Wolfgang K. Seiler

Höhere Mathematik II

Vorlesung an der Universität Mannheim
im Wintersemester 2006/2007

Dieses Skriptum entsteht parallel zur Vorlesung und soll mit möglichst geringer Verzögerung verteilt werden. Es ist in seiner Qualität auf keinen Fall mit einem Lehrbuch zu vergleichen; insbesondere sind Fehler bei dieser Entstehensweise nicht nur möglich, sondern **sicher**. Dabei handelt es sich sicherlich nicht immer nur um harmlose Tippfehler, sondern auch um Fehler bei den mathematischen Aussagen.

Das Skriptum sollte daher mit Sorgfalt und einem gewissen Mißtrauen gegen seinen Inhalt gelesen werden; falls Sie Fehler finden, teilen Sie mir dies bitte persönlich oder per e-mail (seiler@math.uni-mannheim.de) mit, oder informieren Sie Ihren Tutor. Auch wenn Sie Teile des Skriptums unverständlich finden, bin ich für entsprechende Hinweise dankbar.

Falls genügend viele Hinweise eingehen, werde ich von Zeit zu Zeit Berichtigungen und Verbesserungen ausgeben.

Kapitel 3
Harmonische Analyse und Integraltransformationen

Selbst der unmusikalischste Zuhörer erkennt sofort, ob ein Ton, beispielsweise zur Note „g“, auf einer Geige oder auf einer Trompete gespielt wurde – obwohl es sich in beiden Fällen um dieselbe Note mit derselben Frequenz 192 Hertz handelt. Der Grund dafür dürfte allgemein bekannt sein: Die verschiedenen Musikinstrumente produzieren zum selben Grundton verschiedene Obertöne. Anhand der Verhältnisse zwischen den Stärken dieser Obertöne (und auch deren zeitlicher Variation) identifiziert unser Ohr die uns vertrauten Instrumente – auch wenn uns die Verhältnisse selbst quantitativ nicht bewußt sind. Umgekehrt werden bei der Synthese von Tönen etwa durch eine Soundkarte reine Schwingungen erzeugt und kombiniert.

Genau wie unser Ohr reagieren auch elektrische Schaltungen in unterschiedlicher Weise auf verschiedene Frequenzen: Legt man beispielsweise an eine Spule mit OHMschem Widerstand R und Induktivität L eine Gleichspannung U_0 an, so fließt ein Strom der Stärke I_0, für den nach dem OHMschen Gesetz gilt: $U_0 = RI_0$. Ersetzt man aber die Gleichspannung durch eine Wechselspannung der Kreisfrequenz ω, so gilt für die Amplitude I_0 des dann fließenden Wechselstroms $U_0 = \sqrt{R^2 + (\omega L)^2}I_0$, der Widerstand hängt also ab von der Frequenz.

Ersetzt man die Spule durch einen Kondensator mit Kapazität C und Widerstand (der Zuleitungen) R, so fließt beim Anlegen einer Gleichspannung natürlich überhaupt kein Strom: Der Kondensator lädt sich einfach auf. Legt man aber eine Wechselspannung der Kreisfrequenz ω an, fließt ein Strom der Amplitude

$$I_0 = U_0 \Big/ \sqrt{R^2 + \left(\frac{1}{\omega C}\right)^2};$$

auch hier haben wir also wieder einen von der Frequenz abhängigen Widerstand.

Wechselstromkreise haben zwar mehr mit Technischer Informatik zu tun als Geigen und Trompeten; sie gehören aber doch eher zum Arbeitsgebiet eines Elektroinstallateurs als zu dem eines Informatikers. Natürlich fließen auch in einem Computer Ströme, aber mit reinen Wechselströmen kann man fast genauso wenig anfangen wie mit einem Computer, in dem überall ein konstanter Gleichstrom fließt: Elektronische Informationsverarbeitung lebt von schnell und unregelmäßig variierenden Strömen. Diese fließen allerdings durch genau die Schaltelemente, von denen wir gerade gesehen haben, daß ihr Verhalten stark von der Frequenz abhängt.

