Abspielen beschreibt die Nadelspitze N einen kreisbogenförmigen Weg (vgl. Abb. 3.4). Die Längsrichtung des Tonabnehmers schließt mit der Rillentangente den tangentialen Spurfehlerwinkel γ ein.*

$$\begin{aligned} l &: Abstand\ \overline{AN} \\ a &: Abstand\ \overline{AM} \\ r &: Abstand\ \overline{MN} \\ \beta &: Kröpfungswinkel \end{aligned}$$

Die Parameter a, l, β sind als konstant angenommen, und $\gamma = \gamma(r)$. Beschränkungen für r und a:

$$\begin{aligned} R_\mathrm{i} \le r &\le R_\mathrm{a} \\ |a - l| &< R_\mathrm{i} \end{aligned} \tag{3.1}$$

Bei einer 30 cm-Langspielplatte gilt nach DIN IEC98 (ehemals DIN-45547):

$$R_\mathrm{i} = 5{,}75\ \mathrm{cm},\ \ R_\mathrm{a} = 14{,}6\ \mathrm{cm}.$$

a) Man zeige: Für den Spurfehlerwinkel gilt

$$\gamma(r; a, l, \beta) = -\beta + \arcsin \frac{r^2 + l^2 - a^2}{2rl}, \tag{3.2}$$

Definitionsbereich ist (3.1).

b) Man bestimme die Extrema von γ in Abhängigkeit von r (geometrische Bedeutung?).
Die durch γ hervorgerufene Verzerrung ist bei Nadeln von elliptischem Querschnitt proportional zu

$$k(r; a, l, \beta) = \frac{\gamma}{r}.$$

Die Parameter a, l, β müssen so bestimmt werden, dass k auf (3.1) minimiert wird. Eine Näherung für das optimale a, l, β bestimme man wie folgt:

c) Durch Abbrechen der Taylor-Entwicklung von arcsin *nach dem ersten Glied ergibt sich eine Näherung $\tilde{k}(r; a, l, \beta)$. Man bestimme die Extrema r_1, r_2, r_3 von $\tilde{k}(r)$ auf (D).*

d) Durch Gleichsetzen der Funktionswerte

$$\tilde{k}(r_2) = -\tilde{k}(r_1) = -\tilde{k}(r_3)$$

stelle man zwei Gleichungen

$$\beta = f_1(l, a), \quad f_2(l, a) = 0$$

auf.

e) Welche Zahlenwerte ergeben sich für l und β, wenn die Einbautiefe mit $a = 19\,\text{cm}$ festgelegt wird?

Spurfehlerwinkel

Um eine Beziehung für den Spurfehlerwinkel γ zu erhalten, betrachten wir das Dreieck mit den Eckpunkten M,A,N und den Seitenlängen a, l, r. Für den Innenwinkel φ am Eckpunkt N folgt aus dem Kosinussatz für Dreiecke

$$\cos\varphi = \frac{r^2 + l^2 - a^2}{2lr} \quad \text{für} \quad \varphi := \frac{\pi}{2} - \beta - \gamma\,. \tag{3.3}$$

Mit

$$\gamma = \frac{\pi}{2} - \beta - \arccos\frac{r^2 + l^2 - a^2}{2rl} = -\beta + \arcsin\frac{r^2 + l^2 - a^2}{2rl}$$

ergibt sich die angegebene Formel (3.2). Ein Beispiel für γ zeigt die Abb. 3.5, hier ist β so gewählt, dass γ im Bereich $R_\text{i} \le r \le R_\text{a}$ zwei Nullstellen hat.

Ein stationärer Wert des Spurfehlerwinkels $\gamma(r)$ berechnet sich aus

$$0 = \frac{\partial\gamma}{\partial r} = \frac{r^2 - l^2 + a^2}{r\sqrt{(2rl)^2 - (r^2 + l^2 - a^2)^2}}.$$

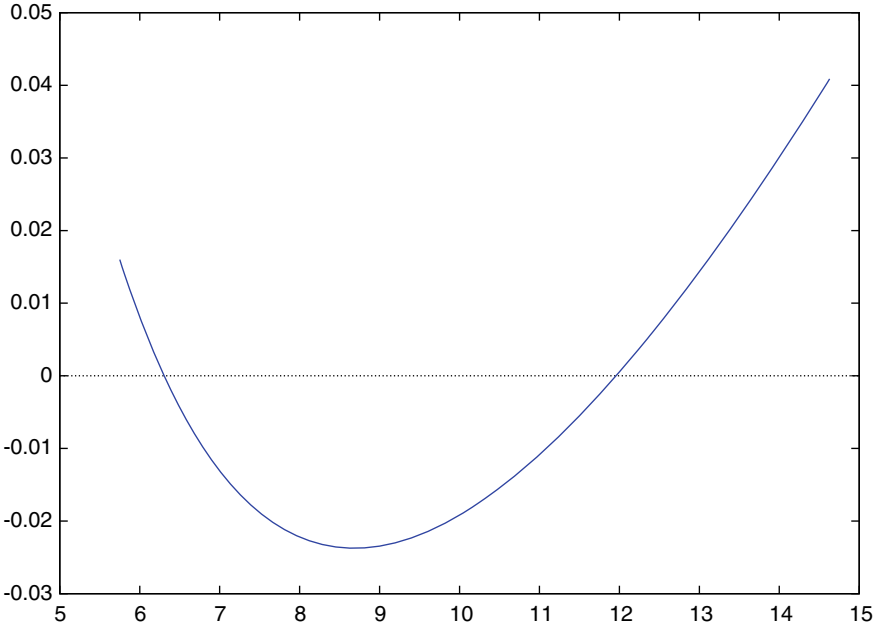

Abb. 3.5 Tangentialer Spurfehlerwinkel $\gamma(r)$ über der r-Achse, γ im Bogenmaß, für $a = 19$, $l = 20{,}89$, $\beta = 0{,}4524$ (25,92°)

Wegen $0 < \varphi < \pi$ und wegen (3.3) gilt $|\cos\varphi| < 1$ und $|r^2 + l^2 - a^2| < 2rl$. Also kann der Nenner nicht verschwinden, und auch das Argument des arcsin nicht den Wert 1 annehmen. Die Gleichung $0 = \partial\gamma/\partial r$ ist erfüllt, wenn

$$r^2 + a^2 = l^2 \quad \text{bzw.} \quad r = \sqrt{l^2 - a^2}$$

gilt. Nach dem Satz des Pythagoras bedeutet dies, dass γ extremal wird, wenn der Winkel bei M ein rechter Winkel ist. Wegen

$$r \gtrless \sqrt{l^2 - a^2} \quad \Rightarrow \quad \frac{\partial\gamma}{\partial r} \gtrless 0$$

liegt für $r = \sqrt{l^2 - a^2}$ und $l > a$ ein Minimum vor, der Wert des Minimums ist

$$\begin{aligned}\gamma(\sqrt{l^2 - a^2}) &= -\beta + \arcsin\frac{\sqrt{l^2 - a^2}}{l} \\ &= -\beta + \arccos\frac{a}{l}.\end{aligned}$$

Das Optimierungsziel ist es, den Spurfehlerwinkel γ klein zu halten. Eine Änderung der Werte von a, l, β bedeutet anschaulich eine Verschiebung des Graphen der Funktion $\gamma(r; a, l, \beta)$ in der (r, γ)-Ebene. Die konvexe Gestalt des Graphen bleibt dabei qualitativ erhalten (Abb. 3.5). Insbesondere bewirkt eine Veränderung des Kröpfungswinkels β eine Verschiebung der Kurve nur in γ-Richtung, das Auftreten von Nullstellen von γ kann so erzwungen werden.

Wenn der Spurfehlerwinkel $|\gamma|$ über das ganze r-Intervall klein sein soll, dann müssen die Parameter a, l, β offensichtlich wie folgt gewählt werden: $\gamma(r)$ muss im Intervall $R_\mathrm{i} < r < R_\mathrm{a}$ zwei Nullstellen haben. Dann liegt das Minimum an $r = \sqrt{l^2 - a^2}$ im Inneren des Intervalls, mit einem negativem Wert $\gamma(\sqrt{l^2 - a^2})$, und an den Rändern $r = R_\mathrm{i}$, $r = R_\mathrm{a}$ gibt es positive Maxima $\gamma(R_\mathrm{i})$, $\gamma(R_\mathrm{a})$. Der Spurfehlerwinkel γ ist klein, wenn die absoluten Werte der drei Extrema ungefähr gleich groß sind.

Mit dieser Überlegung ist bereits eine grobe Bestimmung günstiger Parameterwerte möglich.

Zur **Optimierung** kann eine *black-box*-Routine verwendet werden, die im günstigen Fall ein globales Minimum liefert. Wir versuchen hier mit einer analytischen Überlegung eine Minimierung durchzuführen. Es sei betont, dass der Zugang eine vernünftige Heuristik ist, aber zunächst keine Garantie der Optimalität beinhaltet.

Verzerrung

Zur Optimierung eines Schallplatten-Musikgenusses ist jedoch letztlich nicht der Spurfehlerwinkel γ zu minimieren, sondern die von ihm bewirkte Verzerrung. In Abb. 3.6 wird ein elliptischer Querschnitt der Abtastnadel vorausgesetzt. Das bewirkt im Fall $\gamma \neq 0$ eine Voreilung v einer Rillenflanke. Die Verzerrung kann als proportional zur Voreilung v angenommen werden, relativ zur Datenlänge s. Die durch

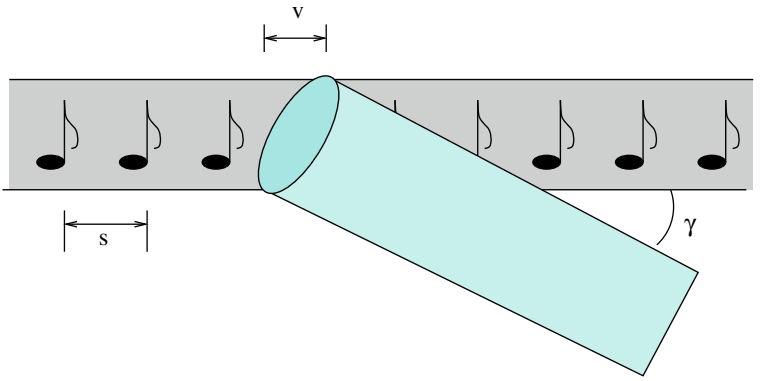

Abb. 3.6 schematisch: eine Rille (grau) einer Schallplatte, mit einem Tonarm (blau) und abtastender elliptischer Nadel, von oben betrachtet. Die Größe v beschreibt die Voreilung der einen Rillenflanke gegenüber der anderen Flanke, und s ist die Länge eines Datenpaketes. Der Quotient v/s ist ein Maß für die Verzerrung

einen Spurfehlerwinkel $\gamma \neq 0$ hervorgerufene Voreilung v der Abspielnadel auf einer Rillenflanke ist ungefähr proportional zu γ (wegen $\tan\gamma \approx \gamma$ für kleine $|\gamma|$).

Andererseits ist die Auswirkung der Voreilung relativ zur Datenlänge s zu sehen. Letztere ist proportional zum Radius r. Da die Schallplatte mit konstanter Winkel-Geschwindigkeit dreht, müssen die Informationen in einer inneren Rille wegen des geringen inneren Umfangs erheblich dichter geschnitten sein als außen. Die Datenlänge s etwa eines Musiktaktes ist proportional zu r. Die Informationsdichte (=Informationsmenge/Spurlänge) ist demnach proportional zu $\frac{1}{r}$. Also ist die Verzerrung proportional zur Informationsdichte $\times$ Voreilung, zu

$$\frac{1}{r} \cdot \gamma(r).$$

Deswegen dient die Funktion

$$k(r; a, l, \beta) := \frac{1}{r}\,\gamma(r; a, l, \beta) \tag{3.4}$$

als Maß für die Verzerrung, unter der Annahme eines elliptischen Querschnitts der Nadel.

Das Problem der Minimierung der Verzerrung lässt sich nun wie folgt formulieren: Fixiere einen Wert für a oder l; bestimme die zwei freien der drei Parameter β, l, a derart, dass

$$F(a, l, \beta) := \max_{R_\mathrm{i} \le r \le R_\mathrm{a}} |k(r; a, l, \beta)|,$$

k aus (3.4), minimal wird (Tschebyschow-Approximation[1]). Da $k(r)$ von ähnlicher Gestalt ist wie $\gamma(r)$, liegen im interessierenden Bereich ebenfalls ein Minimum

[1] andere Schreibweisen auch Tschebyscheff, engl. Chebyshev.

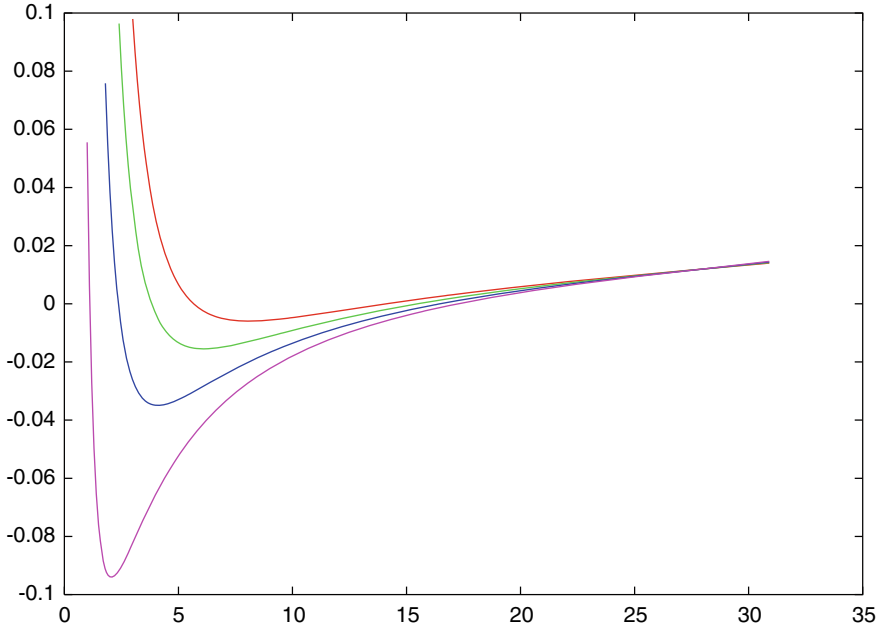

Abb. 3.7 Verzerrung $k(r)$ für $a = 19$, $\beta = 0{,}49$ (28,1°) und $l = 19{,}5$ (violett), $l = 20{,}0$ (blau), $l = 20{,}5$ (grün), $l = 21{,}0$ (rot), ohne Rücksicht auf ein sinnvolles Intervall für r. $k(r)$ ist nicht konvex. Es gibt Werte von l, sodass kein inneres Minimum für $R_\mathrm{i} \le r \le R_\mathrm{a}$ existiert

sowie zwei Randmaxima vor. Dies zeigt die Berechnung einer Kurvenschar $k(r)$ in Abhängigkeit vom Parameter l (Abb. 3.7).

Die Überlegungen zur Optimierung von $k(r)$ gelten analog wie bei $\gamma(r)$: Wenn es möglich ist, die beiden Gleichungen

$$\begin{aligned} k(r_2) &= -k(R_\mathrm{i}) \\ k(R_\mathrm{i}) &= k(R_\mathrm{a}) \end{aligned} \tag{3.5}$$

zu erfüllen, dann dürfte eine minimale Verzerrung gefunden sein. Der unbekannte Radius r_2, für den das Minimum von $k(r)$ angenommen wird, ist definiert durch die Gleichung

$$\frac{\partial k(r_2)}{\partial r} = 0.$$

Das sind zusammen drei Gleichungen für die drei Unbekannten r_2, β und a oder l. Nach Ausführung der Differenziation von (3.4) lautet die dritte Gleichung

$$-\frac{k(r)}{r} + \frac{r^2 - l^2 + a^2}{r^2\sqrt{(2rl)^2 - (r^2 + l^2 - a^2)^2}} = 0. \tag{3.6}$$

Die Lösung der drei Gl. (3.5)/(3.6) erfolgt zweckmäßig mit dem Newton-Verfahren.

Näherung

Um die Konvergenz des Newton-Verfahrens zu erleichtern, muss eine gute Näherung für die Lösung dieser drei Gleichungen bestimmt werden. Hierzu lösen wir vereinfachte Gleichungen, die sich durch Linearisierung der Verzerrungsfunktion ergeben. Wegen

$$\arcsin x = x + \frac{x^3}{6} + \cdots, \quad |x| < 1$$

ist

$$\tilde{k} := \frac{1}{r}\left(-\beta + \frac{r^2 + l^2 - a^2}{2rl}\right) = -\frac{\beta}{r} + \frac{1}{2l}\left(1 + \frac{1^2 - a^2}{r^2}\right)$$

eine Näherung für die Verzerrung k. Der Radius $\tilde{r}_2$, für den $\tilde{k}(r)$ minimal wird, kann bei dieser einfachen Beziehung explizit ausgerechnet werden: Die Ableitung

$$\frac{\partial \tilde{k}}{\partial r} = \frac{1}{r^2}\left(\beta - \frac{l^2 - a^2}{lr}\right)$$

hat die Nullstellen bei

$$\tilde{r}_2 = \frac{l^2 - a^2}{l\beta},$$

hier liegt ein Minimum vor. Bei vernünftiger Wahl der Parameter l, a, β treten Randmaxima auf für R_i und R_a.

Damit hat man die drei Extremwerte

$$\begin{aligned}
\tilde{k}(R_\mathrm{i}) &= -\frac{\beta}{R_\mathrm{i}} + \frac{1}{2l}\left(1 + \frac{l^2 - a^2}{R_\mathrm{i}^2}\right) \qquad \text{Randmaximum,} \\
\tilde{k}(\tilde{r}_2) &= \frac{1}{2l} - \frac{l\beta^2}{2(l^2 - a^2)} \qquad \text{inneres Minimum,} \\
\tilde{k}(R_\mathrm{a}) &= -\frac{\beta}{R_\mathrm{a}} + \frac{1}{2l}\left(1 + \frac{1^2 - a^2}{R_\mathrm{a}^2}\right) \qquad \text{Randmaximum.}
\end{aligned}$$

Durch Umformungen der zwei verbliebenen Gleichungen

$$\begin{aligned}
\tilde{k}(\tilde{r}_2) &= -\tilde{k}(R_\mathrm{i}) \\
\tilde{k}(R_\mathrm{i}) &= \tilde{k}(R_\mathrm{a})
\end{aligned}$$

gewinnt man einfache Beziehungen zwischen den zu bestimmenden Parametern β, a oder l:

Aus der zweiten Gleichung erhält man

$$\beta\left(\frac{1}{R_\mathrm{a}} - \frac{1}{R_\mathrm{i}}\right) = \frac{l^2 - a^2}{2l}\left(\frac{1}{R_\mathrm{a}^2} - \frac{1}{R_i^2}\right),$$

und, nach β aufgelöst,

$$\beta = c_1 \frac{l^2 - a^2}{2l},$$

wobei

$$c_1 := \frac{R_\mathrm{i} + R_\mathrm{a}}{R_\mathrm{i} R_\mathrm{a}}$$

eine von den Radien abhängige Konstante ist.

Umformung der ersten Gleichung $\tilde{k}(\tilde{r}_2) + \tilde{k}(R_\mathrm{i}) = 0$ ergibt mit

$$\begin{aligned} \frac{1}{l} &= \frac{\beta^2}{4} \frac{2l}{l^2 - a^2} + \frac{\beta}{R_\mathrm{i}} - \frac{l^2 - a^2}{2l} \frac{1}{R_\mathrm{i}^2} \\ &= \left(\frac{c_1^2}{4} + \frac{c_1}{R_\mathrm{i}} - \frac{1}{R_i^2} \right) \frac{l^2 - a^2}{2l} \end{aligned}$$

eine implizite Gleichung für l und a:

$$l^2 - a^2 = c_2,$$

die Konstante c_2 ist bestimmt durch

$$c_2 := \frac{8}{c_1^2 + \frac{4}{R_\mathrm{i} R_\mathrm{a}}}.$$

Wir fassen zusammen: Durch Vereinfachung der ursprünglich drei Gleichungen haben wir die zwei elementaren Beziehungen

$$\begin{aligned} &l^2 - a^2 = c_2 \\ &l\beta = c_3, \quad \text{mit } c_3 := \tfrac{c_1 c_2}{2} \end{aligned} \tag{3.7}$$

ermittelt, aus denen sich sehr einfach Näherungen für die zwei optimalen Parameter β, a oder l berechnen lassen.

Legt man die DIN-Radien zugrunde ($R_\mathrm{i} = 5{,}75$, $R_\mathrm{a} = 14{,}6$ cm) und berechnet die Konstanten c_1, c_2, c_3, so lauten die beiden Beziehungen (3.7)

$$l^2 - a^2 = 75{,}18, \ \beta l = 9{,}11.$$

Liegt beispielsweise die Einbautiefe mit $a = 19$ cm fest, so ergeben sich aus diesen Formeln für Tonarmlänge und Kröpfungswinkel die Werte

$$l = 20{,}9 \text{ cm}, \ \beta = 0{,}436 \ (25^\circ).$$

Ausgehend von diesen Näherungen wird das volle System

$$\begin{aligned} k(r_2) + k(R_\mathrm{i}) &= 0 \\ k(R_\mathrm{i}) - k(R_\mathrm{a}) &= 0 \\ \frac{\partial k(r_2)}{\partial r} &= 0, \end{aligned} \tag{3.8}$$

die dritte Gleichung aus (3.6), mit dem Newton-Verfahren numerisch gelöst. Die Newton-Iteration liefert bessere Näherungen, für den r_2-Wert (er interessiert hier nicht), und

$$l = 20{,}9\ \mathrm{cm},\ \ \beta = 25{,}9°.$$

Es fällt auf, dass der erste Näherungswert von l mit dem Wert der Newton-Iteration auf wenigstens drei Stellen übereinstimmt; ein mit der Formel

$$l^2 - a^2 = c_2$$

berechneter Wert für a oder l ist also so genau, dass er nicht verbessert werden muss. Dagegen bedarf der aus

$$l\beta = c_3$$

berechnete Wert für β einer Korrektur.

Da eine Veränderung des Kröpfungswinkels nur eine Verschiebung des Funktionsverlaufes bedeutet, kann ein besserer Wert für β sehr leicht auch ohne Lösung des vollen Systems (3.8) bestimmt werden (z. B. grafisch). Der Grund für die unzureichende Genauigkeit, welche die Gleichung $l\beta = c_3$ liefert, ist folgender: Wegen

$$\beta \approx \arcsin x,\ \ x = \frac{x^2 + l^2 - a^2}{2rl}$$

tritt bei der Subtraktion $\beta - \arcsin x$ *Auslöschung* auf, hier führt die Linearisierung zu starkem Genauigkeitsverlust. Beispielsweise ergibt sich für $\beta = 0{,}4708$ (21,35°), $a = 19, l = 21{,}35$ und $r = 7$

$$\begin{aligned} -\beta + \arcsin x &= 0{,}0312 \quad \text{(richtig)} \\ -\beta + x &= 0{,}0104 \quad \text{(völlig falsch).} \end{aligned}$$

Abhängigkeit der Optimierung von R_i

Die Optimierung ist stark von der Wahl des äußeren und insbesondere des inneren Radius abhängig. Die DIN-Norm geht mit dem kleinen Radius $R_\mathrm{i} = 5{,}75$ cm von einem extremen Fall aus, der bei 30 cm-Schallplatten selten erreicht wird. Wählt man stattdessen etwa $R_\mathrm{i} = 7$ cm, so verschieben sich die Resultate beträchtlich:

Beispiel: $R_\mathrm{i} = 7$ cm, $R_\mathrm{a} = 14{,}5$ cm führt bei obiger Näherung auf die beiden Formeln $l^2 - a^2 = 94{,}92$, $\beta l = 10{,}05$. Für $a = 19$ cm folgt nun $l = 21{,}35$ cm, $\beta = 0{,}471$ (27°).

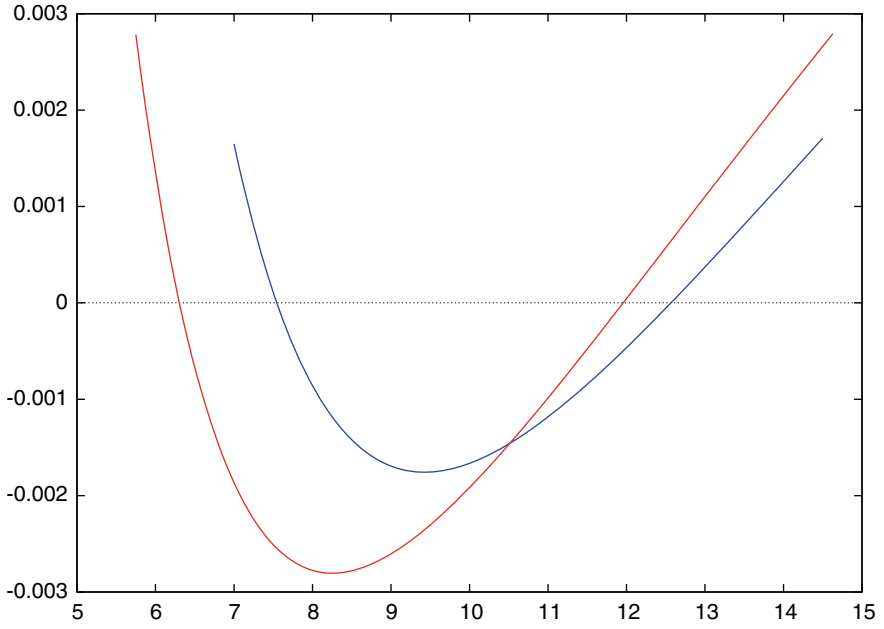

Abb. 3.8 Verzerrung $k(r)$ über der r-Achse, im Bogenmaß, optimiert für einen Innenradius von $R_\mathrm{i} = 7$ cm (in blau), mit $R_\mathrm{a} = 14{,}5$, $a = 19$, $l = 21{,}35$, $\beta = 0{,}49$ (28,1°). Dazu ist geplottet die Optimierung zu $R_\mathrm{i} = 5{,}75$ (in rot) mit $R_\mathrm{a} = 14{,}63$, $a = 19$, $l = 20{,}89$, $\beta = 0{,}452$ (25,92°)

Dieses Beispiel mag als Warnung dienen: Eine sehr genaue Optimierung des Spurfehlerwinkels ist nicht notwendig sinnvoll, für jede Schallplatte ist der optimale Wert ein anderer! In der Abb. 3.8 ist die Verzerrung $k(r)$ gezeigt für $R_\mathrm{i} = 7$ cm optimiert, im Vergleich zu $R_\mathrm{i} = 5{,}75$ cm. Beide sind „optimal“, aber der Unterschied ist bedeutend: Die Optimierung zu $R_\mathrm{i} = 5{,}75$ cm hat eine etwa 50 % stärkere Verzerrung im Vergleich zu $R_\mathrm{i} = 7$ cm.

Abschließend werden in Tab. 3.1 die strengen Lösungen für ausgewählte Werte der Parameter angegeben (erhalten durch Lösen des vollen Gleichungssystems mit Hilfe

Tab. 3.1 Optimale Parameter für $R_\mathrm{i} = 7$ cm, $R_\mathrm{a} = 14{,}5$ cm

a [cm]	l [cm]	β [°]
18,0	20,47	29,43
18,5	20,91	28,75
19,0	21,35	28,10
19,5	21,80	27,48
20,0	22,25	26,88
20,5	22,70	26,30
21,0	23,15	25,75
21,5	23,61	25,22
22,0	24,06	24,71

des Newton-Verfahrens). Bei Bedarf lassen sich Zwischenwerte durch Interpolation ermitteln.

Skatingkraft

Die Skatingkraft steht in engem Zusammenhang mit der hier diskutierten Problemstellung, wir wollen die Größe dieser Kraft abschließend berechnen.

Tangential zur Rillenflanke wirkt auf die Abtastnadel eine Reibungskraft p, welche am Drehpunkt A ein Drehmoment q bewirkt, die *Skatingkraft*. Wir fassen l als Vektor auf (Abb. 3.2) und erhalten für das Vektorprodukt $q = p \times l$

$$\begin{aligned} |q| &= |p|\,|l|\sin(\beta+\gamma) \\ &= |p|\,|l|\frac{r^2+|l|^2-a^2}{2r|l|} \\ &= |p|\frac{r^2+|l|^2-a^2}{2r}. \end{aligned}$$

Die Skatingkraft $|q|$ ist damit proportional zu

$$r + \frac{c_2}{r}.$$

Wegen $l > a$ gilt $|q| \neq 0$, die Skatingkraft weist also nur in eine Richtung.

Stereo-Rundfunk, Amplitudenmodulation

4

Der analoge Stereo-Rundfunk ist nach wie vor stark verbreitet, obwohl eine digitale Alternative (DAB+) entwickelt wurde und nutzbar ist. Mit dem analogen *Multiplexsignal* gelang es, kompatibel zum vorher üblichen Monosignal zu sein. Diese grundlegende Entwicklung ist es wert analysiert zu werden, selbst wenn die analoge Sendung eines Tages nicht mehr eingesetzt werden sollte.

Niederfrequente Schwingungen (Sprache, Musik) benötigen zur drahtlosen Übertragung eine energiereiche hochfrequente Schwingung als Träger. Hierbei wird die Trägerschwingung durch die niederfrequente Information moduliert, beispielsweise durch Modulation der Amplitude.[1] Im Folgenden wird zunächst die Amplitudenmodulation (AM) diskutiert; es folgt danach als wichtige Anwendung die Bildung des Multiplexsignals bei der stereofonen (analogen) Rundfunkübertragung. Beide Themenkreise werden in je einer Aufgabe zusammengefasst. Um einige Begriffe wie Hüllkurve oder Phasenwechsel kennenzulernen, wird empfohlen, zu der Lösung der folgenden Aufgaben saubere Zeichnungen anzufertigen.

4.1 Amplitudenmodulation

Ein niederfrequentes Signal sei mit $S(t)$ bezeichnet, t ist die Zeit. Den Träger $\varphi_T(t)$ nehmen wir in der Form

$$\varphi_T(t) = A \sin(2\pi \omega_T t)$$

[1] zur Modulation der Frequenz siehe Kap. 14. Amplitudenmodulation eher bei Mittel- und Langwelle, Frequenzmodulation bei UKW.

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_4

an, mit „großer“ Frequenz ω_T und Amplitude A. Um die bei der Amplitudenmodulation auftretenden Signale leicht verstehen zu können, konzentrieren wir uns als Beispiel auf die einfache niederfrequente Schwingung

$$S(t) = a\cos(2\pi\omega_N t),$$

mit Frequenz ω_N. Bei Sprache und Musik ist der Bereich der Niederfrequenz etwa

$$20 < \omega_N < 20\,000 \ \text{[Hz]},$$

die Hochfrequenz ω_T des Trägers liegt darüber.

Zur Illustration (Abb. 4.1) einer Amplitudenmodulation nehmen wir (unrealistisch) kleine Parameter an, nämlich $S(t) = \cos(2\pi t)$, also $a = 1$ und $\omega_N = 1$, sowie $A = 3$, und $\varphi_T = A\sin(40\pi t)$.

Aufgabe 1 *Eine amplitudenmodulierte hochfrequente Schwingung sei gegegeben durch*

$$\varphi(t) = (A + a\cos(2\pi\omega_N t))\sin(2\pi\omega_T t).$$

Die Parameter A, ω_T, a, ω_N bedeuten Amplitude und Frequenz der Hoch- bzw. Niederfrequenzschwingung.

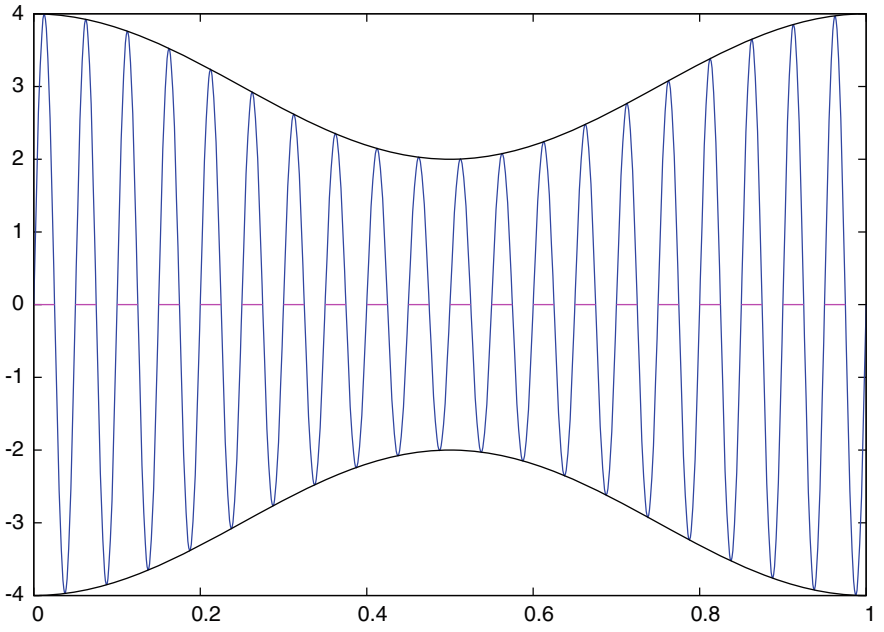

Abb. 4.1 Eine Amplitudenmodulation $\varphi(t) = (A + S(t))\varphi_T(t)$ über der t-Achse für $0 \le t \le 1$. In blau, speziell für $A = 3$, $S(t) = \cos(2\pi t)$, $\varphi_T(t) = A\sin(40\pi t)$, also $\omega_T = 20$. Die Hüllkurven $A + \cos(2\pi t)$ und $-(A + \cos(2\pi t))$ sind eingezeichnet, und die Teilintervalle der t-Achse mit $\varphi_T(t) > 0$ sind violett markiert

a) *Welche Frequenzen strahlt die Sendeantenne aus? (Träger, Seitenbänder)*
b) *Zur Veranschaulichung fertige man eine sorgfältige Zeichnung an von $\varphi(t)$ für $A = 3,\ \omega_\mathrm{T} = 5,\ a = 1,\ \omega_\mathrm{N} = 1$.*
c) *Bei der Doppelseitenband-Übertragung (DSB) wird der Träger $A\sin(2\pi\omega_\mathrm{T}t)$ herausgesiebt, es werden nur die Seitenbänder übertragen:*

$$\varphi_\mathrm{DSB}(t) := \varphi(t) - \varphi_\mathrm{T}(t).$$

Man fertige eine sorgfältige Skizze von φ_DSB an.
d) *Man vergleiche die Vorzeichenverteilung von φ_DSB mit derjenigen von φ. (Phasensprünge werden sichtbar.)*

Die Amplitude A der hochfrequenten Trägerschwingung φ_T wird moduliert mit der speziellen niederfrequenten Schwingung $a\cos(2\pi\omega_\mathrm{N}t)$, das heißt, aus φ_T wird mit Hilfe einer geeigneten Verstärkerschaltung $(A + S(t))\sin(2\pi\omega_\mathrm{T}t)$, hier speziell

$$\varphi(t) = (A + a\cos(2\pi\omega_\mathrm{N}t))\ \sin(2\pi\omega_\mathrm{T}t) \tag{4.1}$$

gebildet. Nach dem Additionstheorem

$$\sin\alpha\cos\beta = \frac{1}{2}\,[\,\sin(\alpha-\beta) + \sin(\alpha+\beta)\,] \tag{4.2}$$

der trigonometrischen Funktionen lässt sich $\varphi(t)$ umformen:

$$\begin{aligned}\varphi(t) &= A\sin(2\pi\omega_\mathrm{T}t) + \frac{a}{2}\,2\cos(2\pi\omega_\mathrm{N}t)\ \sin(2\pi\omega_\mathrm{T}t)\\ &= A\sin(2\pi\omega_\mathrm{T}t) + \frac{a}{2}\,[\,\sin(2\pi\omega_\mathrm{T}t - 2\pi\omega_\mathrm{N}t) + \sin(2\pi\omega_\mathrm{T}t + 2\pi\omega_\mathrm{N}t)\,],\end{aligned}$$

also

$$\varphi(t) = A\sin(2\pi\omega_\mathrm{T}t) + \frac{a}{2}\sin[(\omega_\mathrm{T} - \omega_\mathrm{N})2\pi t] + \frac{a}{2}\sin[(\omega_\mathrm{T} + \omega_\mathrm{N})2\pi t]. \tag{4.3}$$

Diese Summendarstellung (4.3) zeigt, aus welchen Frequenzen sich die amplitudenmodulierte Schwingung zusammensetzt: Neben der Trägerfrequenz ω_T treten noch die „untere Seitenfrequenz“ $\omega_\mathrm{T}-\omega_\mathrm{N}$ und die „obere Seitenfrequenz“ $\omega_\mathrm{T}+\omega_\mathrm{N}$ auf. Da sich die Niederfrequenzen ω_N eines allgemeinen Signals S über ein Band verteilen (ω_N variabel), bilden die Seitenfrequenzen ein oberes und unteres *Seitenband.*

Zum Zeichnen der amplitudenmodulierten Schwingung ist die Produktdarstellung (4.1) geeigneter.[2] Diese Darstellung zeigt, dass der niederfrequente Faktor die Amplitude und damit Hüllkurve des hochfrequenten Faktors $\sin(2\pi\omega_T t)$ ist. Dies wird am Beispiel von Abb. 4.1 deutlich. Man beachte, dass die Berührstellen mit den Hüllkurven nicht die Extremwerte sind, sondern Punkte mit gleicher Tangente.

Um im Empfänger die Originalinformation $S(t)$ zurückzuerhalten, kann die obere Hüllkurve herausgefiltert werden (Abb. 4.1). – Damit keine Missverständnisse auftreten, sei darauf hingewiesen, dass in den Anwendungen die Frequenzen im kHz- oder MHz-Bereich liegen.

Die Wahl obigen Beispiels mit $\varphi_T = A\sin(2\pi\omega_T t)$ und $S(t) = a\cos(2\pi\omega_N t)$ ist gar nicht so speziell. Außer dem Additionstheorem (4.2) gibt es analoge Formeln für die Produkte $\cos\alpha\cos\beta$ und $\sin\alpha\sin\beta$. Deswegen gelten die obigen Aussagen auch für einen Träger $A\cos(2\pi\omega_T t)$ oder ein Signal $S(t) = a\sin(2\pi\omega_N t)$, und für Summen von Sinus- und Kosinus-Schwingungen, und (mit dem Argument der Fourier-Approximation) für allgemeine Signale.

Doppelseitenband-Modulation

Zur Aussendung der Trägerfrequenz muss der Sender eine relativ hohe Leistung aufbringen. Um weniger Energie zu benötigen, wird der Träger herausgesiebt. Bei der Doppelseitenband(DSB)-Modulation[3] werden nur die beiden Seitenbänder übertragen:

$$\varphi_{DSB}(t) := \varphi(t) - \varphi_T(t) = S(t)\sin(2\pi\omega_T t)$$

Bei φ_{DSB} sind $+S(t)$ und $-S(t)$ die Hüllkurven (Abb. 4.2). Die violett markierten Teile der t-Achse illustrieren, dass an den Nullstellen der Hüllkurve Phasenwechsel auftreten. Die „Phase“ wird hier durch das Vorzeichen des Trägers φ_T bestimmt. Vor dem ersten Phasenwechsel an $t = \frac{1}{4}$ in Abb. 4.2 sind die Vorzeichen von φ_T und φ_{DSB} gleich; beide schwingen in die gleiche „Richtung“. In t-Intervallen übereinstimmender Phase ist die *obere* Hüllkurve das Originalsignal. Im Teilintervall $\frac{1}{4} < t < \frac{3}{4}$ sind die Vorzeichen von φ_T und φ_{DSB} verschieden, sie sind nicht „in Phase“. Bis zum nächsten Phasenwechsel an $t = \frac{3}{4}$ ist das Originalsignal durch die *negative* obere Hüllkurve gegeben. Die Phasenwechsel werden also durch das Aussieben des Trägers verursacht. Und die Phasenwechsel entscheiden im Empfänger, welche Hüllkurve das Originalsignal wiedergibt.

[2] Man wird dazu neigen, es sich mit einer Grafik-Routine bequem zu machen. Aber eine Handskizze mit dem niedrigen Wert $\omega_T = 5$ ist nicht schwer, sehr lehrreich, und auch im digitalen Zeitalter wärmstens zu empfehlen. Hierzu zeichne zuerst die beiden Hüllkurven, d. h. den niederfrequenten Faktor mit beiden Vorzeichen, danach markiere die Nullstellen des hochfrequenten Faktors.

[3] auch: Zweiseitenband-Modulation.

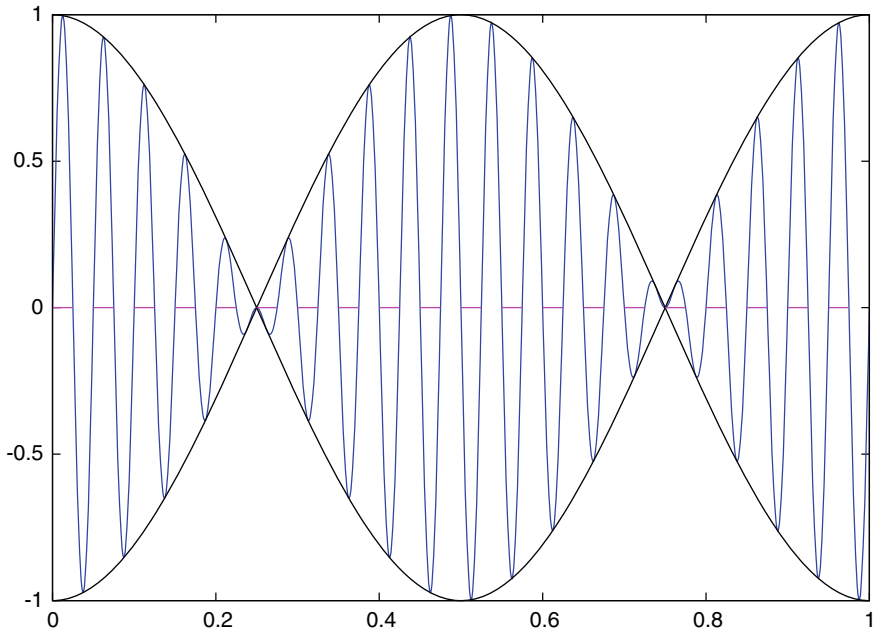

Abb. 4.2 Das Doppelseitenband (DSB) zu Abb. 4.1: $\varphi_{\mathrm{DSB}}(t) = \varphi(t) - \varphi_{\mathrm{T}}(t) = S(t) \cdot \varphi_{\mathrm{T}}(t)$ über t für $\omega_{\mathrm{T}} = 20$, mit Phasenwechseln an den Nullstellen der Hüllkurven $\cos(2\pi t)$ und $-\cos(2\pi t)$

Bei unserem einfachen Beispiel ist

$$\begin{aligned}\varphi_{\mathrm{DSB}}(t) &= a\cos(2\pi\omega_{\mathrm{N}}t)\ \sin(2\pi\omega_{\mathrm{T}}t)\\ &= \frac{a}{2}\sin[(\omega_{\mathrm{T}}-\omega_{\mathrm{N}})2\pi t] + \frac{a}{2}\sin[(\omega_{\mathrm{T}}+\omega_{\mathrm{N}})2\pi t].\end{aligned}$$

Die speziell gewählten Zahlenwerte führen in der Abb. 4.2 dazu, dass genau an den Nullstellen der Hüllkurve auch der Träger eine Nullstelle hat – eine sehr spezielle Situation von doppelten Nullstellen. Bei einem allgemeineren Niederfrequenzsignal kann die DSB-Schwingung zusätzliche Nullstellen gegenüber der Trägerschwingung haben. Die Abb. 4.3 illustriert für die Zahlenwerte von Aufgabe 1 (u. a. $\omega_{\mathrm{T}} = 5$), dass φ_{DSB} als Summe der beiden Seitenbänder erhalten wird.[4]

Abschließend sei als Ergänzung vermerkt, dass es außer der DSB-Modulation noch die Einseitenband-Modulation gibt, bei der nicht der Träger, sondern ein Seitenband herausgesiebt wird. Vorteil: Jeder Sender benötigt nur die halbe Frequenzbreite.

[4] Erwähnt sei das „umgekehrte“ Phänomen der *Schwebung:* Additionstheoreme wie (4.2) zeigen, dass die Summe von zwei Tönen mit ähnlicher Frequenz hörbar ist als Ton mit der halben Differenzfrequenz. Die Reduzierung oder Steuerung von Schwebungen ist eine Grundlage für das Stimmen von Musikinstrumenten.

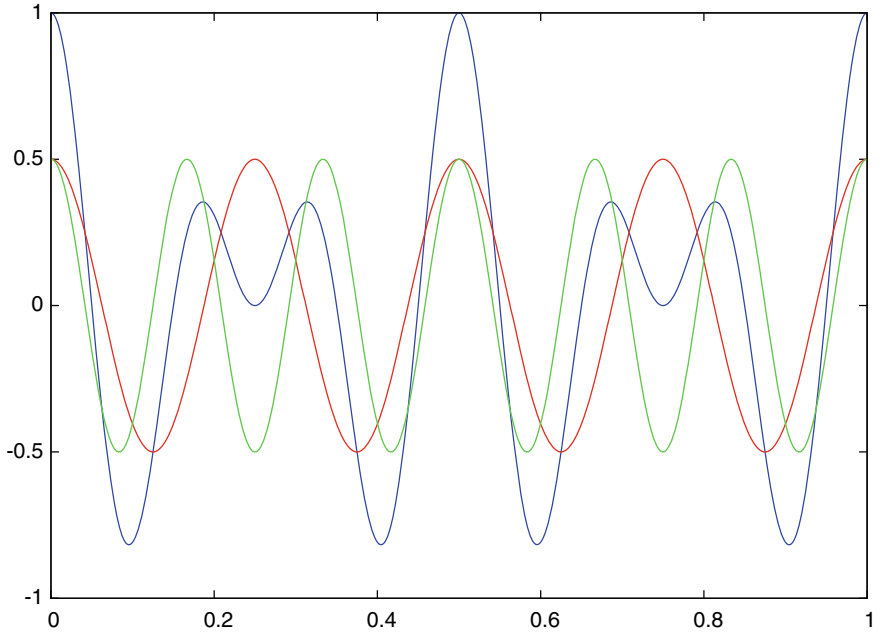

Abb. 4.3 Zu Aufgabe 1d mit $\omega_T = 5$: $\varphi_{DSB}(t)$ (in blau) als Summe der Seitenbänder $\frac{1}{2}\sin(8\pi t)$ und $\frac{1}{2}\sin(12\pi t)$ (rot und grün) für $0 \le t \le 1$. Frage an die Leser: Wo wären im Bild die Hüllkurven zu zeichnen?

4.2 Stereo-Signal

Multiplexsignal

Nach diesen Überlegungen zum Doppelseitenband sind wir vorbereitet, die Verschlüsselung von Stereosendungen beim Rundfunk zu verstehen. Wir verwenden die folgenden Bezeichnungen:

φ_L : Signal vom linken Mikrofon
φ_R : Signal vom rechten Mikrofon
φ_T : Trägerschwingung mit Frequenz 38 kHz

Mit diesen Schwingungen werden gebildet:

1) Summensignal: $\varphi_{L+R} = \varphi_L + \varphi_R$
2) Differenzsignal: $\varphi_{L-R} = \varphi_L - \varphi_R$
3) Mit dem Differenzsignal φ_{L-R} wird der Träger φ_T (Ultraschallbereich) amplitudenmoduliert. Aus dem entstandenen Frequenzgemisch wird die Trägerschwingung φ_T wieder herausgefiltert (DSB-Modulation).
4) Das verbliebene Frequenzgemisch wird zum Summensignal φ_{L+R} addiert.
5) Dazu wird noch eine Schwingung φ_P mit der halben Trägerfrequenz addiert (*Pilotton* mit 19 kHz).

Das so erhaltene Frequenzgemisch heißt *Multiplexsignal.* Mit dem Multiplexsignal wird die hochfrequente Senderwelle frequenzmoduliert.

Für zwei wiederum sehr einfach angenommene spezielle Eingangssignale untersucht die folgende Aufgabe 2 das Multiplexsignal.

Aufgabe 2

a) *Für die speziellen Signale*
$\varphi_L = \sin(2\pi\omega_L t)$
$\varphi_R = \sin(2\pi\omega_R t)$
$\varphi_T = A\sin(2\pi\omega_T t)$
$\varphi_P = b\sin(2\pi\frac{\omega_T}{2}t)$
ermittle man das Multiplexsignal.

b) *Welche Frequenzen enthält das Multiplexsignal, wenn Summen- und Differenzsignal Frequenzen bis 15 kHz enthalten?*

c) *Für* $\omega_L = 2$, $\omega_R = 1$, $\omega_T = 20$, $A = 1$, $b = 0$ *fertige man für* $0 \le t \le 1$ *eine sorgfältige Skizze des Multiplexsignals an.*

Die Figuren in Abb. 4.4 zeigen die für die Aufgabe angenommen einfachsten Eingangssignale. Die Abb. 4.5 zeigt das Differenzsignal, und Abb. 4.6 das Summensignal.

Nun zur Analyse der Aufgabe. Aus den Signalen vom linken und vom rechten Mikrofon, φ_L und φ_R, werden zunächst das Summensignal

$$\varphi_{L+R}(t) = \sin(2\pi\omega_L t) + \sin(2\pi\omega_R t)$$

und das Differenzsignal

$$\varphi_{L-R}(t) = \sin(2\pi\omega_L t) - \sin(2\pi\omega_R t)$$

gebildet. Das Summensignal ist die Information, welche von Monoempfängern (Geräte ohne Stereodecoder) wiedergegeben wird. Nach der Bildung von Summen-

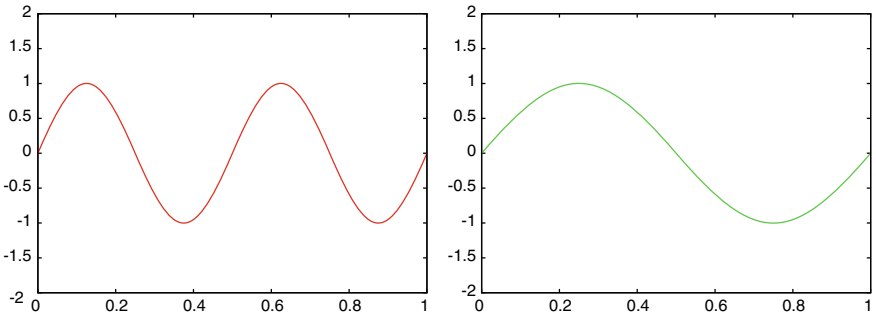

Abb. 4.4 Einfache niederfrequente Eingangssignale, links $\varphi_L(t) = \sin(2\pi 2t)$ und rechts $\varphi_R(t) = \sin(2\pi t)$, beide über der t-Achse für $0 \le t \le 1$

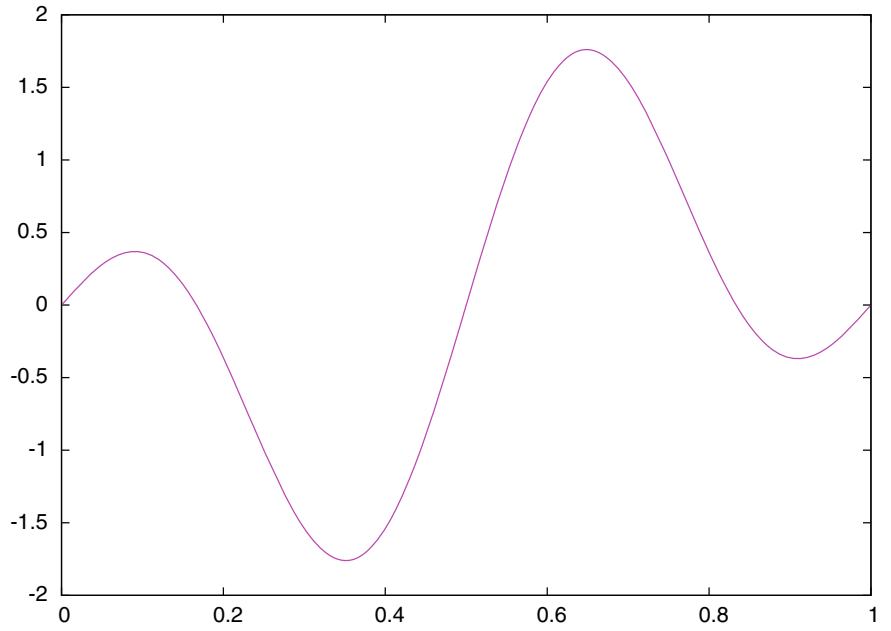

Abb. 4.5 Differenzsignal $\varphi_{\mathrm{L-R}}(t) = \sin(4\pi t) - \sin(2\pi t)$

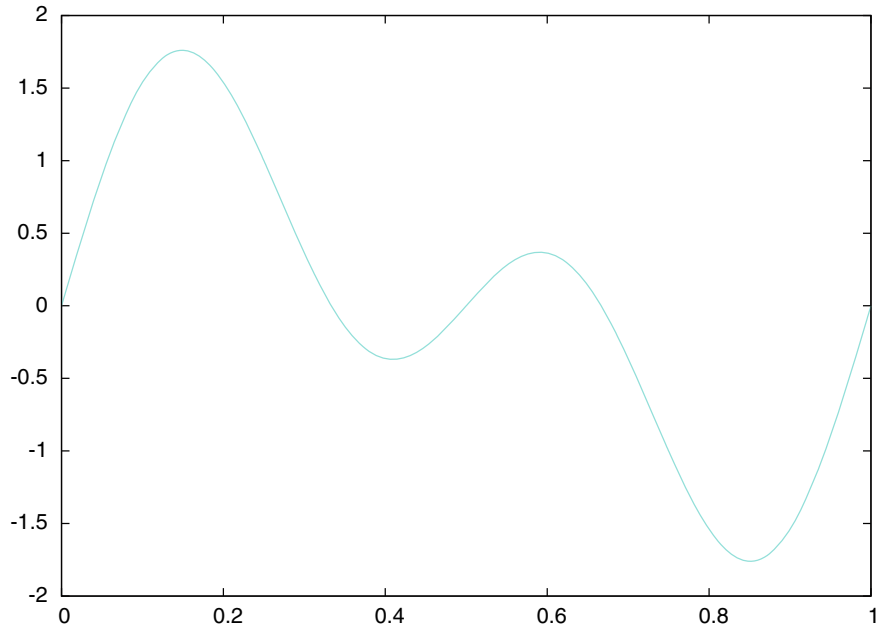

Abb. 4.6 Summensignal $\varphi_{\mathrm{L+R}}(t) = \sin(4\pi t) + \sin(2\pi t)$

und Differenzsignal wird die Hilfsträger-Schwingung (38 kHz) mit dem Differenzsignal amplitudenmoduliert:

$$\begin{aligned}
&(A + \varphi_{\mathrm{L-R}}) \sin(2\pi\omega_{\mathrm{T}}t) \\
&= A \sin(2\pi\omega_{\mathrm{T}}t) + \sin(2\pi\omega_{\mathrm{T}}t)\ \sin(2\pi\omega_{\mathrm{L}}t) - \sin(2\pi\omega_{\mathrm{T}}t)\ \sin(2\pi\omega_{\mathrm{R}}t) \\
&= A \sin(2\pi\omega_{\mathrm{T}}t) + \frac{1}{2}\{\cos[(\omega_{\mathrm{T}} - \omega_{\mathrm{L}})2\pi t] - \cos[(\omega_{\mathrm{T}} + \omega_{\mathrm{L}})2\pi t]\} \\
&\quad - \frac{1}{2}\{\cos[(\omega_{\mathrm{T}} - \omega_{\mathrm{R}})2\pi t] - \cos[(\omega_{\mathrm{T}} + \omega_{\mathrm{R}})2\pi t]\}.
\end{aligned}$$

Beim gebräuchlichen Pilottonverfahren wird zusätzlich der sogenannte Pilotton φ_{P} mit der halben Trägerfrequenz 19 kHz übertragen. Der Pilotton wird vom Sender phasenstarr zum Träger gesendet, damit hieraus im Empfänger der Träger exakt rückgewonnen werden kann. Nach dem Heraussieben des Trägers ist das Multiplexsignal komplett:

$$\begin{aligned}
\varphi(t) = \sin(2\pi\omega_{\mathrm{L}}t) + \sin(2\pi\omega_{\mathrm{R}}t) + b\sin(2\pi\frac{1}{2}\omega_{\mathrm{T}}t) \\
+ \frac{1}{2}\cos[(\omega_{\mathrm{T}} - \omega_{\mathrm{L}})2\pi t] - \frac{1}{2}\cos[(\omega_{\mathrm{T}} + \omega_{\mathrm{L}})2\pi t] \\
- \frac{1}{2}\cos[(\omega_{\mathrm{T}} - \omega_{\mathrm{R}})2\pi t] + \frac{1}{2}\cos[(\omega_{\mathrm{T}} + \omega_{\mathrm{R}})2\pi t].
\end{aligned}$$

Das Stereo-Multiplexsignal ist also die Summe von 7 Schwingungen mit den 7 Frequenzen

$$\omega_{\mathrm{L}},\ \omega_{\mathrm{R}},\ \frac{\omega_{\mathrm{T}}}{2},\ \omega_{\mathrm{T}} - \omega_{\mathrm{L}},\ \omega_{\mathrm{T}} - \omega_{\mathrm{R}},\ \omega_{\mathrm{T}} + \omega_{\mathrm{L}},\ \omega_{\mathrm{T}} + \omega_{\mathrm{R}}.$$

Vor der Bildung des Multiplexsignals filtert ein Tiefpassfilter die Frequenzen oberhalb von 15 kHz heraus, also verbleiben die Frequenzen ω_{L} und ω_{R} jeweils im Intervall

$$0 < \omega \leq 15\ \mathrm{kHz}.$$

Demzufolge liegen mit $\omega_{\mathrm{T}} = 38$ kHz die unteren Seitenfrequenzen im Band

$$23\ \mathrm{kHz}\ \leq \omega \leq 38\ \mathrm{kHz},$$

und die oberen Seitenfrequenzen liegen im Band

$$38\ \mathrm{kHz}\ \leq \omega \leq 53\ \mathrm{kHz}.$$

Die Lücke

$$15\ \mathrm{kHz}\ < \omega < 23\ \mathrm{kHz}$$

Abb. 4.7 Frequenzen des Hauptteils des Stereo-Multiplexsignals in kHz. Summensignal mit Frequenzen ≤ 15 kHz, Seitenbänder mit Frequenzen oberhalb 23 kHz; Pilotton mit 19 kHz. Die RSB-Seitenbänder bei 57 kHz sind nicht im Bild

lässt genügend Platz für die Unterbringung des Pilottons und stellt keine allzu großen Forderungen an die Steilheit des Tiefpassfilters, d. h. an die Striktheit obiger Ungleichungen.

Insgesamt umfassen die Frequenzen der Stereoinformation des Multiplexsignals also den Frequenzbereich

$$\begin{aligned} 1-15 \text{ kHz} \quad & \text{(Summensignal)} \\ 19 \text{ kHz} \quad & \text{(Pilotton)} \\ 23-53 \text{ kHz} \quad & \text{(DSB} - \text{verschlüsseltes Differenzsignal).} \end{aligned}$$

Das Frequenzspektrum des Multiplexsignals kann grafisch wie in Abb. 4.7 dargestellt werden. Jetzt ist auch klar, warum die Frequenz ω_P des Pilottons in der Größenordnung von 38 kHz liegen muss: So passen die Bänder des auf 15 kHz gestutzten Summensignals mit den beiden Seitenbändern und der Pilottonfrequenz gut zusammen.[5] Auch oberhalb von 53 kHz werden noch Frequenzen genutzt: so liegen bei der dreifachen Pilottonfrequenz $\omega = 57$ kHz weitere Informationen, wie RDS (Radio-Daten-System) mit zwei vergleichsweise schmalen Seitenbändern. Mit diesem breiten Frequenzgemisch (Abb. 4.7) wird dann die Trägerwelle des Senders (UKW zwischen 87,5 und 108 MHz) frequenzmoduliert.

Zum Zeichnen des Multiplexsignals lassen wir den Pilotton weg, er würde das Charakteristische verwischen. Zu skizzieren ist demnach

$$\varphi = \varphi_{L+R} + \varphi_{L-R} \cdot \varphi_T.$$

φ_{L+R} und φ_{L-R} jeweils als Produkte geschrieben, ergibt sich mit Hilfe von (4.2) und den gewählten Zahlen das spezielle Multiplexsignal

$$\varphi(t) = 2\cos(\pi t)\ \sin(3\pi t) + 2\cos(3\pi t)\ \sin(\pi t)\ \sin(40\pi t).$$

Ausgehend von dieser Produktdarstellung lassen sich φ_{L+R} und φ_{L-R} bequem über ihre Hüllkurven konstruieren. Der Faktor $\cos(3\pi t)\sin(\pi t)$ beispielsweise hat Nullstellen bei $t = \frac{1}{6}, \frac{1}{2}, \frac{5}{6}$, sowie bei 0 und 1. Das gleiche gilt dann auch für das

[5] Die Dämpfung von Frequenzen oberhalb 15 kHz durch einen Tiefpassfilter unterdrückt Frequenzen in der Nähe von 15 kHz weitgehend. Die Lücke ist so breit, dass das Summensignal sicher vom unteren Seitenband des Differenzsignals getrennt ist und so Übersprechen verhindert wird. Der Pilotton hat nur maximal 10 % des Modulationshubs.

Produkt $\varphi_{L-R}\varphi_T$. Diese drei Schwingungen, aus denen sich unser vereinfachtes Multiplexsignal zusammensetzt, sind in Abb. 4.8 einzeln skizziert, unter Aufgreifen der Farbwahl der vorigen Abbildungen. Zum Zeichnen kann auch die Eigenschaft

$$\varphi_{L+R}(t + \frac{1}{2}) = \varphi_{L-R}(t)$$

hilfreich sein, die sich für die hier betrachteten Signale φ_L und φ_R nachweisen lässt.

Die Addition beider Schwingungen aus Abb. 4.8 liefert das gewünschte Multiplexsignal. Das Zeichnen wird stark vereinfacht, wenn man noch die Hüllkurven des Multiplexsignals berechnet und in der Skizze verwendet:

Durch Umformen ergibt sich zunächst

$$\varphi = \varphi_L + \varphi_R + \varphi_L\varphi_T - \varphi_R\varphi_T = \varphi_L \cdot (1 + \varphi_T) + \varphi_R \cdot (1 - \varphi_T).$$

Es folgen die Abschätzungen

$$2\varphi_R \le \varphi \le 2\varphi_L \text{ im Fall } \varphi_R \le \varphi_L,$$
$$2\varphi_L \le \varphi \le 2\varphi_R \text{ im Fall } \varphi_L \le \varphi_R.$$

Die ursprünglichen Schwingungen φ_L und φ_R (mit Faktor 2) sind also gerade die Hüllkurven des Multiplexsignals (Abb. 4.9).

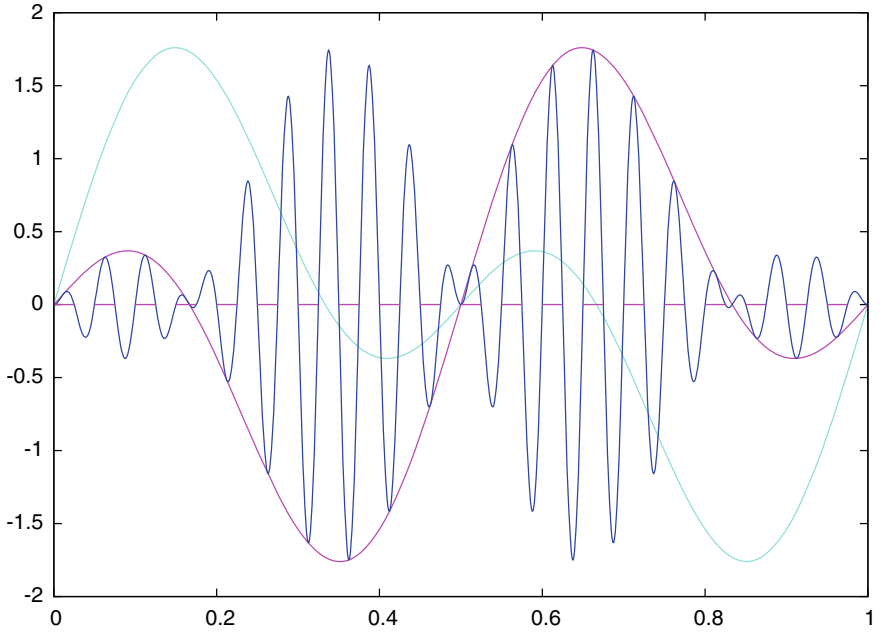

Abb. 4.8 Elemente des Multiplexsignals über der t-Achse: in blau $\varphi_{L-R} \cdot \varphi_T$, in magenta φ_{L-R}, in cyan φ_{L+R} für das Beispiel von Aufgabe 2

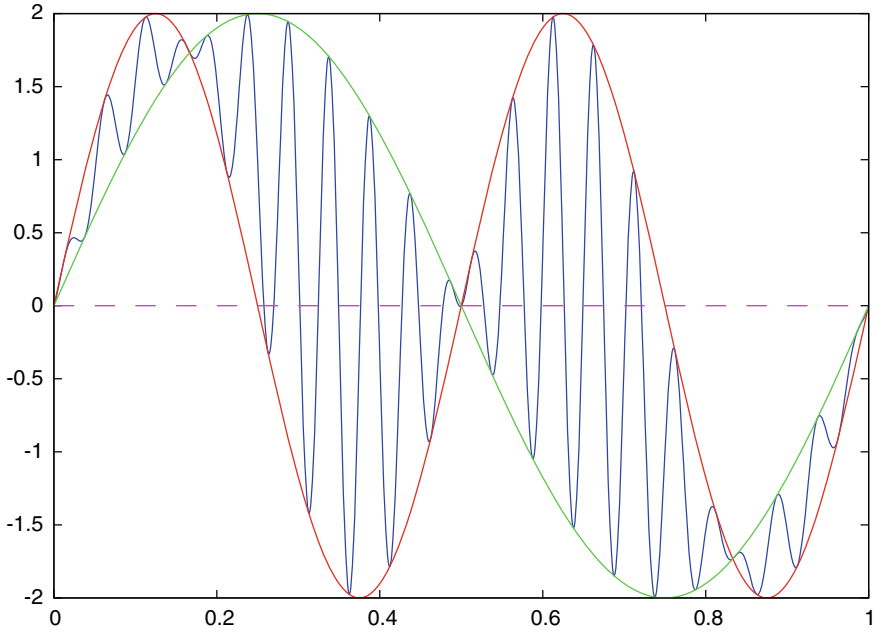

Abb. 4.9 Multiplex-Signal $\varphi_{L+R} + \varphi_{L-R} \cdot \varphi_T$ in blau (kein φ_P). Die Eingangssignale φ_L und φ_R von Abb. 4.4 sind als Hüllkurven erkennbar

Schon bei Abb. 4.8 werden die Phasenwechsel deutlich, die hier bei $t = 0,\ 1/6,\ 1/2,\ 5/6,\ 1$ liegen. In Abb. 4.9 zeigt sich, dass diese Phasenwechsel genau da liegen, wo sich die Hüllkurven schneiden. Das hat eine wichtige technische Konsequenz: Wie schon erwähnt, sind die Hüllkurven gerade die ursprünglichen Eingangssignale. Bei der Decodierung des Multiplexsignals im Empfänger kann man durch Herausfiltern der Hüllkurven die Informationen φ_L und φ_R zurückgewinnen. Je nach Phasenlage ist φ_L (bzw. φ_R) gerade die obere oder die untere Hüllkurve.

Dieses Herausfiltern der Hüllkurven ist nicht die einzige Methode zur Decodierung des Multiplexsignals. Eine andere Möglichkeit ist es, das Differenzsignal φ_{L-R} zurückzumodulieren. Anschließend können durch Addition und Subtraktion die Originalsignale zurückgewonnen werden:

$$\varphi_{L+R} + \varphi_{L-R} = 2\varphi_L,$$
$$\varphi_{L+R} - \varphi_{L-R} = 2\varphi_R.$$

Zum Abschluss noch ein eher historischer Hinweis: *Quadrophonie*-Aufzeichnungen auf Schallplatten können nach dem gleichen Prinzip verschlüsselt werden (CD-4 Verfahren): Auf beiden Rillenflanken wird je ein Multiplexsignal eingepresst. Jedes Multiplexsignal enthält durch ein Summensignal und ein verschlüsseltes Differenzsignal (30 kHz-Träger) zwei Informationen.

Literatur

Zu dem Thema gibt es zahlreiche Literatur, zum Beispiel

Plaßmann, W., Schulz, D. (Hrsg.): Handbuch Elektrotechnik. Springer Vieweg, Wiesbaden (2016)
Meinke, H.H., Gundlach, F.W.: Taschenbuch der Hochfrequenztechnik. Springer, 5. Aufl. (1992)
Lindner, H., Brauer, H., Lehmann, C.: Taschenbuch der Elektrotechnik und Elektronik, Hanser (2018)

Digitale Tonaufzeichnung

5

Digitalisierung

In den Kapiteln über Tonarm und Stereo-Rundfunk diskutierten wir Medien der analogen Welt. Unser Leben und unsere Natur sind analog, aber Datenverarbeitung wird heute weitgehend digital durchgeführt. Zu den immensen Vorteilen digitaler Daten gehört ihre langfristig stabile Speicherung und ihre schnelle Verarbeitung in Computern. Nun muss zuerst die analoge Welt *digitalisiert* werden, und nach der digitalen Verarbeitung oder Speicherung wieder zurückübersetzt werden in unsere analoge Welt. Der Prozess der Digitalisierung nebst seiner Umkehrung ist vielfältig entwickelt und auf die jeweilige Anwendung zugeschnitten. Wichtige Felder dieses großen Komplexes sind die Digitale Bildverarbeitung und die Digitale Signalverarbeitung.

Die Anforderungen sind jeweils unterschiedlich. Ein Tonsignal ist im Prinzip ein „endlos" langer kontinuierlicher Datenfluss, variierend mit der Zeit. Dagegen ist ein Bild begrenzt und ortsabhängig. Beim Tonsignal wird auf die Eigenschaften des menschlichen Hörvermögens Rücksicht genommen, beim Bild bestimmen Sehvermögen und Anforderungen an die Auflösung. Die angewendete Mathematik ist verschieden. Unterschiedliche mathematische Gesichtspunkte werden hier in getrennten Kapiteln behandelt: Die Bildverarbeitung ist Thema der beiden nächsten Kap. 6 und 7, und die Farb-Kodierung wird in Kap. 15 diskutiert. In diesem Kap. 5 konzentrieren wir uns auf die Digitale Tonaufzeichnung, speziell im Hinblick auf Audio-CDs (Compact Disk). Einige der zugrunde liegenden Prinzipien sind auch bei anderen Formen der Digitalisierung bedeutsam.

Diskretisierung des Tonsignals

Einem Tonsignal, etwa vom Mikrofon aufgezeichnet, entspricht eine zeitabhängige elektrische Spannung $U(t)$. Diese ist zunächst analog in doppelter Weise: Die Zeit t variiert kontinuierlich, und die Spannung U ebenfalls. Beide Variable, t und U,

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_5

müssen digitalisiert werden. Letztlich werden diese beiden Variablen in Folgen von Nullen und Einsen übersetzt, also in digitale Bits. Der Reichtum eines empfindsamen Musikstücks steckt dann in einer armseligen Kette aus 0 und 1. Nach Rückübersetzung in die analoge Welt und in ein erneut stetiges Signal $U(t)$ soll der Nutzer von der Art der Digitalisierung nach Möglichkeit nichts hören. Mit diesem Ziel werden die Algorithmen der Digitalisierung und der Verstetigung konstruiert.

Die folgende Aufgabe soll den Prozess der Diskretisierung und der Digitalisierung erläutern.

Aufgabe *Bei der digitalen Tonaufzeichnung wird die stetige Tonspannung mit der Frequenz $\omega = 44{,}1$ kHz abgetastet. Dabei wird $U(t)$ näherungsweise durch die diskreten Tonspannungen der stückweise konstanten Funktion*

$$\hat{U}(t) := U\left(\frac{1}{\omega}\text{floor}\,(\omega t)\right)$$

ersetzt.[1] *Jede einzelne Tonspannung von $\hat{U}(t)$ wird durch eine Dualzahl dargestellt. Bei der PCM-Technik (Pulse-Code-Modulation) stehen meist 16 Bits zur Verfügung: Die Differenz zwischen maximaler Tonspannung $U_{\max}$ und minimaler Tonspannung $U_{\min}$ kann also in 2^{16} gleichabständige Spannungsstufen aufgeteilt werden. Diese 16 Bits erlauben die Codierung eines Dynamikbereiches von etwa 98 dB.*

Aufgabe:

a) *Wie groß sind die Abstände der Spannungstufen bei $U_{\max} = 1$ Volt?*
b) *Wie lautet die binäre Darstellung einer Tonspannung von 0,001 Volt?*
Hinweis: Das Verhältnis V zweier Spannungen wird auf einer logarithmischen Skala in dB (Dezibel) gemessen:

$$V_{[\text{dB}]} = 20\,\log_{10}\left(\frac{U_{\max}}{U_{\min}}\right) \tag{5.1}$$

Abtastung

Zur Beschreibung der Pulsamplitudenmodulation diskretisieren wir zunächst die Zeit in äquidistanter Weise. Das heißt, wir betrachten statt eines kontinuierlichen Ablaufs von t nur diskrete Zeitpunkte $t_0,\ t_1,\ t_2, \ldots$, an denen das Signal U abgetastet wird. Der äquidistante Abstand $t_{k+1} - t_k$ (k natürliche Zahlen) entspricht der Periode $\frac{1}{\omega}$ einer Schwingung $\sin(2\pi\omega t)$. Entsprechend setzen wir

$$t_{k+1} - t_k = \frac{1}{\omega_{\text{A}}},$$

[1] Die floor-Funktion ordnet jeder Zahl x die größte ganze Zahl z mit $z \leq x$ zu. Die floor-Funktion wird auch als „Abrundungsfunktion“ bezeichnet. Andere Bezeichnungen: entier(x) , oder $\lfloor x \rfloor$.

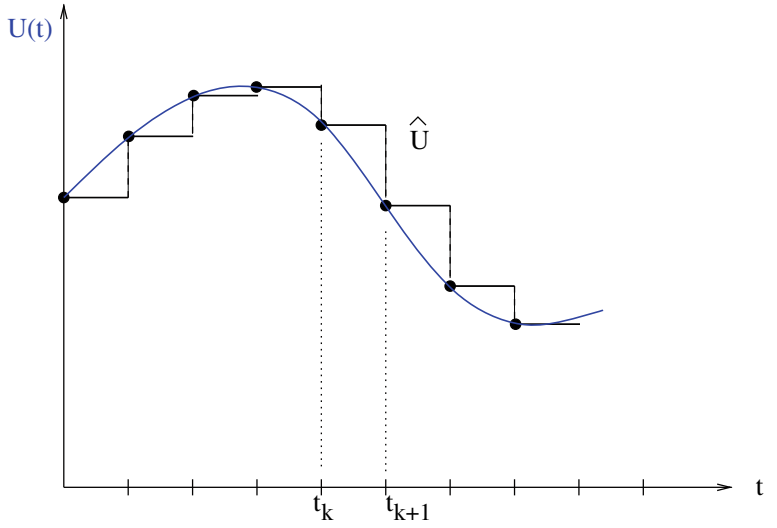

Abb. 5.1 Ein stetiges Signal U (blau) wird durch eine Treppenfunktion $\hat{U}$ approximiert

wobei ω_A die Abtastfrequenz ist, im Folgenden auch kurz mit ω bezeichnet. Also

$$t_k := k\,\frac{1}{\omega}.$$

Ersetzt man die stetige Tonspannungs-Funktion U durch eine Treppenfunktion $\hat{U}$ der t-Feinheit $\frac{1}{\omega}$ (vergleiche Abb. 5.1), so ist $\hat{U}$ gegeben durch

$$\hat{U}(t) := U(t_k) \text{ für } t_k \le t < t_{k+1}.$$

Es gelten die äquivalenten Beziehungen

$$\begin{aligned} t_k &\le t < t_{k+1} \\ \omega t_k &\le \omega t < \omega t_{k+1} \\ k &\le \omega t < k+1 \end{aligned}$$

und somit

$$k = \text{floor}(\omega t).$$

Die diskreten Zeitwerte lassen sich demnach darstellen als

$$t_k = \frac{1}{\omega}\,\text{floor}\,(\omega t),$$

und die Treppenfunktion ist

$$\hat{U}(t) = U\left(\frac{1}{\omega}\,\text{floor}\,(\omega t)\right).$$

Bei der praktischen Realisierung des Abtastens von U ist die Treppenfunktion $\hat{U}$ nicht in der Strenge dieser Definition verfügbar. In der Abtastschaltung wird das Eingangssignal U in Abständen $\frac{1}{\omega}$ in einem Schaltelement gespeichert, damit genügend Zeit zur Codierung des momentanen Abtastwertes bleibt. Bei Einsatz eines Kondensators werden wegen der Ladezeit die „senkrechten" Flanken von $\hat{U}$ (gestrichelt in Abb. 5.1) etwas schräg sein. Die hier definierte Treppenfunktion $\hat{U}$ ist eine Idealisierung.

Wahl der Abtastfrequenz

Die Wahl der Abtastfrequenz ω_A ist etwas knifflig, und nicht überall wird mit der gleichen Frequenz gearbeitet. Von grundlegender Bedeutung ist hierbei das *Abtasttheorem* nach Nyquist und Shannon. Dieses Gesetz besagt, dass beim Abtasten eines Signals keine Information verloren geht, wenn die Abtastfrequenz wenigstens doppelt so groß ist wie die höchste Frequenz des abzutastenden Originals.

Nehmen wir an, dass die Frequenzen des Originals U durch eine Bandbreite B begrenzt sind, oder durch einen Tiefpassfilter so begrenzt werden. Dann muss nach dem Abtasttheorem

$$2B < \omega_A$$

gelten. Anderenfalls, wenn in U Frequenzen größer als die halbe Abtastfrequenz $\frac{1}{2}\omega_A$ auftreten, können Alias-Effekte auftreten: Die Abtastwerte lassen dann andere Interpretationen als das Original U zu; das muss als Fehler angesehen werden. Um eine hohe Klangtreue von $\hat{U}$ zu ermöglichen, arbeitet man bei CD-Tonträgern mit der Bandbreite $B = 20\,\text{kHz}$. Also muss eine Abtastfrequenz von $\omega_A > 40\,\text{kHz}$ gewählt werden. Andererseits, um das Datenvolumen nicht zu groß werden zu lassen, wählt man ω_A nur wenig größer als $2B$. Der tatsächliche Wert von ω_A wurde bestimmt auf Grund von Forderungen, dass das Audiosystem kompatibel zu gebräuchlichen Videosystemen sein sollte. Die Wahl $\omega_A = 44{,}1$ garantiert also, dass sämtliche Informationen des auf 20 kHz bandbegrenzten Signales U in $\hat{U}$ enthalten sind.[2]

Puls-Code-Modulation

Wir haben oben mit der Pulsamplitudenmodulation die Zeit diskretisiert, aber noch nicht die Werte von U; noch ist jeder Wert $U(t_k)$ innerhalb eines gewissen Intervalles möglich. Bei der Puls-Code-Modulation (PCM) werden auch die möglichen Werte von U diskretisiert und digitalisiert. Diese Quantisierung teilt den Bereich möglicher Spannungswerte in eine bestimmte Anzahl von (z. B. gleichabständigen) Spannungsstufen auf. Hierzu stehen bei der PCM-Technik für Audio-CDs

$$2^{16} = 65\,536$$

[2] Beim Radio-Stereosignal mit $B = 15\,\text{kHz}$ genügt eine Abtastfrequenz von 32 kHz, in der Fernsprechtechnik mit $B = 3{,}4$ tastet man mit 8 kHz ab. Der Audio-CD-Standard ist im *Red Book* festgelegt. Die Wahl $\omega_A = 44{,}1$ kann Frequenzen bis 22 kHz aufnehmen.

Stufen für einen Spannungsbereich von etwa 98 dB zur Verfügung. Statt des aktuellen Wertes von $\hat{U}$ wird die zugehörige Spannungs-*Stufe* aufgezeichnet. Hierzu wird nicht der Spannungswert, sondern die Position der Stufe (eine Zahl zwischen 0 und 65 535) als 16-stellige Dualzahl dargestellt.

Die Diskretisierung in zwei Richtungen (Zeit t und Spannung U) kann man sich grafisch als Gitter vorstellen, das über den Graph der stetigen Funktion U gelegt wird (Abb. 5.2). Statt des aktuellen $U(t)$-Wertes wird als Näherung $\hat{U}$ eine der nächstliegenden Gittermaschen aufgezeichnet. Es liegt auf der Hand, dass das Gitter so fein gewählt werden kann, dass man die Differenz von $\hat{U}$ zu U nicht mehr sieht. Dies ist das grafische Analogon zur oben besprochenen akustischen Digitalisierung. Abtastfrequenz ω_A und Bitzahl 16 sind derart, dass das menschliche Ohr keinen Unterschied zwischen der wahren Tonspannung U und der Näherung $\hat{U}$ wahrnimmt. Insofern gilt die Speicherung von Audio-CDs als verlustfrei.

Bei einem Dynamikbereich von 98 dB hat man nach (5.1) als Spannungsverhältnis zwischen maximaler und minimaler Tonspannung

$$\frac{U_{\max}}{U_{\min}} = 10^{\frac{98}{20}}.$$

Die Differenz ΔU der 2^{16} gleichabständigen Spannungsstufen ist

$$\Delta U = \frac{U_{\max} - U_{\min}}{2^{16}} = U_{\max}\, 2^{-16}\,(1 - 10^{-\frac{98}{20}}).$$

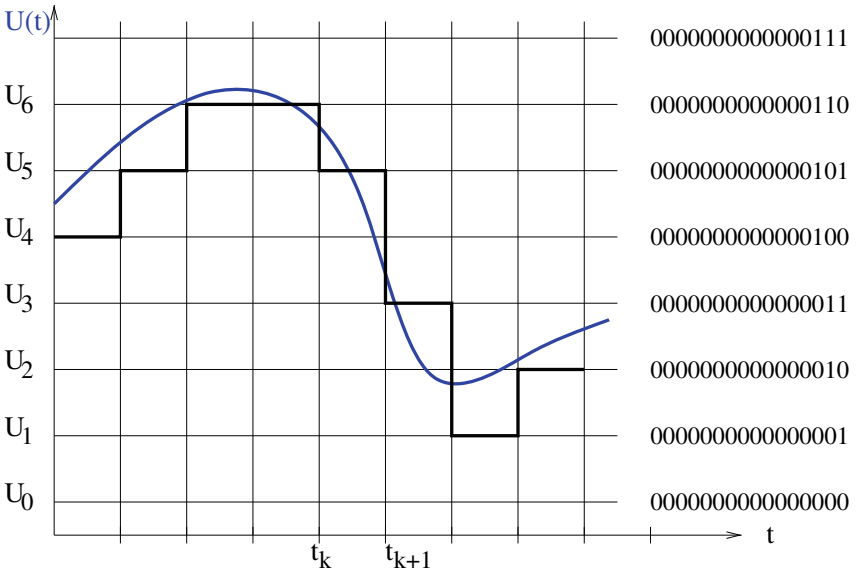

Abb. 5.2 Ein stetiges Signal U (blau) und seine PCM-Approximation (schwarz); das PCM-Gitter ist skizziert. $U_0 = U_{\min}$, $U_i = U_{\min} + i\,\Delta U$, $i = 1, 2, \ldots$, mit ΔU für den Abstand zwischen zwei Spannungsstufen. Rechts die binäre 16-stellige Tonstufe

Speziell für $U_{\max} = 1$ Volt erhält man

$$U_{\min} = 0{,}00001259 \text{ Volt}$$
$$\Delta U = 0{,}00001526 \text{ Volt}$$

(gerundet). Für die Nummer n der Spannungsstufe einer gegebenen Spannung $\hat{U}$ gilt wegen

$$U_{\min} + n\Delta U \leq \quad \hat{U} \quad < U_{\min} + (n+1)\Delta U$$
$$n \leq \frac{\hat{U} - U_{\min}}{\Delta U} < n + 1$$

die Beziehung

$$n = \text{floor}\left(\frac{\hat{U} - U_{\min}}{\Delta U}\right).$$

Für die spezielle Spannungsstufe $\hat{U} = 0{,}001$ erhält man

$$n = \text{floor}\,(64{,}7) = 64 = 2^6.$$

Die binäre Darstellung dieser Tonspannung ist

$$0000000001000000.$$

Speicheraufwand
Pro Sekunde sind also

$$44\,100 \cdot 16$$

binäre Informationen aufzuzeichnen. Für Stereoaufzeichnungen verdoppelt sich die Zahl, die Übertragungsfrequenz beträgt dann

$$1\,411\,200 \text{ Hz},$$

d. h. pro Sekunde sind etwa 1.4 Mio. Bits zu übertragen oder zu speichern. Das bedeutet für einen CD-Tonträger mit einer Stunde Spielzeit die enorme Speicherdichte von mehreren Milliarden Binärinformationen allein für das Nutzsignal (Musik, Sprache). Damit ist die Speicherkapazität der Audio-CD noch keineswegs erschöpft, weitere digitale Daten dienen dem Fehlerschutz, der Fehlererkennung und dem Bedienungskomfort.

Schematische Zusammenfassung
Eingangssignal $U(t)$
↓
Tiefpassfilter: Frequenzen oberhalb 20 kHz werden abgeschnitten.
↓

Abtastung: Diskretisierung der Zeitachse (kein Informationsverlust)
↓
Quantisierung: Diskretisierung der Spannungs-Achse
↓
Codierung: Binärsignal wird gebildet.
↓
Sicherung: Maßnahmen für Fehlerschutz, wie Einfügen von Prüfbits
↓
CD-Tonträger

Verlustbehaftete Digitalisierung
Wie oben dargestellt, gilt die Audio-CD-Digitalisierung als verlustfrei. Diese schöne Eigenschaft wird durch riesige Datenmengen erkauft. Lässt man geringe Einbußen an Genauigkeit und Tontreue zu, dann können enorme Einsparungen an Datenvolumen erreicht werden. Hierbei werden Erkenntnisse der *Psychoakustik* berücksichtigt, die aussagen, in welchen Bereichen ein Verlust an Qualität hörbar ist. Das Wissen über die menschliche Hör-Wahrnehmung erlaubt eine Reduzierung der Daten derart, dass die Einbuße an Qualität kaum merkbar ist. Ein weitverbreitetes verlustbehaftetes Verfahren ist das MP3-Verfahren.[3]

[3] MP3 steht für *MPEG1 Layer III,* ein Standard, der 1993 von der *Moving Picture Experts Group* festgelegt wurde. Ein weiteres verlustbehaftetes Verfahren ist *Dolby Digital*.

Bild- und Daten-Struktur

6

Riesige Datenmengen sind heute zu verarbeiten, und ihre Menge nimmt stetig zu. Die zentrale Frage dabei ist: *Was ist das Wesentliche?* Gibt es eine Struktur in den Daten, kann man ein Muster erkennen? Kann man die Datenmenge ohne gravierenden Informationsverlust reduzieren?

Eine wichtige Anwendung ist die Bildverarbeitung. Zum Beispiel möchte man digital aufgenommene Fotos möglichst effizient speichern. Hier gibt es mit JPEG[1] eine weitverbreitete Methode der Datenkomprimierung, auf die wir im folgenden Kap. 7 eingehen.

Im Folgenden wollen wir nicht nur Daten komprimieren, sondern den Gehalt beispielsweise eines Bildes in geeigneter Weise strukturieren. Eine erkannte Struktur kann dann auch der Bild-Erkennung dienen. Das wird nicht nur bei Fotos angewendet, sondern auch bei allgemeinen Datensätzen. Zum Beispiel können Aktienkurse analysiert werden (Abb. 6.1); dazu später mehr.

Die in diesem Kapitel vorgestellte Analyse ermittelt die *Hauptkomponenten.*[2] Dabei werden mit Eigenvektoren und Eigenwerten wichtige Elemente der Linearen Algebra verwendet, und zusätzlich scharfsinnige Algorithmen der Numerischen Mathematik. Außerdem spielen Aspekte der Stochastik und der Optimierung hinein, also ein breites Spektrum mathematischer Methoden.

Bilder und Daten als Zufallsinformation

Als Datensatz nehmen wir eine Menge von M Tupeln/Punkten. Ein Beispiel sind die täglichen DAX-Notierungen von Aktiengesellschaften, etwa der drei Gesellschaften

[1] JPEG, Joint Photographic Experts Group.
[2] PCA, principal component analysis.

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_6

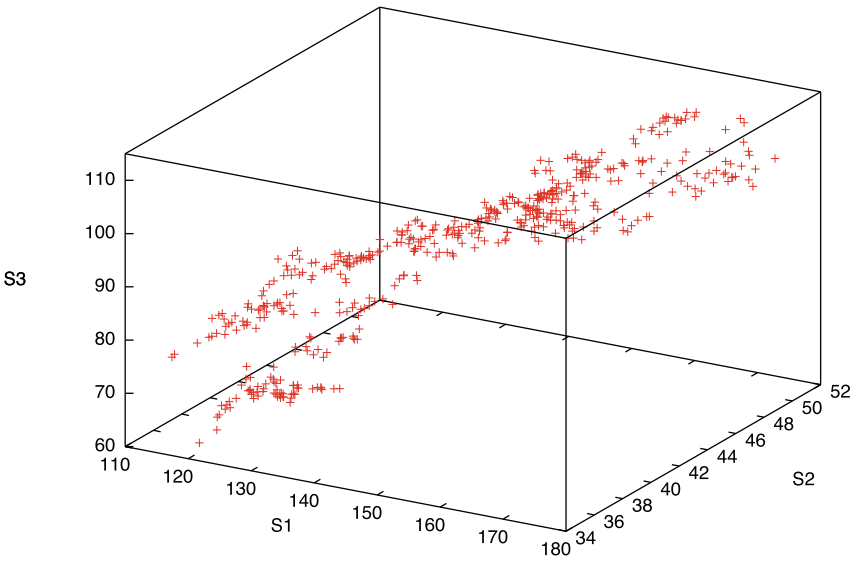

Abb. 6.1 Drei DAX-Aktienkurse: Allianz (S1), BMW (S2), und HeidelbergCement (S3); 500 Handelstage ab dem 5.Nov.2005 (in rot). Die Punktewolke legt nahe, dass eine gewisse Struktur zugrunde liegen könnte

S1, S2, S3 an 500 aufeinanderfolgenden Handelstagen von Abb. 6.1.[3] Diese Notierungen bilden 500 drei-dimensionale Tupel, die als Punkte in Abb. 6.1 wiedergegeben sind. Solche Tupel sind zufallsabhängig, können aber gewisse Abhängigkeiten oder Korrelationen aufweisen. Ein anderes Beispiel sind die Farbwerte der Pixel eines Fotos. Auch ein Bild kann als Zufallsprodukt gesehen werden, jedenfalls aus Sicht des bildverarbeitenden Algorithmus. Bei Abb. 6.1 ist der Zufalls-Charakter offensichtlich auf Grund der stochastischen Quelle. Die Information eines oder mehrerer Tupel oder Pixel ist dann Realisierung einer Zufallsvariablen.