Um auch solche Situationen berechenbar zu machen, müssen wir einen beliebigen Stromverlauf in eine *Summe reiner Wechselströme* zerlegen, so wie man auch den Ton eines Musikinstruments in seine Grundschwingung und die Oberschwingungen zerlegen kann. Wir werden in diesem Kapitel als erstes sehen, daß man jede (halbwegs vernünftige) *periodische* Funktion beliebig genau durch Summen reiner Schwingungen annähern kann; die entsprechende Konstruktion bezeichnet man als *harmonische Analyse.* Sie gestattet es, auch für ein komplizierteres Signal dessen Verhalten in einer (linearen) Schaltung zu berechnen: Wir müssen einfach jede der reinen Schwingungen, aus denen es zusammengesetzt ist, für sich betrachten und die Ergebnisse aufsummieren.

Für nichtperiodische Funktionen wird sich zeigen, daß hier zwar keine Zerlegung in diskrete Grund- und Oberschwingungen mehr möglich ist, daß es aber ein *kontinuierliches Frequenzspektrum* gibt, mit dem man genauso arbeiten kann, wenn man die Beiträge der einzelnen Frequenzen nicht mehr summiert, sondern aufintegriert.

Bevor wir uns mit diesen Zerlegungen beschäftigen, brauchen wir zunächst einige Vorbereitungen über komplexe Funktionen.

§ 1: Funktionen einer komplexen Veränderlichen

a) Wozu komplexe Zahlen

Funktionen einer Veränderlichen werden in der Technik typischerweise dazu eingesetzt, um die Zeitabhängigkeit physikalischer Größen auszudrücken: So kann beispielsweise ein Wechselstrom der Amplitude I_0 und der Kreisfrequenz ω durch die Gleichung $I(t) = I_0 \sin \omega t$ beschrieben werden. Schaut man allerdings in Lehrbücher der Elektrotechnik, so findet man dort oft stattdessen die Formel

$$I(t) = I_0 e^{i\omega t} .$$

(Tatsächlich schreiben Elektrotechniker natürlich $e^{j\omega t}$, denn im Gegensatz zu Mathematikern und Physikern bezeichnen sie $\sqrt{-1}$ nicht mit i, sondern mit j.)

Auf den ersten Blick erscheint dies unsinnig: Was soll man sich beispielsweise unter einem Strom von $4 - 2i$ Milliampère vorstellen? Einen solchen Strom gibt es natürlich nicht. Tatsächlich ist die obige Gleichung so zu interpretieren, daß ihr Imaginärteil den tatsächlichen Strom beschreibt, während der Realteil ignoriert wird. Auf Grund der EULERschen Formeln

$$e^{i\omega t} = \cos \omega t + i \sin \omega t, \quad \cos \omega t = \mathfrak{Re}\, e^{i\omega t} \quad \text{und} \quad \sin \omega t = \mathfrak{Im}\, e^{i\omega t}$$

beschreibt also auch dieser Formalismus einen tatsächlich fließenden Strom $I_0 \sin \omega t$. (Einige Bücher verwenden die Konvention, daß nur der Realteil zählt und der Imaginärteil ignoriert wird; in diesem Fall würde $I(t) = I_0 e^{i\omega t}$ den Strom $I_0 \cos \omega t$ beschreiben.)

Der Sinn dieser Vorgehensweise liegt in mindestens zwei rechnerischen Vorteilen: Zum ersten sind Additionsregeln für trigonometrische Funktionen, vor allem wenn man sie mehrfach anwenden muß, ziemlich unangenehm, wohingegen wir für die Exponentialfunktion, egal ob mit reellen oder komplexen Argumenten, stets mit der einfachen Regel $e^{x+y} = e^x e^y$ rechnen kann.

Der zweite Vorteil wird offensichtlich, wenn wir Wechselstromnetzwerke betrachten, die nicht nur Widerstände, sondern auch Spulen und

Kondensatoren enthalten: Geht ein variabler Strom durch eine Spule der Induktivität L, so wird die Spannung $U(t) = L\dot{I}(t)$ induziert. In reeller Beschreibung ist also bei einem Wechselstrom $I(t) = I_0 \sin \omega t$

$$U(t) = LI_0\omega \cos \omega t .$$

Beim komplexen Ansatz mit $I(t) = I_0 e^{i\omega t} = I_0\big(\cos \omega t + i \sin \omega t\big)$ können wir dagegen einfach mit $i\omega L$ multiplizieren, denn

$$i\omega L \cdot I_0 e^{i\omega t} = i\omega L \cdot I_0\big(\cos \omega t + i \sin \omega t\big) = LI_0\omega(-\sin \omega t + i \cos \omega t)$$

hat als Imaginärteil genau die gerade berechnete Funktion $U(t)$. Während wir im Reellen also stets auch die Zeitabhängigkeit der Stromstärke im Auge behalten müssen, reicht es bei komplexer Darstellung, einfach die (komplexen) „Amplituden“ zu betrachten.