Beispiel: Ein Bild von 400 × 400 Pixeln wird aufgeteilt in 100^2 *„Kacheln" zu je* $4^2 = 16$ *Pixeln. Ordnet man jeder Kachel einen Vektor* $\tilde{x} \in \mathbb{R}^{16}$ *zu, dann ergibt sich das Bild durch* 100^2 *Ziehungen der Zufallsvariablen.*

Die Vektoren $\tilde{x}$ der M $(> n)$ Daten bilden eine Punktewolke im $\mathbb{R}^n$, zum Beispiel mit $n = 3$ in Abb. 6.1, $n = 2$ in Abb. 6.2, oder $n = 16$ bei obigem Beispiel. Eine Punktewolke ist im Allgemeinen nicht regellos verteilt, sondern hat eine Struktur oder ein Muster. Eine solche Struktur deutet sich zum Beispiel in Abb. 6.1 an, und ist illustriert in der schematischen Abb. 6.2. Die Punkte häufen sich vielleicht an einer Stelle, jedenfalls ist ihr Mittelpunkt (Mittelwert, Erwartungswert) $\mu := \mathsf{E}(\tilde{x})$,

[3] Die Auswahl von drei Aktiengesellschaften in Abb. 6.1 ist willkürlich, deswegen neutral mit S1, S2, S3 bezeichnet.

definiert.[4] Die Abstände x der Punkte vom Mittelpunkt der Daten μ sind Vektoren im $\mathbb{R}^n$, sie enthalten die Struktur. Die Bezeichnungen zusammengefasst:

$$\begin{aligned} \tilde{x} &\in \mathbb{R}^n \quad \text{Zufallsvariable, Daten} \\ \mu &:= \mathsf{E}(\tilde{x}) \quad \text{Erwartungswert, Mittelwert der Daten} \\ x &:= \tilde{x} - \mu \quad \text{Abweichung vom Mittelwert} \end{aligned}$$

Haben diese Vektoren x eine oder mehrere „Hauptrichtungen“, welche die Struktur der Daten charakterisieren? Ein solcher spezieller Richtungsvektor sei mit v bezeichnet, und wir nehmen ihn als auf euklidische Einheitslänge normiert an. Das Skalarprodukt $x^{tr}v$ hat eine wichtige geometrische Bedeutung:

$$x^{tr}v = \text{Länge der Projektion von } x \text{ auf } v \text{ wenn } \|v\|_2 = 1.$$

Dabei bezeichnet ^{tr} die Transposition; $x^{tr}v$ ist also „Zeile mal Spalte“, eben das euklidische Skalarprodukt.[5] Eine Hauptrichtung v zeichnet sich dadurch aus, dass die Länge der Projektionen der Zufallsvektoren x auf v maximal ist (Abb. 6.2). Damit haben wir die folgende Optimierungsaufgabe:

Gesucht ist ein Vektor v, normiert mit $\|v\|_2 = 1$, so dass für die Zufallsvariable $(x^{tr}v)^2$ gilt: $\mathsf{E}((x^{tr}v)^2)$ ist maximal!

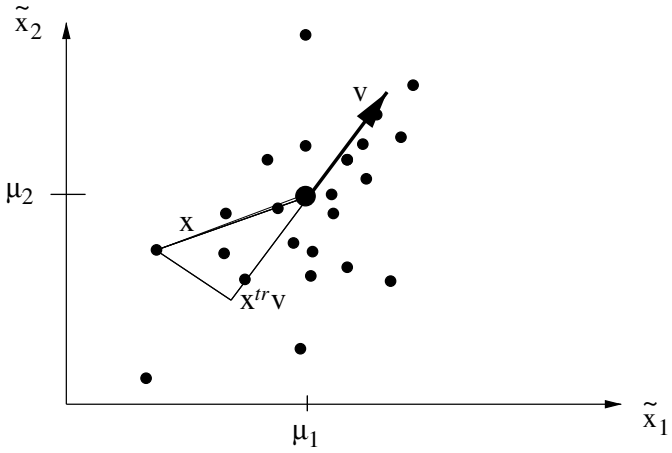

Abb. 6.2 schematisch im Fall $n = 2$: Datenwolke; jedem (kleinen) Punkt entspricht ein Datenvektor $\tilde{x}$. Erwartungswert μ (dicker Punkt), ein Richtungsvektor x, und die Projektion $x^{tr}v$ auf einen Vektor v

[4] Bei Vektoren und Matrizen wird der Erwartungswert komponentenweise genommen, also $\mu \in \mathbb{R}^n$.
[5] x ist Spaltenvektor, x^{tr} ist Zeilenvektor, $\|v\|_2 := (\sum_i v_i^2)^{1/2}$ und $x^{tr}v = \sum_i x_i v_i$, wenn x_i und v_i die Komponenten der Vektoren x und v bezeichnen, für $i = 1, \ldots, n$.

Hier tritt die Kovarianzmatrix $\Sigma := \mathsf{E}((\tilde{x} - \mu)(\tilde{x} - \mu)^{tr})$ auf, denn

$$\mathsf{E}((x^{tr} v)^2) = \mathsf{E}((x^{tr} v)^{tr} (x^{tr} v)) = \mathsf{E}(v^{tr} x x^{tr} v) = v^{tr} \mathsf{E}(x x^{tr}) v = v^{tr} \Sigma v.$$

Anschaulich bedeutet diese Maximierung: In Richtung v liegt besonders viel Bild-Information. In der Kovarianz-Matrix $\Sigma \in \mathbb{R}^{n \times n}$ stecken unter anderem die Korrelationen. Sie ist nach Definition symmetrisch und positiv definit. Also müssen ihre Eigenwerte λ positiv sein.

Lösung der Maximierungsaufgabe
Zu maximieren ist also $v^{tr} \Sigma v$ unter der Nebenbedingung $\|v\|_2^2 - 1 = 0$. Das Prozedere kennen wir aus der Analysis: Die Nebenbedingung wird mit einem Lagrange-Multiplikator l angekoppelt. Zu maximieren ist dann

$$\Phi(v) := v^{tr} \Sigma v + l(v^{tr} v - 1).$$

Der Gradient verschwindet an einem (inneren) Maximum:

$$0 = \operatorname{grad} \Phi = 2\Sigma v + 2lv.$$

Es folgt

$$\Sigma v = -lv \tag{6.1}$$

und damit ein Eigenwertproblem der Linearen Algebra. Der Lösungsvektor v der Maximierungsaufgabe ist Eigenvektor der Kovarianzmatrix Σ mit Eigenwert $\lambda = -l$, und es folgt mit (6.1)

$$\mathsf{E}((x^{tr} v)^2) = v^{tr}(-lv) = -l = \lambda > 0. \tag{6.2}$$

Festzuhalten ist: Der zum größten Eigenwert λ gehörende Eigenvektor v maximiert die Varianz $\mathsf{E}((x^{tr} v)^2)$.

Da die Matrix Σ symmetrisch ist, sind die Eigenvektoren zu paarweise verschiedenen Eigenwerten orthogonal. Die Eigenwerte seien nach ihrer Größe nummeriert, λ_1 ist der größte Eigenwert.[6] Die Eigenvektoren seien

$$v^{(1)}, \ldots, v^{(n)},$$

in der Nummerierung nach Größe der Eigenwerte, d. h. $v^{(1)^{tr}} \Sigma v^{(1)} = \lambda_1$ ist maximal. Also ist die anfangs gesuchte Hauptrichtung v gegeben durch den Eigenvektor $v^{(1)}$ zum größten Eigenwert λ_1 der Kovarianzmatrix Σ.

[6] Den Sonderfall eines mehrfachen Eigenwertes lassen wir hier der Einfachheit halber weg.

Transformation
Da die Eigenvektoren ein Orthonormalsystem bilden, ist die Matrix aus den Spalten

$$B := (v^{(1)}, \ldots, v^{(n)})$$

orthogonal. Es gilt also $B^{-1} = B^{tr}$, oder $BB^{tr} = I$, wobei I die Einheitsmatrix ist.[7] Damit können die Datenvektoren $\tilde{x}$ auch so geschrieben werden:

$$\tilde{x} = \mu + Ix = \mu + BB^{tr}x. \tag{6.3}$$

Das Produkt $B^{tr}x$ ist der Spaltenvektor

$$y := B^{tr}x = \begin{pmatrix} v^{(1)^{tr}}x \\ \cdot \\ \cdot \\ \cdot \\ v^{(n)^{tr}}x \end{pmatrix},$$

dessen Komponenten $y_i = v^{(i)^{tr}}x$ $(i = 1, \ldots, n)$ jeweils Skalarprodukte der Eigenvektoren $v^{(i)}$ mit x sind. Das verbleibende Produkt By ist

$$By = \sum_{i=1}^{n} v^{(i)} y_i \,,$$

insgesamt gilt also nach (6.3)

$$\tilde{x} = \mu + By = \mu + \sum_{i=1}^{n} v^{(i)} \cdot (v^{(i)^{tr}}x)\,, \tag{6.4}$$

die Daten $\tilde{x}$ ergeben sich durch Linearkombination der Eigenvektoren von Σ. Die Größenordnung der Koeffizienten $y_i = v^{(i)^{tr}}x$ kennen wir aus (6.2)

$$\mathsf{E}(y_i^2) = \lambda_i \;\geq \mathsf{E}(y_{i+1}^2).$$

Wenn für ein k der Eigenwert λ_k die kleineren Eigenwerte dominiert im Sinne von $\lambda_k \gg \lambda_{k+1}$, dann ist

$$\mu + \sum_{i=1}^{k} v^{(i)}(v^{(i)^{tr}}x) \tag{6.5}$$

eine Näherung für $\tilde{x}$. Wenn k klein ist, dann benötigt die Näherung (6.5) im Gegensatz zu (6.4) nur wenig Speicherplatz. Speziell für $k = 2$ spannen $v^{(1)}$ und $v^{(2)}$ durch

[7] Zur Matrix B ist B^{tr} ist die transponierte Matrix, also an der Hauptdiagonale gespiegelt.

(6.5) eine Ebene auf. Für große n kann es eventuell mehrere solche k geben. Und der Test auf Dominanz $\lambda_k \gg \lambda_{k+1}$ sollte relativ sein, also skaliert beispielsweise durch die Summe der n Eigenwerte.

Eine Transformation wie (6.4) taucht in ähnlicher Form in mehreren Bereichen der Angewandten Mathematik auf, beispielsweise unter den Namen Hauptachsen-Transformation, oder Hotelling-Transformation. Die Eigenwerte (mit jeweils zugehörendem Eigenvektor) heißen auch *principal components.* Die Matrix B vermittelt eine Transformation der Daten auf ein rechtwinkliges Koordinatensystem. Der Eigenwert λ_2 maximiert die Varianz orthogonal zu $v^{(1)}$, der Eigenwert λ_3 orthogonal zur durch $v^{(1)}$, $v^{(2)}$ aufgespannten Ebene, u.s.w.

Anwendung: Aktienkurse
In Abb. 6.1 sind von drei Aktien 500 Aktienkurs-Tripel dargestellt. Eine Frage ist es, ob die drei gewählten Aktien miteinander in Beziehung stehen, ihre Dynamik etwa eine gemeinsame Struktur bildet. Ein Hintergrund ist die Aussage von Markowitz, dass man ein Portfolio von Aktien so bilden sollte, dass Diversifikation (Risikostreuung) gesichert ist, um das Risiko kleiner zu halten.[8] Die Eigenwerte der Kovarianzmatrix sind (auf ganze Zahlen gerundet)

$$401, \quad 26, \quad 3.$$

Relativ gesehen, beansprucht λ_1 etwa 93 % der Summe der Eigenwerte, und λ_2 beansprucht 6 %. Die Eigenwerte klaffen also auseinander, ein Hinweis auf eine zugrundeliegende Struktur. Der Größenvergleich lässt erwarten, dass die Kursdynamik bereits mit $k = 1$ durch die Gerade

$$\mu + v^{(1)} y, \quad y \in \mathbb{R}$$

gut wiedergegeben wird, erst recht mit $k = 2$. Die Abb. 6.3 bestätigt das visuell. Der dominierende Eigenvektor $v^{(1)}$ definiert die blaue Gerade. Die durch $v^{(1)}$, $v^{(2)}$ aufgespannte Ebene ist in grün eingezeichnet. Die roten Punkte liegen nicht genau auf der grünen Ebene, sondern nur relativ nahe. Analog (in der Abbildung offensichtlich) gilt das für die blaue Gerade. Wären die Eigenwerte in ähnlicher Größenordnung, dann hätte sich keine Struktur ergeben.

Aus der Analyse könnte man den etwas gewagten Schluss ziehen, dass am 501. Handelstag die Kurse wahrscheinlich wiederum in der Nähe dieser Geraden oder Ebene liegen. Für eine praktische Portfolio-Analyse bedeutet ein starkes Auseinanderklaffen der Eigenwerte einen Hinweis auf hohe Korrelationen zwischen den Aktien und eine zu geringe Diversifikation, und ein zu hohes Risiko.

[8] Tatsächlich steckt die Dynamik eines Kapitalmarktes eher in den *relativen Veränderungen,* den *returns,* entsprechend müssten die Markt-Preise noch aufbereitet werden. In unseren Ausführungen geht es um die Methodik, noch nicht um ein marktgerechtes Urteil. Die Erkenntnisse von Markowitz (1952) wurden 1990 mit einem Nobel-Preis gewürdigt.

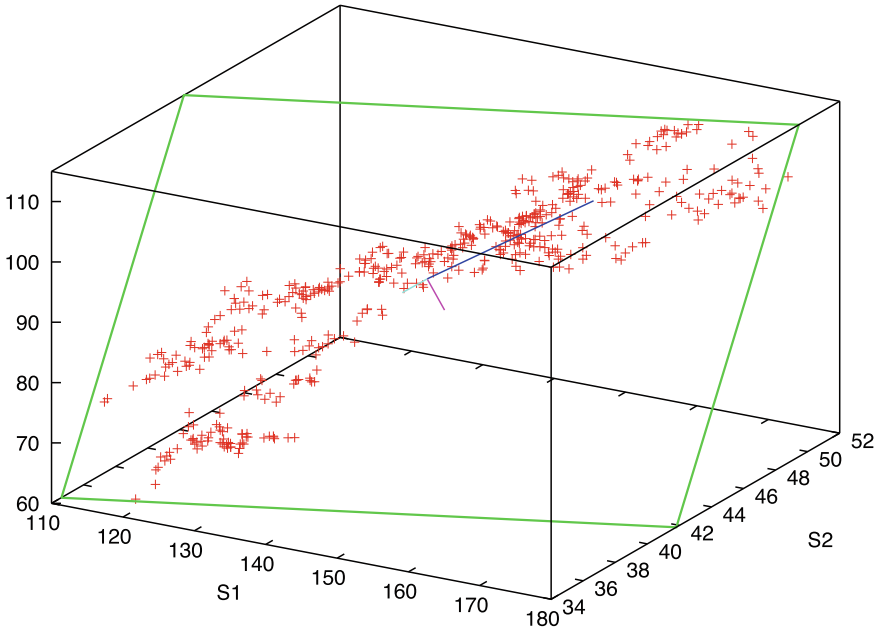

Abb. 6.3 wie Abb. 6.1, jetzt aber mit den Ergebnissen der Hauptkomponenten-Analyse: Die Eigenwerte der Kovarianzmatrix sind 400,8; 25,8; 2,73. Die Eigenvektoren $v^{(1)}$, $v^{(2)}$, $v^{(3)}$ sind die drei Geradenstücke, jeweils vom Mittelwert μ ausgehend und mit $\sqrt{\lambda}$ skaliert. Die grün markierte Ebene wird von $v^{(1)}$, $v^{(2)}$ aufgespannt (die aufspannenden Vektoren in blau und magenta), der Vektor $v^{(3)}$ (cyan) weist aus der Ebene heraus

Anwendung: Bildkompression

Die Näherung (6.5) eignet sich allgemein zur Kompression von Daten, und damit auch speziell zur Bildkompression. Allerdings gibt es zur Bildkompression leistungsfähigere Methoden, wie JPEG (Kap. 7). Trotzdem illustrieren wir die Anwendung auf Bildkompression mit den zwei Bildern der Abb. 6.4 (Original) und 6.5 (datenreduziert).

In ganz ähnlicher Weise wie die Hauptkomponenten-Analyse arbeitet die Singulärwert-Analyse (hier nicht erklärt). Die $(l \times m)$-Matrix der Bildinformation kann als Summe von gewichteten Eigenbildern dargestellt werden, die Gewichte sind die *Singulärwerte*. Ähnlich wie in (6.5) wird die Summe nach k Termen abgebrochen. Die Speichererersparnis ist erheblich, wenn $k \ll l, m$ ist. Im Hinblick auf die große Anzahl von Pixeln in Abb. 6.4 bedeutet $k = 100$ nominell eine große Einsparung.[9]

[9] Durch die Überlagerung von JPEG- und EPS-Formaten fällt die tatsächliche Ersparnis in der Darstellung von Abb. 6.5 eher gering aus.

Abb. 6.4 Y-Information des Farbbildes Abb. 15.2, siehe Kap. 15. Das Bild hat 2710×2919 Pixel, also $l = 2919$ und $m = 2710$

Algorithmen

Die Berechnung von Eigenwerten und Eigenvektoren erfolgt mit scharfsinnigen numerischen Algorithmen, etwa mit dem QR-Verfahren. Noch raffinierter sind die Algorithmen der Singulärwert-Zerlegung. Für diese Algorithmen verweisen wir auf die Literatur. Die Singulärwert-Zerlegung zu den Daten der Abb. 6.4 wurde mit MATLAB berechnet.

Abb. 6.5 Reduzierte Daten von Abb. 6.4 ($k = 100$)

Literatur

zu PCA:

Jolliffe, I.T.: Principal Component Analysis. Second Edition. Springer, New York (2002)

zu Algorithmen der Linearen Algebra (z.B. zum QR-Algorithmus und zur Singulärwert-Zerlegung):

Golub, G.H., Van Loan, C.F.: Matrix Computations. Fourth Edition. John Hopkins University Press, Baltimore (2013)

zur Risiko-Streuung bei Aktien

Markowitz, H.M.: Portfolio Selection. J. of Finance **7**:77–91 (1952)

Bildkompression und JPEG

7

Ein hochauflösendes digitales Foto besteht aus Tausenden von Pixeln in beiden Richtungen (x-Achse horizontal, y-Achse vertikal). Damit sind pro Bild Millionen von Informationen aufgenommen, Helligkeits- und Farbinformationen für jedes Pixel. Das erfordert Kompressions- und Speichermethoden, die das Bild kompakter darstellen, dabei aber Verluste an Information und Bildqualität möglichst nicht sichtbar werden lassen. Hier hat JPEG, die *Joint Photographic Experts Group,* einen Standard etabliert. Die JPEG-Methodik besteht aus einem breiten Spektrum von Maßnahmen und Tricks. Dabei ist die wesentliche mathematische Grundlage die *Diskrete Kosinus-Transformation,* kurz als DCT abgekürzt. Der JPEG-Zugang ist ein wunderbares Beispiel für das Zusammenwirken mehrerer Disziplinen, wie Mathematik, Visualisierung, und Computertechnologie.

Die Pixel eines digitalen Bildes seien in der Horizontalen mit i oder k indiziert und in der Vertikalen mit j oder l, und eine Bild-Information am Pixel (i, j) wird hier mit $f_{i,j}$ bezeichnet. Dieser Wert f definiert die Helligkeit am Pixel (i, j) oder einen der beiden Farbwerte[1]. Im Folgenden nehmen wir eine dieser drei Komponenten, betrachten also f als Skalar. Die Werte der $f_{i,j}$ können wir uns als eine riesengroße Matrix vorstellen. Um effiziente Algorithmen zu ermöglichen, wird die DCT auf kleine Blöcke von Pixeln angewendet. Hierzu wird das Bild in Blöcke von jeweils 8×8 Pixeln aufgeteilt. Jeder Block hat also 64 Werte $f_{i,j}$, für $i, j = 0, \ldots, 7$, und charakterisiert einen kleinen Teil des Gesamtbildes. Einen beliebigen solchen Block werden wir im Folgenden diskutieren.

Die Aufteilung der Bildinformation in Blöcke ist auch vorteilhaft, um bei einem Video *Bewegung* zu erkennen. Hierzu werden zwei im Zeitablauf aufeinanderfolgende Einzelbilder und ihre Blöcke verglichen. Die Zahlenwerte der Blöcke lassen

[1] Zur Farb-Kodierung siehe Kap. 15.

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_7

erkennen, ob sich ein Bild-Gegenstand (sein Block) in seiner Position verändert und damit bewegt hat. Die Grundlagen von JPEG spielen auch bei den Video-Formaten M-JPEG (*motion* JPEG) und MPEG-1 eine Rolle.

Diskrete Kosinus-Transformation

Die *Diskrete Kosinus-Transformation* ordnet den $(f_{i,j})$-Werten Koeffizienten $c_{i,j}$ zu. Und mit diesen Koeffizienten kann die Bildinformation als Summe von Kosinus-Termen dargestellt werden, letzteres ist die *inverse* Diskrete Kosinus-Transformation. Die diskrete Kosinus-Transformation spezialisiert sich hier für das 8×8 große Feld zu

$$c_{i,j} := \frac{1}{4}\,\alpha_i\,\alpha_j \sum_{k=0}^{7}\sum_{l=0}^{7} f_{k,l}\,\cos\frac{(2k+1)i\pi}{16}\,\cos\frac{(2l+1)\,j\pi}{16} \tag{7.1}$$

für $i, j = 0, \ldots, 7$, wobei

$$\alpha_i := \begin{cases} \frac{1}{\sqrt{2}} & \text{falls } i = 0, \\ 1 & \text{sonst.} \end{cases} \tag{7.2}$$

Die Formel (7.1) definiert die zweidimensionale DCT[2]. Die Koeffizienten $c_{i,j}$ für $i, j = 0, \ldots, 7$ bilden wie die $f_{i,j}$ eine (8×8)-Matrix. In diesen Koeffizienten steckt die ganze Information, denn mit der inversen diskreten Kosinus-Transformation werden die ursprünglichen $(f_{i,j})$-Werte wieder erhalten, als Summe von Kosinus-Werten,

$$f \;\longrightarrow\; c \;\longrightarrow\; f\,.$$

Die inverse Transformation spezialisiert sich hier im (8×8)-Fall zu

$$f_{i,j} = \frac{1}{4}\sum_{k=0}^{7}\sum_{l=0}^{7} \alpha_k\,\alpha_l\,c_{k,l}\,\cos\frac{(2i+1)k\pi}{16}\,\cos\frac{(2j+1)l\pi}{16} \tag{7.3}$$

mit α_k, α_l wie in (7.2) definiert.

Bildaufbau

Was nützt uns DCT für die Bildkompression? Die Koeffizienten $c_{i,j}$ aus (7.1) haben die schöne Eigenschaft, dass sie in absoluter Größe klein werden für wachsende i, j. Der Wert $c_{0,0}$ dominiert, er kennzeichnet eine mittlere Helligkeit des Blocks. Wie (7.1) für $i = j = 0$ zeigt, ist $c_{0,0}$ die Summe sämtlicher $f_{i,j}$, mit Faktor $\frac{1}{8}$, also ein Mittelwert. Damit ist $c_{0,0}$ die dominierende Zahl für die Bildinformation des Blocks. Wegen der Abnahme der $|c_{i,j}|$ für wachsende i und j werden in der Doppelsumme

[2] Allgemeiner am Ende dieses Kapitels.

(7.3) nicht alle 64 Terme benötigt, um die Bildinformation in ausreichender Qualität wiederzugeben. Statt (7.3) werden für $i, j = 0, \ldots, 7$ die Näherungen

$$\overline{f}_{i,j} := \frac{1}{4} \sum_{k=0}^{M} \sum_{l=0}^{M} \alpha_k \, \alpha_l \, c_{k,l} \, \cos \frac{(2i+1)k\pi}{16} \, \cos \frac{(2j+1)l\pi}{16} \tag{7.4}$$

berechnet für ein $M < 7$. Es werden also statt der ursprünglichen $8^2 = 64$ Werte $f_{i,j}$ nur M^2 Werte $c_{i,j}$ $(i, j = 1, \ldots, M)$ benötigt, um daraus 64 Approximationen $\overline{f}_{i,j}$ für die $f_{i,j}$ zu erhalten, für alle $i, j = 0, \ldots, 7$. Die durch DCT erreichte Speicherersparnis ist also $8^2 - M^2$. Sollte etwa $M = 2$ ausreichen, dann erreicht man durch DCT eine Ersparnis von 60/64, also etwa 93 %, und bei $M = 4$ noch 75 %. Das ist noch nicht das ganze Kompressions-Potential von JPEG, aber ein wesentlicher Teil. Für die Auswertung obiger Formeln gibt es schnelle Algorithmen, verwandt mit der schnellen Fourier-Transformation. Natürlich ist für $M < 7$ die Näherung (7.4) im Vergleich zu (7.3) wegen fehlender Summanden verlustbehaftet.

Die Darstellung (7.4) erlaubt einen allmählichen progressiven Bildaufbau. Man beginnt mit $M = 0$, und verfeinert das Bild mit den zusätzlichen Summanden von $M = 1$, ehe man mit dem für die angeforderte Qualität notwendigen M endet. Die für $M = 0$ oder $M = 1$ erhaltene Näherung erlaubt eine grobe Vorschau auf das Bild.

Beispiel
Das Potential der Kompression soll nun an einem Beispiel illustriert werden. Die Abb. 7.1 deutet die 8×8 Pixel in der Ebene an, wie ein (nicht eingezeichnetes) Schachbrett. In der Mitte jedes der 64 Quadrate in der Ebene wird willkürlich ein Funktionswert f definiert durch

$$f_{i,j} = 0{,}5 + 0{,}1 \cdot (i+1) \cdot (j+1) \text{ für } i, j = 0, \ldots, 7. \tag{7.5}$$

Zu diesem Beispiel ermitteln wir mit (7.1) die (8×8)-Matrix der Koeffizienten $c_{i,j}$. Diese 8×8 Elemente der c-Matrix sind näherungsweise (gerundet)

$$\begin{pmatrix}
20{,}2 & -8{,}2 & \epsilon & -0{,}857 & \epsilon & -0{,}256 & \epsilon & -0{,}0645 \\
-8{,}2 & 4{,}15 & \epsilon & 0{,}434 & \epsilon & 0{,}129 & \epsilon & 0{,}0327 \\
\epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon \\
-0{,}857 & 0{,}434 & \epsilon & 0{,}0454 & \epsilon & 0{,}0135 & \epsilon & 0{,}0034 \\
\epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon \\
-0{,}256 & 0{,}129 & \epsilon & 0{,}0135 & \epsilon & 0{,}00404 & \epsilon & 0{,}001 \\
\epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon \\
-0{,}0645 & 0{,}0327 & \epsilon & 0{,}0034 & \epsilon & 0{,}001 & \epsilon & 0{,}00026
\end{pmatrix}$$

Dabei bedeuten die „ϵ“ symbolhaft Zahlen nahe der Null, hier geschuldet der speziellen Symmetrie des Beispiels (7.5). Die jeweiligen Zahlenwerte der ϵ repräsentieren Rundungsfehler, sie sind hier ohne Interesse. Die Abb. 7.2 zeigt, dass bereits für

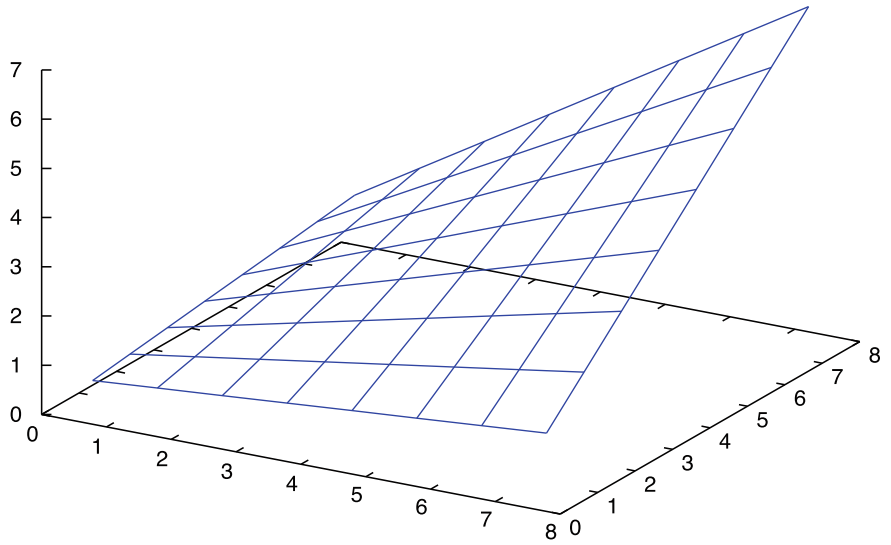

Abb. 7.1 8×8 Werte f von Beispiel (7.5). Die 8×8 Pixel eines Blocks sind durch die waagerechte Ebene repräsentiert. Die f-Werte sind auf der senkrechten Achse aufgetragen. Punkte $(i + \frac{1}{2}, j + \frac{1}{2}, f_{i,j})$ für $i, j = 0, \ldots, 7$ sind durch ein Netz von Geraden verbunden, um den Eindruck einer Fläche zu generieren

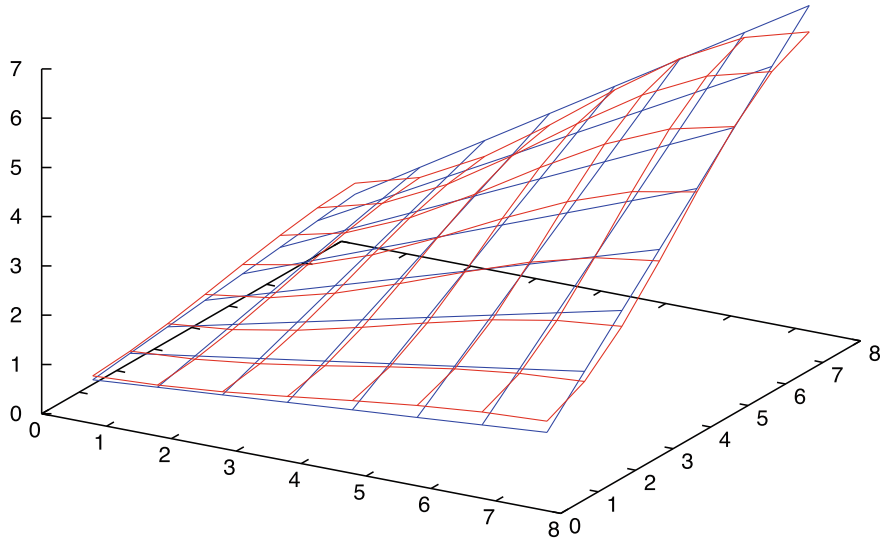

Abb. 7.2 in blau die originalen f-Werte wie Abb. 7.1, in rot die $\overline{f}$-Näherungswerte für $M = 2$

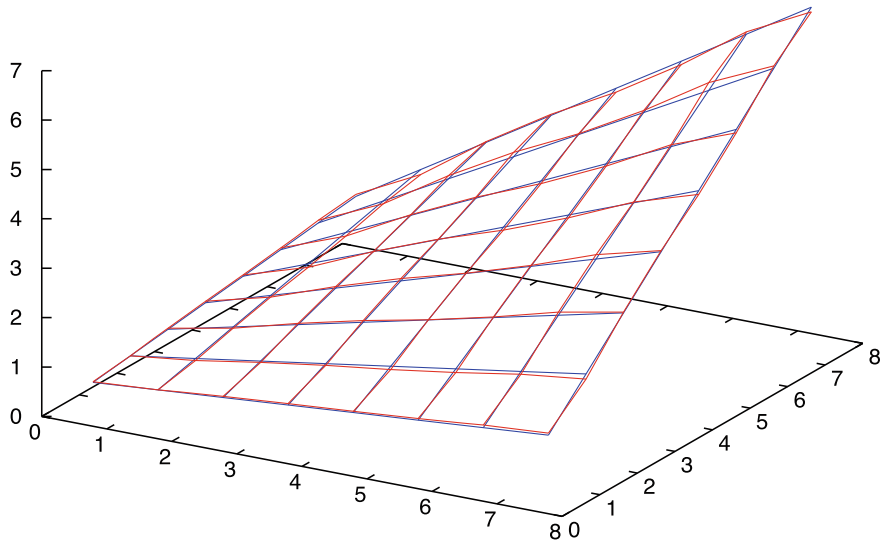

Abb. 7.3 in blau die originalen f-Werte wie Abb. 7.1, in rot die $\overline{f}$-Näherungswerte für $M = 4$

$M = 2$ eine brauchbare Näherung erreicht wird, und in Abb. 7.3 mit $M = 4$ ist praktisch kein Unterschied zum Original mehr erkennbar.

Weitere Kompression

Die oben beschriebene Diskrete Kosinus-Transformation ist der mathematische Kern der JPEG-Kompression. Viele weitere Maßnahmen sorgen für zusätzliche Kompression. Da sie weniger mathematisch sind, seien sie hier nur kurz skizziert.

Die Farbwerte ändern sich weniger von Pixel zu Pixel als die Helligkeitswerte. Deswegen brauchen die Komponenten der Farbwerte in jeder Richtung nur für jedes zweite Pixel gespeichert zu werden. Hierzu werden Makro-Blöcke gebildet aus 16×16 Pixeln. Die Helligkeit dieser 16^2 Pixel wird in vier (8×8)-Blöcken abgelegt, und jede der beiden Farbinformationen in je einem (8×8)-Block. Damit ist jeder Makro-Block auch in seiner Farbinformation mit sechs skalaren (8×8)-Blöcken vollständig ausreichend definiert. Die Farbinformation hat so einen Faktor 4 eingespart.[3]

Die in (7.1) definierten c-Werte werden auf ganze Zahlen quantisiert, dabei werden die Koeffizienten „unten rechts“ ($i + j$ „groß“) in der c-Matrix als weniger relevant angesehen und geringer gewichtet. Nach Abrundung auf die nächste ganze Zahl sind viele der quantisierten Binärzahlen $\in \{0, \ldots, 255\}$ Null, und somit ist die resultierende Matrix dünn besiedelt. So lässt sich die c-Information welche aktiv ist, mit wenigen Bits speichern. Entsprechend der Größen-Verteilung der quantisierten Zah-

[3] Auch der Standard MPEG-1 arbeitet mit diesen Makro-Blöcken.

len in der Matrix erfolgt die Speicherung in einem Zick-Zack, beginnend „oben links“ mit der Position $(0, 0)$, und dann weiter $(0, 1), (1, 0), (2, 0), (1, 1), (0, 2), (0, 3), (1, 2), (2, 1), (3, 0)$, und so fort, in Diagonalen (i, j) mit $i + j = 1, 2, \ldots$. Die Quantisierung ist der wesentliche Grund für Verluste an Qualität, mehr als die Reduzierung in (7.4).

Eine weitere Ersparnis ergibt sich, wenn nicht nicht jeder der (8×8)-Pixel-Blöcke unabhängig vom Nachbarblock DCT-transformiert wird. Die meist geringen Veränderungen von Block zu Block legen es nahe, nur die Differenzen zu speichern.

Allgemeine Diskrete Kosinus-Transformation
Als Ergänzung zu der für JPEG wichtigen Transformation (7.1) hier noch die allgemeine Transformation: Die zweidimensionale Diskrete Kosinus-Transformation für eine Matrix

$$F := \begin{pmatrix} f_{0,0} & \cdots & f_{0,N-1} \\ \vdots & & \vdots \\ f_{N-1,0} & \cdots & f_{N-1,N-1} \end{pmatrix}$$

mit N^2 Werten $f_{i,j}$ $(i, j = 0, \ldots, N-1)$ ist

$$c_{i,j} := \frac{1}{\sqrt{2N}} \alpha_i \alpha_j \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} f_{k,l} \cos \frac{(2k+1)i\pi}{2N} \cos \frac{(2l+1)j\pi}{2N}$$

mit α wie in (7.2) definiert, und Rücktransformation analog wie oben in (7.3).

Da diese DCT für Bildverarbeitung eine so große Rolle spielt, lohnt ein kleiner Einblick in die Struktur, hier als Aufgabe formuliert. Für die folgende Aufgabe definieren wir die Vektoren

$$v(i) := \begin{pmatrix} \cos \frac{i\pi}{2N} \\ \cos \frac{3i\pi}{2N} \\ \vdots \\ \cos \frac{(2N-1)i\pi}{2N} \end{pmatrix}$$

für $i = 0, \ldots, N-1$. Der transponierte(Zeilen-)Vektor ist $v(i)^{tr}$.

Aufgabe:
a) Für $N = 2$ zeige

$$c_{i,j} = \frac{1}{2} \alpha_i \alpha_j \; v(i)^{tr} F v(j)$$

b) Für $N = 3$ zeige, dass die Vektoren $v(0), v(1), v(2)$ zueinander orthogonal sind.

Wir rechnen für $N = 2$ ganz elementar Zeile×Matrix×Spalte,

$$\begin{aligned} v(i)^{tr} F v(j) &= \begin{pmatrix} \cos\frac{i\pi}{4} \\ \cos\frac{3i\pi}{4} \end{pmatrix}^{tr} \begin{pmatrix} f_{0,0} & f_{0,1} \\ f_{1,0} & f_{1,1} \end{pmatrix} \begin{pmatrix} \cos\frac{j\pi}{4} \\ \cos\frac{3j\pi}{4} \end{pmatrix} \\ &= \begin{pmatrix} \cos\frac{i\pi}{4} \\ \cos\frac{3i\pi}{4} \end{pmatrix}^{tr} \begin{pmatrix} f_{0,0}\cos\frac{j\pi}{4} + f_{0,1}\cos\frac{3j\pi}{4} \\ f_{1,0}\cos\frac{j\pi}{4} + f_{1,1}\cos\frac{3j\pi}{4} \end{pmatrix} \\ &= f_{0,0}\cos\frac{i\pi}{4}\cos\frac{j\pi}{4} + f_{0,1}\cos\frac{i\pi}{4}\cos\frac{3j\pi}{4} \\ &\quad + f_{1,0}\cos\frac{3i\pi}{4}\cos\frac{j\pi}{4} + f_{1,1}\cos\frac{3i\pi}{4}\cos\frac{3j\pi}{4}, \end{aligned}$$

womit die Behauptung a) gezeigt ist.
Für b) und $N = 3$ stellen wir die drei Vektoren auf:

$$v(0) = \begin{pmatrix} \cos 0 \\ \cos 0 \\ \cos 0 \end{pmatrix}, \quad v(1) = \begin{pmatrix} \cos\frac{\pi}{6} \\ \cos\frac{3\pi}{6} \\ \cos\frac{5\pi}{6} \end{pmatrix}, \quad v(2) = \begin{pmatrix} \cos\frac{\pi}{3} \\ \cos\pi \\ \cos\frac{10\pi}{6} \end{pmatrix}.$$

Die konkreten Zahlenwerte sind

$$v(0) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad v(1) = \begin{pmatrix} \frac{1}{2}\sqrt{3} \\ 0 \\ -\frac{1}{2}\sqrt{3} \end{pmatrix}, \quad v(2) = \begin{pmatrix} \frac{1}{2} \\ -1 \\ \frac{1}{2} \end{pmatrix}.$$

Die drei Skalarprodukte

$$v(0)^{tr} v(1) = 0, \quad v(0)^{tr} v(2) = 0, \quad v(1)^{tr} v(2) = 0$$

dieser drei Vektoren miteinander ergeben jeweils 0, also sind die drei Vektoren $v(0), v(1), v(2)$ orthogonal.

Analoge Resultate gelten für allgemeine N.

Literatur

Ahmed, N., Natarajan, T., Rao, K.R.: Discrete Cosine Transform. IEEE Transactions on Computers, Januar (1974)
Strang, G.: The Discrete Cosine Transform. SIAM Review **41**, 135–147 (1999)

Navigation mit Filtern

8

Positionsbestimmungen finden überall statt, zum Beispiel im engeren Sinne im Verkehr. Ein einfaches Szenario soll als Illustration dienen: Ein Schiffsweg[1] an einem geradlinigen Ufer entlang. Auf dem Ufer gibt es drei Beobachtungsstationen, die gleichzeitig den Winkel zu einem zu beobachtenden Objekt messen, illustriert in Abb. 8.1. Drei gemessene Richtungswinkel bedeuten drei Geraden, die sich im Allgemeinen paarweise schneiden. Das vorläufige Messresultat besteht also aus drei Punkten, oder dem zugehörigen Dreieck (vergrößert in Abb. 8.3). Wo in dem Dreieck, oder in der Nähe des Dreiecks, befindet sich nun das Objekt, für dessen Position wir uns interessieren?[2]

Die Größe eines solchen Dreiecks hängt von vielen Umständen ab, wie unklare Sicht, ausgedehntes Objekt, Fehler im Messgerät, schlechte Synchronisation der Beobachtungen, oder Ablesefehler. Die genaue Größe dieser Unsicherheiten und Ungenauigkeiten ist unbekannt. Man muss sie als Zufallsvariable ansehen und Wahrscheinlichkeitsverteilungen annnehmen. Die Frage ist, wie kann man eine wahrscheinliche Position des Objektes aus der Wolke der Ungenauigkeiten heraus*filter*n? Diese Problemstellung ist nicht nur für Häfen, Flughäfen oder Satellitenbahnen wichtig, sondern auch bei Beobachtungen von Zuständen allgemeiner Systeme.

Wir nehmen drei Beobachter B_j $(j = 1, 2, 3)$ an, ihre gemessenen Winkel seien φ_j, gemessen gegen die x_1-Achse, auf der die Beobachter sitzen, mit den Abständen l_j zum Nullpunkt. Die Variablen unseres Beispiel-Szenarios sind in Abb. 8.2 illustriert. Die Position des Gegenstandes G sei (x_1, x_2). Nach elementarer Geometrie

[1] oder Straße mit zu beobachtendem Lastwagen, oder ...

[2] Wenn es mehr als drei Messungen gibt, resultieren auch mehr Schnittpunkte, und nicht nur ein Dreieck. Um das Wesentliche herauszufiltern, bleiben wir hier bei drei Messungen.

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_8

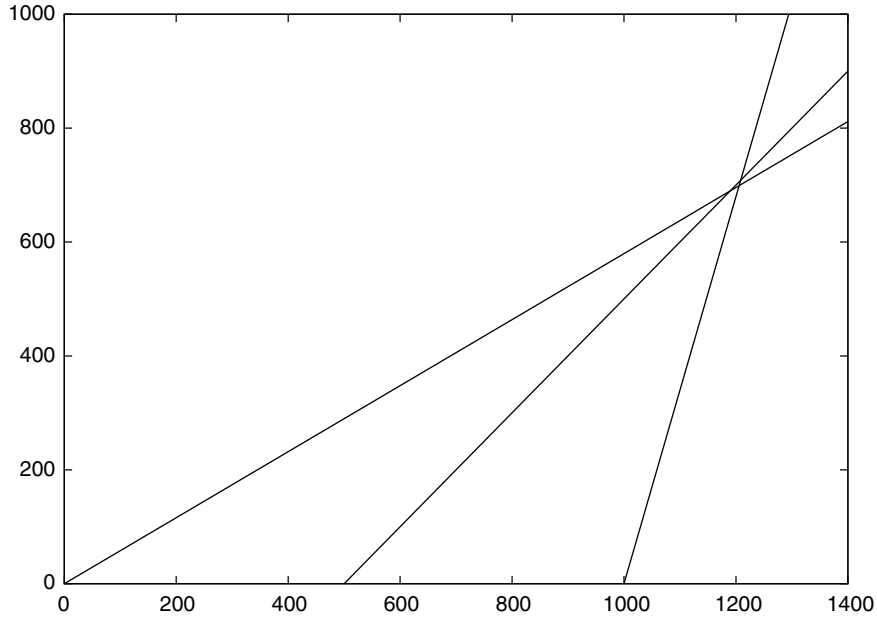

Abb. 8.1 In einer (x_1, x_2)-Ebene sitzen drei Beobachter auf der x_1-Achse (die waagerechte Achse) an den Positionen 0, 500 und 1000, mit jeweils einem Messwinkel in Richtung auf ein Objekt, dessen Lage in der Ebene zu bestimmen ist

gilt

$$\varphi_j = \arctan \frac{x_2}{x_1 - l_j} + u_j.$$

Dabei steht u_j für die Ungenauigkeiten. Mit der Funktion

$$f_j(x) := \arctan \frac{x_2}{x_1 - l_j}, \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{8.1}$$

ist das Gleichungssystem

$$\begin{aligned} \varphi_1 &= f_1(x) + u_1 \\ &\vdots \\ \varphi_m &= f_m(x) + u_m \end{aligned} \tag{8.2}$$

zu lösen, bei unserem illustrierenden Beispiel mit $m = 3$. Damit steuern wir bereits auf eine allgemeine Systematik zu, mit dem Vektor $x = (x_1, \ldots, x_n)^{tr}$, im Beispiel mit $n = 2$. Wenn wir die Messwerte (hier die φ_j) allgemein auch mit y_j bezeichnen, und für $j = 1, \ldots, m$ zu einem m-Vektor y montieren, ebenso wie die f_j und die u_j, dann lautet (8.2) in Vektorschreibweise kompakt

$$y = f(x) + u. \tag{8.3}$$

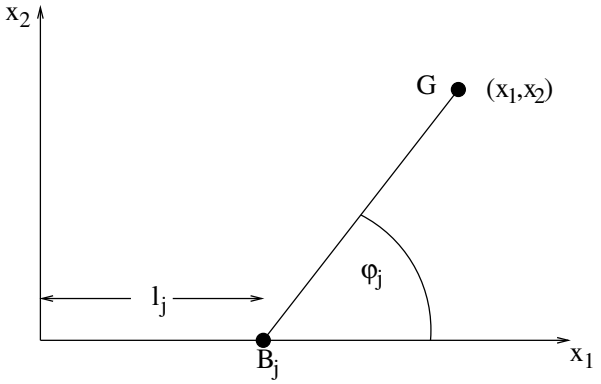

Abb. 8.2 In einer (x_1, x_2)-Ebene beobachtet ein Beobachter B_j an Position $(l_j, 0)$ einen Gegenstand G (hier punktförmig) unter dem Winkel φ_j

Das System

Die unbekannten u_j lassen wir erstmal so stehen, und konzentrieren uns auf die Lösung von (8.2)/(8.3). Die Funktion f ist im Allgemeinen nichtlinear, so auch die Funktion (8.1). Deswegen wird grundsätzlich iterativ vorgegangen. Wegen der Ungenauigkeiten u_j, auch wegen einer möglichen Bewegung des Objektes G, wird die Iteration zunächst nur aus einem Schritt bestehen: Eine Näherung oder Schätzung der Position von G, eventuell grafisch ermittelt (Abb. 8.1), sei mit dem Vektor $\bar{x}$ bezeichnet. Das Ziel ist es, einen verbesserten Wert x für die Position zu erhalten. Der Schritt von der Start-Näherung $\bar{x}$ zu einer besseren Näherung x ist diese eine Iteration (die natürlich wiederholt werden kann).

Für den Verbesserungsschritt wird f in (8.2) um $\bar{x}$ linearisiert. Bis auf Terme höherer Ordnung gilt

$$f_j(x) \approx f_j(\bar{x}) + (\text{grad}\, f_j(\bar{x}))^{tr}(x - \bar{x})$$

für $j = 1, \ldots, m$, umso genauer, je näher $\bar{x}$ an x liegt. Im Spezialfall $n = 2$ heißt das

$$f_j(x) \approx f_j(\bar{x}) + \left(\frac{\partial f_j(\bar{x})}{\partial x_1}, \frac{\partial f_j(\bar{x})}{\partial x_2} \right) \begin{pmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \end{pmatrix}.$$

Damit sind wir schon bei einer Matrix-Schreibweise. Mit den Bezeichnungen

$$A := \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}, \quad x := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \bar{x} := \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_n \end{pmatrix},$$

$$u := \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}, \quad y := \begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}, \quad \bar{y} := \begin{pmatrix} f_1(\bar{x}) \\ \vdots \\ f_m(\bar{x}) \end{pmatrix},$$

lässt sich das linearisierte System kompakt als Vektorgleichung schreiben:

$$y - \bar{y} = A(x - \bar{x}) + u. \tag{8.4}$$

Oder, mit den Differenzen $\Delta y := y - \bar{y}$ und $\Delta x := x - \bar{x}$,

$$\Delta y = A\,\Delta x + u.$$

Da die Messung in y steckt, und $\bar{y}$ leicht auszurechnen ist, kann Δy als bekannt angesehen werden. Der Korrekturvektor Δx ist unbekannt. Die Matrix A ist eine Rechteck-Matrix, mit m Zeilen und n Spalten.

Das Vorgehen
Die Gl. (8.4) ist kein Standard-Gleichungssystem für (zum Beispiel) den Gaußschen Algorithmus: Die Matrix A ist nicht quadratisch, und u ist unbekannt. Was nun? Lösen wir (8.4) nach u auf, dann motiviert der Wunsch nach nur geringen Fehlern den Gedanken, u „klein“ zu haben, also den Vektor $\Delta y - A\,\Delta x$ nahe dem Nullvektor. Ein denkbares Ziel wäre die Minimierung des größten Absolutbetrages der Komponenten, das erfordert allerdings aufwendige Methoden. Der Rechenaufwand ist geringer, wenn man die Summe der Quadrate minimiert. In Norm-Schreibweise kann dieses Ziel so formuliert werden:

$$\text{Minimiere } \| \Delta y - A\,\Delta x \|_2^2,$$

über alle x, wobei $\| \ \|_2$ die euklidische Norm ist[3]. Als Skalarprodukt geschrieben, bedeutet das

$$\text{Minimiere } (\Delta y - A\,\Delta x)^{tr}(\Delta y - A\,\Delta x), \tag{8.5}$$

über alle x. Das ist die Methode der kleinsten Quadrate, es wird sozusagen das Quadrat des Fehlers minimiert.

In der Version (8.5) findet keine Gewichtung statt zwischen den Komponenten u_j, welche die Fehler der Messung y_j zusammenfassen. Eine Gewichtung kann in (8.5) mit einer Diagonalmatrix D eingeführt werden: Zu minimieren wäre dann $u^{tr}Du$ statt $u^{tr}u$, also

$$\text{Minimiere } (\Delta y - A\,\Delta x)^{tr} D(\Delta y - A\,\Delta x)\ !$$

Zur Wahl der gewichtenden Diagonalmatrix D bedient man sich der Wahrscheinlichkeiten. Erinnert sei an die Kovarianzmatrix Σ, hier zum Vektor u.[4] Die Matrix Σ enthält Informationen, wie groß die Varianzen von u_j sind, und wie die verschiedenen

[3] $\|u\|_2^2 := \sum_{j=1}^{m} u_j^2 = u^{tr}u$ wobei ^{tr} die Transposition bedeutet: u^{tr} ist Zeilenvektor; siehe auch Kap. 6.

[4] $\Sigma := \mathsf{E}(uu^{tr}) - \mathsf{E}(u)(\mathsf{E}(u))^{tr}$, wobei $\mathsf{E}(u)$ den Erwartungswert der Zufallsvariablen u bezeichnet. Typischerweise wird Normalverteilung angenommen, und $\mathsf{E}(u) = 0$.

Tab. 8.1 Messungen von Abb. 8.1: Positionen l in [m] und Richtungswinkel φ und y; Zahlenbeispiel von [Bryson & Ho (1969)]

l	φ	y [in Bogenmaß]	σ^2
0	30,1°	0,525344	0,01
500	45,0°	0,785398	0,01
1000	73,6°	1,28456	0,04

Ungenauigkeiten u_j korreliert sind. In der Praxis weiß man oft wenig über die Werte in der Kovarianzmatrix, deswegen ist eine häufige und vereinfachende Annahme, dass die Messungen unkorreliert sind. In dem Fall ist Σ eine Diagonalmatrix,

$$\Sigma = \text{diag}\,(\sigma_1^2, \ldots, \sigma_m^2),$$

σ_j^2 ist die Varianz von u_j, für $j = 1, \ldots, m$. Je kleiner σ^2, umso genauer die Messung! Und ein großer Wert von σ^2 bedeutet, dass die Messung stark verrauscht ist. Das legt nahe, die Messung y_j mit $\frac{1}{\sigma_j^2}$ zu gewichten,[5] oder $D = \Sigma^{-1}$. Also lautet die Aufgabe:

$$\text{Minimiere } J(x) := (\Delta y - A\,\Delta x)^{tr}\,\Sigma^{-1}(\Delta y - A\,\Delta x) \text{ über alle } x. \tag{8.6}$$

Standard-Analysis weist den weiteren Weg: Der Gradient von J muss bei einem Minimum verschwinden, mit

$$\text{grad}\, J(x) = -2A^{tr}\,\Sigma^{-1}(\Delta y - A\Delta x).$$

Das heißt

$$A^{tr}\,\Sigma^{-1}(\Delta y - A\Delta x) = 0,$$

oder

$$A^{tr}\,\Sigma^{-1}A\,\Delta x = A^{tr}\,\Sigma^{-1}\Delta y. \tag{8.7}$$

Nun sind die oben angesprochenen Probleme gelöst: Das Gleichungssystem (8.7) ist von Standard-Form, mit quadratischer $n \times n$ Matrix $A^{tr}\,\Sigma^{-1}A$ und dem n-Vektor $A^{tr}\,\Sigma^{-1}\Delta y$ der rechten Seite.[6]

Das Beispiel

Zur Motivation haben wir oben bereits ein richtungweisendes Beispiel eingeführt mit drei Beobachtungsstellen und zugehörigen drei Messwinkeln (Abb. 8.1, 8.2). Dieses Beispiel wird nun konkretisiert, mit Zahlen unterlegt, und numerisch mit Hilfe des eben beschriebenen Vorgehens gelöst. Die Zahlenwerte der Positionen l der Messpunkte und der Messwinkel φ sind in der Tab. 8.1 aufgeführt. Für die Rechnung benötigen wir die Winkel im Bogenmaß, sowie angenommene Werte der Varianz σ^2. Auch diese Angaben stehen in der Tab. 8.1.

[5] Ein $\sigma_j = 0$ würde bedeuten, dass diese Gleichung fehlerfrei ist, und abgespalten werden kann.
[6] Das Gleichungssystem kann schlecht konditioniert sein, deswegen Lösung zum Beispiel mit orthogonalen Transformationen.

Aufgabe 1 *Für die Zahlen nach Tab. 8.1 und* $\bar{x} = \begin{pmatrix} 1210 \\ 700 \end{pmatrix}$ *berechne man die Matrizen* A *und* $A^{tr}\,\Sigma^{-1}A$, *sowie die Vektoren* $\bar{y}$ *und* $A^{tr}\,\Sigma^{-1}\Delta y$.

Der Vektor $\bar{x} = \begin{pmatrix} 1210 \\ 700 \end{pmatrix}$ kann zum Beispiel das Resultat einer vorhergehenden Messung sein. Für $\bar{y}$ resultiert

$$\bar{y} = f(\bar{x}) = \begin{pmatrix} \arctan \frac{\bar{x}_2}{\bar{x}_1 - 0} \\ \arctan \frac{\bar{x}_2}{\bar{x}_1 - 500} \\ \arctan \frac{\bar{x}_2}{\bar{x}_1 - 1000} \end{pmatrix} = \begin{pmatrix} \arctan \frac{700}{1210} \\ \arctan \frac{700}{710} \\ \arctan \frac{700}{210} \end{pmatrix} = \begin{pmatrix} 0{,}52447 \\ 0{,}778306 \\ 1{,}27934 \end{pmatrix}.$$

Der Vektor y beinhaltet die Messung, hier die Winkel φ (Tab. 8.1). Im Bogenmaß ist das

$$\Delta y = y - \bar{y} = \begin{pmatrix} 0{,}000874 \\ 0{,}007092 \\ 0{,}00522 \end{pmatrix}.$$

Die Matrix $A(x)$ der partiellen Ableitungen ist

$$A(x) = \begin{pmatrix} -\frac{x_2}{x_1^2 + x_2^2} & \frac{x_1 - 0}{x_1^2 + x_2^2} \\ -\frac{x_2}{(x_1 - 500)^2 + x_2^2} & \frac{x_1 - 500}{(x_1 - 500)^2 + x_2^2} \\ -\frac{x_2}{(x_1 - 1000)^2 + x_2^2} & \frac{x_1 - 1000}{(x_1 - 1000)^2 + x_2^2} \end{pmatrix},$$

und $\bar{x}$ eingesetzt gilt

$$A(\bar{x}) = \begin{pmatrix} -3{,}582212 \cdot 10^{-4} & 6{,}19211 \cdot 10^{-4} \\ -7{,}041545 \cdot 10^{-4} & 7{,}14214 \cdot 10^{-4} \\ -1{,}310616 \cdot 10^{-3} & 3{,}93185 \cdot 10^{-4} \end{pmatrix}.$$

Aus den Daten σ der Tabelle ergibt sich die gewichtende Matrix Σ^{-1},

$$\Sigma^{-1} = \begin{pmatrix} 100 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 25 \end{pmatrix},$$

und damit[7] die Matrix

$$A^{tr}\,\Sigma^{-1}A = \begin{pmatrix} 1{,}05358 \cdot 10^{-4} & -0{,}85356 \cdot 10^{-4} \\ -0{,}85356 \cdot 10^{-4} & 0{,}932172 \cdot 10^{-4} \end{pmatrix}$$

[7] Nun ja, mit Handrechnung bzw. Taschenrechner mühsam, aber eine gute Übung. Soviel Stellen wie hier angegeben, braucht man nicht, dies soll nur einen besseren Vergleich ermöglichen für die Programmierer. (Ergebnisse gerundet)

des Gleichungssystems, sowie die rechte Seite

$$A^{tr}\,\Sigma^{-1}\,\Delta y = \begin{pmatrix} -7{,}01834 \cdot 10^{-4} \\ 6{,}11994 \cdot 10^{-4} \end{pmatrix}.$$

Die Lösung des Gleichungssystems (8.7) ist schließlich

$$\Delta x = \begin{pmatrix} -5{,}20025 \\ 1{,}80354 \end{pmatrix}.$$

Damit lautet der korrigierte Vektor $x = \bar{x} + \Delta x$

$$x = \begin{pmatrix} 1205 \\ 702 \end{pmatrix}.$$

Die Abb. 8.3 zeigt den berechneten Punkt x in dem Dreieck, das durch die Beobachtungsgeraden gebildet ist.

Ausblick
Oben haben wir einen einfachen Fall mit drei Messungen betrachtet. In vielen Anwendungen, vor allem bei sich bewegenden Objekten, werden viele Messungen erzeugt. Dann wäre es unwirtschaftlich, die gesamte Rechnung von Beginn für

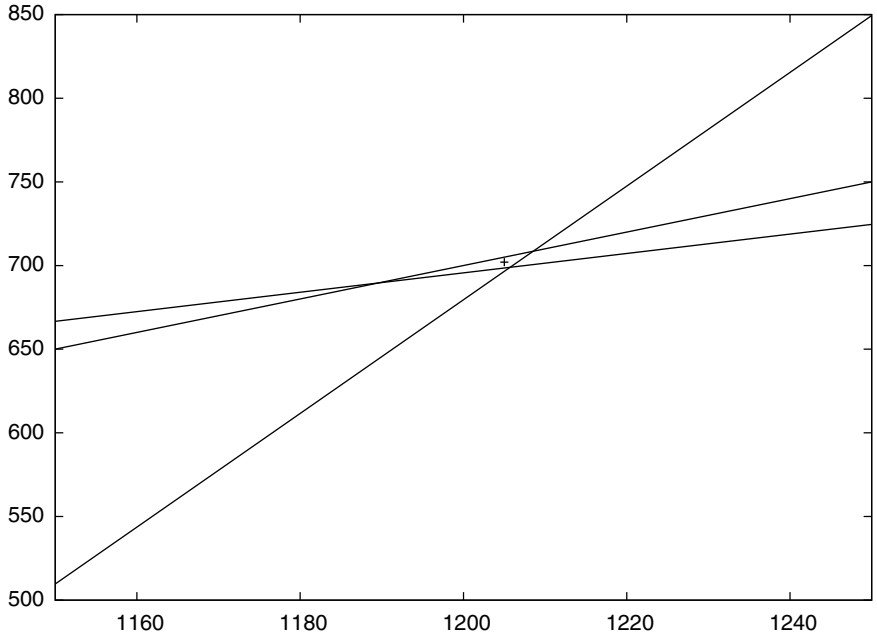

Abb. 8.3 (x_1, x_2)-Ebene, Detail von Abb. 8.1. Der berechnete Vektor x ist durch ein Kreuz markiert

entsprechend großes m erneut durchzuführen. Stattdessen gibt es Rekursionen, die einen großen Teil des bereits Berechneten wiederverwenden.

Interessant ist das vor allem, wenn ein sich bewegendes Objekt (zum Beispiel ein Flugzeug) einem dynamischen Gesetz gehorcht, etwa physikalischen Bewegungsgleichungen. Dann ist das x nicht nur Resultat einer Minimierung, sondern vorab muss bei jeder Messung die Bewegungsgleichung berücksichtigt werden. Vektoren $x^{(k)}$ sind dann zu verstehen als Position zum Zeitpunkt t_k, für $k = 0, 1, 2, \ldots$ Ein solches Modell kann im einfachsten Fall durch ein lineares Gesetz beschrieben werden wie in der folgenden Aufgabe:

Aufgabe 2 *Ein Fahrzeug bewegt sich auf einem geradlinigen Weg, seine Position zum Zeitpunkt t ist $s(t)$. Für die Vektorfunktion*

$$x(t) := \begin{pmatrix} s(t) \\ \dot{s}(t) \end{pmatrix}$$

stelle man die Bewegungsgleichung in der Form

$$x(t + \Delta t) = F \cdot x(t) + g(t)$$

dar, für kleines Δt. Dabei bezeichnet $\dot{s}$ die Ableitung nach t, F ist eine (2×2)-Matrix. Man bestimme die Matrix F sowie die Funktion $g(t)$.

In dieser Aufgabe haben wir ein System mit einem Freiheitsgrad vor uns, repräsentiert durch s. Taylor-Entwicklung für genügend glattes s besagt

$$\begin{aligned} s(t + \Delta t) &= s(t) + \Delta t\, \dot{s}(t) + \tfrac{1}{2}\Delta t^2\, \ddot{s}(t) + T.h.O. \\ &= (1,\ \Delta t) \begin{pmatrix} s(t) \\ \dot{s}(t) \end{pmatrix} + \tfrac{1}{2}\Delta t^2\, \ddot{s}(t) + T.h.O., \end{aligned} \tag{8.8}$$

die Terme höherer Ordnung in Δt sind zu „T.h.O." zusammengefasst. In (8.8) ist

$$(1, \Delta t) \begin{pmatrix} s \\ \dot{s} \end{pmatrix}$$

„Zeile mal Spalte" das Skalarprodukt. Analog wie (8.8) gilt für $\dot{s}$

$$\begin{aligned} \dot{s}(t + \Delta t) &= \dot{s}(t) + \Delta t\, \ddot{s}(t) + \tfrac{1}{2}\Delta t^2\, \dddot{s}(t) + T.h.O. \\ &= (0,\ 1) \begin{pmatrix} s(t) \\ \dot{s}(t) \end{pmatrix} + \Delta t\, \ddot{s}(t) + \tfrac{1}{2}\Delta t^2\, \dddot{s}(t) + T.h.O. \end{aligned} \tag{8.9}$$

Die Gl. (8.8) und (8.9) tragen in den rechten Seiten bereits den Vektor x in sich. Damit lassen sich beide Gleichungen zu einer Vektor-Gleichung zusammenfassen:

$$\begin{pmatrix} s(t + \Delta t) \\ \dot{s}(t + \Delta t) \end{pmatrix} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} s(t) \\ \dot{s}(t) \end{pmatrix} + \ddot{s}(t) \begin{pmatrix} \frac{1}{2}\Delta t^2 \\ \Delta t \end{pmatrix} + T.h.O.$$

Bezeichnet man die Beschleunigung mit $a := \ddot{s}$, und die Matrix F mit

$$F := \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix},$$

dann lautet das System

$$x(t + \Delta t) = F\, x(t) + a(t) \begin{pmatrix} \frac{1}{2}\Delta t^2 \\ \Delta t \end{pmatrix} + T.h.O.,$$

und die Funktion g ist von der Form

$$g(t) = a(t) \begin{pmatrix} \frac{1}{2}\Delta t^2 \\ \Delta t \end{pmatrix} + T.h.O.$$

Wird das Fahrzeug aus der Ferne beobachtet, und sind die Beobachtungen unsicher, dann kann g, also der Beschleunigungsterm mit den Termen höherer Ordnung, als eine Zufallsvariable v interpretiert werden.

Mit $x^{(k)} := x(t_k)$ kann das System in der Form

$$x^{(k+1)} = F_k\, x^{(k)} + v \tag{8.10}$$

geschrieben werden. Die Gl. (8.10) ist linear und eine einfache Form einer Bewegungsgleichung. Nun kommt die Beobachtung zum Zeitpunkt t_k hinzu, mit einer Mess-Gleichung analog (8.2)/(8.3). Damit lautet das System

$$\begin{aligned} x^{(k+1)} &= F_k\, x^{(k)} + v, \\ y^{(k)} &= f(x^{(k)}) + u. \end{aligned}$$

Dabei ist F_k eine Matrix, und v der zufällige Fehlervektor, als unabhängig von u angenommen. Eine anspruchsvolle Aufgabe, von höchster Wichtigkeit in den Anwendungen, ist die Schätzung der Positionen $x^{(k)}$ mit Hilfe der Messungen $y^{(k)}$, wiederum rekursiv. Eine häufig verwendete Lösungsmethode geht auf Kálmán zurück, deswegen wird hier auch vom Kalman-Filter gesprochen.

Literatur

für Beispiele und eine Erklärung des Kalman-Filters siehe

Bryson, A.E., Ho, Y.-C.: Applied Optimal Control. Optimization, Estimation, and Control. Ginn and Company, Waltham (1969)

Crassidis, J.L., Junkins, J.L.: Optimal Estimation of Dynamical Systems. Chapman & Hall, Boca Raton (2011)

Kalman, R.E.: A new approach to linear filtering and prediction problems. J. Basic Eng. **82**(1), 35–45 (1960)

Eigenschaften und Rekursion z.B. in §2.5 von

Strang, G.: Introduction to Applied Mathematics. Wellesley-Cambridge Press, Wellesley (1986)

Berechnung des Sinus

9

Computer sind aus dem täglichen Leben nicht mehr wegzudenken. Wie selbstverständlich wird auf einem Taschenrechner oder in einem Rechenprogramm (Algorithmus) etwa der SINUS aufgerufen, ohne dass man sich überlegt, wie eine solche Funktion berechnet werden kann. Einerseits muss das Ergebnis bei einem Rechner mit zum Beispiel 7 Dezimalstellen auch auf 7 dezimale Stellen richtig sein, andererseits soll der Rechenaufwand so gering sein wie möglich.[1] Beide Forderungen müssen für einen weiten Definitionsbereich erfüllt sein, man benötigt $\sin(600)$ ebenso genau und schnell wie $\sin(-0{,}04)$.

Der Sinus ist eine 2π-periodische Funktion. Im Prinzip genügt es also, diese Funktion in einem Intervall $a \le y < a + 2\pi$ (a beliebig) berechnen zu können, wenn man nur jedem x außerhalb dieses Intervalls das „richtige" y mit gleichem sin-Wert zuordnen kann. Eine derartige Reduktion des Definitionsbereiches wird im Folgenden für $0 \le y < 2\pi$ durchgeführt.

Aufgabe 1 *Es sei x gegeben im Definitionsbereich $-\infty < x < \infty$. Man konstruiere eine Funktion $y = f(x)$ mit Wertebereich $0 \le y < 2\pi$ derart, dass gilt:*

$$\sin x = \sin y.$$

Hinweis: Man verwende die Funktion $v = \text{floor } u$.

Die floor-Funktion ordnet jeder Zahl u die größte ganze Zahl v zu, für die $v \le u$ gilt. Der Graph dieser Funktion ist der einer Treppenfunktion, wir haben sie bereits

[1] Heutige Rechner und ihre Algorithmen berechnen typischerweise 16 oder mehr Dezimalstellen. Um Methoden kompakter darzustellen, beschränken wir uns hier auf 7 Stellen.

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_9

in Kap. 5 verwendet. Zur Konstruktion der Funktion $y = f(x)$ überlegt man sich, dass es für jedes x eine ganze Zahl n gibt mit

$$2\,\pi\,n \le x < 2\,\pi\,(n+1).$$

Hierzu äquivalent ist

$$n \le \frac{x}{2\pi} < n+1.$$

Das ist gerade die Definition der floor-Funktion, demnach ist n bestimmt durch

$$n = \text{floor}\left(\frac{x}{2\pi}\right).$$

Weitere äquivalente Umformungen obiger Ungleichung sind

$$\begin{aligned} 0 &\le \frac{x}{2\pi} - n < 1 \\ 0 &\le x - 2\pi n < 2\pi. \end{aligned}$$

Vergleicht man diese Beziehung mit dem Ziel, so sieht man die Lösung:

$$y = f(x) := x - 2\pi n = x - 2\pi \cdot \text{floor}\left(\frac{x}{2\pi}\right).$$

Für diesen Wert y bestätigen die Additionstheoreme des Sinus die verlangte Beziehung

$$\sin y = \sin\left(x - 2\pi \cdot \text{floor}\,\frac{x}{2\pi}\right) = \sin x.$$

Die Funktion $f(x)$ ermöglicht bei beliebigem x die Reduktion der Berechnung des Sinus auf das Intervall $0 \le y < 2\pi$.

Wie wir sehen werden, ist zur Konstruktion eines effektiven Algorithmus die Reduktion auf das Intervall

$$-\frac{\pi}{2} \le y < 3\frac{\pi}{2}$$

günstiger. Durch anschließende weitere Reduktionen des Definitionsbereiches wird es möglich, sich bei der Berechnung des Sinus mit dem „kleinen“ Intervall

$$-\frac{\pi}{6} \le y \le \frac{\pi}{6}$$

zu begnügen (Abb. 9.1) .

Aufgabe 2 *Für einen Rechner entwerfe man einen Algorithmus zur Berechnung von* $\sin x$, x *aus dem Intervall* $-\infty < x < \infty$.

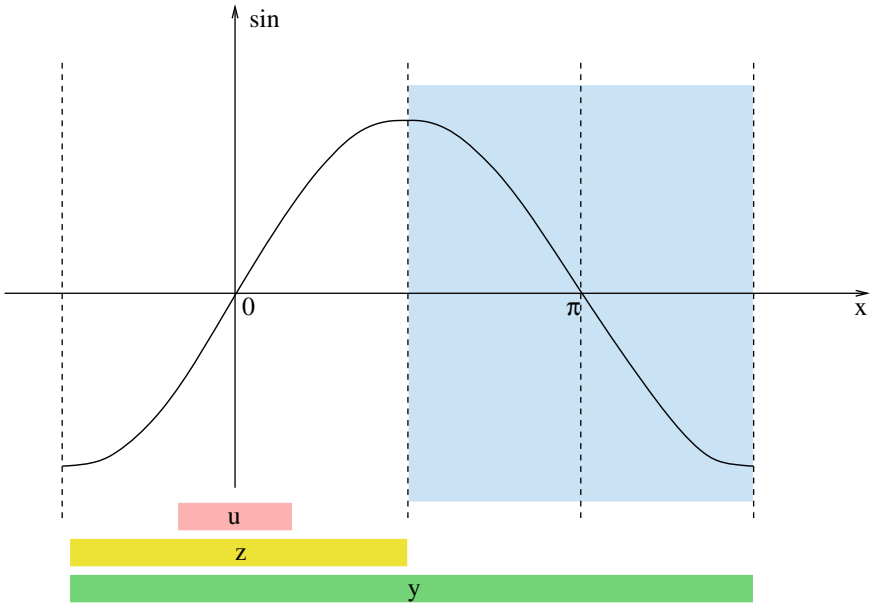

Abb. 9.1 schematisch: zur Berechnung des Sinus, verschiedene Definitionsbereiche, $-\frac{\pi}{2} \le y \le \frac{3\pi}{2}$, $-\frac{\pi}{2} \le z \le \frac{\pi}{2}$, $-\frac{\pi}{6} \le u \le \frac{\pi}{6}$

Anleitung: Für gegebenes x führe man zunächst

1.) mit Hilfe der Funktion floor (x) *eine Intervallreduktion auf das Intervall* $-\frac{\pi}{2} \le y < \frac{3\pi}{2}$ *durch, derart, dass* $\sin x = \sin y$.

2.) Man bestimme z *aus* $-\frac{\pi}{2} \le z \le \frac{\pi}{2}$ *derart, dass* $\sin z = \sin y$ *gilt.*

3.) Mit Hilfe von

$$\sin z = 3\left(1 - \frac{4}{3}\sin^2 u\right)\sin u, \quad u := \frac{z}{3}$$

führe man die Berechnung von $\sin z$ *auf die von* $\sin u$ *(mit* $-\frac{\pi}{6} \le u \le \frac{\pi}{6}$*) zurück. Wieviele Glieder der Taylor-Entwicklung müssen berücksichtigt werden, damit das Resultat auf 7 Dezimalstellen genau ist?*

Nachdem wir mit der ersten Aufgabe eine Intervallreduktion bereits geübt haben, verfahren wir jetzt analog.

Erste Reduktion

Eine ganze Zahl k ist derart zu bestimmen, dass $y = x - 2\pi k$ im Intervall

$$-\frac{\pi}{2} \le y < \frac{3\pi}{2}$$

liegt. Die äquivalenten Umformungen

$$\begin{aligned} -\frac{\pi}{2} &\le x - 2\pi k < \frac{3\pi}{2} \\ 0 &\le x - 2\pi k + \frac{\pi}{2} < 2\pi \\ 0 &\le \frac{x + \pi/2}{2\pi} - k < 1 \\ k &\le \frac{x + \pi/2}{2\pi} < k + 1 \end{aligned}$$

zeigen, dass

$$y := x - 2\pi \cdot \operatorname{floor}\left(\frac{x + \pi/2}{2\pi}\right)$$

die gewünschte Reduktion vermittelt.

Zweite Reduktion
Der Sinus ist symmetrisch zur Geraden $x = \pi/2$, deswegen genügt eine Hälfte des y-Intervalls. Setzt man

$$z := \begin{cases} y & , \text{falls } -\frac{\pi}{2} \le y \le \frac{\pi}{2} \\ \pi - y & , \text{falls } \frac{\pi}{2} < y, \end{cases}$$

so liegt z im Intervall

$$-\frac{\pi}{2} \le z \le \frac{\pi}{2}.$$

Der Sinus nimmt hier noch sämtliche Werte an,

$$-1 \le \sin z \le 1.$$

Um den Sinus im z-Intervall zu berechnen, wollen wir die Taylor-Entwicklung

$$\sin z = z - \frac{z^3}{3!} + \frac{z^5}{5!} \mp \dots$$

verwenden. Diese Potenzreihe ist um $z = 0$ entwickelt und konvergiert daher schnell nur in der Intervallmitte (d. h. um $z = 0$). Da man sich aus Gründen der Effizienz nur auf wenige Glieder der Taylor-Entwicklung beschränkt, muss man sich mit der Intervallmitte begnügen.

Dritte Reduktion
Aus den Additionstheoremen folgt die Beziehung

$$\sin(3u) = 3\sin u - 4\sin^3 u.$$

Beschränkt man sich mit u auf das Intervall

$$-\frac{\pi}{6} \le u \le \frac{\pi}{6},$$

und setzt $z = 3u$, so lassen sich mit Hilfe der Beziehung

$$\sin z = 3\left(1 - \frac{4}{3}\sin^2\frac{z}{3}\right)\sin\frac{z}{3}$$

alle Werte des Sinus berechnen.

Nach drei Reduktionen des Definitionsbereiches $x \to y \to z \to u$ mit $\sin x = \sin y = \sin z$ (Abb. 9.1) steht nun die eigentliche Berechnung des Sinus an.

Berechnung mit Taylor-Entwicklung
Zu berechnen ist $\sin u$ für $-\frac{\pi}{6} \le u \le \frac{\pi}{6}$, wobei wir uns hier exemplarisch auf eine Genauigkeit von 7 dezimalen Stellen beschränken. Wir probieren, ob die vier Terme des Polynoms

$$p(u) := u - \frac{u^3}{3!} + \frac{u^5}{5!} - \frac{u^7}{7!} \tag{9.1}$$

der abgebrochenen Taylor-Entwicklung für die verlangte Genauigkeit ausreichen. Hierzu wird das Restglied der Taylor-Entwicklung abgeschätzt. Die obigen vier Terme sind tatsächlich sogar acht Glieder der Taylor-Entwicklung, da alle geraden Potenzen verschwinden. Für den absoluten Fehler R_8 gilt die Abschätzung (mit $0 < |\zeta| < |u|$)

$$|R_8| = \frac{|\sin^{(9)}(\zeta)||u|^9}{9!} = \frac{|\cos\zeta||u|^9}{9!} \le \frac{|u|^9}{9!}.$$

Wenn 7 Dezimalstellen korrekt sein sollen, dann muss der *relative* Fehler kleiner als 10^{-7} sein. Der relative Fehler ist

$$\frac{|p(u) - \sin u|}{|\sin u|} = \frac{|R_8|}{|\sin u|}.$$

Wegen $|\sin u| > 0{,}95|u|$ im betrachteten u-Bereich gilt für solche u die Abschätzung

$$\frac{|p(u) - \sin u|}{|\sin u|} < \frac{|R_8|}{0{,}95|u|} < \frac{1{,}1|u|^8}{9!} \le \frac{1{,}1}{9!}\left(\frac{\pi}{6}\right)^8 \approx 0{,}18 \cdot 10^{-7} < 10^{-7}.$$

Die vier Terme von (9.1) genügen also für eine Genauigkeit von 7 Dezimalstellen.[2] Eine analoge Rechnung zeigt, dass drei Terme nicht ausreichen. Für einen angenommenen fiktiven 7-stelligen Rechner gilt im betrachteten Intervall

$$\sin u \doteq u - \frac{u^3}{3!} + \frac{u^5}{5!} - \frac{u^7}{7!}.$$

Die Reihenfolge der Rechenoperationen wird durch das Horner-Schema festgelegt. Die inversen Fakultäten sind dabei fertig berechnete Konstanten. So ist die Rechenvorschrift

$$\begin{aligned} v &= u^2 \\ \sin u &\doteq u \cdot (1 + v \cdot (-0{,}16666666 + v \cdot (0{,}0083333333 - v \cdot 0{,}0001984127))) \end{aligned}$$

in einem 7-stelligem Rechenwerk eine erste wirtschaftliche Möglichkeit, $\sin u$ auszuwerten.

Die Überlegungen seien zusammengefasst in einem ersten Algorithmus:

1. Algorithmus zur Berechnung von sin x

1.) $y := x - 2\pi \ \text{floor}\left(\dfrac{x + \pi/2}{2\pi}\right)$

2.) $u := 0{,}33333333 \cdot \begin{cases} y, & \text{falls } -\frac{\pi}{2} \le y \le \frac{\pi}{2} \\ \pi - y, & \text{falls } \frac{\pi}{2} < y \end{cases}$

3.) $v := u \cdot u$

4.) $\xi := u\,(1 + v\,(-0{,}16666666 + v\,(0{,}83333333_{10^{-2}} - v\,0{,}1984127_{10^{-3}})))$

5.) $\sin x \doteq (3 - 4\xi^2)\xi.$

Natürlich sind auch die Zahlenwerte der ersten Reduktion als feste Konstanten anzugeben. Der Aufwand einer Sinus-Auswertung mit diesem Algorithmus beträgt 7 Additionen und 11 Multiplikationen.

Bisher haben wir die Genauigkeit bei der Berechnung von $\sin(u)$ studiert. Stillschweigend wurde dabei vorausgesetzt, dass die Reduktion exakte Resultate liefert. Leider ist dies für „grosse“ Werte von $|x|$ nicht gewährleistet. Die Subtraktion (im Rechenwerk mit endlicher Stellenzahl!) bei der Berechnung von y löscht Genauigkeit aus insbesondere für Argumente $x \approx n\,2\pi$, n eine ganze Zahl. Als Schutzmaßnahme gegen diese Genauigkeitsprobleme wird der x-Bereich begrenzt, so ist $\sin(600)$ auf einem Taschenrechner vielleicht nicht erlaubt.

[2] Mit dem Leibniz-Kriterium für alternierende monotone Reihen lassen sich Fehlerabschätzungen auch für höhere Genauigkeiten ermitteln.

Häufig kommt es vor, dass der Sinus viele tausendmal in einem Programm auszuwerten ist. Es ist daher wesentlich, mit wie vielen Rechenoperationen ein Algorithmus auskommt. Bei dem oben aufgeführten ersten Algorithmus lassen sich noch 3 Multiplikationen einsparen. Die Beziehung

$$\frac{y}{2\pi} = \frac{x}{2\pi} - \text{floor}\left(\frac{x}{2\pi} + 0{,}25\right)$$

legt nahe, x, y und z durch

$$\tilde{x} = \frac{x}{2\pi}, \quad \tilde{y} = \frac{y}{2\pi}, \quad \tilde{z} = \frac{z}{2\pi}$$

zu ersetzen (spart eine Operation). Die Multiplikation mit $1/3$ kann in die Konstanten gesteckt werden, mit deren Hilfe ξ berechnet wird. Eine dritte Multiplikation wird eingespart, wenn man den vierten Schritt umformt zu

$$\sin x = (\sqrt[3]{4}\sin u) \cdot \left[\frac{3}{\sqrt[3]{4}} - (\sqrt[3]{4}\sin u)^2\right].$$

Im dritten Schritt wird nun statt $\sin u$

$$\sqrt[3]{4}\ \sin\frac{z}{3}$$

berechnet, der Faktor $\sqrt[3]{4}$ wird ebenfalls in die Konstanten gesteckt:

$$\begin{aligned}
\sqrt[3]{4}\ \sin\frac{z}{3} &= \sqrt[3]{4}\ \sin\left(\tilde{z}\frac{2\pi}{3}\right) \\
&= \sqrt[3]{4}\,\tilde{z}\frac{2\pi}{3}\left(1 - \frac{\tilde{z}^2}{3!}\left(\frac{2\pi}{3}\right)^2 + \frac{\tilde{z}^4}{5!}\left(\frac{2\pi}{3}\right)^4 - \dots\right) \\
&= \tilde{z}\left(\sqrt[3]{4}\frac{2\pi}{3} - \tilde{z}^2\frac{\sqrt[3]{4}}{3!}\left(\frac{2\pi}{3}\right)^3 + \tilde{z}^4\frac{\sqrt[3]{4}}{5!}\left(\frac{2\pi}{3}\right)^5 - \dots\right) \\
&= \tilde{z}\,(c_0 - \tilde{z}^2 c_1 + \tilde{z}^4 c_2 - \dots).
\end{aligned}$$

Der Algorithmus kommt nun mit 8 Multiplikationen und 7 Additionen aus:

2. Algorithmus

$$
\begin{aligned}
&\tilde{x} := x \cdot 0{,}15915494 \quad \left(= \frac{x}{2\pi}\right)\\
&\tilde{y} := \tilde{x} - \operatorname{floor}(\tilde{x} + 0{,}25)\\
&\tilde{z} := \begin{cases} \tilde{y} & \text{falls } \tilde{y} \le 0{,}25 \\ 0{,}5 - \tilde{y} & \text{falls } \tilde{y} > 0{,}25 \end{cases}\\
&v := \tilde{z}\tilde{z}\\
&w := \tilde{z}(c_0 + v(-c_1 + v(c_2 - vc_3)))\\
&\sin x \doteq w(d - ww)
\end{aligned}
$$

mit den Konstanten

$$
\begin{aligned}
c_0 &= 0{,}332464499_{10^1}\\
c_1 &= 0{,}243058747_{10^1}\\
c_2 &= 0{,}53308748\\
c_3 &= 0{,}5567579_{10^{-1}}\\
d &= 0{,}188988158_{10^1}.
\end{aligned}
$$

Andere Methoden
Die Verwendung einer Taylorreihe ist nicht die einzige Möglichkeit zur Berechnung des Sinus. Wählt man die Tschebyschow-Entwicklung anstelle der Taylor-Entwicklung, so erhält man ein anderes Polynom, das bessere Approximationseigenschaften besitzt: Die Koeffizienten der Tschebyschow-Entwicklung nehmen schneller ab, für eine vorgegebene Genauigkeitsforderung genügen daher weniger Terme, vgl. etwa Sauer & Szabó.

Neben der Verwendung von abgebrochenen Reihen ist auch ein anderer Zugang attraktiv: Die Approximation von Funktionen mit Hilfe von vorab berechneten und gespeicherten Wertetabellen. Darauf aufbauend, können gewünschte Werte mit Interpolation, *Shift-and-Add*-Techniken oder Integration äußerst effizient berechnet werden.

Als Anregung sei die simultane Berechnung von $\sin x$ und $\cos x$ betrachtet. Zunächst berechnet man für gegebenes x ($|x| \le \frac{\pi}{2}$) die Darstellung

$$x \doteq \sum_{k=0}^{n} \xi_k \lambda_k \quad \text{mit } \lambda_k = \arctan(2^{-k}), \quad \xi_k = +1 \text{ oder } \xi_k = -1,$$

d. h. die Vorzeichen ξ_k sind zu bestimmen (die λ_k sind feste Zahlenwerte). Für dieses x gilt mit der imaginären Einheit $\mathrm{i} = \sqrt{-1}$:

$$\begin{aligned}
\cos x + \mathrm{i} \sin x = \mathrm{e}^{\mathrm{i}x} &\doteq \exp \sum_{k=0}^{n} \mathrm{i}\, \xi_k\, \lambda_k \\
&= \prod_{k=0}^{n} \exp(\mathrm{i}\xi_k \lambda_k) = \prod_{k=0}^{n} (\cos(\xi_k \lambda_k) + \mathrm{i} \sin(\xi_k \lambda_k)) \\
&= \prod_{k=0}^{n} (\cos \lambda_k + \mathrm{i}\xi_k \sin \lambda_k) \\
&= \prod_{k=0}^{n} \cos \lambda_k \cdot (1 + \mathrm{i}\xi_k \tan \lambda_k) \\
&= \prod_{k=0}^{n} \cos \lambda_k \cdot \prod_{k=0}^{n} (1 + \mathrm{i}\xi_k 2^{-k}) \\
&= \text{Konstante} \cdot \prod_{k=0}^{n} (1 + \mathrm{i}\xi_k 2^{-k}).
\end{aligned}$$

Die Berechnung des Produktes lässt sich schnell mit einer Laufanweisung durchführen. Hierbei werden Real- und Imaginärteil des Produktes (also $\cos x$ und $\sin x$) ohne echte Multiplikation berechnet: Ausdrücke der Form $\xi_k 2^{-k} b$ sind bei einer Binärzahl b lediglich „Verschiebungen“ der Mantisse mit eventueller Umkehrung des Vorzeichens. Der Index n hängt von der Stellenzahl des Rechners ab.

Literatur

zur numerischen Berechnung von Funktionen:

Bulirsch, R., Stoer, J. : Darstellung von Funktionen in Rechenautomaten. In: Sauer, R., Szabó, I. (Hrsg.), Mathematische Hilfsmittel des Ingenieurs. Bd. III. Springer, Berlin (1968)

Ein Mikroprogramm zur simultanen Berechnung von Sinus und Kosinus in

Spaniol, O.: Arithmetik in Rechenanlagen. Teubner, Stuttgart (1976)

Moderne Computermethoden in

Muller, J.-M.: Elementary Functions, Algorithms and Implementation. Second Edition. Birkhäuser, Boston (2006)

Herzschlag

10

Modelle für das Verhalten des menschlichen Herzschlags werden intensiv studiert. Ein gesundes Herz reagiert auf Anstrengung oder Aufregung. Die Beschleunigung des Herzschlags nach einem Adrenalin-Ausstoß sowie die Verlangsamung des Herzschlags auf Grund gewisser Nervenimpulse müssen verstanden werden, um bessere künstliche Herzen zu konstruieren, welche genügend flexibel auf derartige Belastungen reagieren können. Neben dem „normalen" Verhalten des Herzens wird auch pathologisches Verhalten modelliert, wie das Auftreten von irregulären Schlägen. Das allgemeine Verständnis des Herzschlags sowie die Entwicklung von Steuermöglichkeiten und Mikroprozessoren für künstliche Herzen oder Herzschrittmacher hängen entscheidend davon ab, welche Modelle von Biologen, Medizinern und Mathematikern entwickelt und gelöst werden können.

Dieses Kapitel diskutiert ein extrem einfaches Modell für die Dynamik des Herzschlags[1]. Das Modell konzentriert sich auf die Bewegung einer Herz-Muskelfaser, wie sie auf den elektrochemischen Impuls der Herzschrittmacherwelle reagiert, und wie die Muskel-Kontraktion vom Blutdruck abhängt. Dieses „lokale" Modell erhebt keinen Anspruch, das gesamte Herz des Menschen zu modellieren, und ist insoweit weit vom heutigen Wissensstand entfernt. Es deckt aber wichtige qualitative Verhaltensweisen auf und erklärt die periodisch wiederkehrende Kontraktion einer Muskelfaser.

Mit einer ersten Aufgabe werden zunächst die Variablen definiert, und der Ruhezustand des Herzschlags diskutiert.

[1] nach E.C. Zeeman (1977).

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_10

Aufgabe 1 *Der Herzschlag des Menschen kann qualitativ durch das folgende System von Differenzialgleichungen beschrieben werden:*

$$\dot{x} = -k \cdot X(x, y), \quad X(x, y) := x^3 - bx + y, \tag{10.1}$$

$$\dot{y} = Y(x, y), \quad Y(x, y) := x - x_0. \tag{10.2}$$

Variable:

$x(t)$: Länge einer Herzmuskelfaser zum Zeitpunkt t (plus einer Konstanten), $x(t)$ beschreibt den Herzschlag. x_0 ist die Länge dieses Herzmuskels im Ruhestand Diastole.

$y(t)$: elektrochemischer Impuls, Steuergröße.

b: Blutdruck, vereinfacht als konstant angenommen, Steuergröße.

k: Parameter zur Steuerung der Schnelligkeit der Bewegung.

Zahlenwerte zur Illustration: $b = 1,\ x_0 = 1,1,\ k = 100$ oder $k = 10$.

a) Durch $X(x, y) = 0$ ist in der (x, y)-Ebene eine Kurve K definiert. Man zeichne den Graphen von K.

b) Das System hat eine Gleichgewichtslage (x_s, y_s) (d. h. $X(x_s, y_s) = Y(x_s, y_s) = 0$), die stabil ist und in der Abb. 10.1 mit S bezeichnet ist. Man ermittle diesen Ruhezustand des Systems.

c) Man diskutiere das qualitative Verhalten der Bahnkurven in einer Umgebung dieser Gleichgewichtslage.

Ehe wir Lösungen des Differenzialgleichungs-System (10.1)/(10.2) diskutieren, versuchen wir eine erste Bewertung. Die Differenzialgleichung (10.2) für y beschreibt

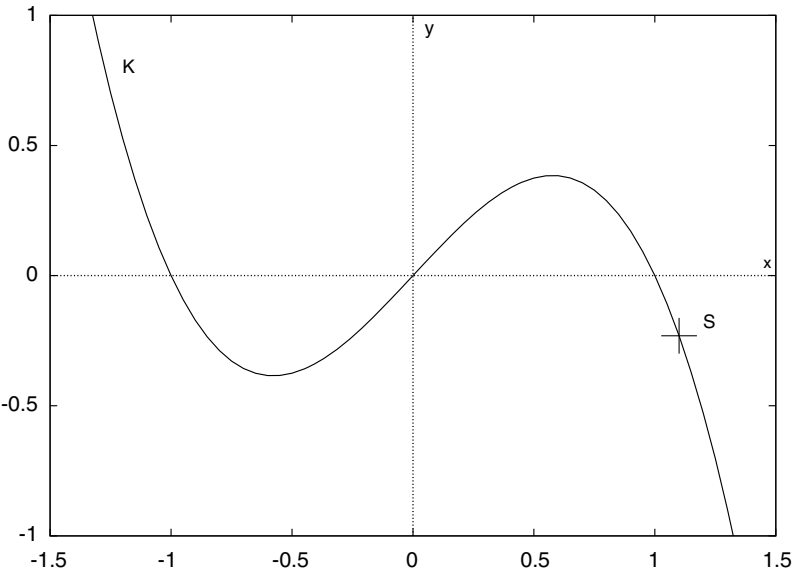

Abb. 10.1 (x, y)-Phasenebene mit Kurve K; entlang K gilt $\dot{x} = 0$; Gleichgewichtslage und Attraktor bei S

keine besondere Dynamik, sondern lediglich eine Rückführung auf einen Grundzustand: y nimmt zu ($\dot{y} > 0$) wenn die Muskelfaser gedehnt ist ($x > x_0$), und nimmt ab im umgekehrten Fall. Mit diesem schlichten Sachverhalt wird keine autonome periodische Bewegung der Muskelfaser eintreten. Dazu wird zusätzlich zu (10.1)/(10.2) eine externe Kraft notwendig, die Herzschrittmacherwelle.

Lokales Lösungsverhalten nahe S
In der (x, y)-Ebene ist durch

$$x^3 - bx + y = 0$$

eine Kurve K definiert. Entlang dieser Kurve K gilt $\dot{x} = 0$ (der Punkt bedeutet die Ableitung nach der Zeit t). Für $b = 1$ ist K das kubische Polynom $y(x) = x - x^3$. Zum Skizzieren benötigte Größen des Polynoms sind:

$$\begin{aligned}
&\text{Nullstellen bei } x = 0,\ x = \pm 1,\ \text{mit}\\
&\text{Steigungen } y'(0) = 1,\ y'(\pm 1) = -2,\\
&\text{Extrema } (y' = 0) \text{ für}\\
&(x, y) = \left(\pm\frac{1}{\sqrt{3}},\ \pm\frac{2}{3\sqrt{3}}\right) = (\pm 0{,}577,\ \pm 0{,}385).
\end{aligned}$$

Demzufolge hat die Kurve K die in der Abb. 10.1 gezeichnete Form.

Der Gleichgewichtspunkt[2] (x_s, y_s) muss auf K liegen. Aus $\dot{y} = 0$ folgt $x_s = x_0$ und $y_s = x_0(1 - x_0^2)$, bei den gewählten Zahlenwerten also

$$(x_s, y_s) = (1{,}1, -0{,}231).$$

Dieser Punkt S in der (x, y)-Ebene ist die einzige Ruhelage.

Um das qualitative Verhalten der Bahnkurven in einer Umgebung des Gleichgewichts zu studieren, wird das Differenzialgleichungssystem (10.1), (10.2) um (x_s, y_s) linearisiert: In der Nähe von (x_s, y_s) („lokal“) gilt für die Lösungen (Bahnkurven, Trajektorien) $x(t)$, $y(t)$

$$\begin{aligned}
\begin{pmatrix}\dot{x}\\ \dot{y}\end{pmatrix} &\approx \begin{pmatrix}-k(3x_s^2 - b) & -k\\ 1 & 0\end{pmatrix}\begin{pmatrix}x - x_s\\ y - y_s\end{pmatrix}\\
&= \begin{pmatrix}-263 & -100\\ 1 & 0\end{pmatrix}\begin{pmatrix}x - 1{,}1\\ y + 0{,}231\end{pmatrix} =: A\begin{pmatrix}x - 1{,}1\\ y + 0{,}231\end{pmatrix}.
\end{aligned}$$

Die Matrix A enthält die partiellen Ableitungen der Funktionen $X(x, y)$ und $Y(x, y)$, ausgewertet an (x_s, y_s). Zur Diskussion dieses linearen Systems berechnet man die Eigenwerte λ der Matrix A aus

$$0 = \lambda^2 + 263\,\lambda + 100.$$

[2] auch *stationärer Punkt, Ruhelage* oder *kritischer Punkt.* Im Hinblick auf die spezielle Lösung der Differenzialgleichung auch als *stationäre Lösung* bezeichnet.

Die Wurzeln λ_1, λ_2 dieser quadratischen Gleichung

$$\lambda_1 = -0{,}38078$$
$$\lambda_2 = -262{,}62$$

sind beide reell und negativ. Also ist (x_s, y_s) ein stabiler Knoten; zumindest benachbarte Bahnkurven werden von (x_s, y_s) angezogen.

Unklar ist noch, in welcher Weise die Trajektorien auf (x_s, y_s) zulaufen. Um dies zu klären, müssen die Eigenvektoren $v^{(i)}$ (zum Eigenwert λ_i, $i = 1, 2$) der Matrix A bestimmt werden. Die Trajektorien verhalten sich dann lokal wie

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} \approx \begin{pmatrix} x_s \\ y_s \end{pmatrix} + c_1 v^{(1)} e^{\lambda_1 t} + c_2 v^{(2)} e^{\lambda_2 t}, \tag{10.3}$$

c_1 und c_2 sind vom Ausgangspunkt der Trajektorie abhängige reelle Zahlen.[3]

Bezeichnen v_1 und v_2 die beiden Komponenten eines Eigenvektors v, so lautet das homogene Gleichungssystem $(A - \lambda I)v = 0$ komponentenweise

$$(-263 - \lambda)v_1 - 100v_2 = 0$$
$$v_1 - \lambda v_2 = 0.$$

Die zweite dieser Gleichungen lässt die gesuchten Eigenvektoren bequem ablesen:

$$\text{Wähle } v_2 = 1, \text{ dann gilt } v_1 = \lambda.$$

Damit lauten die Eigenvektoren von A

$$v^{(1)} = \begin{pmatrix} -0{,}38078 \\ 1 \end{pmatrix}, \quad v^{(2)} = \begin{pmatrix} -262{,}62 \\ 1 \end{pmatrix}.$$

In Abb. 10.2 sind die durch $v^{(1)}$ und $v^{(2)}$ gegebenen Eigengeraden eingezeichnet. Nach (10.3) würde eine Trajektorie, die auf einer Eigengeraden startet ($c_1 = 0$ oder $c_2 = 0$), auf dieser Geraden entlang laufen in Richtung des stabilen Gleichgewichtes S. Da $|\lambda_2|$ erheblich größer ist als $|\lambda_1|$, verläuft die Bewegung entlang der durch $v^{(2)}$ definierten Eigengeraden mit viel höherer Geschwindigkeit als auf der anderen Eigengeraden. Dies ist durch die Größe der Pfeile in Abb. 10.2 angedeutet.

Die Richtung der $v^{(1)}$-Eigengeraden fällt im stationären Punkt S mit der Richtung der Kurve K zusammen. Wegen $|\lambda_1| \ll |\lambda_2|$ schmiegen sich die Trajektorien an K

[3] zur Erinnerung: Bei einem linearen Differenzialgleichungssystem $\begin{pmatrix} \dot{p} \\ \dot{q} \end{pmatrix} = A \begin{pmatrix} p \\ q \end{pmatrix}$ mit Skalaren $p(t), q(t)$ führt der Ansatz $\begin{pmatrix} p \\ q \end{pmatrix} = e^{\lambda t} v$ auf das Gleichungssystem $(A - \lambda I)v = 0$, I ist die Einheitsmatrix.

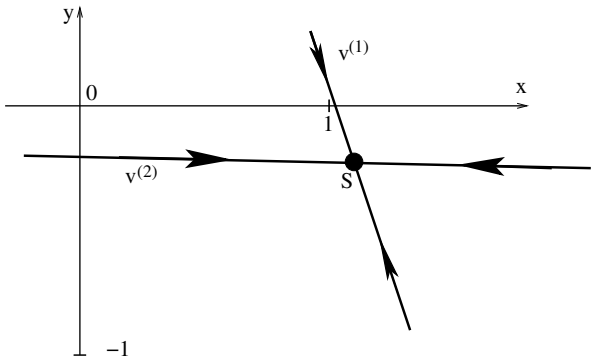

Abb. 10.2 Gleichgewichtslage und Attraktor bei S, mit Eigenvektoren/Eigengeraden

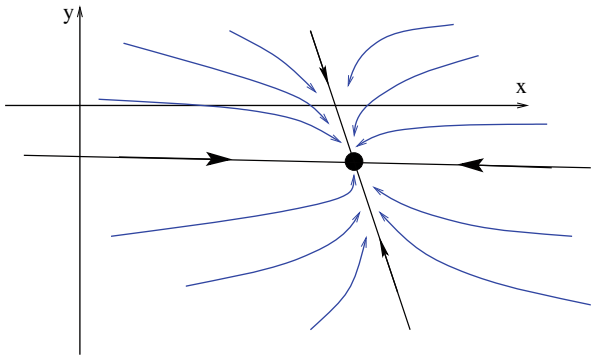

Abb. 10.3 (x, y)-Phasenebene, qualitatives lokales Verhalten von Trajektorien (blau) nahe des stabilen Gleichgewichts S (schematisch und stark vergrößert)

an, vergleiche die Abb. 10.3. Das lokale Verhalten der Trajektorien um den stabilen Knoten S ist in dieser Abbildung stark vergrößert wiedergegeben.

Globales Lösungsverhalten
Soweit die Analyse des lokalen Bahnkurven-Verhaltens nahe S. Nun folgt die Analyse des *globalen* Verhaltens. Es geht darum, das in Abb. 10.4 skizzierte Lösungsverhalten des Systems (10.1)/(10.2) zu begründen.

Aufgabe 2 *Gegeben sind die Differenzialgleichungen einer Herzmuskelfaser (10.1), (10.2), mit* $b = 1$, $x_0 = 1{,}1$, $k = 100$. *Die Abb. 10.4 zeigt qualitativ globale Lösungsverläufe.*

a) Man begründe das Lösungsverhalten, ohne die Differenzialgleichungen zu lösen. Das System befinde sich im Gleichgewichtspunkt $(x_\mathrm{s}, y_\mathrm{s})$. *Durch eine äußere Einwirkung (Herzschrittmacherwelle) wird* y *von* y_s *bis auf* y_max *angehoben; die*

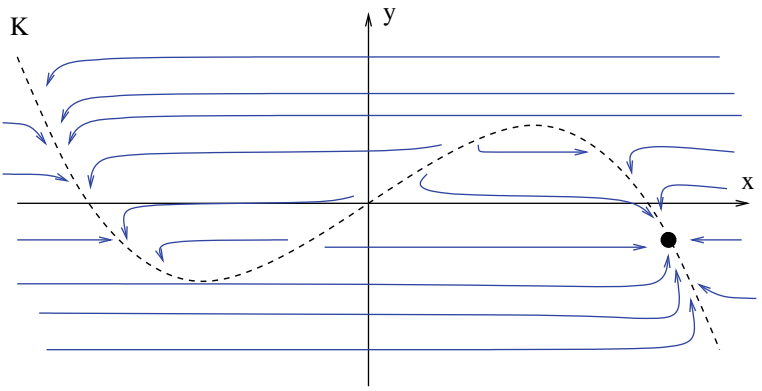

Abb. 10.4 (x, y)-Phasenebene, schematische Skizze des globalen Lösungsverhaltens. Die durch $X(x, y) = 0$ definierte Kurve K ist gestrichelt eingezeichnet, die stationäre Lösung (x_s, y_s) als fetter Punkt

Variable x wird währenddessen durch ihre Differenzialgleichung $\dot{x} = -kX(x, y)$ *bestimmt. Wenn* y_{max} *erreicht ist, endet die äußere Einwirkung.*

b) Man überlege sich mit Hilfe der Abb. 10.4, welchen Weg das System $(x(t), y(t))$ *zurück zum stabilen Ruhezustand* (x_s, y_s) *nimmt und zeichne diesen Weg. (Die sprunghafte Kontraktion und Entspannung des Herzmuskels werden sichtbar.)*

c) Man überlege sich b) für $b = 1{,}5$ *(Herzinfarkt bei zu hohem Blutdruck) und zeichne auch hier einen Zyklus von* (x_s, y_s) *zurück zu* (x_s, y_s).

Wir denken uns die (x, y)-Ebene aufgeteilt in drei Bereiche (Abb. 10.5): In einem Streifen in der „Mitte“ entlang und beiderseits der Kurve K gilt $\dot{x} \approx 0$, also Bahnkurven mehr oder weniger parallel zur y-Achse. Außerhalb des Streifens ist $X(x, y)$

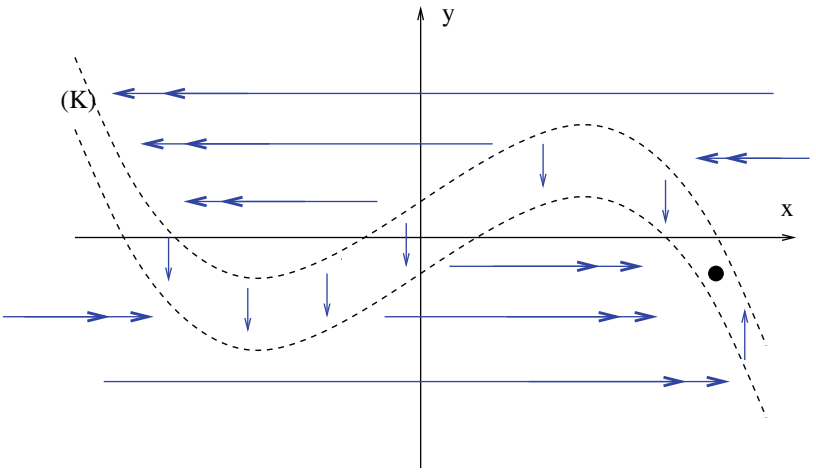

Abb. 10.5 Aufteilung der (x, y)-Ebene in drei Bereiche mit völlig unterschiedlichem Lösungsverhalten (schematisch und vereinfacht)

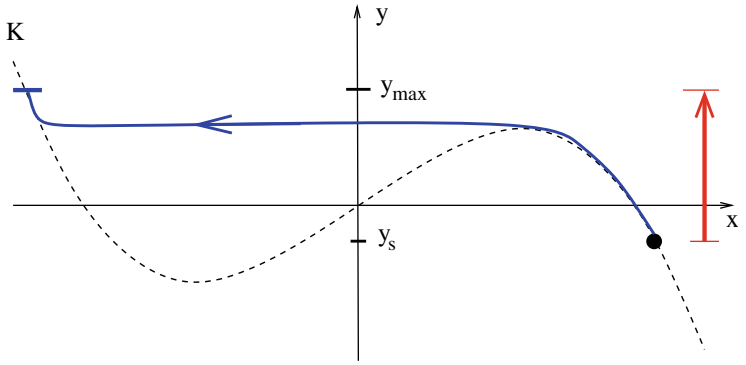

Abb. 10.6 Erste Phase: Herzschrittmacherwelle (rot) und Kontraktion der Muskelfaser (blau) als Antwort des Systems

deutlich genug von 0 verschieden. Der große Wert des Faktors k sorgt dafür, dass die Bahnkurven mehr oder weniger parallel zur x-Achse verlaufen, hier ist $|\dot{x}|$ groß im Vergleich zu $|\dot{y}|$. Oberhalb des Streifens gilt $X(x, y) > 0$, und wegen des Faktors $-k$ ist $\dot{x} < 0$, geht die Bewegung nach „links“, also in Richtung fallender x-Werte. Unterhalb des Streifens ist es umgekehrt. Dies ist schematisch in der Abb. 10.5 wiedergegeben. Je größer k ist, umso schmaler ist der Streifen.

Für die folgende Überlegung teilen wir die Kurve K und den Streifen um die Kurve K in drei Teile auf: Das Mittelstück mit positiver Steigung des Polynoms und die beiden äußeren Äste mit negativer Steigung (Abb. 10.5). Offenbar ist der mittlere Teil instabil: Jeder Punkt (x, y) aus diesem Mittelstück wird weggerissen hin zu den äußeren Ästen von K. Damit ist das globale Lösungsverhalten hinlänglich charakterisiert, jede Trajektorie endet schließlich im stabilen Knoten (x_s, y_s). Und ohne eine äußere Einwirkung würde sie dort bleiben.

Durch die äußere Einwirkung der Herzschrittmacherwelle wird y angehoben, bei den von uns gewählten Zahlenwerten auf $y_{max} = 0{,}5$ (Abb. 10.6). Während dieser äußeren Einwirkung auf das System bleibt vorübergehend nur die x-Differenzialgleichung (10.1) aktiv. Solange die Herzschrittmacherwelle den Wert von y bestimmt, ist die Differenzialgleichung (10.2) außer Kraft. Das ist die erste Phase:

1. Phase (Kontraktion)
x bestimmt durch $\dot{x} = -kX(x, y)$,
$y \longrightarrow y_{max}$ durch äußere Einwirkung.

Das Systemverhalten in dieser ersten Phase ist in der Abb. 10.6 wiedergegeben. Zunächst wird (x, y) aus der Ruhelage herausgezogen. Infolge des ziemlich „waagerechten Drucks“ bleibt (x, y) noch in der Nähe des rechten Astes von K. Sobald (x, y) den instabilen Bereich von K erreicht (y hat hier seinen maximalen Wert noch nicht angenommen) springt die Bewegung schnell zum linken äußeren Ast von K. Hier erfolgt noch ein kleines Stück langsamer Bewegung, bis mit dem maximalen Wert

von $y = y_{\max}$ die Kontraktion des Herzmuskels abgeschlossen ist. Siehe hierzu die Abb. 10.6.

Wenn die äußere Einwirkung auf y endet, folgt die zweite Phase:

2. Phase (Entspannung)
x und y werden beide durch ihre Differenzialgleichungen bestimmt.

Die Trajektorie des „Rückwegs" gehorcht dem globalen Lösungsverhalten, wie es durch die beiden Differenzialgleichungen (10.1)/(10.2) definiert ist (Abb. 10.7). Zusammen erklären die Kontraktions- und die Entspannungsphase das qualitative Bild von Abb. 10.4. Nachdem der Herzmuskel in der Herzpause eine kurze Zeit im stabilen Gleichgewicht S verharrt, kann der Herzschlag mit dem Einsetzen der ersten Phase erneut beginnen.

Wir haben nun die Dynamik der Herzbewegung im Wesentlichen erfasst. Das Modell bietet aber noch mehr: Der Blutdruck b, bisher als konstant angenommen, kann variiert werden. Für den Wert $b = 1{,}5$ ändert die Kurve K ihre Gestalt, die Durchbiegung des Polynoms wird im Mittelteil größer. Die Nullstellen des kubischen Polynoms liegen jetzt bei $x = \pm\sqrt{3/2} = \pm 1{,}225$, die Extremwerte bei $(x, y) = (\pm 0{,}707,\ \pm 0{,}707)$. Die Zahlenwerte sind derart, dass das Maximum von K, die „Schulter", oberhalb des Niveaus $y_{\max}$ liegt (Abb. 10.8). Im Verlauf der Herzschrittmacherwelle wird die Schulter nicht mehr erreicht und es erfolgt kein Sprung, keine Kontraktion. Zu hoher Blutdruck bedeutet in diesem Modell ein Anheben der Schulter von K über $y_{\max}$ hinaus, also

$$\frac{2}{3}\, b \sqrt{\frac{b}{3}} > y_{\max}.$$

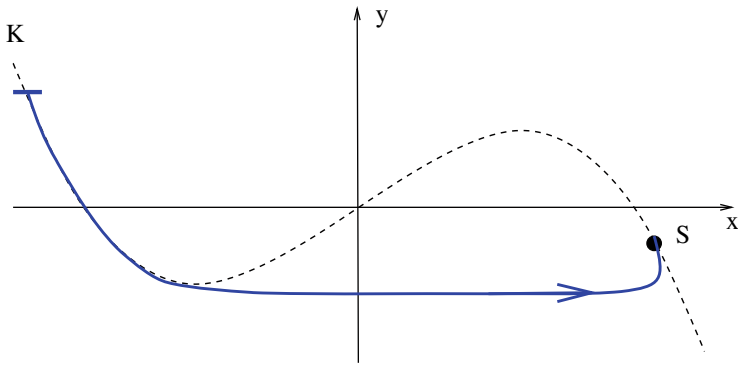

Abb. 10.7 Zweite Phase: Entspannung des Herzmuskels

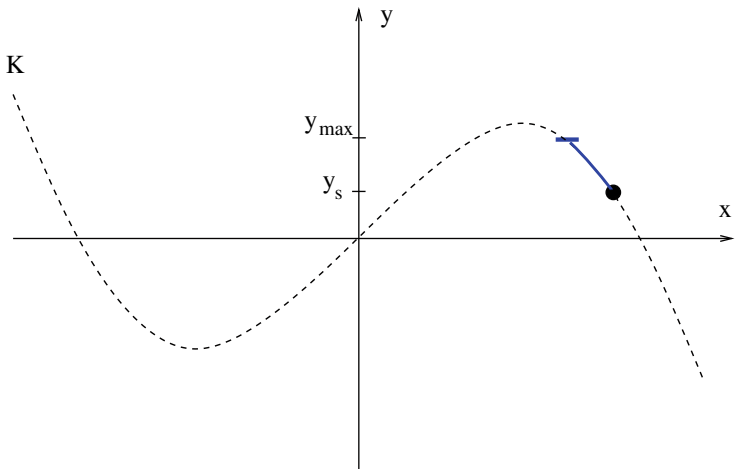

Abb. 10.8 zu hoher Blutdruck: keine Kontraktion

Für

$$b > \left(\frac{3\sqrt{3}}{3}\, y_{\max}\right)^{2/3}$$

kontrahiert der Herzmuskel nicht, er führt nur eine sehr schwache Bewegung aus.[4]

Fasst man b als kontinuierliche Variable auf, so wird durch

$$0 = X(x, y, b) = x^3 - bx + y$$

statt einer Kurve eine Fläche im dreidimensionalen (x, y, b)-Raum definiert (Abb. 10.9). Über einen Teilbereich der (y, b)-Ebene (in der Form eines Horns, *cusp*) schieben sich Teile der Fläche übereinander. Die Fäche kann als „langsame Mannigfaltigkeit“ bezeichnet werden, entlang der Fläche ist die Bewegung in x-Richtung langsam. Die Fläche hat zwei Kanten mit „senkrechter“ Steigung; hier treten Sprünge von der oberen Teilfläche auf die untere auf (und umgekehrt), entsprechend den Abb. 10.5 und 10.6. Anhand dieser Fläche von Abb. 10.9 können verschiedene Arten von Herzbewegungen erklärt werden:

b_1: Herzvorkammer, geringer Blutdruck, geringe Kontraktion;
b_2: Hauptkammer, normaler Blutdruck, starke Kontraktion;
b_3: zu hoher Blutdruck, keine Kontraktion.

[4] bei den hier gewählten willkürlichen Zahlen.

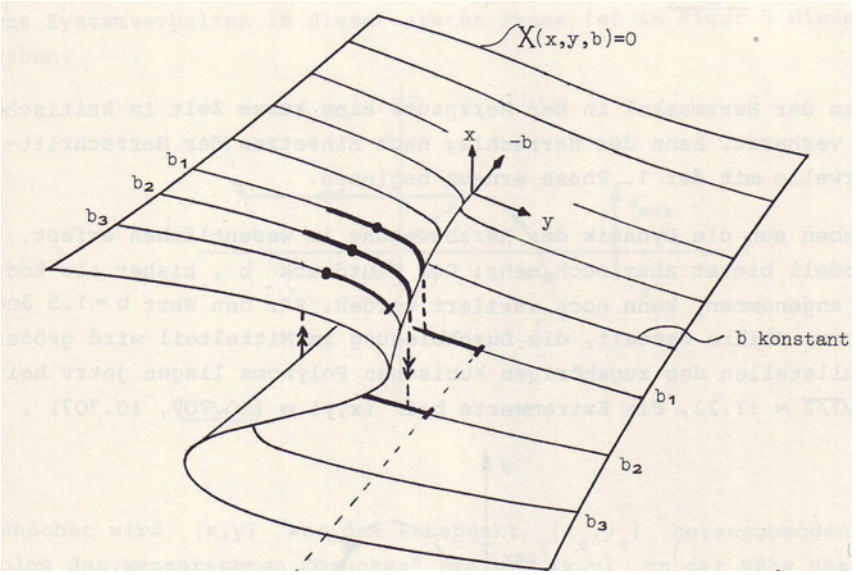

Abb. 10.9 „Cusp-Katastrophe“, b variabel, Fläche $X(x, y, b) = 0$ im (x, y, b)-Raum. (Quelle: R. Seydel, R. Bulirsch: Vom Regenbogen zum Farbfernsehen. Springer 1986)

Unterhalb der kritischen Blutdruck-Grenze bewirkt ein höherer Blutdruck einen kräftigeren Herzschlag. Phänomene, die sich mit einer solchen Fläche beschreiben lassen, heißen auch *Cusp-Katastrophe.*

Bisher haben wir qualitativ diskutiert, mit entweder schnellen oder mit langsamen Bewegungen. In Wirklichkeit haben die Lösungen der Differenzialgleichungen auch Übergangsbereiche, keine explizit „waagerechte“ Bewegung. Um diese realistischeren Bewegungen zu sehen, haben wir das Differenzialgleichungs-System numerisch integriert. Diese Illustration findet sich in der Abb. 10.10. Um die Bewegung etwas weicher aussehen zu lassen, ist hier ein kleinerer Wert von k gewählt, $k = 10$.

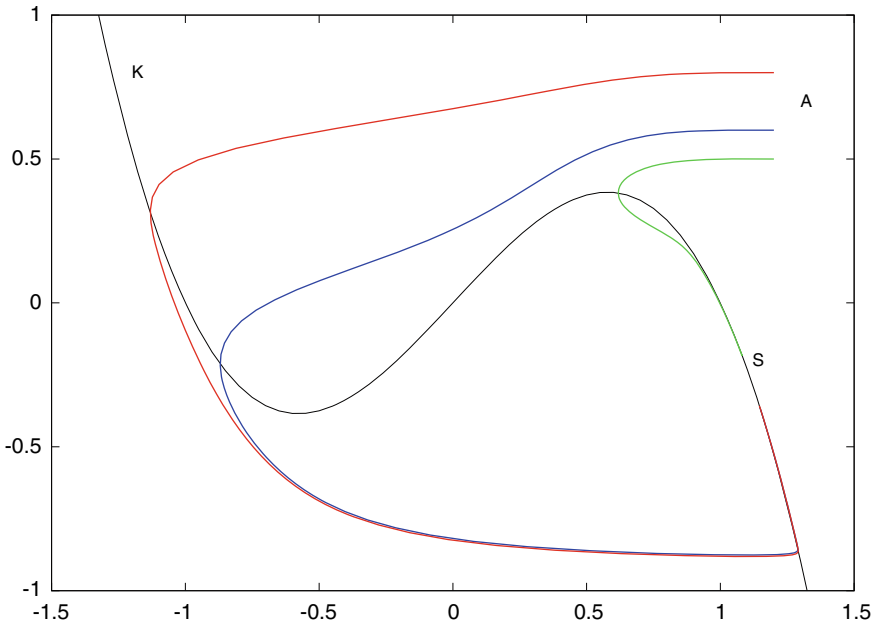

Abb. 10.10 (x, y)-Ebene. Drei numerisch berechnete Trajektorien $(x(t), y(t))$ und Lösungen von (10.1)/(10.2) für $k = 10$; Start jeweils im Bereich A. Schwarze Kurve K: die kubische Parabel $X(x, y) = 0$. Stabile Gleichgewichtslage und Attraktor bei S. Die grüne Trajektorie führt keinen vollen Herzschlag aus

Literatur

Die Fallstudie basiert auf dem Modell von

Zeeman, E.C.: Catastrophe Theory. Addison-Wesley, London (1977)

Als Beispiel für ein anspruchsvolleres Modell sei verwiesen auf

Beeler, G.W., Reuter, H.: Reconstruction of the action potential of ventricular myocardial fibres. J. Physiol. **268**:177–210 (1977)

Nervenimpulse

11

Von ähnlicher Bedeutung wie Herzschlagmodelle sind Modelle, welche die Ausbreitung von Nervenimpulsen beschreiben. Unter anderem dienen einige Nervenmodelle auch der Beschreibung des Herzschlags, dann nämlich, wenn angenommen wird, dass die mechanische Kontraktion einer Muskelfaser durch eine Nervenreizung ausgelöst wird. Mathematisch werden Nervenimpulse meist mit partiellen Differenzialgleichungen modelliert, deren Lösung aufwendige numerische Methoden erfordern. Ein Prototyp ist das Modell von Hodgkin und Huxley. Unter zusätzlichen vereinfachenden Annahmen resultiert das Modell von FitzHugh. Letzteres wird durch gewöhnliche Differenzialgleichungen beschrieben, wir diskutieren es in der vorliegenden Fallstudie.

Elektrische Spannungsschwankungen spielen für die Ausbreitung von Nervenimpulsen eine wesentliche Rolle. Im Ruhezustand sind die Neuronen innen negativ geladen. Dies wird insbesondere durch eine im Inneren im Vergleich zum Äußeren niedrige Konzentration von Na^+-Ionen verursacht. Ein entsprechendes Konzentrationsgefälle von Natrium von der äußeren zur inneren Seite der Nervenmembran, ermöglicht durch geringere Permeabilität für Na^+, wird durch Ionen-Pumpen aufrecht erhalten. Im Falle einer Nervenreizung „öffnet“ sich die Membran vorübergehend (kurzzeitige hohe Na^+-Permeabilität) und lässt Na^+-Ionen einströmen. Hierdurch ändert sich die Ladungsverteilung, und es fließt ein Strom. Nach Absinken der Na^+-Permeabilität nimmt im Inneren des Neurons die Na^+-Konzentration ab und es stellt sich erneut das Ruhepotenzial ein.

Dieser Wechsel von Depolarisation und Repolarisation, hier vereinfacht dargestellt, läuft in wenigen Millisekunden ab, und kann als einzelne Schwingung des elektrischen Potenzials angesehen werden. Bei Andauern der Nervenreizung wiederholt sich dieser Vorgang periodisch. Die Amplitude der Schwingung ist nicht proportional zur Intensität der Nervenreizung: Der Nerv spricht nicht an, wenn die Reizung schwächer als ein gewisser Schwellenwert ist; hier bleibt das Potenzial *stationär*

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_11

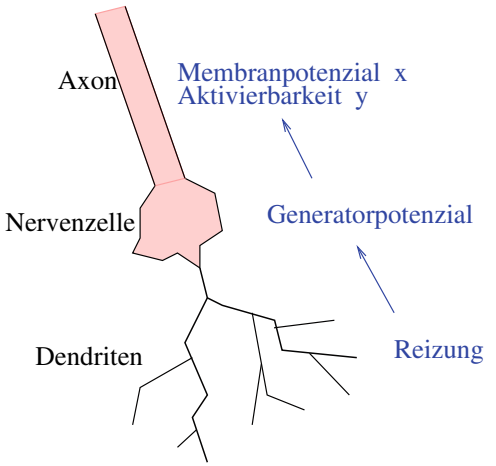

Abb. 11.1 Nerven-Reizung bewirkt ein Generatorpotenzial; dies beeinflusst die Membran des Axons, insbesondere die Durchlässigkeit (Aktivierbarkeit) für Ionen

auf seinem Ruhe-Niveau. Erst wenn die Nervenreizung diesen kritischen Schwellenwert überschreitet, wird das Ruhepotenzial instabil, und es setzt die Schwingung mit sprunghaft voller Amplitude ein (Spannungsschwankung etwa 100 mV).

In der folgenden Aufgabe wird ein solches Schwellenverhalten zwischen Ruhe und (Ladungs-)Bewegung simuliert. Das Modell ist qualitativer Natur, die Zahlenwerte lassen keine unmittelbare Deutung zu. Beim Größenvergleich des Generatorpotenzials γ ist zu bedenken, dass γ negativ ist.[1] (Abb. 11.1).

Aufgabe

Erklärung: Die Intensität der Nervenreizung bewirkt ein unterschiedliches Generatorpotenzial γ in den Dendriten der Nervenzellen. Abhängig von dem Wert des Generatorpotenzials γ kann das Membranpotenzial der Nervenfaser (Axon) zwei Grundzustände annehmen: Ist γ größer als ein Schwellenwert γ_0 ($\gamma_0 < \gamma < 0$, geringe Reizung), so ist das Membranpotenzial stationär. Im Fall $\gamma < \gamma_0 < 0$ (Reizung genügend groß[2]), so besteht das Membranpotenzial aus periodischen Impulsen.

Mathematisches Modell:

Das Potenzialverhalten von Nervenfasern wird qualitativ durch die Lösungen der Differenzialgleichungen

$$\dot{x} = 3\left(y + x - \frac{1}{3}x^3 + \gamma\right) \tag{11.1}$$

$$\dot{y} = -\frac{1}{3}\left(x - \frac{7}{10} + \frac{8}{10}y\right) \tag{11.2}$$

[1] $\gamma > \gamma_0$ bedeutet demnach, dass der absolute Wert des Potenzials γ *schwächer* ist als $|\gamma_0|$

[2] Allerdings sollte im betrachteten Modell $|\gamma|$ nicht zu groß sein, siehe Abb. 11.2.

beschrieben. Hierbei bedeuten

$$x(t) : Membranpotenzial\ zum\ Zeitpunkt\ t,$$
$$y(t) : Aktivierbarkeit\ der\ Nervenzelle,$$
$$\gamma : Generatorpotenzial.$$

Aufgabe:

a) *Man berechne die stationären Lösungen (Punkte* (x, y) *in der Phasenebene mit* $\dot{x} = \dot{y} = 0$*) der Differenzialgleichungen in Abhängigkeit von* γ*. Welches ist ein Schwellenwert* γ_0*, der Bereiche stabiler und instabiler stationärer Punkte trennt? (Hinweis: Es ist hier sinnvoll,* γ *durch* x *auszudrücken.)*
b) *Man skizziere in einer* (γ, x)*-Ebene die stationären Punkte; hierbei markiere man den instabilen Bereich.*

Stabile Ruhelagen
Um die stationären Punkte (x_s, y_s) des Systems zu berechnen, wird $\dot{x} = \dot{y} = 0$ gesetzt. Aus $\dot{y} = 0$ folgt

$$y = \frac{7}{8} - \frac{5}{4}x,$$

aus $\dot{x} = 0$ ergibt sich

$$\gamma = \frac{x^3}{3} - x - y = \frac{x^3}{3} + \frac{x}{4} - \frac{7}{8}.$$

$\gamma(x)$ ist monoton ($\mathrm{d}\gamma/\mathrm{d}x = x^2 + \frac{1}{4} > 0$), also entspricht bei dieser Gleichung jedem γ-Wert ein eindeutiger x-Wert, berechenbar etwa mit dem Newton-Verfahren. Damit ist implizit die stationäre Lösung in Abhängigkeit vom Generatorpotenzial γ gegeben,

$$x_s = x_s(\gamma), \quad y_s = y_s(\gamma).$$

Wir bleiben bei der bequemeren expliziten Darstellung

$$\gamma(x_s), \quad y_s = y_s(x_s).$$

Nachdem nun die Lage der stationären Punkte bestimmt ist, folgt die Diskussion des qualitativen Lösungsverhaltens in der Umgebung von (γ, x_s, y_s). Hierzu werden die Eigenwerte der Matrix der Linearisierung benötigt. Das an einer stationären Lösung linearisierte System lautet

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} \approx \begin{pmatrix} 3 - 3x_s^2 & 3 \\ -1/3 & -8/30 \end{pmatrix} \begin{pmatrix} x - x_s \\ y - y_s \end{pmatrix}.$$

Die Eigenwerte λ_1, λ_2 der Matrix

$$\begin{pmatrix} 3\alpha & 3 \\ -1/3 & -8/30 \end{pmatrix}, \quad \alpha := 1 - x_s^2$$

sind die Lösungen der quadratischen Gleichung

$$\begin{aligned} 0 &= (3\alpha - \lambda)\left(-\frac{8}{30} - \lambda\right) + 1 \\ &= \lambda^2 + \lambda\left(\frac{4}{15} - 3\alpha\right) + \left(1 - \frac{4}{5}\alpha\right). \end{aligned}$$

Man erhält

$$\begin{aligned} \lambda_{1,2} &= \frac{1}{2}\left(3\alpha - \frac{4}{15}\right) \pm \frac{1}{2}\sqrt{\Delta}, \quad \text{mit} \\ \Delta &:= 9\alpha^2 + \frac{8}{5}\alpha - \frac{884}{225}. \end{aligned}$$

Um den Typ der stationären Lösung zu bestimmen ($\lambda_{1,2}$ reell oder komplex?), muss die Vorzeichenverteilung der Diskriminante Δ untersucht werden. Die Nullstellen von Δ sind

$$\alpha_1 = \frac{26}{45}, \ \alpha_2 = -\frac{34}{45}.$$

$\Delta < 0$ bedeutet

$$\begin{aligned} -\frac{34}{45} < \alpha = 1 - x_s^2 &< \frac{26}{45}, \\ \frac{19}{45} < \quad x_s^2 \quad &< \frac{79}{45}. \end{aligned}$$

Demnach gibt es zwei Bereiche von x_s, für welche wegen $\Delta < 0$ die Eigenwerte $\lambda_{1,2}$ komplex sind. Diese beiden Bereiche sind gegeben durch

$$\sqrt{\frac{19}{45}} < |x_s| < \sqrt{\frac{79}{45}},$$

hier liegen also Strudel vor. Außerhalb dieser Bereiche sind die stationären Punkte Knoten. Entscheidend ist die Stabilität, sie ist durch das Vorzeichen des Realteils bestimmt. Die folgenden äquivalenten Umformungen kennzeichnen den Bereich der

Stabilität für Strudelpunkte:

$$\begin{aligned} 3\alpha - \frac{4}{15} &< 0 \\ \alpha = 1 - x_s^2 &< \frac{4}{45} \\ x_s^2 &> \frac{41}{45} \\ |x_s| &> \sqrt{\frac{41}{45}}. \end{aligned}$$

Die Zahlenwerte von Interesse[3] sind

$$\begin{aligned} \sqrt{19/45} &= 0{,}6498 \\ \sqrt{41/45} &= 0{,}9545 \\ \sqrt{79/45} &= 1{,}325. \end{aligned}$$

Zusammenfassend werden in der folgenden Aufstellung die 7 Bereiche von x_s mit unterschiedlichem qualitativen Verhalten aufgelistet:

$$\begin{aligned} |x_s| &< \sqrt{19/45} && \text{instabiler Knoten} \\ \sqrt{19/45} < |x_s| &< \sqrt{41/45} && \text{instabiler Strudel} \\ \sqrt{41/45} < |x_s| &< \sqrt{79/45} && \text{stabiler Strudel} \\ \sqrt{79/45} < |x_s| & && \text{stabiler Knoten} \end{aligned}$$

Nur für stabile stationäre Punkte ist das Membranpotenzial in Ruhelage. Die zur Stabilitätsgrenze $x_s^2 = 41/45$ gehörenden γ_0-Werte sind

$$\gamma_0 = x_s \left(\frac{x_s^2}{3} + \frac{1}{4} \right) - \frac{7}{8} = \begin{cases} -0{,}3465 \\ -1{,}4035 \end{cases} .$$

Diese beiden Werte von γ_0 sind die Schwellenwerte, an denen Stabilität der Ruhelage verloren geht oder gewonnen wird. Was tatsächlich passiert, werden wir unten sehen.

Für die Skizze der stationären Punkte in der (γ, x)-Ebene (Abb. 11.2) berücksichtigen wir, dass $\gamma(x)$ eine kubische Parabel mit Punktsymmetrie ist:

$$\begin{aligned} &\gamma + \frac{7}{8} = \frac{x^3}{3} + \frac{x}{4} =: \gamma^*(x) \quad \text{mit} \\ &\gamma^*(-x) = -\gamma^*(x), \end{aligned}$$

[3] wie immer, gerundet.

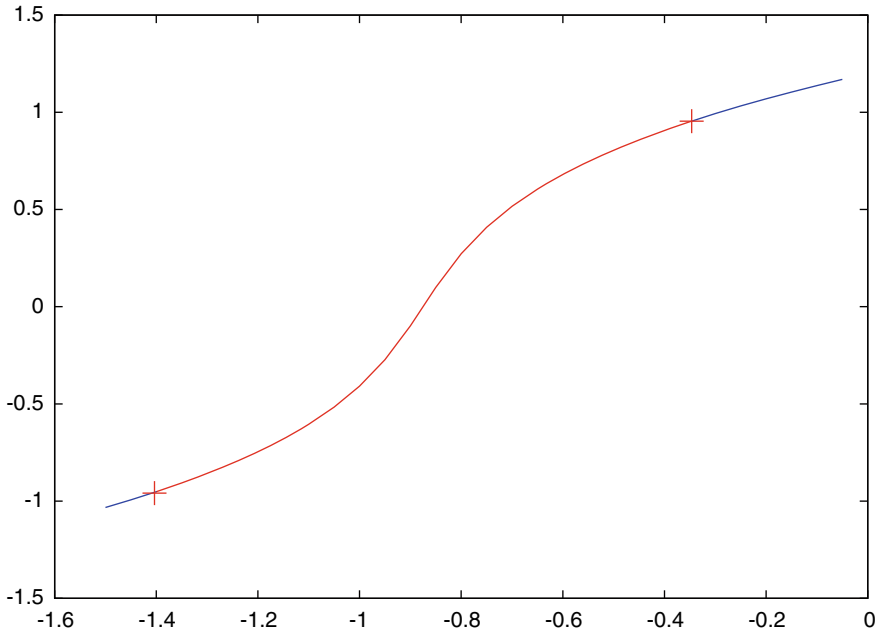

Abb. 11.2 Die Werte des stationären x_s sind hier auf der senkrechten Achse aufgetragen über der (waagerechten) γ-Achse; rot: instabile stationäre Lösungen, blau: stabile stationäre Lösungen; Kreuze: Hopf-Bifurkationspunkte. Aufgetragen sind die Resultate der Analyse; für die Anwendung des Modells ist allerdings nur die rechte Hälfte relevant, für $|\gamma|$ nicht zu groß

also liegt Punktsymmetrie zu $(\gamma,\ x) = (-7/8,\ 0)$ vor. In der Abb. 11.2 sind die beiden Schwellenwerte γ_0 mit Kreuzen gekennzeichnet. Bezug zur Anwendung hat der Wert $\gamma_0 = -0{,}3465$: Denn wenn kein Nervenreiz erfolgt, ist $\gamma = 0$. Aufkommender Nervenreiz bedeutet hier wachsendes negatives Potenzial, also ein γ-Wert, der von 0 in Richtung -0,3465 driftet, und darüber hinaus.[4]

Grenzzykel

Bisher haben wir die stationären Punkte und ihr Stabilitätsverhalten berechnet. Lösungen $x(t)$, $y(t)$ des nichtlinearen Differenzialgleichungssystems werden in den Bereichen stabiler Ruhelagen ($\gamma > -0{,}3465$, $\gamma < -1{,}4035$) gegen die stationären Lösungen konvergieren

$$x(t) \to x_s, \quad y(t) \to y_s,$$

für wachsende $t \to \infty$. Im instabilen Bereich streben Lösungen aus der Umgebung der stationären Punkte heraus. Im Folgenden soll der Frage nachgegangen werden, wohin diese Trajektorien streben. Das wird am Ende dieses Kapitels mit numerischer Simulation erforscht. Zunächst wollen wir nachweisen, dass Grenzzykel existieren.

[4] Auch wenn Werte $\gamma < -0{,}5$ für die Nervenreiz-Interpretation des Modells vermutlich ohne Belang sind, werden wir für die Analyse keine Einschränkung von γ vornehmen.

Wir schneiden aus der (x, y)-Ebene ein „großes“ Rechteck heraus (Abb. 11.3) und untersuchen, in welcher Richtung Bahnkurven $x(t)$, $y(t)$ den Rand überqueren, von außen nach innen, oder umgekehrt. Das Rechteck, in dessen Inneren der stationäre Punkt liegen muss, setzen wir an als

$$R := \{\, (x, y) \mid \ |x| \le A,\ |y| \le B \,\}.$$

Für zum Beispiel $A = 10$, $B = 20$ werden die Vorzeichen von $\dot{x}, \dot{y}$ an den vier Randstücken berechnet, jeweils für den Bereich

$$-1{,}4035 \le \gamma \le -0{,}3465$$

instabiler Ruhelagen:

$$\begin{aligned}
x = 10,\ -20 \le y \le 20: &\quad \dot{x} = 3(y + 10 - 1000/3 + \gamma) < 0\\
x = -10,\ -20 \le y \le 20: &\quad \dot{x} = 3(y - 10 + 1000/3 + \gamma) > 0\\
-10 \le x \le 10,\ y = 20: &\quad \dot{y} = -(x - 0{,}7 + 16)/3 < 0\\
-10 \le x \le 10,\ y = -20: &\quad \dot{y} = -(x - 0{,}7 - 16)/3 > 0.
\end{aligned}$$

Hier zeigt es sich, dass Trajektorien den Rand des gewählten Rechtecks nur von außen nach innen überqueren.[5] Demnach müssen die vom stationären Punkt wegstrebenden Trajektorien in der Nähe bleiben. Schematisch ist dieser Sachverhalt in der Abb. 11.3 illustriert. Da es im farbig getönten Bereich keine weitere stationäre Lösung gibt, werden die Trajektorien gegen einen stabilen Grenzzykel konvergieren. Dies bedeutet das Auftreten einer periodischen Lösung, ohne dass eine äußere Einwirkung vorliegt.

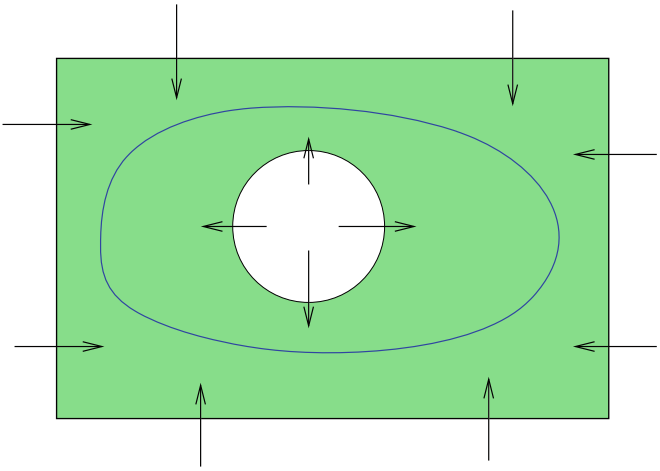

Abb. 11.3 in der (x, y)-Ebene: Grenzzykel (schematisch, in blau) zu (11.1)/(11.2). Die weiße Kreisfläche stellt eine kleine Umgebung um die instabile stationäre Lösung dar

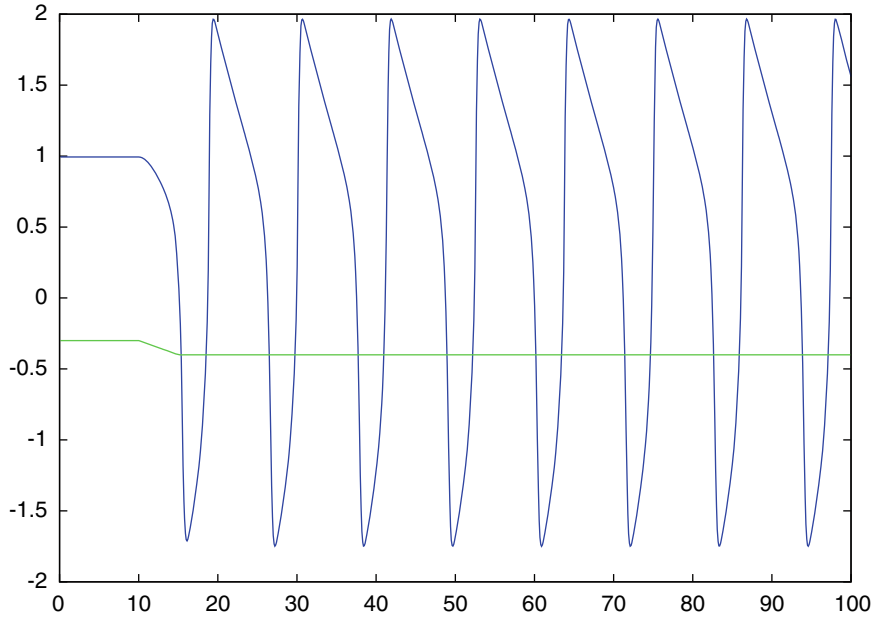

Abb. 11.4 Waagerechte Achse t, senkrechte Achse x: Membranpotenzial $x(t)$ (in blau). Angenommenes Szenario: Für $0 \leq t \leq 10$ bleibt x stationär auf seinem stabilen Ruhepotenzial-Wert für $\gamma = -0{,}3$. Ab $t = 10$ verändert sich das Generatorpotenzial γ (grün) von $-0{,}3$ zu $-0{,}4$, und überschreitet dabei den Schwellenwert $\gamma_0 = -0{,}3465$. Als Folge beginnt das Membranpotenzial plötzlich periodisch an zu „feuern“

Numerische Simulation

Nun wollen wir den Grenzzykel sehen. Mit einem numerischen Integrator[6] integrieren wir das System (11.1)/(11.2), ausgehend von $t = 0$ mit passenden Anfangswerten in der (x, y)-Phasenebene. Dabei simulieren wir die Realität insofern, als wir den Nervenreiz erhöhen. Für $0 \leq t \leq 10$ halten wir das Generatorpotenzial fest auf dem Wert $\gamma = -0{,}3$, für den nach unserer Analyse ein stabiler Ruhepunkt vorliegt, von dem wir die Integration starten. Die $(x(t), y(t))$-Antwort des Systems bestätigt die Analyse: x und y bleiben auf diesem Niveau (Abb. 11.4). Danach, für $10 \leq t \leq 15$, verstärken wir den Nervenreiz, das heißt, wir verstärken γ auf den Wert -0,4, also in den Bereich, in dem die stationäre Lösung instabil ist. Und ganz plötzlich setzt sprunghaft eine periodische Lösung ein! Die Abb. 11.4 zeigt $x(t)$.

[5] Das Rechteck kann viel kleiner gewählt werden, eine Aufgabe für interessierte Leser!
[6] beispielsweise einfach ein Runge-Verfahren.

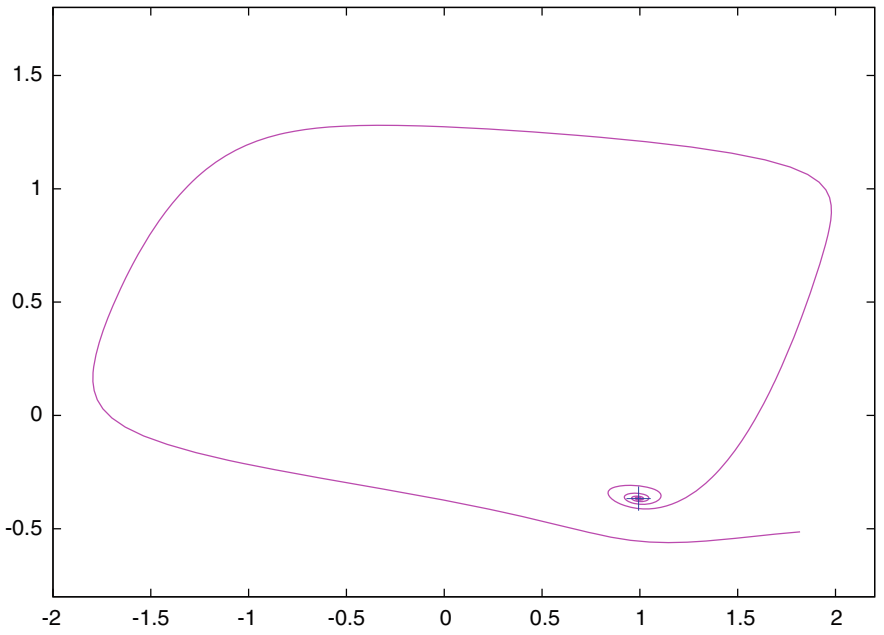

Abb. 11.5 (x, y)-Phasenebene für $\gamma = -0{,}3$: Trajektorien werden strudelförmig von der stabilen Ruhelage (blaues Kreuz) angezogen, eine Trajektorie ist gezeigt (in magenta) mit Startpunkt $(x, y) = (1{,}8,\ -0{,}5)$

Bistabilität

Jetzt fehlt noch eine Erklärung, warum bei obigem Experiment die Lösung $(x(t), y(t))$ so blitzartig die Struktur ändert, von Ruhe zu heftiger Schwankung des Membranpotenzials. Dazu sei die Lösungsstruktur untersucht, für konstanten Parameter γ im Bereich des Experiments von Abb. 11.4. Die Abbildungen 11.5 und 11.6 zeigen Trajektorien für die Randpunkte der Simulation, also für $\gamma = -0{,}3$ und $\gamma = -0{,}4$. Bei näherer Analyse stellt sich heraus, dass es in diesem γ-Intervall einen sehr schmalen Bereich gibt mit *Bistabilität.* Im Intervall

$$-0{,}34647 \leq \gamma \leq -0{,}33685$$

existieren gleichzeitig ein stabiler Grenzzykel, ein stabiler Ruhepunkt, und dazwischen eine instabile periodische Lösung. Das wird in Abb. 11.7 exemplarisch für den Parameterwert $\gamma = -0{,}337$ gezeigt. Im Zusammenhang mit dem Vorliegen der Bistabilität spricht man von einem *harten Stabilitätsverlust,* und bei dem Schwellenwert $\gamma = -0{,}34647$ von einem *Bifurkationspunkt,* speziell ein Hopf-Bifurkationspunkt. Bei dem harten Stabilitätsverlust muss sich der Grenzzykel nicht langsam aufbauen, sondern er existiert bereits in voller Größe. Deswegen kann die Trajektorie augenblicklich auf den Grenzzykel aufspringen, wie in der Abb. 11.4

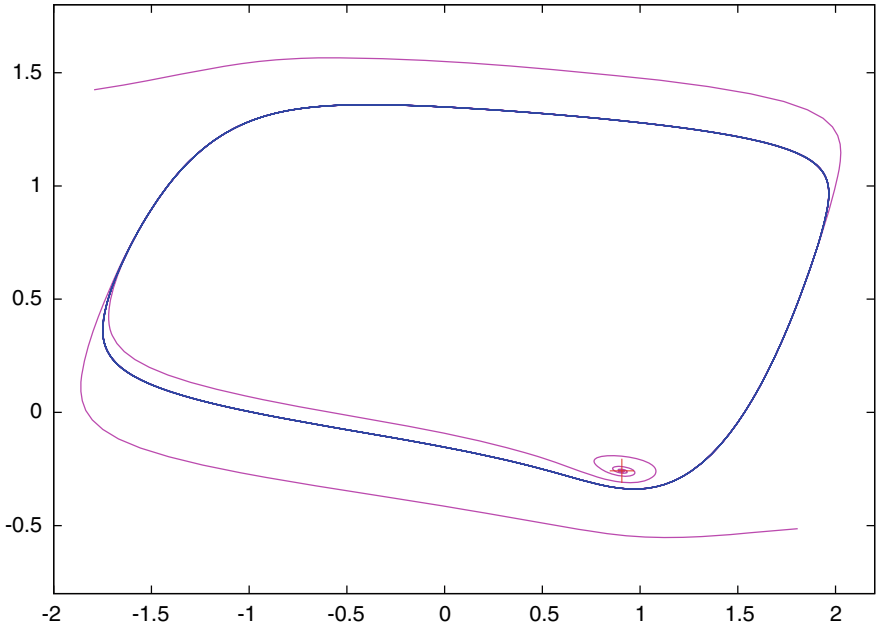

Abb. 11.6 (x, y)-Phasenebene für $\gamma = -0{,}4$: Ein stabiler Grenzzykel (blau), der instabile stationäre Punkt (rotes Kreuz), und drei Trajektorien (magenta), Umlaufsinn jeweils im Uhrzeigersinn. Der stabile Grenzzykel wirkt sehr attraktiv auf die Trajektorien

gezeigt. Das entspricht dem bei Nervenimpulsen beobachteten Verhalten. Ist der Nervenreiz kleiner als ein Schwellenwert (hier $\gamma > -0{,}34647$), dann passiert „nichts“. Bei Überschreiten des Schwellenwertes (hier $\gamma < -0{,}34647$) ist die periodische Oszillation des Membranpotenzials sofort da mit voller Amplitude. Solche Schwellenwerte[7] (Bifurkationspunkte) spielen bei vielen Anwendungen eine zentrale Rolle.

[7] In anderem Zusammenhang werden solche Umschlagpunkte auch *Kipppunkte* genannt.

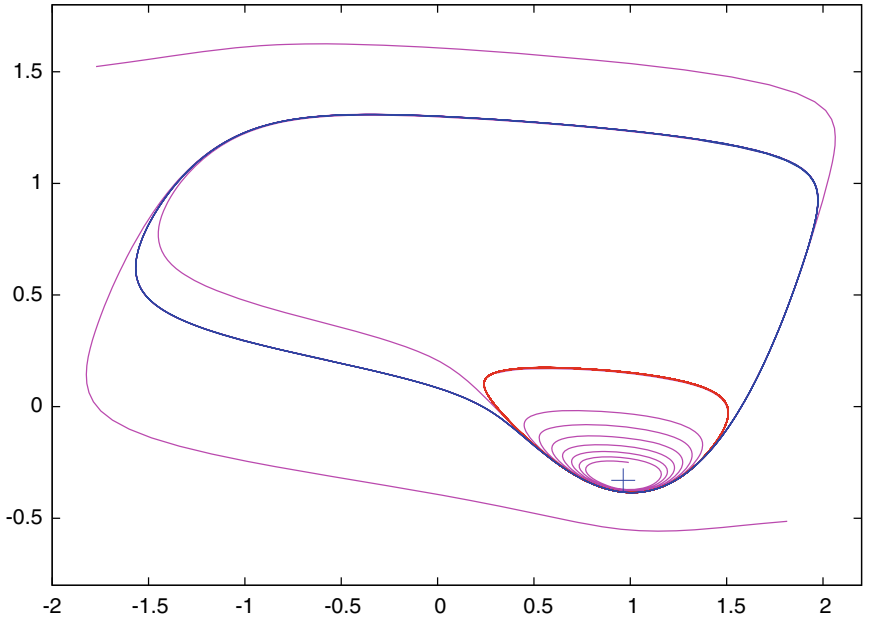

Abb. 11.7 (x, y)-Phasenebene für $\gamma = -0{,}337$. Eine instabile periodische Lösung (in rot) trennt die beiden Einzugsbereiche für Trajektorien, die gegen den Attraktor *Ruhelage* (blaues Kreuz) oder gegen den Attraktor *Grenzzykel* (blaue Kurve) streben. Drei Trajektorien sind eingezeichnet (in magenta), Bewegung jeweils im Uhrzeigersinn. Von der „innersten" Trajektorie, welche gegen die stabile Ruhelage strebt, ist für bessere Übersicht nur ein Stück eingetragen

Literatur

das ursprüngliche Nervenmodell:

Hodgkin, A.L., Huxley, A.F.: A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. **117**:500–544 (1952)

das hier verwendete vereinfachte Modell:

FitzHugh, R.: Impulses and physiological states in theoretical models of nerve membrane. Biophys. J. **1**:445–466 (1961)

zu Bifurkationen und Bistabilität, und zum Hodgkin&Huxley-Modell:

Seydel, R.: Practical Bifurcation and Stability Analysis. Third Edition. Interdisciplinary Applied Mathematics, Bd.5, Springer, New York (2010)

Populations-Dynamik

12

Die beiden Modelle zu Herzschlag und Nervenimpuls beschreiben dynamisches Verhalten aufgrund von Gesetzen der Biologie und damit der Naturwissenschaften. Modelle zur Dynamik von Populationen greifen eher auf Annahmen zu sozialem Verhalten zurück. Versucht man etwa, die Ausbreitung einer Epidemie zu modellieren, so werden Kenntnisse über die Vermehrung von Viren ergänzt durch Hypothesen zum Verhalten von Infizierten.

Dieses Kapitel beschreibt zunächst ein einfaches Modell einer Epidemie. Es möge als Anregung dienen, damit zu experimentieren, die definierenden Gleichungen geeignet zu modifizieren, Annahmen abzuschwächen und Daten anzupassen. Der zweite Teil dieses Kapitels thematisiert den Einfluss von Werbung auf Wanderungsbewegungen. Als Beispiel modellieren wir Migrationsverhalten von Studierenden zwischen verschiedenen Studienrichtungen. Erkenntnisse, die durch Simulation solcher Modelle gewonnen werden, sind typischerweise von qualitativer Art.

12.1 Ein einfaches Epidemie-Modell

Der Hintergrund des im Folgenden vorgestellten Modells ist eine Epidemie unter Menschen, verursacht durch einen Virus. Die Anzahl der Infizierten zum Zeitpunkt t sei als kontinuierliche Variable aufgefasst und mit $x(t)$ bezeichnet. Die Gesamtzahl der Population $\hat{x}$ wird als konstant angenommen. Also sollte für das Modell

$$0 \le x(t) \le \hat{x} \tag{12.1}$$

gelten. Der Ansteckungs-Mechanismus legt einen Zuwachs $\dot{x} = \frac{\mathrm{d}x}{\mathrm{d}t}$ proportional zur aktuellen Anzahl x nahe. Das bedeutet für die zeitliche Änderung von x ein exponentielles Wachstum, mit $\dot{x} = \tilde{\alpha} x$ für einen Parameter $\tilde{\alpha}$. Allerdings kann exponentielles

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_12

Wachstum wegen (12.1) nur lokal stattfinden, eine Sättigung ist zwingend. Eine Modifikation des exponentiellen Wachstums vermittelt die *logistische Gleichung*

$$\dot{x} = \alpha x(\hat{x} - x). \tag{12.2}$$

Der Parameter α repräsentiert die Ansteckungsrate. Die Dynamik von (12.2) ist die Grundlage für die folgenden Überlegungen.

Zunächst normieren wir x auf eine Einheitspopulation mit $\hat{x} = 1$. Um den griffigeren Prozentsatz zu verwenden, setze $p := x \cdot 100$. Diskretisieren wir die Zeit t mit Zeit-Abständen Δt, und ersetzen $\dot{x}$ durch einen Differenzenquotienten, so ergibt sich als erste Version eines groben diskreten Modells die Differenzengleichung

$$\frac{p(t + \Delta t) - p(t)}{\Delta t} = \frac{\alpha}{100}\, p(t)\,(100 - p(t)). \tag{12.3}$$

Hier repräsentiert $p(t)$ den prozentualen Anteil der Infizierten zum Zeitpunkt t. Um die tägliche Änderung zu betrachten, setze $\Delta t = 1$. Dann würden nach (12.3) täglich

$$\tilde{a}_1(t) := \frac{\alpha}{100}\, p(t)\,(100 - p(t))$$

Prozent der Individuen angesteckt.

Dieser Ansatz ist noch zu grob. Denn die Infizierten gesunden wieder und sind dann vielleicht immun. Wir nehmen als optimistisches Szenario an, dass alle Infizierten wieder gesund werden, sagen wir nach 30 Tagen. Und die Genesenen seien dann immun. Nach dieser optimistischen Annahme fallen $a_2(t) := \tilde{a}_1(t - 30)$ viele Personen zum Zeitpunkt t wieder aus der Krankenbilanz heraus. Um die Immunisierung zu beschreiben, sei noch die Variable $g(t)$ eingeführt, der Prozentsatz von Gesunden, die noch infiziert werden können. Es gilt $g(0) = 100 - p(0)$, dabei steht $t = 0$ für „heute“. Da die Infizierten und die Genesenen aus der Dynamik wieder herausfallen, verändert sich $\tilde{a}_1$ zur Anzahl a_1 der tatsächlich Infizierten: Vor jedem Iterationsschritt $t \to t + \Delta t$ ist der Anteil der Infizierbaren

$$g(t) := g(t - 1) - a_1(t - 1) \tag{12.4}$$

für $t > 0$. Mit diesem g ist die Ansteckung

$$a_1(t) := \frac{\alpha}{100}\, p(t)\,(g(t) - p(t)) \tag{12.5}$$

Prozent. Alles zusammen ergibt sich die Differenzengleichung

$$p(t + 1) = p(t) + \frac{\alpha}{100}\, p(t)\,(g(t) - p(t)) \; - a_1(t - 30). \tag{12.6}$$

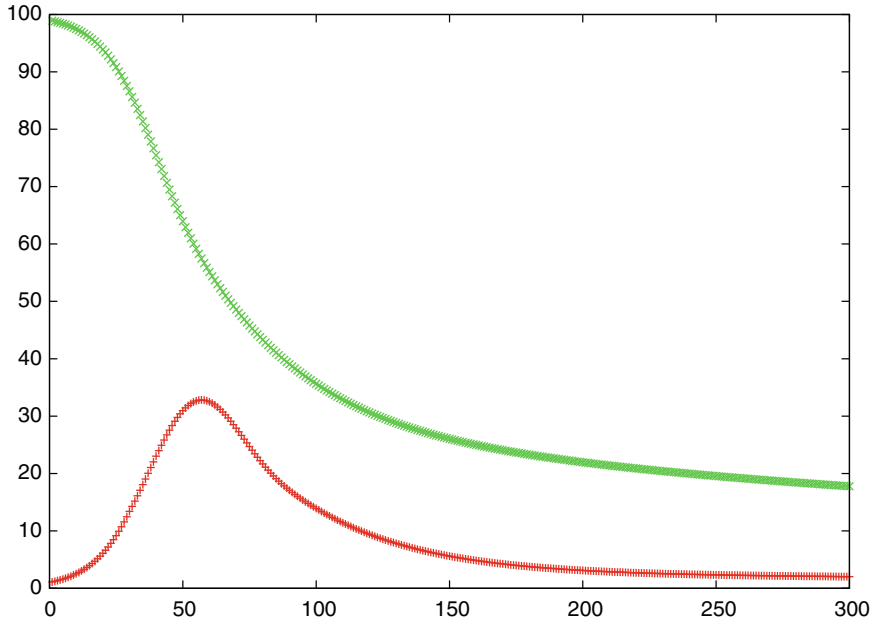

Abb. 12.1 Die „Kurven“ sind jeweils aus 300 Kreuzen zusammengesetzt, für 300 Tage. Für die Rate $\alpha = 0{,}1$ im Modell (12.6) zeigt die rote Kurve den Prozentsatz $p(t)$ der Infizierten, und die grüne Kurve den Prozentsatz $g(t)$ der Infizierbaren, $t = 1, 2, \ldots, 300$; Anfangsrate der Infizierten: 1 %

Eine Simulation[1] des Modells (12.4)/(12.5)/(12.6) benötigt einen repräsentativen Wert für die Rate α. Konkrete Fallzahlen liegen zur Covid-19-Epidemie vor. Zum Beispiel gab es an zwei aufeinanderfolgen Tagen $x_0 = 28049$ und $x_1 = 31195$ Infizierte. Aus diesen Daten lässt sich ein Schätzwert $\tilde{\alpha}$ wie folgt gewinnen: Für die lokale Annahme

$$x(t) = x_0 \exp(\tilde{\alpha} t)$$

ergibt sich mit

$$\tilde{\alpha} = \frac{1}{t} \log \frac{x(t)}{x_0}$$

für $t = 1$ der Wert $\tilde{\alpha} = 0{,}1063$. Das legt für einen numerischen Test den Wert $\alpha = 0{,}1$ als repräsentativ nahe. Die Dynamik der Abb. 12.1 mit $\alpha = 0{,}1$ zeigt in der ersten Phase eine exponentielle Zunahme der Anzahl der Infizierten, und nach zunehmender Immunisierung ein Abnehmen; das sieht qualitativ plausibel aus. – Das obige einfache Modell hat Potenzial und lädt zu Verfeinerungen und zum Experimentieren ein.

[1] „Simulation“ bedeutet eine numerische Auswertung der Modellgleichung mit dem Ziel, Erkenntnisse zu gewinnen.

12.2 Modell einer Verteilung von Studierenden

Die Dynamik von Populationen spielt eine immense Rolle auch in gesellschaftlichen Strömungen. Als Beispiel betrachten wir Migration an einer Universität: Hinsichtlich der Wahl von Studienfächern, konzentrieren sich Studierende häufig auf beliebtere Themen und lassen andere Studienrichtungen „links“ liegen. Die Gründe können vielfältig sein, wie Anwendungsnähe des Fachs, sympathische Dozenten, Meinung Anderer oder Höhe der Ansprüche. Entscheidungen werden beeinflusst durch Kommunikationsverhalten, Marktmechanismen wie Werbung, erkannte Risiken und Psychologie. Im Folgenden diskutieren wir ein entsprechendes Modell.

Das Modell
Ein Studienfach biete zwei Optionen der Spezialisierung an. Die Anzahl der Studierenden von Option 1 zum Zeitpunkt t sei $X_1(t)$, und derjenigen von Option 2 sei mit $X_2(t)$ bezeichnet. Die Studierenden der jeweiligen Gruppe stehen im Austausch und kommunizieren die jeweiligen Vor- und Nachteile der beiden Optionen. Im Laufe der Zeit können Wanderungsbewegungen stattfinden, etwa von Option 1 zu Option 2. Die Veränderung von $X_i(t)$ $(i = 1, 2)$ in einem Zeitintervall Δt sei mit ΔX_i bezeichnet. Die Migration von der X_1-Gruppe zur X_2-Gruppe von Individuen ist cX_1, wobei c die Wirkung der Kommunikation oder Werbung widerspiegelt, also $c = a(t)X_2(t)$ für ein geeignetes a. Damit ist der Übergang von X_1 zu X_2 durch einen nichtlinearen Term $a(t)X_1(t)X_2(t)$ beschrieben. Der Faktor $a(t)$ misst die Kontaktrate und den Effekt von Kommunikation und Werbung.

Durch neu hinzukommende Studenten und durch erfolgreiche Abschlussprüfungen ändert sich die Anzahl der Individuen. Die Anzahl der im Zeitintervall Δt Hinzukommenden seien mit λ bezeichnet, als konstant angenommen, und die Raten der erfolgreichen Abschlüsse seien d_1 und d_2. Zusätzlich verabreden wir, dass Neuzugänge nur die Option 1 wählen können, und die Wahlmöglichkeit von Option 2 (etwa ein Vertiefungsfach) erst zu einem späteren Zeitpunkt möglich ist.[2] Natürlich können Studierende von Option 2 enttäuscht sein und zu Option 1 zurückkehren. Die Rückflussrate bezeichnen wir mit β. Damit ist das Modell durch die beiden folgenden Gleichungen beschrieben:

$$\begin{aligned} \Delta X_1 &= X_1(t_{j+1}) - X_1(t_j) = \lambda - a(t_j)X_1(t_j)X_2(t_j) + \beta X_2(t_j) - d_1 X_1(t_j) \\ \Delta X_2 &= X_2(t_{j+1}) - X_2(t_j) = a(t_j)X_1(t_j)X_2(t_j) - \beta X_2(t_j) - d_2 X_2(t_j). \end{aligned} \tag{12.7}$$

Zugänge und Abgänge erfolgen zu diskreten Zeitpunkten t_j, $j = 0, 1, 2, \ldots$, etwa semesterweise. So kann t_j das Ende des j-ten Semesters und den Anfang des $(j+1)$-ten Semesters meinen, und Δt die Dauer eines Semesters. Die Iteration zu (12.7)

[2] Motivation: X_1 könnte zum Beispiel die Menge aller Mathematik-Studenten an einer Universität repräsentieren, vor und nach dem Bachelor, und X_2 die Anzahl derjenigen, die nach dem Bachelor in ein anderes Fach wechseln.

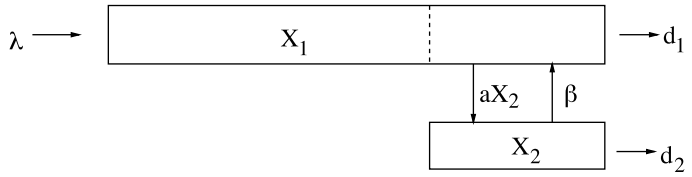

Abb. 12.2 Schematischer Aufbau des Modells (12.7)

ist

$$X_i(t_{j+1}) = X_i(t_j) + \Delta X_i(t_j), \quad \text{für } i = 1, 2 \text{ und } j = 0, 1, \ldots .$$

Für X_1, X_2 sind einige Restriktionen zu berücksichtigen. So ist natürlich

$$X_1(t) \geq 0, \;\; X_2(t) \geq 0 \tag{12.8}$$

zu fordern. Zum Studium von Option 2 seien nur Bachelor-Absolventen zugelassen. Wenn $\gamma X_1(t)$ den Anteil der Studierenden bezeichnet, die zu Option 2 wechseln dürfen, dann haben wir für die Migration zu X_2 die weitere Beschränkung

$$a(t)X_1(t)X_2(t) \leq \gamma X_1(t). \tag{12.9}$$

Die Abb. 12.2 fasst das Modell symbolisch zusammen.

Technische Annahmen
Zur einfacheren Beurteilung von Konvergenz seien kontinuierliche Anzahlen angenommen, $X_i \in \mathbb{R}$. Über $a(t)$ haben wir noch nicht verfügt. Es gelte $a(t_j) \geq 0$, und für einen maximalen Wert $a_{\max}$ gelte $a(t_j) \leq a_{\max}$. Wir setzen konstant $a(t_j) = a_{\max}$ wenn die Schranken (12.8) und (12.9) nicht aktiv sind. Um für numerische Experimente realistische Zahlen zu erhalten, gehen wir von 10 Semestern als mittlere Studiendauer aus ($d_1 = 0{,}1$), und nehmen ungefähr drei Semester an für einen erfolgreichen Abschluss der Option 2 ($d_1 = 0{,}3$). Der Anteil der graduierten Studenten sei 40 %; die Konstante β wird willkürlich gesetzt. Die gewählten Parameter sind

$$d_1 = 0{,}1, \;\; d_2 = 0{,}3, \;\; \gamma = 0{,}4, \;\; \beta = 0{,}1. \tag{12.10}$$

Diese willkürlich gesetzten Zahlen dienen nur zur Illustration. Frei sind noch die Parameter λ und $a_{\max}$.

Einige Experimente
Bezeichnung auch kürzer: $y_1^j := X_1(t_j)$, $y_2^j := X_2(t_j)$. Als Startwert wählen wir $X_1(t_0) = 200$, $X_2(t_0) = 20$, und iterieren (12.7) über 40 Semester, für die Wahl $\lambda = 50$, $a = 0{,}002$. Die zeitliche Abfolge $X_1(t)$ und $X_2(t)$ aus dieser Simulation ist in Abb. 12.3 wiedergegeben; Abb. 12.4 zeigt das zugehörige (X_1, X_2)-Phasendiagramm. Man erkennt, dass sich die Anzahlen der Studierenden der Option 1 und der Option 2 zunächst stark ändern, und dann in eine stationäre Situation einschwenken

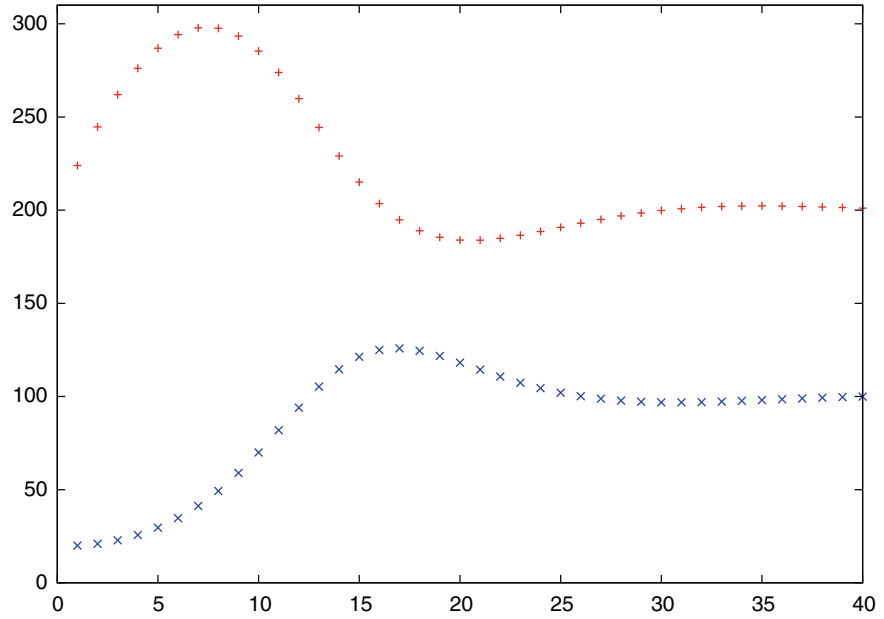

Abb. 12.3 Iteration von (12.7), $y_1(t_j)$ (X_1, rot), $y_2(t_j)$ (X_2, blau), jeweils über 40 Semester, $j = 1, \ldots, 40$, für $\lambda = 50$, $a = 0{,}002$. Die Werte konvergieren gegen stationäre Lagen

mit konstanten X_1, X_2. Für den Wert $a = 0{,}002$ sind die Schranken (12.8) und (12.9) nicht aktiv. Wählt man größere Werte von a, wird die Schranke (12.9) wichtig.

Fixpunkte

Interessant ist das Verhalten der stationären Lagen für verschiedene Werte der Parameter a und λ. Für die durch (12.7) vermittelte Abbildung sind die stationären Werte *Fixpunkte*. Diese lassen sich durch einfache Handrechnung ermitteln: Fixpunkte ergeben sich, wenn die Werte ΔX in (12.7) verschwinden. Hier gibt es zwei Fixpunkte. Ein Fixpunkt ist

$$X_1 = \frac{\lambda}{d_1}, \quad X_2 = 0; \tag{12.11}$$

der andere für $X_2 \neq 0$ ist

$$X_1 = \frac{\beta + d_2}{a}, \quad X_2 = \frac{1}{d_2}\left(\lambda - \frac{d_1}{a}(\beta + d_2)\right). \tag{12.12}$$

Beide Fixpunkte sind Funktionen des Parameters λ. Offensichtlich gilt für den speziellen Parameterwert

$$\lambda_0 := \frac{d_1}{a}(\beta + d_2),$$

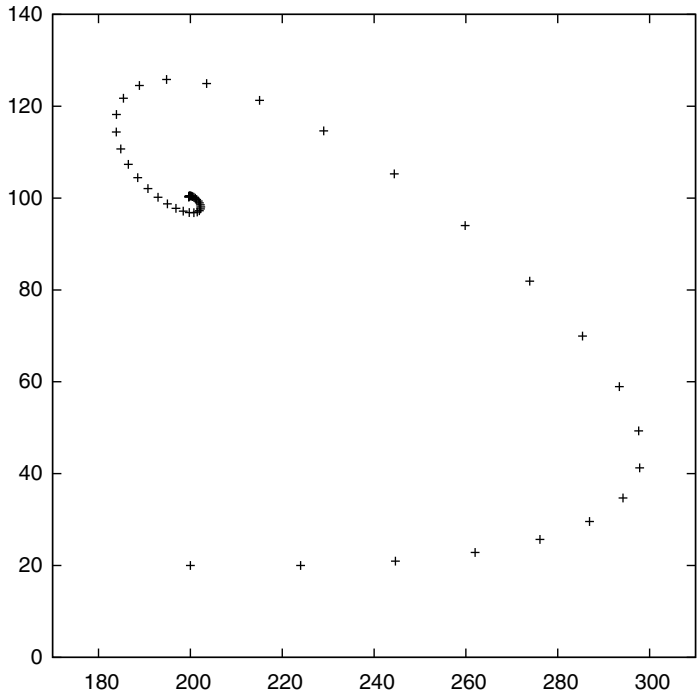

Abb. 12.4 Iteration von (12.7), Phasen-Diagramm der (y_1^j, y_2^j)-Werte für $j = 0, 1, \ldots, 40$ und $\lambda = 50$, $a = 0{,}002$. Start bei $(y_1^0, y_2^0) = (200,\ 20)$

dass beide Fixpunkte identisch sind.[3]

Eine Stabilitätsanalyse basiert auf der Jacobi-Matrix zur Abbildung (12.7)

$$J = \begin{pmatrix} -aX_2 - d_1 & -aX_1 + \beta \\ aX_2 & aX_1 - \beta - d_2 \end{pmatrix}.$$

Den ersten Fixpunkt eingesetzt, ergibt

$$J = \begin{pmatrix} -d_1 & \beta - \lambda \frac{a}{d_1} \\ 0 & \frac{a}{d_1}(\lambda - \lambda_0) \end{pmatrix}.$$

Die Diagonalelemente einer Dreiecksmatrix sind ihre Eigenwerte. Hier sind die Eigenwerte für $\lambda < \lambda_0$ negativ. Dies zeigt, dass der Fixpunkt (12.11) mit $X_2 = 0$ stabil ist für $\lambda < \lambda_0$. Umgekehrt ist der Fixpunkt (12.12) stabil für $\lambda > \lambda_0$.

Diskussion

Zur Diskussion sei auf die Bedeutung der Variablen hingewiesen: $X_2 = 0$ bedeutet,

[3] eine Bifurkation.

dass es keine Studierenden gibt für Option 2. Um diese Studienrichtung zu erhalten, muss die Zahl der Studienanfänger λ größer sein als der Schwellenwert λ_0. Damit der Schwellenwert nicht zu groß wird und

$$\lambda > \frac{d_1}{a}(\beta + d_2) \tag{12.13}$$

bleiben kann, bieten sich (außer der Erhöhung von λ) die folgenden Maßnahmen an:

- die Kommunikation verbessern (a vergrößern), oder
- die Bedingungen der Option 2 verbessern (damit β kleiner wird), oder
- weniger Abschlüsse (d verringern).

Anschaulich sind solche Maßnahmen naheliegend, aber hier folgen diese qualitativen Schlüsse aus der Stabilitätsanalyse und dem Bemühen, λ_0 klein zu halten. Bemerkenswert ist die Existenz eines Schwellenwertes λ_0.

Von Interesse ist noch die Frage, ob beliebig hohe Ausgaben für Werbung etwas nützen, wenn also a sehr groß ist. In diesem Fall ist die Beschränkung (12.9) aktiv, und $aX_1X_2 = \gamma X_1$ kann in die Gl. (12.7) eingesetzt werden. Direkt folgt die „Randlösung“ $X_2 = \frac{\gamma}{a}$. Durch weitere Umformungen erhält man eine Beziehung zwischen a und λ, sodass der „freie“ Zweig (12.12) die Randlösung $X_2 = \frac{\gamma}{a}$ trifft. Das Kriterium ist

$$a\lambda = \beta d_1 + \gamma d_2 + d_1 d_2.$$

Im Hinblick auf die Forderung (12.13) beschränkt dies den freien Zweig (12.12) auf das Intervall

$$d_1(\beta + d_2) < a\lambda < d_1(\beta + d_2) + \gamma d_2.$$

Folglich nützt es hier nichts, die Ausgaben für Werbung a beliebig zu steigern. Für die gewählten Parameterwerte (12.10) und $a = 0{,}002$ folgt aus diesen Ungleichungen das Intervall $20 < \lambda < 80$, der Schwellenwert ist 20. Eine Simulation mit beispielsweise $\lambda = 15$ Studienanfängern bestätigt die Analyse: Der Studiengang der Option 2 stirbt aus, X_2 konvergiert gegen 0.

Literatur

Das zweite Modell wurde motiviert durch

Feichtinger, G.: Limit Cycles in Dynamic Economic Systems. Annals of Operations Research **37**:313–344 (1992)

Schwingungsverhalten eines Oszillators

13

Elektrische Schwingkreise können mit Hilfe geeigneter Rückkopplungs-Schaltungen entdämpft werden. Auf diese Weise lassen sich periodische Spannungsschwankungen konstanter Amplitude erzeugen. Ein Beispiel einer solchen elektrischen Schaltung zeigt die Abb. 13.1. Die Schwingung im LCR-Kreis steuert mit der Rückkopplung über die Induktivität L selbst die Energiezufuhr, und es kann eine selbsterregte Schwingung entstehen.

Ähnlich wie bei den Nervenimpulsen können diese Wechselspannungen durch eine nichtlineare Differenzialgleichung zweiter Ordnung beschrieben werden. Während wir in der vorigen Fallstudie die zugehörigen periodischen Lösungen numerisch berechnet haben, wollen wir hier mit einer analytischen Methode eine Näherung für die Wechselspannung konstruieren. – Die folgende Aufgabe kann bereits bei Kenntnis der Bernoullischen Differenzialgleichung (bzw. Trennung der Variablen) gelöst werden; die zugrunde liegende Näherungsmethode wird hinterher erläutert.

Aufgabe 1

Ein LCR-Schwingkreis ist an ein Rückkopplungsnetzwerk mit Verstärker angeschlossen. Die Spannung U dieses selbsterregten Oszillators genüge der Differenzialgleichung

$$LC\ddot{U} + RC\dot{U} + U = \dot{U}CS\exp(-U^2), \tag{13.1}$$

dabei ist der Parameter S die Steilheit. Man berechne die Oszillatorkreisfrequenz $\omega = 2\pi\nu$ und untersuche das qualitative Verhalten des Oszillators wie folgt:

a) *Man setze*

$$U(t) = A(t) \cdot \cos\omega t \tag{13.2}$$

und berechne $\dot{U}$ bzw. $\ddot{U}$. Für den technisch interessanten Fall ist es gestattet, bei $\dot{U}$ das Glied mit $\dot{A}$ und bei $\ddot{U}$ das Glied mit $\ddot{A}$ zu vernachlässigen. Mit

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_13

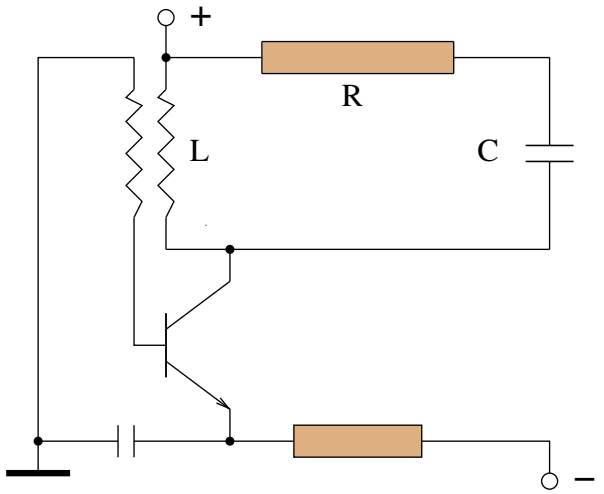

Abb. 13.1 Schaltung eines LCR-Schwingkreises mit Ohmschem Widerstand R, Induktivität L und Kapazität C, und durch Rückkopplung über einen Transistor gesteuerte Energiezufuhr

diesen Ausdrücken für $U, \dot{U}, \ddot{U}$ *gehe man in (13.1) ein (hierbei ist es zulässig,* $\exp(-U^2)$ *durch die Näherung* $1 - \frac{A^2}{4}$ *zu ersetzen) und zeige, dass wegen der linearen Unabhängigkeit von* $\sin \omega t$ *und* $\cos \omega t$

$$\omega = \frac{1}{\sqrt{LC}}$$

sein muss, und dass $A(t)$ *der folgenden Differenzialgleichung genügt:*

$$\dot{A} = \frac{S - R}{2L} \cdot A - \frac{S}{8L} \cdot A^3. \tag{13.3}$$

b) *Man bestimme die Lösung* $A(t)$ *mit* $A(0) = A_0 > 0$.

c) *Man bestimme* $\lim_{t \to \infty} A(t)$ *und unterscheide dabei die Fälle* $S < R, S = R, S > R$. *Wie groß muss* S *gewählt werden, damit der Oszillator stabil schwingt?*

d) *Es sei* $S = 100, L = 4, C = 10^{-4}$. *Man skizziere* $U(t)$ für $0 \leq t \leq \pi$ *in den drei Fällen*

$$(\alpha) \quad R = 96, \ A_0 = 2/3$$
$$(\beta) \quad R = 96, \ A_0 = 1/4$$
$$(\gamma) \quad R = 104, \ A_0 = 1/2$$

Methode der langsam veränderlichen Amplitude

Der Lösungsvorschlag zu dieser Aufgabe beruht auf der „Methode der langsam veränderlichen Amplitude“ nach Van der Pol. Bei der autonomen Differenzialgleichung (13.1) genügt es, die Spannung $U(t)$ in (13.2) als Produkt aus Amplitude A und

Oszillation $\cos \omega t$ anzusetzen.[1] A oszilliert im Vergleich zu $\cos \omega t$ kaum, also sind $\dot{A}$ und $\ddot{A}$ klein in Relation zu A und ω. Konkret sei Folgendes angenommen:

$$|\dot{A}| \ll |A|\omega \quad \text{und} \quad |\ddot{A}| \ll |A|\omega^2.$$

Daher können die Ableitungen

$$\begin{aligned} \dot{U} &= \dot{A}\cos \omega t - \omega A \sin \omega t \\ \ddot{U} &= \ddot{A}\cos \omega t - 2\omega \dot{A} \sin \omega t - \omega^2 A \cos \omega t \end{aligned}$$

vereinfacht werden zu

$$\begin{aligned} \dot{U} &\approx -\omega A \sin \omega t \\ \ddot{U} &\approx -2\omega \dot{A} \sin \omega t - \omega^2 A \cos \omega t, \end{aligned} \tag{13.4}$$

ohne dass eine wesentliche Ungenauigkeit eingeschleppt wird. Ist man nur an dem asymptotischen Verhalten des Oszillators interessiert, so kann man hier bereits $\dot{A} = 0$ setzen, und damit die folgende Rechnung verkürzen. Da wir am Einschwingverhalten interessiert sind, verwenden wir die vereinfachten Ausdrücke aus (13.4) für $\dot{U}$ und $\ddot{U}$ mit $\dot{A} \neq 0$ und setzen sie zusammen mit (13.2) in die Differenzialgleichung ein. Es ergibt sich

$$\begin{aligned} &- LC2\omega \dot{A} \sin \omega t - LC\omega^2 A \cos \omega t - RC\omega A \sin \omega t + A \cos \omega t \\ &\quad = -SC\omega A \sin \omega t \cdot \exp(-A^2 \cos^2 \omega t). \end{aligned} \tag{13.5}$$

Diese Gleichung ist von der Form

$$\alpha(t) \sin \omega t + \beta(t) \cos \omega t = f(\sin \omega t, \cos \omega t)$$

mit einer nichtlinearen Funktion f auf der rechten Seite.

Die Konstruktion einer Näherung geschieht in der folgenden Weise: Die rechte Seite f wird in eine Fourier-Reihe entwickelt,

$$f(t) = a_1 \sin \omega t + a_2 \sin 2\omega t + a_3 \sin 3\omega t + \ldots$$

(cos-Terme entfallen, da f ungerade ist.) Nach Vernachlässigung der Oberschwingungen wird f ersetzt durch $\tilde{f}$,

$$f(t) \approx \tilde{f}(t) := a_1 \sin \omega t.$$

Durch anschließenden Vergleich der Sinus-Terme wird eine Differenzialgleichung gewonnen, welche die Amplitude näherungsweise beschreibt.

[1] Allgemein $U(t) = A(t)\cos \omega t + B(t) \sin \omega t$. Wegen der fehlenden expliziten Zeitabhängigkeit in (13.1) braucht die Phase und damit der Sinus-Term nicht berücksichtigt zu werden.

Nun die Ausführung dieser Näherung: Mit Hilfe der Potenzreihe der Exponentialfunktion erhält man

$$f(t) = -SC\omega A \sin\omega t \cdot (1 - A^2 \cos^2 \omega t + \frac{1}{2} A^4 \cos^4 \omega t \mp \ldots).$$

Bei kleinen Amplituden A nehmen die Terme der Reihe rasch ab. Für den zweiten Term dieser Reihe formen wir um:

$$\begin{aligned} \sin\omega t \, \cos^2 \omega t &= \sin\omega t - \sin^3 \omega t \\ &= \sin\omega t - \frac{1}{4}(3 \sin\omega t - \sin 3\omega t) \\ &= \frac{1}{4} \sin\omega t + \frac{1}{4} \sin 3\omega t. \end{aligned}$$

Damit lautet die Reihe für f

$$f(t) = -SC\omega A \sin\omega t + SC\omega A^3 \frac{1}{4} \sin\omega t + SC\omega A^3 \frac{1}{4} \sin 3\omega t + A^5 \cdot (\ldots).$$

Für $A^2 \ll 1$ leistet der letzte (A^5)-Term keinen wesentlichen Beitrag, man kann ihn ohne weiteres Kopfzerbrechen vernachlässigen. Gravierender könnte sich auswirken, dass wir auch den Oberschwingungsterm $A^3 \sin 3\omega t$ streichen. Es wird später zu prüfen sein, wie sich die bisher getroffenen mutigen Annahmen und Vernachlässigungen auswirken! Zunächst setzen wir also für die rechte Seite der Gl. (13.5) den Ausdruck

$$\tilde{f}(t) = -SC\omega A \sin\omega t + SC\omega A^3 \frac{1}{4} \sin\omega t.$$

Die so aus (13.5) durch Vereinfachungen erhaltene Beziehung ist von der Form

$$a(t) \sin\omega t + b(t) \cos\omega t = 0 \tag{13.6}$$

mit

$$\begin{aligned} a(t) &= -LC2\omega\dot{A} - RC\omega A + SC\omega A - SC\omega \frac{1}{4} A^3 \\ b(t) &= -LC\omega^2 A + A. \end{aligned}$$

Da die Gl. (13.6) für alle Zeiten t gilt, muss für die niederfrequenten Faktoren a und b

$$a \equiv b \equiv 0$$

gelten. Aus $a = 0$ gewinnt man die Differenzialgleichung (13.3) für die Amplitude

$$\dot{A} = \frac{S - R}{2L} A - \frac{S}{8L} A^3,$$

aus $b = 0$ die Frequenz des Schwingkreises

$$\omega = \frac{1}{\sqrt{LC}}.$$

Lösung der Differenzialgleichung
Die Differenzialgleichung 1. Ordnung (13.3) für A ist eine Bernoulli-Differenzialgleichung, hier hilft (Trennung der Variablen oder) die Substitution

$$A = z^{-1/2}$$

weiter. Differenzieren dieser Relation nach t und Einsetzen liefert eine lineare Differenzialgleichung für z,

$$\dot{z} + \frac{S-R}{L} z = \frac{S}{4L}. \tag{13.7}$$

Die Lösung der homogenen Differenzialgleichung von (13.7) ist

$$\eta(t) = D \exp\left(-\frac{S-R}{L} t\right)$$

mit einer Konstanten D. Variation der Konstanten mit dem Ansatz

$$z(t) = v(t) \exp\left(-\frac{S-R}{L} t\right)$$

ergibt nach Differenzieren und Einsetzen (für $S \neq R$)

$$v(t) = \frac{S}{4(S-R)} \exp\left(\frac{S-R}{L} t\right) + D,$$

und als Lösung der Hilfs-Differenzialgleichung (13.7)

$$z(t) = \frac{S}{4(S-R)} + D \exp\left(-\frac{S-R}{L} t\right).$$

Nach Rücksubstitution erhält man für die Amplitude

$$A(t) = \left[\frac{S}{4(S-R)} + D \exp\left(-\frac{S-R}{L} t\right)\right]^{-1/2}.$$

Die Integrationskonstante D hängt ab vom Anfangswert zur Zeit $t = 0$

$$A_0 = A(0) = \left(\frac{S}{4(S-R)} + D\right)^{-1/2},$$

es gilt also

$$D = \frac{1}{A_0^2} - \frac{S}{4(S-R)}.$$

Damit ist für den Fall $S \neq R$ die Lösung des Anfangswertproblems berechnet. Es bleibt der Fall $S = R$ zu untersuchen, hier lautet die Differenzialgleichung

$$\dot{A} = -\frac{S}{8L} A^3.$$

Trennung der Variablen und Integration ergibt

$$\frac{1}{A^2} = \frac{S}{4L} t + D\,, \quad \text{mit } D = \frac{1}{A_0^2},$$

also

$$A(t) = \left(\frac{S}{4L} t + \frac{1}{A_0^2} \right)^{-1/2}.$$

Hiermit ist die Differenzialgleichung gelöst. Wir fassen die erhaltene Näherung für U zusammen:

$$\tilde{U}(t) := A(t) \cos\left(\frac{1}{\sqrt{LC}} t \right) \tag{13.8}$$

mit

$$A(t) = \left[\frac{S}{4L} t + \frac{1}{A_0^2} \right]^{-1/2} \text{ für } S = R,$$

$$A(t) = \left[\frac{S}{4(S-R)} + \left(\frac{1}{A_0^2} - \frac{S}{4(S-R)} \right) \exp\left(-\frac{S-R}{L} t \right) \right]^{-1/2} \text{ für } S \neq R.$$

Statt $\tilde{U}$ schreiben wir vereinfachend U, auch wenn die Funktion in (13.8) auf Grund der getroffenen Vereinfachungen nicht die exakte Lösung der Differenzialgleichung (13.1) ist. Die Abweichung zur Näherung wird unten untersucht.

Asymptotisches Verhalten

Diese Näherung $U(t)$ von (13.8) charakterisiert einen Oszillator mit Kapazität C, Induktivität L, Widerstand R und Steilheit S. Kapazität und Induktivität wirken sich nur auf die Frequenz aus, während R und S das asymptotische Verhalten des Oszillators bestimmen. Wenn, nach Ablauf des transienten Einschwingvorgangs, die Amplitude bei einem konstanten Wert $\neq 0$ verharrt, dann schwingt der Sender stabil. Zu untersuchen ist also der Grenzwert

$$\lim_{t \to \infty} A(t)$$

Tab. 13.1 Werte der Amplitude $A(t)$ für verschiedene Werte von t in den Fällen (α), (β), (γ) der Aufgabe 1

	0	$\pi/4$	$\pi/2$	π
Fall (α)	0,67	0,48	0,43	0,406
Fall (β)	0,25	0,31	0,35	0,387
Fall (γ)	0,50	0,25	0,15	0,066

für verschiedene Kombinationen von Widerstand R und Steilheit S. Für $R = S$ und für $R > S$ verschwindet die Amplitude nach einiger Zeit. Nur für $R < S$ gilt

$$\lim_{t\to\infty} \exp\left(-\frac{S-R}{L}t\right) = 0$$

und deshalb

$$\lim_{t\to\infty} A(t) = \sqrt{\frac{4(S-R)}{S}} > 0.$$

Demnach schwingt der Oszillator stabil nur für $R < S$.

Einige Zahlenbeispiele sollen das Oszillatorverhalten verdeutlichen. Für die drei in der Aufgabe angegebenen Zahlenkombinationen lauten die zugehörigen Schwingungen

$$(\alpha) \quad U(t) = \frac{2\cos 50t}{\sqrt{25 - 16\exp(-t)}}$$

$$(\beta) \quad U(t) = \frac{2\cos 50t}{\sqrt{25 + 39\exp(-t)}}$$

$$(\gamma) \quad U(t) = \frac{2\cos 50t}{\sqrt{-25 + 41\exp(t)}}.$$

Die Nullstellen der Schwingungen sind

$$t = \frac{\pi}{100}, \frac{3\pi}{100}, \frac{5\pi}{100}, \ldots$$

Einige Zahlenwerte für die Amplitude (=Einhüllende) können der Tab. 13.1 entnommen werden. Die Abb. 13.2 und 13.3 zeigen das Schwingungsverhalten in den Fällen (α) und (β). Im Fall (γ) (Abb. 13.4) läuft die Schwingung aus, hier waren R oder S nicht „richtig“ gewählt. Diese Abbildungen zeigen jeweils transiente Phasen, nämlich das (hier langsame) Einschwingen zum asymptotischen Attraktor.

Abschließend soll untersucht werden, wie gut die verwendete Näherungsmethode die wirkliche Lösung approximiert. Zu diesem Zweck wird durch numerische Integration der Differenzialgleichung (13.1) eine Lösung mit hoher Genauigkeit berechnet (Fall (α), Anfangswert $U(0) = A_0 = 2/3$, $\dot{U}(0) = 0$). In der Abb. 13.5 ist sowohl die Näherung (13.8) als auch die Integration des Anfangswertproblemes für

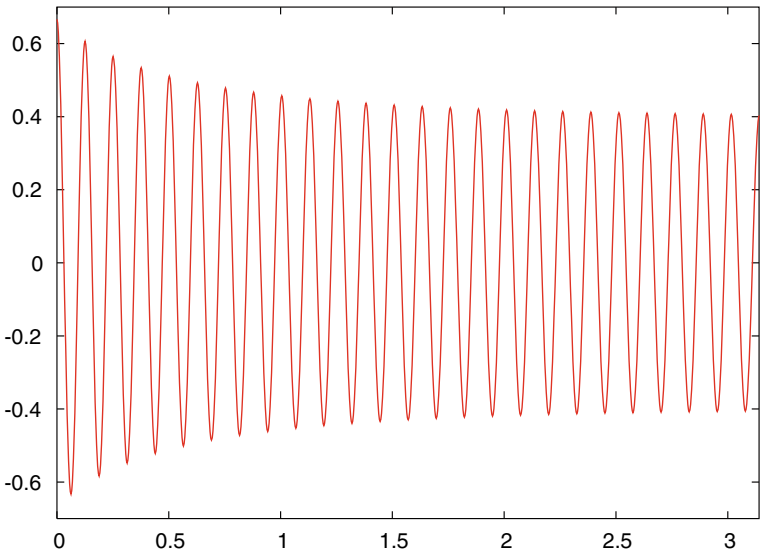

Abb. 13.2 Waagerechte Achse t, senkrechte Achse U: $U(t)$ im Fall (α): $R = 96$, $A_0 = 2/3$; Konvergenz gegen die periodische Oszillation $\frac{2}{5}\cos 50t$

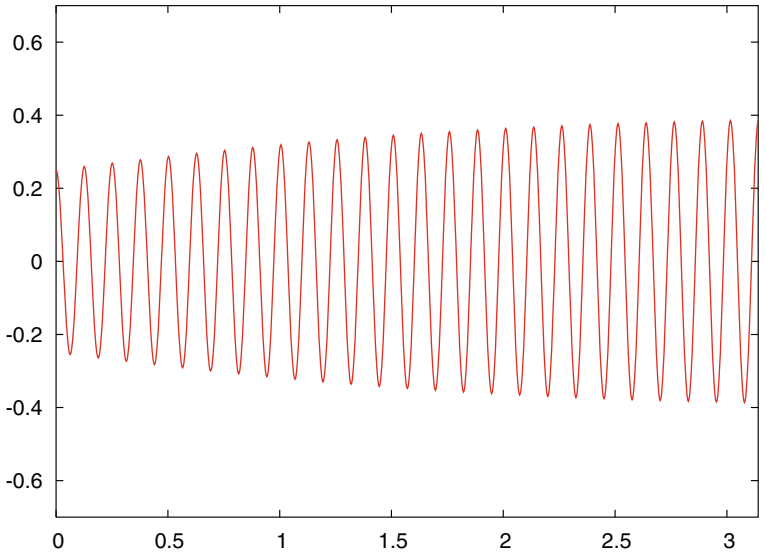

Abb. 13.3 $U(t)$ im Fall (β): $R = 96$, $A_0 = 1/4$; Konvergenz gegen die periodische Oszillation $\frac{2}{5}\cos 50t$

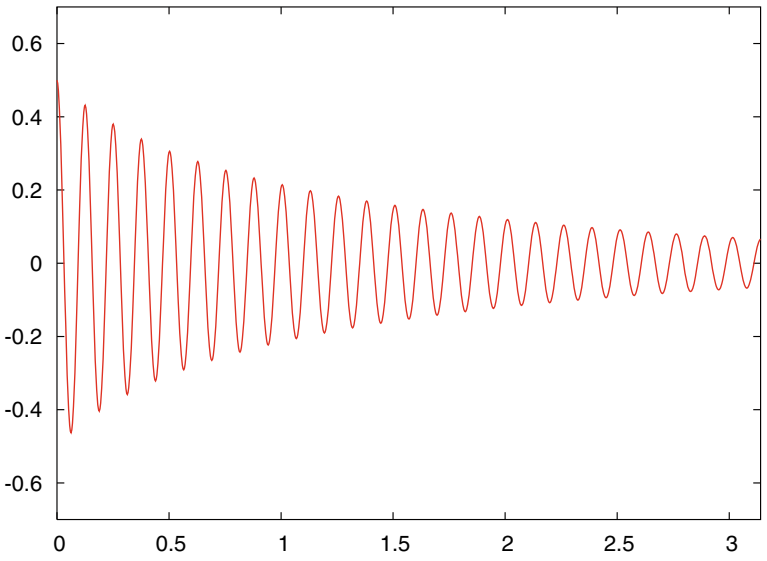

Abb. 13.4 $U(t)$ im Fall (γ): $R = 104$, $A_0 = 1/2$; Konvergenz gegen die stationäre Lösung $U = 0$

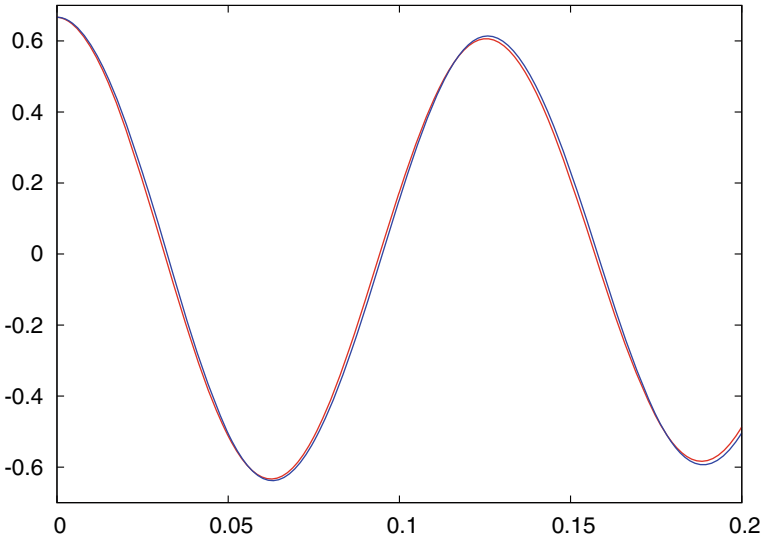

Abb. 13.5 Vergleich zweier Näherungen für $U(t)$ im Fall (α): in rot die analytische Näherung (13.8); in blau eine genaue numerische Näherung

$0 \leq t \leq 0{,}2$ wiedergegeben. Ein Vergleich zeigt, dass die Näherungsmethode trotz der mutigen Annahmen gute Resultate geliefert hat.

Stabilität
Bei dem bisherigen Vorgehen wurde nach der Methode der langsam veränderlichen Amplitude eine Näherungslösung der Differenzialgleichung (13.1) berechnet. Die anschließend durchgeführten Überlegungen zur Stabilität sollen nun mit anderen Methoden unter einem anderen Blickwinkel wiederholt werden.

Aufgabe 2
Gegeben ist die Differenzialgleichung des selbsterregten Oszillators

$$LC\ddot{U} + RC\dot{U} + U = SC\dot{U}\exp(-U^2).$$

a) Man forme die Differenzialgleichung um in ein System von Differenzialgleichungen 1. Ordnung und ermittle eventuelle stationäre Punkte.
b) Man untersuche mögliche stationäre Punkte auf Stabilität und vergleiche das Resultat mit der vorigen Aufgabe.

Setzt man $V := \dot{U}$, so erhält man wegen $\dot{V} = \ddot{U}$ das System von zwei Differenzialgleichungen 1. Ordnung

$$\begin{aligned}\dot{U} &= V\\ \dot{V} &= \frac{S}{L}V\exp(-U^2) - \frac{R}{L}V - \frac{1}{LC}U.\end{aligned}$$

Nur für $U = V = 0$ gilt $\dot{U} = \dot{V} = 0$, d. h. einziger stationärer Punkt des Systems ist

$$(U_\mathrm{s}, V_\mathrm{s}) = (0, 0).$$

Um das Lösungsverhalten in der Umgebung des stationären Punktes zu studieren, wird das System linearisiert. Die partiellen Ableitungen der zweiten Differenzialgleichung nach U und V sind

$$\frac{S}{L}(-2U)\,V\exp(-U^2) - \frac{1}{LC}$$

und

$$\frac{S}{L}\exp(-U^2) - \frac{R}{L}.$$

Den stationären Punkt eingesetzt, hat man

$$\begin{pmatrix}\dot{U}\\ \dot{V}\end{pmatrix} = \begin{pmatrix}0 & 1\\ -\frac{1}{LC} & \frac{S-R}{L}\end{pmatrix}\begin{pmatrix}U\\ V\end{pmatrix} + \text{ Terme höherer Ordnung.}$$

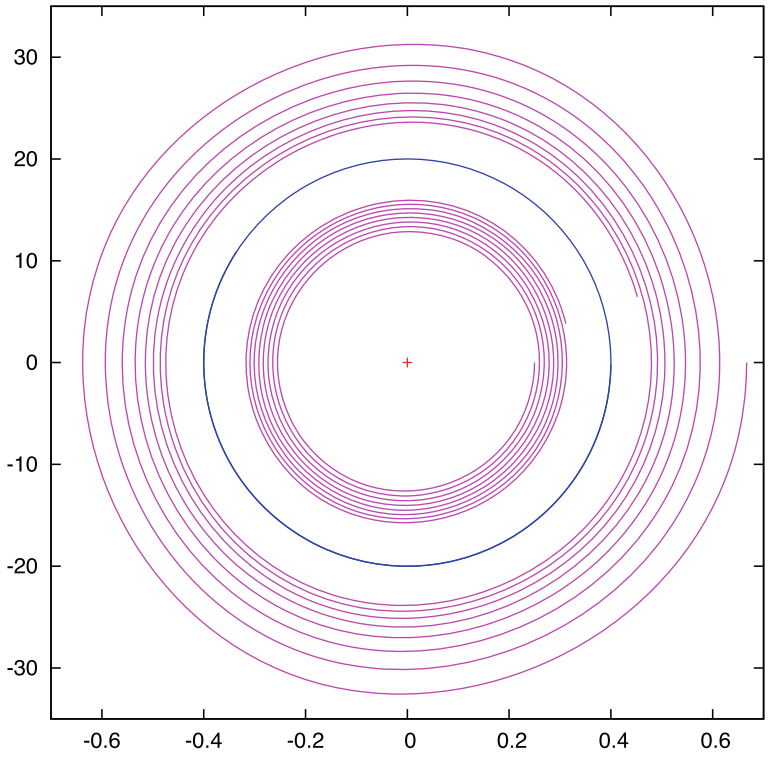

Abb. 13.6 Phasen-Porträt mit Trajektorien-Stücken in der (U, V)-Ebene für $R < S$: der Fall (α) in magenta startet von $(2/3, 0)$ und windet sich nach innen, und (β) ebenfalls in magenta startet von $(1/4, 0)$ und windet sich nach außen gegen den eingezeichneten stabilen Grenzzykel $(U, V) = (\frac{2}{5}\cos 50t, -20\sin 50t)$ (in blau). Eingezeichnet ist auch die instabile Ruhelage $(U, V) = (0, 0)$

Die Eigenwerte λ_1, λ_2 der Matrix sind durch die quadratische Gleichung

$$0 = \lambda^2 - \frac{S-R}{L}\lambda + \frac{1}{LC}$$

bestimmt,

$$\lambda_{1,2} = \frac{1}{2}\left(\frac{S-R}{L} \pm \sqrt{\left(\frac{S-R}{L}\right)^2 - \frac{4}{LC}}\right).$$

Das Vorzeichen der Diskriminante

$$\Delta := \left(\frac{S-R}{L}\right)^2 - \frac{4}{LC}$$

zeigt, welcher Typ von stationärem Punkt vorliegt. Für $\Delta > 0$ sind die Eigenwerte reell und verschieden. Wegen $L > 0$, $C > 0$ gilt $\sqrt{\Delta} < |S-R|/L$, also haben λ_1 und λ_2 gleiches Vorzeichen. Im Fall $\Delta > 0$ ist (U_s, V_s) demnach ein Knoten.

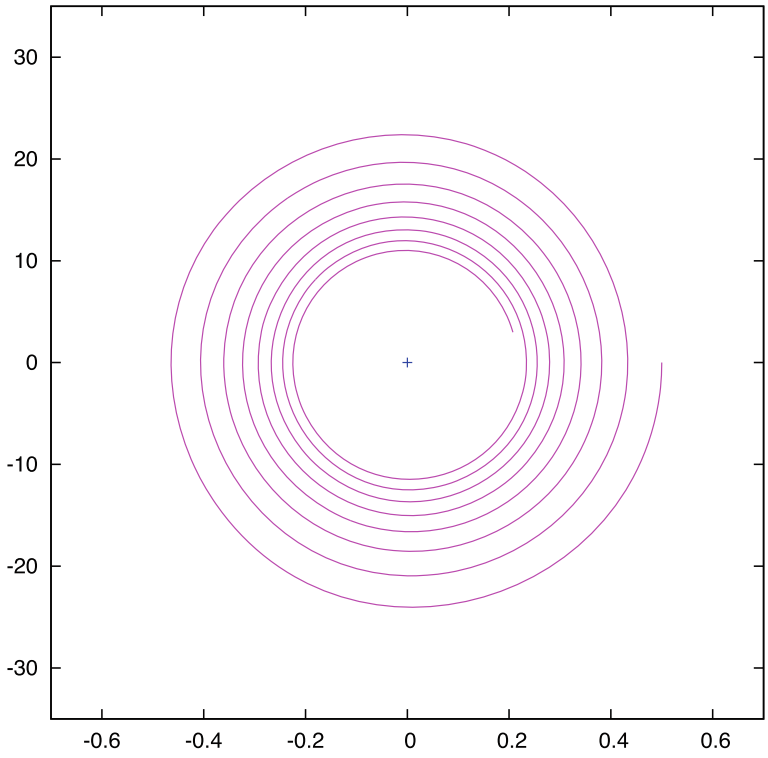

Abb. 13.7 Phasen-Porträt in der (U, V)-Ebene: der Fall (γ) mit $R > S$, Konvergenz gegen die stabile Ruhelage $(U, V) = (0, 0)$

Für $\Delta < 0$ und $S \neq R$ liegt ein Strudel vor (λ_1, λ_2 konjugiert komplex). Die entarteten Fälle $\Delta = 0$ (ausgearteter Knoten) und $\Delta < 0$, $S = R$ (Wirbel) lassen wir hier unberücksichtigt. Die Stabilität hängt wiederum vom Vorzeichen von $S - R$ ab:

$$(0, 0) \text{ stabil für } \quad R > S$$
$$(0, 0) \text{ instabil für } R < S$$

Vergleichen wir nun die Ergebnisse der beiden Vorgehensweisen bzw. Aufgaben. Bei den gewählten Zahlenwerten ($\Delta < 0$) liegen Strudel vor. Auf den ersten Blick scheinen sich die Stabilitätsresultate zu widersprechen. Um Klarheit zu gewinnen, zeichnen wir Phasendiagramme für $R < S$ und $R > S$ in den Abbildungen 13.6 und 13.7.

Für $R < S$ gibt es einen stabilen Grenzzykel und einen instabilen stationären Punkt $(0, 0)$. Der Begriff der Stabilität bezog sich also einmal auf den Grenzzykel, das andere Mal auf die Ruhelage. Die Resultate beider Aufgaben ergänzen sich. Im Fall $R > S$ ist die Ruhelage $(U, V) = (0, 0)$ stabil, es gibt hier keinen stabilen Grenzzykel.

Nun halten wir die Steilheit $S = 100$ fest und variieren den Ohmschen Widerstand R, und beobachten asymptotische Lösungen. Bei dem Parameterwert $R = S$

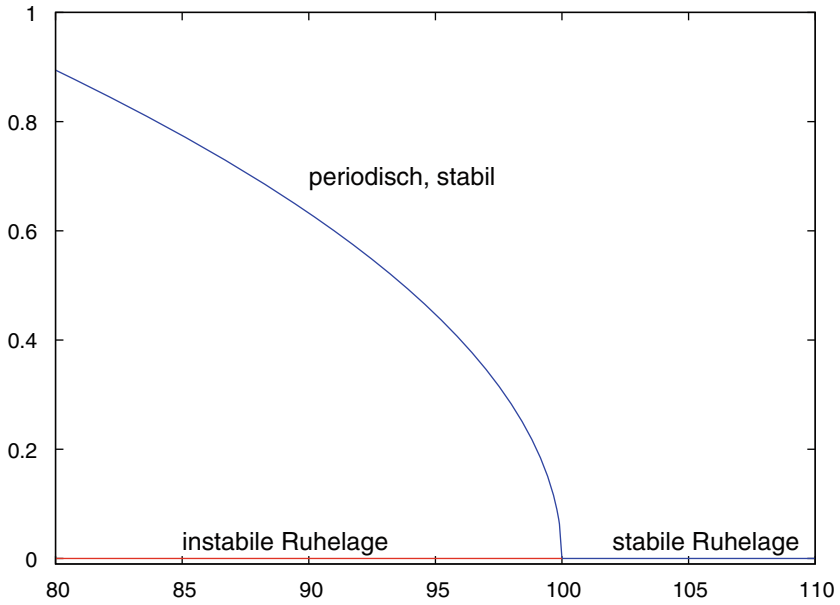

Abb. 13.8 Grenzwert der Amplitude A über dem Parameter Ohmscher Widerstand R; Bifurkation: weicher Stabilitätsverlust an $R = S = 100$ (blau: stabile Lösungen)

ereignet sich ein *Strukturwechsel* (Abb. 13.8). Für alle R gibt es eine stabile Lösung der Differenzialgleichung; für $R > S$ ist die Lösung stationär, und für $R < S$ ist sie periodisch. Die Stabilität wechselt an $R = S$ von dem einen Attraktor-Typ zum anderen. Dieser Übergang bei $R = S$ ist stetig, bei fallendem R baut sich die Amplitude der periodischen Lösung erst langsam auf. Insofern handelt es sich hier um einen *weichen Stabilitätsverlust* der stationären Lösung.[2]

Literatur

Magnus, K.: Schwingungen. Teubner, Stuttgart (1976)

[2] im Gegensatz zum harten Stabilitätsverlust bei Nervensignalen in Kap. 11, wo die Amplitude sofort auf den vollen Wert sprang. Hier wie dort handelt es sich um eine Hopf-Bifurkation: die Geburt einer periodischen Lösung aus einer Ruhelage (Literatur am Ende von Kap. 11).

Frequenzmodulation

14

Bei der Amplitudenmodulation (AM) wird die Amplitude eines Trägers moduliert, Phase und Frequenz bleiben konstant. Die Frequenzmodulation (FM) dagegen hält die Amplitude konstant und variiert die Frequenz (bzw. die Phase), die Frequenz des Trägers schwankt im Takt der Tonfrequenz. Ähnlich wie bei der Amplitudenmodulation treten auch bei der Frequenzmodulation Seitenfrequenzen auf. Im FM-Fall sind die Amplituden der Seitenbänder durch Bessel-Funktionen bestimmt.

Bessel-Funktionen
Die Bessel-Funktionen $J_n(x)$ für $n = 0, 1, 2, \ldots$ sind definiert durch[1]

$$J_n(x) := \frac{1}{\pi} \int_0^\pi \cos\,[x \sin t - nt]\,\mathrm{d}t. \tag{14.1}$$

Aufgabe 1
a) Man zeige für $n = 0, 1, 2, \ldots$ das Verschwinden der beiden folgenden Integrale:

$$\frac{1}{\pi} \int_0^\pi \sin(x \sin t) \sin nt \,\mathrm{d}t = 0 \quad \textit{für gerade } n, \tag{14.2}$$

$$\frac{1}{\pi} \int_0^\pi \cos(x \sin t) \cos nt \,\mathrm{d}t = 0 \quad \textit{für ungerade } n. \tag{14.3}$$

b) Man zeige für $J_n(x)$ aus (14.1)

$$\begin{aligned} J_n(x) &= \tfrac{1}{\pi} \int_0^\pi \sin(x \sin t) \sin(nt)\,\mathrm{d}t \quad \textit{für ungerade } n, \\ J_n(x) &= \tfrac{1}{\pi} \int_0^\pi \cos(x \sin t) \cos(nt)\,\mathrm{d}t \quad \textit{für gerade } n. \end{aligned} \tag{14.4}$$

[1] Bessel-Funktionen 1.Art; die Ordnung n darf auch ganzzahlig sein.

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_14

Zunächst wird das Integral in (14.2) mit Hilfe der Substitution $z = t - \frac{\pi}{2}$ umgeformt zu

$$\frac{1}{\pi}\int_{-\pi/2}^{\pi/2} \sin\left(x\,\sin(z+\frac{\pi}{2})\right)\,\sin\left(nz+\frac{n\pi}{2}\right)\,\mathrm{d}z.$$

Für gerade n, $n = 2m$, erhält man mit Hilfe der Additionstheoreme

$$\begin{aligned}&\frac{1}{\pi}\int_{-\pi/2}^{\pi/2} \sin(x\cos z)\,[\,\sin 2mz\,\cos m\pi + 0\,]\,\mathrm{d}z\\ &= \frac{1}{\pi}(-1)^m \int_{-\pi/2}^{\pi/2} \sin(x\cos z)\,\sin nz\,\mathrm{d}z.\end{aligned}$$

Der Integrand $f(t) := \sin(x\cos t)\sin nt$ ist wegen

$$f(-t) = \sin(x\cos(-t))\sin(-nt) = -f(t)$$

eine ungerade Funktion, demnach verschwindet für beliebige Integrationsgrenzen α das Integral

$$\int_{-\alpha}^{\alpha} \sin(x\cos t)\sin nt\,\mathrm{d}t = 0.$$

Also gilt für geradzahlige n die Gl. (14.2).

Das zweite Integral in (14.3) wird analog umgeformt, n ist hier ungerade, $n = 2m+1$:

$$\begin{aligned}&\int_0^{\pi} \cos(x\,\sin t)\cos nt\,\mathrm{d}t\\ &= \int_{-\pi/2}^{\pi/2} \cos\left(x\,\sin(z+\frac{\pi}{2})\right)\cos\left(nz+\frac{n\pi}{2}\right)\,\mathrm{d}z\\ &= \int_{-\pi/2}^{\pi/2} \cos(x\cos z)\,\left[\,0 - \sin\left((2m+1)z\right)\,\sin\left((2m+1)\frac{\pi}{2}\right)\right]\,\mathrm{d}z\\ &= -\sin(n\frac{\pi}{2})\,\cdot \int_{-\pi/2}^{\pi/2} \cos(x\cos z)\sin nz\,\mathrm{d}z = 0.\end{aligned}$$

Das Integral verschwindet, da wiederum der Integrand eine ungerade Funktion ist, also gilt (14.3).

Die Beziehungen (14.4) für $J_n(x)$ aus (14.1) folgen mit den Additionstheoremen unter Verwendung von (14.2) und (14.3). Für weitere Eigenschaften der Bessel-Funktion sei auf die Literatur verwiesen.

Frequenzmodulation

Die Modulation der Frequenz lässt sich in eine Modulation der Phase überführen, diese zwei Arten von Modulation sind äquivalent. Mit der folgenden Aufgabe wird das bei einer derartigen Modulation entstehende Frequenzgemisch berechnet.

Aufgabe 2
Eine frequenzmodulierte Schwingung ist (vereinfacht) gegeben durch

$$S(t) = \sin(Nt + \alpha \sin t).$$

Es bezeichnen

$$\begin{aligned} \sin Nt &: \text{ unmodulierte Senderwelle, } N \text{ Trägerfrequenz (ganzzahlig)} \\ \sin t &: \text{ Beispiel einer modulierenden Schwingung (Frequenz 1)} \\ \alpha &: \text{ Konstante (Modulationsindex)} \end{aligned}$$

Man zeige: Der Sender strahlt (theoretisch) alle Wellen der Frequenzen $N \pm n$, $n = 0, 1, 2 \ldots$ *aus.*[2]

Anleitung: Additionstheorem; Fourier-Entwicklung der Faktoren; Additionstheoreme in „umgekehrter" Richtung; man verwende die Bessel-Funktionen aus (14.4).

Anwendung des Additionstheorems der trigonometrischen Funktionen auf $S(t) = \sin(Nt + \alpha \sin t)$ ergibt

$$S(t) = \sin Nt \, \cos(\alpha \sin t) + \cos Nt \, \sin(\alpha \sin t). \tag{14.5}$$

Erinnert sei an die Fourier-Koeffizienten, oft mit a_ν, b_ν bezeichnet, für $\nu = 0, 1, 2, \ldots$. Speziell wenn $f(t)$ eine gerade Funktion in t ist, gilt

$$a_0 = \frac{1}{\pi} \int_0^\pi f(t)\,\mathrm{d}t, \; a_\nu = \frac{2}{\pi} \int_0^\pi f(t) \cos \nu t \,\mathrm{d}t, \; b_\nu = 0 \;\; (\nu \geq 1), \tag{14.6}$$

und wenn $f(t)$ ungerade ist:

$$a_\nu = 0 \;\; (\nu \geq 0), \quad b_\nu = \frac{2}{\pi} \int_0^\pi f(t) \sin \nu t \,\mathrm{d}t \;\; (\nu \geq 1). \tag{14.7}$$

Zunächst wird in (14.5) der Faktor $\cos(\alpha \sin t)$, eine gerade Funktion in t, in eine Fourier-Reihe mit Koeffizienten

$$a_\nu = \frac{2}{\pi} \int_0^\pi \cos(\alpha \sin t) \cos \nu t \,\mathrm{d}t$$

entwickelt, für $\nu = 1, 2, \ldots$, und analog $a_0 = \frac{1}{\pi} \int_0^\pi f(t)\,\mathrm{d}t$. In dem Ausdruck für a_ν erkennen wir die Bessel-Funktionen (14.4), es gilt

$$a_\nu = \begin{cases} J_0(\alpha) & \text{falls } \nu = 0, \\ 2J_\nu(\alpha) & \text{falls } \nu \text{ gerade } \geq 2, \\ 0 & \text{falls } \nu \text{ ungerade.} \end{cases}$$

[2] *Beim FM-Rundfunk ist* N *im MHz-Bereich und* α *im kHz-Bereich.*

Damit lautet die Fourier-Reihe

$$\cos(\alpha \sin t) = J_0(\alpha) + 2J_2(\alpha)\cos 2t + 2J_4(\alpha)\cos 4t + \ldots$$

Analog bestimmen sich die Koeffizienten der Fourier-Reihe der ungeraden Funktion $\sin(\alpha \sin t)$ in (14.5):

$$a_\nu = 0 \quad \text{für alle } \nu$$
$$b_\nu = \begin{cases} 0 & \text{falls } \nu \text{ gerade,} \\ 2J_\nu(\alpha) & \text{falls } \nu \text{ ungerade,} \end{cases}$$

also

$$\sin(\alpha \sin t) = 2J_1(\alpha)\sin t + 2J_3(\alpha)\cos 3t + \ldots$$

Setzt man die beiden Reihen zur Schwingung S zusammen, so ergibt sich

$$\begin{aligned} S(t) = J_0 \sin Nt \; &+ 2J_1 \cos Nt \; \sin t \; + \; 2J_2 \sin Nt \; \cos 2t \; + \\ &+ 2J_3 \cos Nt \; \sin 3t + 2J_4 \sin Nt \; \cos 4t + \ldots, \end{aligned}$$

die J_ν jeweils Funktionen vom Modulationsindex α. Die einzelnen Produkte lassen sich aufspalten:

$$\begin{aligned} S(t) = J_0 \sin Nt \; &+ J_1 \, [\, \sin(1-N)t + \sin(N+1)t \,] \\ &+ J_2 \, [\, \sin(N-2)t + \sin(N+2)t \,] \\ &+ J_3 \, [\, \sin(3-N)t + \sin(N+3)t \,] \\ &+ J_4 \, [\, \sin(N-4)t + \sin(N+4)t \,] + \ldots \end{aligned}$$

also

$$S(t) = J_0(\alpha)\sin Nt + \sum_{i=1}^{\infty} J_i(\alpha) \, [\, (-1)^i \sin(N-i)t \; + \; \sin(N+i)t \,]. \qquad (14.8)$$

Die Interpretation dieses Ausdrucks zeigt, dass im Signal S unendlich viele Frequenzen auftreten:

$$N \pm n, \quad n = 0, 1, 2, \ldots$$

Eine frequenz- (oder phasen-) modulierte Schwingung ist also einem breiten Frequenzband gleichwertig. Während bei AM die Seitenbänder begrenzt sind, ist das FM-Frequenzband theoretisch unendlich breit.

Die vorliegende Aufgabe 2 beschränkt sich auf eine Tonfrequenz $\sin \omega t$ mit $\omega = 1$. Für eine allgemeinere Tonfrequenz ω treten die Frequenzen

$$N \pm n\omega, \quad n = 0, 1, 2, \ldots$$

Tab. 14.1 Werte der Bessel-Funktionen $J_\nu(\alpha)$ für $\alpha = 2{,}2$, jeweils nur führende Ziffern, nach Table 9.2 in Abramowitz and Stegun (1968)

ν	$J_\nu(2{,}2)$
0	0,11036
1	0,55596
2	0,39505
3	0,16233
4	$0{,}47647 \cdot 10^{-1}$
5	$0{,}10937 \cdot 10^{-1}$
6	$0{,}20660 \cdot 10^{-2}$
7	$0{,}33195 \cdot 10^{-3}$
8	$0{,}46434 \cdot 10^{-4}$
9	$0{,}57535 \cdot 10^{-5}$

auf. Wie im Kap. 4 beim Stereo-Rundfunk gesehen, nehmen in einem praktisch wichtigen Fall die Frequenzen ω Werte bis 60 kHz an. Während sich bei der einfachen Sinus-Modulation der Aufgabe 2 die Frequenzen symmetrisch um die Trägerfrequenz N verteilen, kann bei einem allgemeineren Frequenzgemisch die Symmetrie ebenso wie die Lücken zwischen den Frequenzen verloren gehen.

Für die praktische Anwendung der FM-Modulation ist von wesentlicher Bedeutung, wie schnell die zu den Seitenfrequenzen gehörenden Amplituden abnehmen. Der Sender-Kanalabstand bedingt eine Beschränkung der Bandbreite der Seitenfrequenzen, das heißt, Frequenzen außerhalb dieser Bandbreite werden abgeschnitten. Um zu sehen, wie sich diese Bandbreitenbeschränkung auswirkt, müssen wir die Amplituden (also die Bessel-Funktionen) näher in Augenschein nehmen. Die Abnahme der Bessel-Funktionen ist an einem Zahlenbeispiel für $\alpha = 2{,}2$ in Tab. 14.1 demonstriert. Dieses Beispiel illustriert, dass die Bessel-Funktionen und damit die Amplituden der Seitenfrequenzen sehr schnell abnehmen (siehe auch Abb. 14.1). Diese starke Abnahme der Seitenfrequenzen ist wesentlich für die Übertragungsqualität: Das Abschneiden der äußersten Frequenzen aufgrund der Beschränkung der Bandbreite des Senders wirkt sich praktisch nicht negativ aus.[3] Nach (14.8) bedeutet eine reduzierte Bandbreite das Ersetzen von $S(t)$ durch die abgebrochene Reihe

$$J_0(\alpha)\sin Nt + \sum_{i=1}^{K} J_i(\alpha)\,[(-1)^i \sin(N-i)t \;+\; \sin(N+i)t],$$

ein zu kleiner Wert von K führt zu hörbaren Verzerrungen.

[3] Ein zusätzliches Stutzen der Bandbreite im Empfänger aufgrund technischer Kompromisse verschlechtert die Wiedergabequalität.

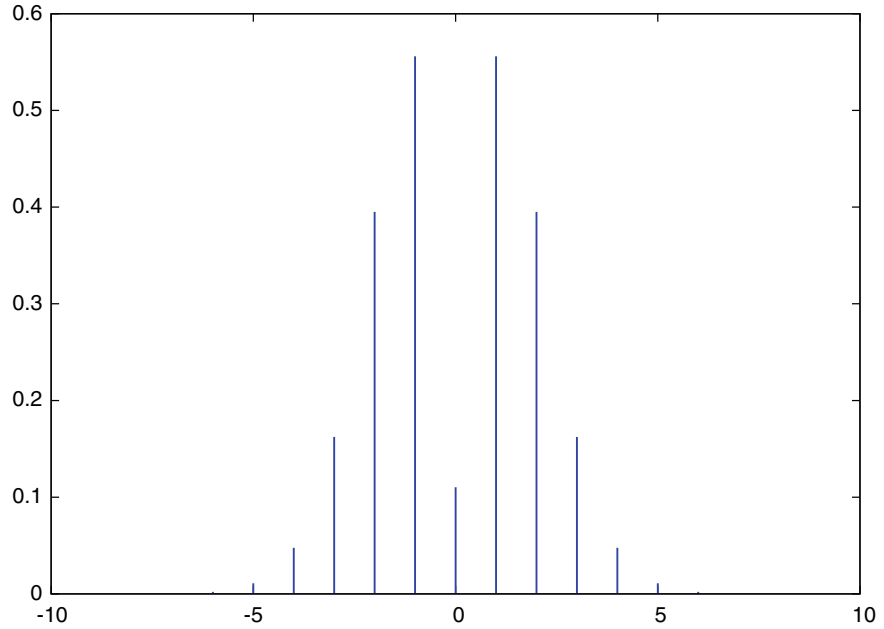

Abb. 14.1 Illustration der Tab. 14.1: Auf der horizontalen Achse ist von den Frequenzen $N \pm n$ die Zahl n aufgetragen, also entspricht die „0“ der Trägerfrequenz N. Auf der vertikalen Achse sind die Amplituden J_n aufgetragen. Die schnelle Abnahme der Amplituden der Seitenfrequenzen wird deutlich

Literatur

zu Bessel-Funktionen und Fourier-Reihen siehe Analysis-Bücher, und

Abramowitz, M., Stegun, I.: Handbook of Mathematical Functions. With Formulas, Graphs, and Mathematical Tables. Dover, New York (1968)

siehe auch, vom National Institute of Standards and Technology,

NIST Digital Library of Mathematical Functions. https://dlmf.nist.gov/

zur Berechnung von Besselfunktionen:

Bulirsch, R., Stoer, J.: Darstellung von Funktionen in Rechenautomaten. in: Sauer, R., Szabó, I.: Mathematische Hilfsmittel des Ingenieurs. Band III. Springer, Berlin (1968)

zu technischen Hinweisen, zum Unterschied Frequenz- und Phasenmodulation, und zu Computer-Animationen konsultiere man das Internet, insbesondere Wikipedia

Farbverschlüsselung und PAL

15

Farbige Darstellung auf Bildschirmen ist ein zentrales Medium unserer Zeit. Beim klassischen Fernsehen wurden kleine rote, grüne und blaue Phosphorscheibchen einer Bildröhre mit einem Elektronenstrahl zum Leuchten angeregt; neuere Plasma- oder LED-Bildschirme verwenden kleine Leuchtstoffelemente oder Leuchtdioden. Die Wirkung der farbigen Leuchtpunkte beruht auf dem Prinzip der *additiven Farbmischung.* Dieser physiologische Prozess im Gehirn mischt und summiert den visuellen Eindruck der Grundfarben R (Rot), G (Grün), B (Blau; Abb. 15.1) zu der gewünschten Farbempfindung. Schon vor mehr als 100 Jahren nutzten dies die *Pointillisten,* eine Gruppe von Malern um Seurat: Mosaikartig gesetzte kleine Farbpunkte aus wenigen Primärfarben erzeugen im Auge bei genügend großem Betrachtungsabstand den gewünschten Eindruck.

Aufnahmekameras erzeugen je einen Farbauszug in rot, grün und blau. Die entsprechenden Signale, beim Fernsehen und bei Filmen von der Zeit t abhängig, seien mit $B(t)$, $G(t)$, $R(t)$ bezeichnet.[1] Aus diesen drei Farbauszügen wird ein Schwarz-Weiß-Signal und eine kodierte Farbinformation gebildet. Die ursprüngliche Intention der Kodierung war Kompatibilität mit der alten Welt des Schwarz-Weiß-Fernsehens und seinem NTSC-System.[2] Obwohl diese Kompatibilität heute keine Rolle mehr spielt, wird das Grundprinzip der Farbverschlüsselung auch in der modernen digitalen Welt verwendet, wie zum Beispiel in der Fotografie bei JPEG und bei MPEG.

Eine Forderung nach Kompatibilität ist mathematisch eine Nebenbedingung, unter der ein System zu optimieren ist. Die Sicherung der Kompatibilität zum Schwarz-Weiß-System (nach NTSC beim analogen Farbfernsehen), bei gleichzeitig hohem Anspruch an die Qualität der Farben, war eine große Herausforderung.

[1] zur Ortsabhängigkeit später
[2] National Television System Commitee

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_15

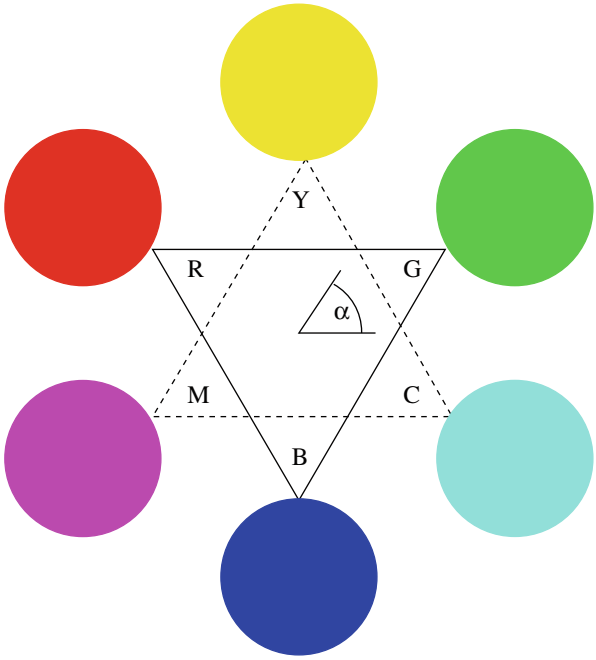

Abb. 15.1 wichtige Grundfarben: R (rot), G (grün), B (blau), C (cyan), M (magenta), Y (yellow/gelb); Mischfarben sind hier nicht gezeigt. RGB bilden die Grundlage für additive Farbmischung, CMY für subtraktive Farbmischung. Der Winkel α spezifiziert den Farbton

Beim Fernsehen gab es mit dem PAL-System[3] einen genialen Fortschritt. Da PAL aus mathematischer Sicht fasziniert, werden wir im letzten Teil dieses Kapitels auch dem PAL-System einen angemessenen Raum geben, obwohl es kaum noch verwendet wird.

15.1 Farbverschlüsselung

Die Ausführungen dieses Abschnitts sind für digitale Farbfotografie und für Farbfernsehen von Bedeutung.

Vom RGB-Modell zum YUV-Modell
Das menschliche Auge ist besonders empfindlich für die Farbe Grün, weniger für Rot, und noch weniger für Blau. Entsprechend den Empfindlichkeiten des Auges werden die Daten R, G, B gewichtet, um hieraus das *Leuchtdichte-Signal* zu bilden. Beim YUV-Modell ist das

$$Y(t) := 0{,}299\,R(t) + 0{,}587\,G(t) + 0{,}114\,B(t)\ , \tag{15.1}$$

[3] Phase Alternation Line

bei anderen Modellen auch mit anderen Konstanten. Die Variable Y steht für die Helligkeit des Bildes und liefert die Schwarz-Weiß-Information.[4] Da dieses Signal Y für Schwarz-Weiß-Fernsehgeräte wegen der Kompatibilität zum NTSC-System ohnehin zu übertragen war, braucht man nur noch zwei Signale zusätzlich, um im Empfänger R, G, B zurückzugewinnen.

Es bleibt das Problem, die Farbart und die Farbsättigung zu verschlüsseln. In Abb. 15.1 sind lediglich die sechs wichtigsten Grundfarben gezeigt. Wir denken uns nun diese Grundfarben eingebettet in einen Ring mit allen Mischfarben. Dabei ergeben die Primärfarben mit den Mischfarben einen kontinuierlichen Farbenkreis. Jede Farbe kann dann mit der Angabe des Winkels α (Abb. 15.1) ausgewählt werden. Auch der Radius des Farbenkreises wird genutzt: er gibt die Farbsättigung an. So ist für den Radius 0 die Farbe völlig entsättigt (weiß), für den vollen Radius liegt dagegen 100 %-ge Sättigung vor. So kann mit zwei Koordinaten U und V in einer gedachten Farbebene eindeutig der Farbton (Winkel α) und die Farbsättigung (Vektorlänge $\sqrt{U^2+V^2}$) definiert werden. Man setzt für das YUV-Modell

$$U(t) := 0{,}493\,(B(t) - Y(t)) \quad \text{und} \quad V(t) := 0{,}877\,(R(t) - Y(t)) \tag{15.2}$$

als die zwei *Farbdifferenzsignale.*[5] Für eine Illustration der Signale U und V siehe das Beispiel der Abb. 15.2 und 15.3.

Eine Abhängigkeit von der Zeit t spielt für die folgenden Ausführungen zur Digitalisierung keine Rolle. Es genügt zunächst, wenn wir ein Pixel zu einem Zeitpunkt betrachten und die Verschlüsselung seiner Helligkeits- und Farbinformation.

In Matrix-Form geschrieben, lautet die YUV-Transformation (15.1)/(15.2)

$$\begin{pmatrix} Y \\ U \\ V \end{pmatrix} = \begin{pmatrix} 0{,}299 & 0{,}587 & 0{,}114 \\ -0{,}147407 & -0{,}289391 & 0{,}436798 \\ 0{,}614777 & -0{,}514799 & -0{,}099978 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}. \tag{15.3}$$

Die Dekodierung, also die Rücktransformation (im Empfänger), in Matrix-Form mit der inversen Matrix geschrieben, ist

$$\begin{aligned} \begin{pmatrix} R \\ G \\ B \end{pmatrix} &= \begin{pmatrix} 1 & 0 & \frac{1}{0{,}877} \\ 1 & -\frac{0{,}114}{0{,}493 \cdot 0{,}587} & -\frac{0{,}299}{0{,}877 \cdot 0{,}587} \\ 1 & \frac{1}{0{,}493} & 0 \end{pmatrix} \begin{pmatrix} Y \\ U \\ V \end{pmatrix} \\ &\approx \begin{pmatrix} 1 & 0 & 1.14 \\ 1 & -0{,}394 & -0{,}581 \\ 1 & 2{,}03 & 0 \end{pmatrix} \begin{pmatrix} Y \\ U \\ V \end{pmatrix}. \end{aligned} \tag{15.4}$$

Mit obigem YUV-Modell arbeitet das Fernsehen. Auch die Verschlüsselung des PAL-Systems (unten beschrieben) arbeitet mit YUV. Die folgenden Ausführungen

[4] *Luminanz,* Lichtstärke. Die Bezeichnung Y in (15.1) hat nichts mit *yellow* zu tun. Die Konstanten in (15.1) sind auf Kathodenstrahlröhren zugeschnitten.

[5] *Chrominanzsignale*.

Abb. 15.2 Ein Farbbild (Quelle: Autor). Die Schwarz-Weiß-Information (Y-Version nach (15.1)) ist in Abbildung 6.4 wiedergegeben; die U- und die V-Komponenten in Abb. 15.3

zur Skalierung sind für das PAL-System ohne Belang, aber wichtig für die Digitalisierung.

Skalierungen als Vorbereitung zur Digitalen Verarbeitung
Aufgrund der technischen Gegebenheiten der Aufnahmegeräte ist von

$$0 \le R \le 1, \quad 0 \le G \le 1, \quad 0 \le B \le 1$$

auszugehen. Für die spätere Digitalisierung sollten auch die Y, U, V-Werte in geeignet normierten Bereichen liegen, zum Beispiel im Einheitsintervall [0, 1]. Dann hat man die Diskretisierungsfehler besser in der Hand. Für eine Normierung sind Skalierungen notwendig.

Zunächst werden die R, G, B-Werte der sogenannten *Gamma-Korrektur* unterworfen: Statt der drei Variablen R, G, B werden korrigierte R', G', B' verwendet,

Abb. 15.3 Zu dem Farbbild von Abb. 15.2 die U- (links) und die V-Information (rechts) nach (15.2)

wobei jede der drei Variablen R, G, B mit einer Potenzfunktion

$$x' = x_{\text{out}} := x_{\text{in}}^{\gamma}$$

transformiert wird; dabei steht x_{in} für jede der drei Variablen R, G, B. Im Fall $\gamma = 1$, dem linearen Fall, ändert sich nichts. Der nichtlineare Fall $\gamma \neq 1$ kann zur Aufhellung oder Abdunklung des Bildes dienen. Bei der Rücktransformation im Empfänger wird die inverse Transformation angewendet. Gründe für eine γ-Korrektur liegen in unterschiedlicher Intensitäts-Wahrnehmung für helle und dunkle Signale, und in den Eigenschaften verschiedener Bildschirme. Aus den Eigenschaften der Potenzfunktion folgt, dass auch nach Anwendung der γ-Korrektur die resultierenden R', G', B'-Werte im Einheitsintervall [0, 1] liegen:

$$0 \le R' \le 1, \quad 0 \le G' \le 1, \quad 0 \le B' \le 1. \tag{15.5}$$

Was für Folgerungen hat diese Normierung für das Y, U, V-Signal?

Aufgabe 1 *Für die Gleichung (15.3) nehme man den Zustand nach einer γ-Korrektur an: Die Werte im Vektor der der rechten Seite seien die R', G', B'-Werte, alle im Intervall* [0, 1].

a) *In welchem Zahlenbereich liegen dann die Y-, U- und die V-Werte?*
b) *Wie kann man U und V skalieren, sodass sie in den Zahlenbereich des Intervalls* $[-0{,}5, +0{,}5]$ *gelangen?*

Die Gl. (15.3) besteht auf der rechten Seite aus drei Skalarprodukten. Das erste ist

$$Y' = 0{,}299\, R' + 0{,}587\, G' + 0{,}114\, B'.$$

Die Verteilung (15.5) bedeutet, dass die (R', G', B')-Werte im Einheitswürfel liegen, und das Skalarprodukt sein Maximum und sein Minimum jeweils an einer Ecke des Würfels annimmt. Für Y' ist sofort klar, dass der maximale Wert für $R' = G' = B' = 1$ angenommen wird mit dem Wert $Y' = 1$, und das Minimum für $R' = G' = B' = 0$. Die Folgerung ist

$$0 \le Y' \le 1.$$

Eine analoge Überlegung für U: Der maximale Wert liegt an der Ecke $(R', G', B') = (0, 0, 1)$ und der minimale an der Ecke $(1, 1, 0)$, und die Folgerung ist

$$-0{,}436798 \le U \le 0{,}436798,$$

ziemlich „krumme“ Werte. Eine weitere Skalierung streckt dieses Intervall auf Einheitslänge, wenn U durch $2 \cdot 0{,}436798 = 0{,}873596$ dividiert wird. Schließlich V: Aus den Werten in (15.3) ergibt sich

$$-0{,}614777 \le V \le 0{,}614777,$$

und die Werte gelangen wiederum in ein Intervall von Einheitslänge, wenn V durch $2 \cdot 0{,}614777 = 1{,}229554$ dividiert wird.

Das entsprechend skalierte Modell hat einen Namen, es heißt YPbPr-Modell. Mit

$$Pb := \frac{U}{0{,}873596}, \quad Pr := \frac{V}{1{,}229554}$$

folgt aus (15.3)

$$\begin{pmatrix} Y' \\ Pb \\ Pr \end{pmatrix} = \begin{pmatrix} 0{,}299 & 0{,}587 & 0{,}114 \\ -0{,}168736 & -0{,}331264 & 0{,}5 \\ 0{,}5 & -0{,}418688 & -0{,}081312 \end{pmatrix} \begin{pmatrix} R' \\ G' \\ B' \end{pmatrix} \tag{15.6}$$

und

$$-0{,}5 \le Pb \le 0{,}5, \quad -0{,}5 \le Pr \le 0{,}5.$$

Verglichen mit der Matrix in (15.3) ergeben sich die Koeffizienten in der zweiten und dritten Zeile der Matrix (15.6) durch Division mit den oben hergleiteten Zahlen. Das YPbPr-Modell nach (15.6) ist noch analog und gilt unabhängig von jeder Diskretisierung, beinhaltet also keine Beschränkung der Auflösung.

YCbCr-Modell

Die digitalisierte Version des YPbPr-Modells ist das YCbCr-Modell. Aus der Variablen Pb wird Cb, und aus Pr wird Cr. Für die Digitalisierung wird jede der drei Größen Y', Pb, Pr mit typischerweise 8 Bit kodiert, also mit $2^8 = 256$ Stufen. Das YCbCr-Modell addiert bei Pb und Pr jeweils den Offset $\frac{1}{2}$, sodass die Variablen im

Wertebereich von Y' liegen. Bei der 8-Bit-Kodierung wird nur die Stufen-Nummer $0, \ldots, 255$ angegeben.[6]

Diese Aspekte sind eher technischer Natur und hier nicht weiter ausgeführt. Das YCbCr-Modell ist der Standard für die Verschlüsselung von Helligkeit und Farbe jedes Pixels, 24 Bit werden benötigt. Auch die Konventionen von JPEG und MPEG arbeiten mit dieser Darstellung. Die Signale Y', Cb, Cr werden beim digitalen Fernsehen getrennt übertragen, und die im Folgenden beschriebenen Techniken entfallen.

15.2 Analoges Fernsehen

Im Folgenden widmen wir uns der Farbverschlüsselung beim klassischen analogen Fernsehen. Nun geht es nicht um einzelne Pixel, sondern um ein ganzes Bild und seine zeitliche Veränderung.[7] Die Digitalisierung beschränkt sich hier auf die Aufteilung eines Fernsehbildes in Zeilen.

Quadratur-Amplitudenmodulation
Die zeitabhängigen Farbdifferenzsignale U und V aus (15.2) werden auf zwei Träger gleicher Frequenz, die um $\frac{\pi}{2}$ phasenverschoben sind, aufmoduliert. Diese beiden Träger sind $\sin 2\pi\omega t$ und $\cos 2\pi\omega t$ mit einer noch zu bestimmenden Trägerfrequenz ω. Die beiden modulierten Schwingungen summiert bilden das Farbsignal[8]

$$\varphi(t) := U(t)\sin 2\pi\omega t + V(t)\cos 2\pi\omega t. \tag{15.7}$$

Diese Summe von zwei Schwingungen gleicher Frequenz lässt sich zusammenfassen zur Schwingung

$$\begin{aligned} &\varphi(t) = \sqrt{U^2+V^2}\,\sin(2\pi\omega t+\alpha),\\ &\text{mit}\quad \tan\alpha = \frac{V(t)}{U(t)}, \end{aligned} \tag{15.8}$$

$U(t) \neq 0$ vorausgesetzt. Da der Farbton α von t abhängt, ist das Argument der Sinus-Funktion eine modulierte Phase. Man kann also das Farbsignal φ auffassen als *eine* Schwingung, bei der sowohl die Amplitude (Farbsättigung) als auch die Phase (Farbton) moduliert sind. Dieses Verfahren wird auch Quadratur-Amplitudenmodulation genannt. Ähnlich wie beim Stereo-Rundfunk die hochfrequente Rauminformation zum Mono-Summensignal tritt, wird beim analogen Farbfernsehen die Farbinformation φ zum Schwarz-Weiß-Signal Y addiert. Es entstehen jeweils Multiplexsignale, mit denen die Senderwelle moduliert wird.

[6] im Studiobereich 10 Bit. Das menschliche Auge kann circa 256 Graustufen voneinander unterscheiden. Beim hochauflösenden HDTV auch andere Koeffizienten in (15.1)/(15.2)/(15.4).
[7] (Es wird komplizierter.)
[8] Die Argumente sind zu verstehen wie in $\cos(2\pi\omega t)$. Eine solch aufwendige Klammerung wird im Folgenden häufig vermieden.

Ortsabhängigkeit
Bisher haben wir uns im Wesentlichen mit der Verschlüsselung der Farbe beschäftigt. Bei einem Film sind Helligkeit und Farbe sowohl von der Zeit t als auch vom Ort auf dem Bild abhängig. Die Farbsignale Y, U, V varieren im Allgemeinen langsam mit Zeit und Ort. Die Abhängigkeit von t haben wir formal bereits in (15.1) und (15.2) vorgesehen. Die Abhängigkeit vom Ort wird über die feste Zeilenstruktur des Fernsehbildes in eine (weitere) Zeitabhängigkeit aufgelöst! Der Kathodenstrahl der Bildröhre flitzt die Zeilen entlang, in ihrem festen Zeilen-Aufbau. So ist die Ortsabhängigkeit über die in schneller Abfolge abgetasteten Zeilen in einer zusätzlichen komplizierten Zeitabhängigkeit enthalten. Darauf geht auch die Wahl der Frequenz ω ein, auch hierin verbirgt sich Mathematik.

Wahl der Frequenz
Die Frequenzspektren von Leuchtdichtesignal Y und Farbsignalen U, V haben wegen der Zeilenstruktur des Fernsehbildes einen periodischen Charakter. Wir gehen aus von einem Bildschirm mit 625 Zeilen, der 25 mal in der Sekunde erneuert wird. Das führt zu einer Bildfrequenz $f_B = 25$ Hz und einer Zeilenfrequenz $f_Z = 15\,625$ Hz ($= 625 \cdot 25$). Die Zeilen sind ähnlich aufgebaut, vereinfacht stellen wir uns eine typische Zeile als eine Schwingung S_Z mit der Zeilenfrequenz f_Z als Grundschwingung vor, also etwa

$$S_Z(t) = a_1 \cos(2\pi f_Z t) + a_2 \cos(2\pi 2 f_Z t) + a_3 \cos(2\pi 3 f_Z t) + \ldots .$$

Die Veränderungen von Zeile zu Zeile kann man interpretieren als Überlagerung einer Schwingung S_B mit Grundfrequenz f_B,

$$S_B(t) = b_1 \cos(2\pi f_B t) + b_2 \cos(2\pi 2 f_B t) + b_3 \cos(2\pi 3 f_B t) + \ldots$$

Vereinfacht deuten wir das Leuchtdichtesignal Y (ebenso wie auch die Farbsignale U und V) als Modulation der Amplitude des Zeilensignals S_Z,

$$Y(t) = S_B(t) \cdot S_Z(t).$$

Dieses Produkt besteht aus der Summe von Produkten der Form

$$\cos(2\pi k f_Z t) \cos(2\pi \nu f_B t),$$

mit natürlichen Zahlen k, ν. Ein solcher Ausdruck kann umgeformt werden zu

$$\frac{1}{2} \cos(2\pi t(k f_Z - \nu f_B)) + \frac{1}{2} \cos(2\pi t(k f_Z + \nu f_B)).$$

Demnach enthält das Y-Signal die Frequenzen

$$k f_Z \pm \nu f_B \quad (k, \nu = 1, 2, 3, \ldots). \tag{15.9}$$

Für $k = 1$ sind dies die Frequenzen

$$15\,625 \pm \nu \cdot 25\,.$$

Aus dieser Überlegung folgt, dass Y, U und V annähernd *Balkenspektren* haben in der Form, wie sie in Abb. 15.4 skizziert ist. Für das Y-Signal (illustriert in Abb. 15.4) arbeitet man mit einer Bandbreite von Frequenzen $0 - 5{,}5$ MHz.

Auch die Frequenzen des Farbsignals φ sind nach der Amplitudenmodulation der Träger in (15.7) „balkenförmig" verteilt. Das folgt für jeden der beiden Summanden in (15.7) mit den gleichen Argumenten wie in Kapitel 4 über Amplitudenmodulation und Stereo-Rundfunk. Die Energiemaxima liegen bei den Frequenzen $\omega \pm k f_Z \pm \nu f_B$, für $k, \nu = 1, 2, \ldots$. Das ist schematisch in Abb. 15.5 gezeigt.

Aus Gründen der Kompatibilität von Schwarz-Weiß-Empfängern und Farbempfängern muss innerhalb der Bandbreite des Y-Schwarz-Weiß-Signals von 5,5 MHz auch die Farbinformation untergebracht werden; letztere hat eine Bandbreite

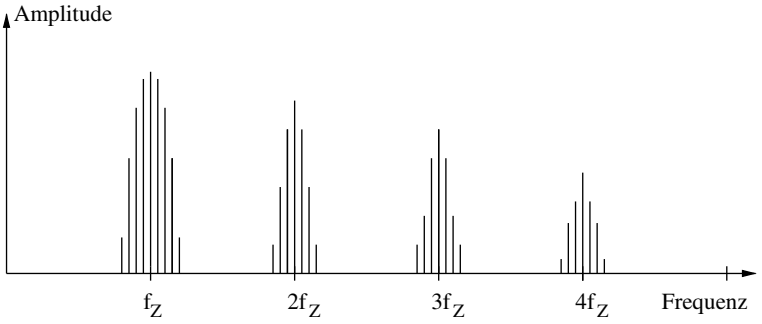

Abb. 15.4 Frequenzen (15.9) von Y und Größenordnung der zugehörigen Amplituden, schematisch

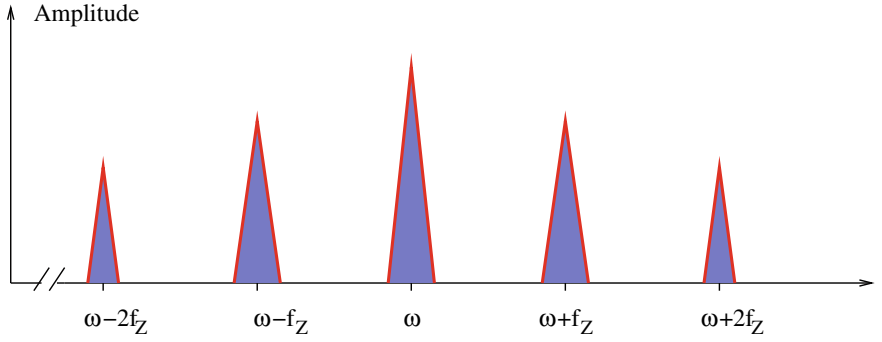

Abb. 15.5 Energie der Frequenzen des Farbsignals φ aus (15.7); die horizontale Achse zeigt die Frequenzen an. Die Energiewerte der diskreten Frequenzen sind hier stark vereinfacht als Dreiecke zusammengefasst

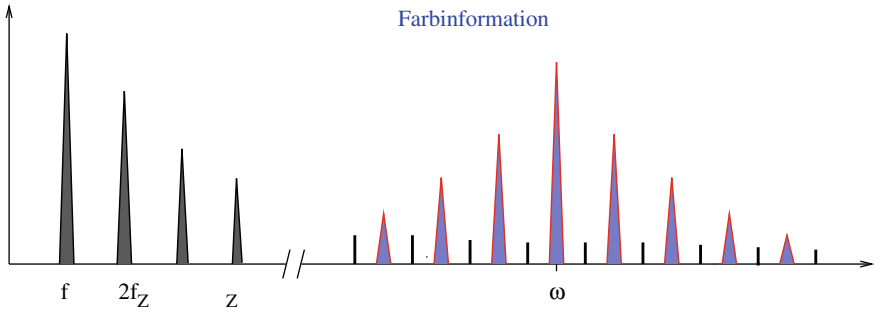

Abb. 15.6 Spektrum des Multiplexsignals $Y + \varphi$, schematisch

von 1,5 MHz.[9] Hierzu bieten sich die „Lücken“ zwischen den Vielfachen der Zeilenfrequenz f_Z an (Abb. 15.4). Exakt die Mitte der Lücken der Frequenzen des SW-Signals sind die ungeraden Vielfachen der halben Zeilenfrequenz $f_Z/2$, also

$$(2n-1)\frac{f_Z}{2}.$$

Man schiebt die Farbinformation in den oberen Bereich des Frequenzbandes, weil dort die Y-Amplituden nur noch schwach sind, und deswegen gegenseitige Störungen verringert werden. Das NTSC-System wählt $n = 284$, also

$$\omega_{\text{NTSC}} := 567 \cdot \frac{15625}{2} = 4\,429\,687{,}5 \text{ Hz}.$$

Die Abb. 15.6 zeigt das Prinzip: Unter Ausnutzung der Kammstruktur der Spektren von Y, U, V gelingt es, durch geeignete Wahl von ω die Frequenzen von SW-Signal und Farbinformation weitgehend zu entkoppeln. Beim PAL-System (eine Variante des NTSC-Systems) hat es sich aus technischen Gründen als sinnvoll gezeigt, die Trägerfrequenz ω geringfügig aus der Lücken-Mitte herauszuschieben:

$$\omega = \omega_{\text{NTSC}} + 25 + \frac{f_Z}{4} = 4\,433\,618{,}75 \text{ Hz}.$$

Farbton und Farbsättigung im NTSC-System
Laufzeitverzögerungen auf dem Übertragungsweg Sender-Empfänger führen zu Störungen des Farbtons. Das wird in der folgenden Aufgabe diskutiert, die das NTSC-System zusammenfasst.

[9] Für den Farbeindruck genügt eine geringere Auflösung als für die Grau-Stufen des Schwarz-Weiß-Signals, da das Auge für Farben weniger empfindlich ist als für Hell-Dunkel-Kontraste.

Aufgabe 2 *Es bezeichnen $R(t)$, $G(t)$, $B(t)$ das Rot-, Grün-, Blau-Signal (Farbauszüge) der Aufnahmekamera und* $\sin 2\pi\omega t$, $\cos 2\pi\omega t$ *die Trägerschwingungen mit* $\omega = 4\,433\,618{,}75$ *Hz.*

Kodierung: Es werden gebildet

(i) Leuchtdichtesignal (Schwarz-Weiß-Bild):

$Y(t) = 0{,}299\,R(t) + 0{,}587\,G(t) + 0{,}114\,B(t)$.

(ii) Farbdifferenzsignale:

$U(t) = 0{,}493\,(B(t) - Y(t))$, $V(t) = 0{,}877\,(R(t) - Y(t))$.

Aus den drei Signalen Y, U, V lässt sich B, G, R zurückgewinnen. Das ist der Kern des NTSC-Systems.

Mit den Farbdifferenzsignalen werden die zwei um $\frac{\pi}{2}$ phasenverschobenen Trägerschwingungen amplitudenmoduliert (DSB-Modulation: der Träger wird ausgesiebt),

$$U(t)\sin 2\pi\omega t \quad \text{und} \quad V(t)\cos 2\pi\omega t\ .$$

Diese beiden Signale zusammen mit Y bilden das Multiplexsignal, in welchem die gesamte Bildinformation enthalten ist.

Multiplexsignal beim PAL-System:

(i) Farbartsignale:

$$\begin{aligned} \varphi(t) &= U(t)\sin 2\pi\omega t + V(t)\cos 2\pi\omega t \\ \psi(t) &= U(t)\sin 2\pi\omega t - V(t)\cos 2\pi\omega t \end{aligned} \tag{15.10}$$

(ii) Multiplexsignal:

$$M(t) = \begin{cases} Y(t) + \varphi(t), & \nu\text{-te Bildzeile } (\nu \text{ ungerade}) \\ Y(t) + \psi(t), & (\nu+1)\text{-te Bildzeile } (PAL\text{-}Zeile, \nu \text{ ungerade}). \end{cases}$$

Mit $M(t)$ wird die hochfrequente Senderwelle amplitudenmoduliert (ESB-Modulation: ein Seitenband wird ausgesiebt).

Farbverfälschungen durch Störungen:

Auf dem Übertragungsweg Sender-Empfänger können Laufzeitverzögerungen τ bei den Farbsignalen auftreten: der Empfänger erhält aus dem Multiplexsignal nur die gestörten Signale

$$\varphi_\tau(t) := \varphi(t-\tau) \quad , \quad \psi_\tau(t) := \psi(t-\tau)$$

zurück. Das verursacht Farbverfälschungen auf dem Bildschirm des Empfängers.

Hierbei können sich verändern

- *der Farbton* $\left(\text{definiert durch den Winkel } \tan\alpha = \frac{V(t)}{U(t)}\right)$, und
- die *Farbsättigung (Abstand zum Nullpunkt* $\sqrt{U^2+V^2}$).

Beim reinen NTSC-System ändert sich im Wesentlichen der Farbton:

a) Unter den Annahmen $U(t-\tau)=U(t)$, $V(t-\tau)=V(t)$ zeige man:
Bei Bezugnahme auf die Trägerschwingungen der Referenzträger $\sin 2\pi\omega t$, $\cos 2\pi\omega t$ *interpretiert der Empfänger den Laufzeitunterschied* τ *als veränderte Amplituden* U_1, V_1, U_2, V_2 *der Farbartsignale:*

$$\begin{aligned}\varphi_\tau(t) &= U_1(t,\tau)\sin 2\pi\omega t + V_1(t,\tau)\cos 2\pi\omega t\\ \psi_\tau(t) &= U_2(t,\tau)\sin 2\pi\omega t - V_2(t,\tau)\cos 2\pi\omega t.\end{aligned} \tag{15.11}$$

b) U_1, V_1 und U_2, V_2 *entstehen aus* U, V *durch Koordinatentransformation (Drehung um den Winkel* $2\pi\omega\tau$*: Änderung des Farbtons). Man ermittle den Drehwinkel.*

c) Man zeige

$$\sqrt{U^2+V^2}=\sqrt{U_1^2+V_1^2}=\sqrt{U_2^2+V_2^2}$$

(keine Abschwächung der Farbsättigung)

Im Vergleich mit den hochfrequenten Trägerschwingungen $\sin 2\pi\omega t$ und $\cos 2\pi\omega t$ sind die Farbinformationen $U(t)$ und $V(t)$ niederfrequenter, und deswegen robuster gegen Laufzeitverzögerungen $t \to t-\tau$. Dies drücken wir aus durch die vereinfachenden Annahmen

$$U(t-\tau)=U(t), \quad V(t-\tau)=V(t).$$

Mit Hilfe der Additionstheoreme der trigonometrischen Funktionen und nach Einführung der Abkürzung $\epsilon := 2\pi\omega\tau$ erhält man

$$\begin{aligned}\varphi_\tau(t) :&= \varphi(t-\tau)\\ &= U(t)\sin(2\pi\omega t - 2\pi\omega\tau) + V(t)\cos(2\pi\omega t - 2\pi\omega\tau)\\ &= U(t)\sin 2\pi\omega t\ \cos\epsilon - U(t)\cos 2\pi\omega t\ \sin\epsilon\\ &\quad + V(t)\cos 2\pi\omega t\cos\epsilon + V(t)\sin 2\pi\omega t\sin\epsilon\\ &= [\,U(t)\cos\epsilon + V(t)\sin\epsilon\,]\ \sin 2\pi\omega t\\ &\quad + [\,-U(t)\sin\epsilon + V(t)\cos\epsilon\,]\ \cos 2\pi\omega t\\ &= U_1(t,\tau)\sin 2\pi\omega t + V_1(t,\tau)\cos 2\pi\omega t,\end{aligned}$$

wobei die U_1 und V_1 durch die Ausdrücke in eckigen Klammern definiert sind. Ganz analog berechnen wir

$$\begin{aligned}\psi_\tau(t) :&= U(t)\sin(2\pi\omega t - \epsilon) - V(t)\cos(2\pi\omega t - \epsilon)\\ &= [\,U(t)\cos\epsilon - V(t)\sin\epsilon\,]\ \sin 2\pi\omega t\\ &\quad - [\,U(t)\sin\epsilon + V(t)\cos\epsilon\,]\ \cos 2\pi\omega t\\ &= U_2(t,\tau)\sin 2\pi\omega t - V_2(t,\tau)\cos 2\pi\omega t.\end{aligned}$$

Damit sind die durch Laufzeitverzögerungen veränderten Amplituden der Farbartsignale ermittelt. Diese gestörten Amplituden lassen sich in Matrixschreibweise wie folgt ausdrücken:

$$\begin{pmatrix} U_1 \\ V_1 \end{pmatrix} = \begin{pmatrix} \cos\epsilon & \sin\epsilon \\ -\sin\epsilon & \cos\epsilon \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix} \quad (\varphi - \text{Zeile}), \tag{15.12}$$

$$\begin{pmatrix} U_2 \\ V_2 \end{pmatrix} = \begin{pmatrix} \cos\epsilon & -\sin\epsilon \\ \sin\epsilon & \cos\epsilon \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix} \quad (\psi - \text{Zeile}). \tag{15.13}$$

Wegen $\cos(\epsilon) = \cos(-\epsilon)$ und $-\sin(\epsilon) = \sin(-\epsilon)$ sind diese Transformationen Drehungen um die Winkel $-\epsilon$ und $+\epsilon$. Da der Winkel den Farbton festlegt, interpretiert der Empfänger die Laufzeitverzögerungen als Farbverfälschungen. Wegen des negativen Vorzeichens von V in der ψ-Zeile verschieben sich die Farbtöne bei beiden Zeilen in die gleiche Richtung. Das lässt sich wie folgt berechnen:

$$\begin{aligned} \varphi_\tau(t) &= \sqrt{U_1^2 + V_1^2} \cdot \left[\frac{U_1}{\sqrt{U_1^2 + V_1^2}} \sin 2\pi\omega t + \frac{V_1}{\sqrt{U_1^2 + V_1^2}} \cos 2\pi\omega t \right] \\ &= \sqrt{U_1^2 + V_1^2} \cdot \sin(2\pi\omega t + \alpha - \epsilon) \quad \text{mit } \tan(\alpha - \epsilon) = \frac{V_1}{U_1}, \end{aligned}$$

denn nach (15.12) und (15.8) gilt

$$\frac{V_1}{U_1} = \frac{-U\sin\epsilon + V\cos\epsilon}{U\cos\epsilon + V\sin\epsilon} = \frac{-\tan\epsilon + \frac{V}{U}}{1 + \frac{V}{U}\tan\epsilon} = \frac{\tan\alpha - \tan\epsilon}{1 + \tan\alpha\tan\epsilon} = \tan(\alpha - \epsilon)$$

und analog mit (15.13)

$$\psi_\tau(t) = \sqrt{U_2^2 + V_2^2} \cdot \sin(2\pi\omega t - \alpha - \epsilon), \quad \text{mit } \tan(\alpha + \epsilon) = \frac{V_2}{U_2}.$$

Grafisch sind diese Drehungen in Abb. 15.7 dargestellt.

Aus diesen Überlegungen folgt, dass bei der Drehung die Vektorlänge (Farbsättigung) erhalten bleibt. Eine einfache Rechnung bestätigt das:

$$\begin{aligned} U_1^2 + V_1^2 &= U^2\cos^2\epsilon + V^2\sin^2\epsilon + 2UV\sin\epsilon\cos\epsilon \\ &\quad + U^2\sin^2\epsilon + V^2\cos^2\epsilon - 2UV\sin\epsilon\cos\epsilon \\ &= U^2 + V^2, \end{aligned}$$

und analog für $U_2^2 + V_2^2$.

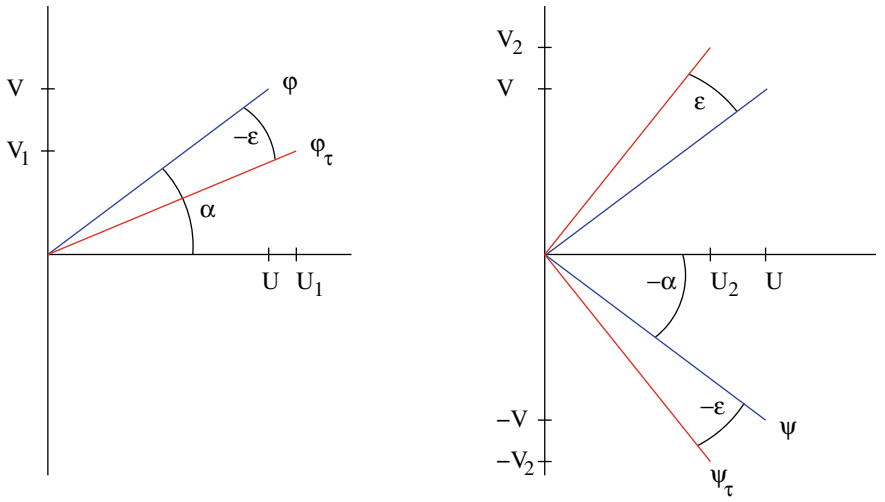

Abb. 15.7 Drehung der Farbton-Information: Die unverfälschten Signale nach (15.10) in blau, die gestörten Signale nach (15.11) in rot. Links: die φ-Zeile, rechts: die ψ-Zeile; α: der korrekte Farbton, ϵ: die Farbton-Verfälschung. Im rechten Bild im ersten Quadranten ist das Signal ψ_τ^* von Aufgabe 3 angedeutet

15.3 PAL-System

In Aufgabe 2 wurde bereits eine von Zeile zu Zeile abwechselnde Polung der V-Komponente erwähnt. Diese wechselnde V-Polung in (15.10) wird beim PAL-Fernsehen vorteilhaft zur Verringerung des Farbfehlers ausgenutzt. Exemplarisch wird dies in der folgenden Aufgabe an einer φ-Zeile besprochen. Zur Erhaltung des richtigen Farbtons wird zusätzlich zur aktuellen φ-Zeile auch die vorherlaufende ψ-Zeile benötigt. Zur Speicherung wird diese „alte“ ψ-Information in der Vorzeile in eine Verzögerungsleitung gegeben, an deren Ende die ψ-Information der vorigen Zeile gleichzeitig mit dem aktuellen φ-Signal zur Verfügung steht.

Aufgabe 3 (PAL-Prinzip)
Durch geeignete Schaltungen im Empfänger können die Farbartsignale $\psi_\tau(t)$ um die Dauer einer Bildzeile (64 μs) verzögert und transformiert werden in $\psi_\tau^*(t)$, mit

$$\psi_\tau^*(t) := U_2(t - 64\mu s, \tau) \sin 2\pi\omega t + V_2(t - 64\mu s, \tau) \cos 2\pi\omega t.$$

(Beachte das + Zeichen im Vergleich zu (15.10).)
a) Man zeige: Unter den Annahmen

$$U_2(t, \tau) = U_2(t - 64\mu s, \tau), \quad V_2(t, \tau) = V_2(t - 64\mu s, \tau)$$

(d. h. nur vernachlässigbare Farbänderung von Zeile zu Zeile) erhält man durch Bildung von

$$\frac{1}{2}\,(\,\varphi_\tau(t) + \psi_\tau^*(t)\,)$$

das ungestörte Signal $\varphi(t)$ zurück mit lediglich um den Faktor $\cos\epsilon$, $\epsilon = 2\pi\omega\tau$, abgeschwächter Farbamplitude.
(Umwandlung störender Farbtonverfälschungen in kaum wahrnehmbare Farbabschwächung)

b) Wie groß ist die Farbabschwächung (Entsättigung) bei $\epsilon = 20°$ (i.A. größter in der Praxis vorkommender Laufzeitunterschied) ?

Die V-Komponente des „alten" ψ_τ-Signals der Vorzeile (aus der Verzögerungsleitung) wird also zurückgepolt, es ergibt sich ψ_τ^*. Wie bei der Lösung von Aufgabe 2 festgestellt wurde, haben die Fehler von φ_τ und ψ_τ die gleiche Richtung (vergleiche Abb. 15.7). Durch das Zurückpolen von V bei Bildung von ψ_τ^* erhält der Fehler die entgegengesetze Richtung. Bei Addition der Signale φ_τ und ψ_τ^* mit ihren entgegengesetzten Fehlern heben sich die Farbfehler weitgehend auf:

$$\begin{aligned}\varphi_\tau(t) + \psi_\tau^*(t) &= [U_1(t,\tau) + U_2(t,\tau)]\sin 2\pi\omega t + [V_1(t,\tau) + V_2(t,\tau)]\cos 2\pi\omega t\\ &= [2U(t)\cos\epsilon]\sin 2\pi\omega t + [2V(t)\cos\epsilon]\cos 2\pi\omega t\\ &= 2\cos\epsilon\cdot\varphi(t).\end{aligned}$$

Demnach hat das Signal

$$\frac{1}{2}\,(\,\varphi(t) + \psi_\tau^*(t)\,) = \cos\epsilon\cdot\varphi(t),$$

unter den Voraussetzungen

$$U_2(t,\tau) = U_2(t - 64\mu s,\tau),\quad V_2(t,\tau) = V_2(t - 64\mu s,\tau),$$

den gleichen Farbton α wie das gesendete Signal $\varphi(t)$. Lediglich die Amplitude (Farbsättigung) ist mit dem Faktor $\cos\epsilon$ versehen, also abgeschwächt. Dieses Prinzip des PAL-Systems sei noch einmal zusammengefasst:

> Die Umpolung der V-Komponente vor der Übertragung und das Zurückpolen danach ermöglichen eine Verlagerung des Farbfehlers vom Farbton weg (wo er sehr stört) hin zur Farbsättigung (z.B. $\cos 20° = 0{,}94$). Der Betrachter nimmt eine solche Entsättigung kaum wahr.

Grafisch deutet sich der PAL-Kunstgriff schon teilweise in Abb. 15.7 an, diese Abbildung wird in Abb. 15.8 fortgesetzt.

In der dritten Aufgabe wurde der Farbfehler einer φ-Zeile diskutiert. In analoger Weise kann der Farbfehler der ψ-Zeile behandelt werden: So hat der Farbton des Signals

$$\frac{1}{2}\,(\,\psi_\tau + \varphi_\tau^*\,)^*$$

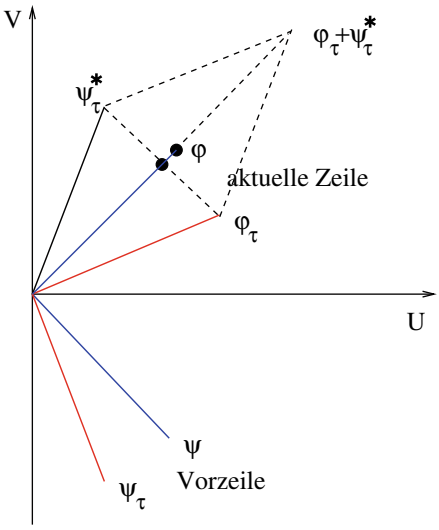

Abb. 15.8 Das PAL-System geometrisch: Das unverfälschte Signal in blau, das gestörte in rot. Die beiden Punkte markieren die Originalfarbe sowie die geringfügig entsättigte Farbe

die gewünschte Eigenschaft. Wenn also die V-Komponente der vorhergehenden φ_τ-Zeile in φ^*_τ umgepolt wurde, muss abschließend noch einmal umgepolt werden, um die richtige Phasenlage zu erhalten.

PAL in komplexen Variablen
Mit Hilfe von komplexen Größen[10] lässt sich das PAL-Prinzip knapp darstellen. Das bringt keine neuen Erkenntnisse gegenüber der oben diskutierten Darstellung mit reellen Zahlen, macht aber das Potenzial komplexer Zahlen deutlich.

Aufgabe 4 *Die Übertragung der zwei Farbdifferenzsignale $U(t)$, $V(t)$ durch DSB-Modulation zweier um $\frac{\pi}{2}$ phasenverschobener Trägerschwingungen lässt sich formal als Übertragung eines komplexen Signals deuten: Es seien*

$$\Phi(\alpha) = U + \mathrm{i}V = \sqrt{U^2+V^2}\,\mathrm{e}^{\mathrm{i}\alpha} \quad \text{und} \quad \Psi(\alpha) = \overline{\Phi(\alpha)}$$

die Farbartsignale, die wir hier als von t unabhängig betrachten. Der Farbton α ist durch $\tan\alpha = \frac{V}{U}$ *bestimmt. Im Multiplexsignal wird in der ν-ten Bildzeile das komplexe Farbsignal Φ und in der $(\nu+1)$-ten Bildzeile (PAL-Zeile) das konjugiert-komplexe Farbsignal Ψ übertragen (ν ungerade). Auf dem Übertragungsweg Sender-Empfänger können Laufzeitverzögerungen τ auftreten. Diese Laufzeitverzögerungen verursachen störende Farbverfälschungen auf dem Bildschirm.*
a) Man zeige, dass sich die gestörten Signale

$$\Phi_\epsilon := \Phi(\alpha-\epsilon) \quad \text{und} \quad \Psi_\epsilon := \Psi(\alpha+\epsilon), \quad \epsilon = 2\pi\omega\tau,$$

[10] i ist die komplexe Einheit.

mit Hilfe einer komplexen Größe $C(\epsilon)$ darstellen lassen als

$$\Phi_\epsilon = C(\epsilon)\Phi(\alpha),\ \Psi_\epsilon = C(\epsilon)\Psi(\alpha).$$

b) Bildet man im Empfänger das Signal

$$\tfrac{1}{2}\,(\,\Phi_\epsilon + \overline{\Psi_\epsilon}\,),$$

erhält man das ungestörte Signal Φ in abgeschwächter Form zurück.

Die komplexen Größen Φ_ϵ, Ψ_ϵ und $\overline{\Psi_\epsilon}$ entsprechen den reellen Größen φ_τ, ψ_τ und ψ_τ^*. Man erhält durch einfache Umformungen

$$\Phi_\epsilon = \sqrt{U^2+V^2}\,\mathrm{e}^{\mathrm{i}(\alpha-\epsilon)} = \mathrm{e}^{-\mathrm{i}\epsilon}\,\Phi(\alpha)$$

und

$$\begin{aligned}\Psi_\epsilon &= \overline{\Phi(\alpha+\epsilon)} = \sqrt{U^2+V^2}\,\mathrm{e}^{-\mathrm{i}(\alpha+\epsilon)}\\ &= \mathrm{e}^{-\mathrm{i}\epsilon}\,\overline{\Phi(\alpha)} = \mathrm{e}^{-\mathrm{i}\epsilon}\,\Psi(\alpha).\end{aligned}$$

Die Drehung um den Winkel $-\epsilon$ steckt in dem Faktor $C(\epsilon) := \mathrm{e}^{-\mathrm{i}\epsilon}$. Das ungestörte Signal erhält man wie folgt:

$$\begin{aligned}\tfrac{1}{2}\,(\,\Phi_\epsilon + \overline{\Psi_\epsilon}\,) &= \tfrac{1}{2}(\mathrm{e}^{-\mathrm{i}\epsilon}\,\Phi(\alpha) + \mathrm{e}^{\mathrm{i}\epsilon}\,\overline{\Psi(\alpha)})\\ &= \tfrac{1}{2}(\mathrm{e}^{-\mathrm{i}\epsilon} + \mathrm{e}^{\mathrm{i}\epsilon})\,\Phi(\alpha)\\ &= \cos\epsilon\ \Phi(\alpha).\end{aligned}$$

Die Interpretation ist natürlich die gleiche wie bei der reellen Darstellung: Der Farbton (α) bleibt erhalten, die Farbsättigung wird geringer um den Faktor $\cos\epsilon$.

Literatur

Bruch, W.: Vom Farbensehen zum Farbfernsehen. Bild der Wissenschaft (1966) 525–535

Anhang

A

Verwendete Mathematische Methoden und Begriffe

	Kapitel	Mathematische Methoden und Begriffe
1	Regenbogen	elementare Geometrie, Differenzialrechnung
2	Kreiskolbenmotor	Parameterdarstellung
3	Plattenspieler	elementare Geometrie, Minimierung, Differenzialrechnung, Newtonverfahren
4	Stereo-Rundfunk	Trigonometrie
5	Digitale Tonaufzeichnung	Binärzahlen
6	Bild- und Datenstruktur	Eigenwerte, elementare Stochastik
7	Bildkompression	Matrizen
8	Navigation mit Filtern	Optimierung, Matrizen
9	Berechnung des Sinus	Trigonometrie, Taylorreihe
10	Herzschlag	qualitative Methoden bei Differenzial-gleichungen
11	Nervenimpulse	Differenzialgleichungen, Stabilität, periodische Lösungen
12	Populations-Dynamik	Differenzengleichungen
13	Oszillator-Schwingungen	Differenzialgleichungen, Methode von Van der Pol, Fourierreihe, Stabilität
14	Frequenzmodulation	Besselfunktionen, Fourierentwicklung, Trigonometrie
15	Farbfernsehen und PAL	Trigonometrie, Koordinatentransformation, komplexe Zahlen

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4_A

Allgemeine Literaturhinweise
Spezielle Literaturhinweise finden sich am Schluss der jeweiligen Fallstudien. Allgemein zur Mathematik mit ihren Anwendungen und numerischen Methoden gibt es viele hervorragende Bücher. In der folgenden Liste kann nur eine kleine Auswahl berücksichtigt werden:

Literatur

Freund, R.W., Hoppe, R.W.: Stoer/Bulirsch: Numerische Mathematik 1, 10th edn. Springer, Berlin (2007)
Golub, G.H., Van Loan, C.F.: Matrix Computations, 4th edn. John Hopkins University Press, Baltimore (2013)
Logan, J.D.: Applied Mathematics. Fourth Edition. Wiley (2013)
Meyberg, K., Vachenauer, P.: Höhere Mathematik, (Bd. 1: 6. Aufl., Bd. 2: 4. Aufl.). Springer (2001)
Sauer, R., Szabó, I.: Mathematische Hilfsmittel des Ingenieurs. 4 Bände, insbesondere Teil III. Springer, Berlin (1968)
Schwarz, H.-R., Köckler, N.: Numerische Mathematik, (8. Aufl.). Vieweg & Teubner (2011)
Stoer, J., Bulirsch, R.: Numerische Mathematik 2, 5th edn. Springer, Berlin (2005)
Strang, G.: Introduction to Applied Mathematics. Wellesley-Cambridge Press, Wellesley (1986)

Stichwortverzeichnis

R. Seydel, *Höhere Mathematik im Alltag*,
https://doi.org/10.1007/978-3-662-64049-4