Ähnlich ist es bei Kondensatoren: Hier fließt bei Kapazität C und Ladung $Q(t)$ des Kondensators der Strom $I(t) = \dot{Q}(t)$. Falls dies ein reiner Wechselsstrom ist, können wir ihn – in reeller Form – als $I(t) = I_0 \sin \omega t$ schreiben, und die Spannung zwischen den beiden Platten des Kondensators ist

$$U(t) = \frac{Q(t)}{C} = \frac{1}{C}\int I(t)\,dt = \frac{1}{C}\int I_0 \sin \omega t\,dt = -\frac{I_0 \cos \omega t}{\omega C} ,$$

der Imaginärteil von

$$\frac{1}{i\omega C} \cdot I_0 e^{i\omega t} = \frac{1}{i\omega C} \cdot I_0(\cos \omega t + i \sin \omega t) = \frac{I_0}{\omega C} \cdot (\sin \omega t - i \cos \omega t) .$$

Auch hier müssen wir also bei komplexer Darstellung nur mit der festen Zahl $\frac{1}{i\omega C}$ multiplizieren statt wie im Reellen zu integrieren.

Rein formal kann man somit im komplexen Kalkül, der sogenannten *komplexen Zeigerrechnung*, Induktivitäten und Kapazitäten als rein imaginäre „Widerstände“ hinschreiben und mit diesen genauso rechnen, wie man es bei Gleichstromnetzen mit nur OHMschen Widerständen gewohnt ist. Zusammen mit den auch in Wechselstromnetzen allgegenwärtigen OHMsche Widerständen, für die das klassische OHMsche Gesetz gilt, hat man somit insgesamt formal einen komplexen Widerstand, die sogenannte *Impedanz*. Sein Betrag beschreibt den auf die Amplituden des Wechselstroms bezogenen Widerstand, sein Argument die Phasenverschiebung.

Die KIRCHHOFFschen Gesetze gelten auch für die komplexe Beschreibung von Strömen und Spannungen, insbesondere gelten für die Parallel- und Serienschaltung von Impedanzen genau die Regeln, die man von den OHMschen Widerständen her gewohnt ist, und man kann Ströme und Spannungen in Wechselstromnetzwerken mit nur passiven Bauelementen genauso berechnen wie bei Gleichstromnetzwerken, die nur Widerstände enthalten. Der einzige Unterschied besteht darin, daß man nun lineare Gleichungssysteme mit *komplexen* Koeffizienten lösen muß.

Bei einer rein reellen Beschreibung müßte man statt dessen Differentialgleichungssysteme betrachten, was – wie wir im nächsten Kapitel sehen werden – erheblich aufwendiger ist. (Bei kompliziertenen Schaltungen, die auch aktive Bauteile enthalten, gibt es dazu allerdings keine Alternative mehr.)

Auch die eingangs erwähnten Formeln für die Spannungsamplituden in einem RL- *bzw.* RC-Kreis lassen sich durch komplexe Zeigerrechnung leicht erklären: Im RL-Kreis ist die Impedanz gleich $R + i\omega L$; bei einem Wechselstrom $I(t) = I_0 e^{i\omega t}$ ist die Spannung also in komplexer Darstellung $U(t) = (R + i\omega L) \cdot I_0 e^{i\omega t}$. Da e^{ix} für reelles x stets den Betrag eins hat, ist der Betrag von $U(t)$ gleich dem der komplexen Zahl $(R + i\omega L)I_0$, also $\sqrt{R^2 + \omega^2 L^2}\, I_0$.

Warum ist dieser Betrag gleich der Amplitude des Imaginärteils? Für die Funktion $Ae^{i\omega t}$ mit komplexem A (hier ist $A = (R + i\omega L) \cdot I_0$), können wir A auch in Polarkoordinaten schreiben als $A = |A| \cdot e^{i\psi_0}$. Dann ist

$$Ae^{i\omega t} = |A| \cdot e^{i(\omega t + \psi)} = |A| \cdot \big(\cos(\omega t + \psi) + i \sin(\omega t + \psi)\big) ,$$

der Imaginärteil hat also in der Tat Amplitude $|A|$. Bei reeller Rechnung kämen wir zwar zum selben Ergebnis, aber wir müßten in diesem Fall die Amplitude der Funktion

$$\omega L I_0 \cos \omega t + R I_0 \sin \omega t$$

berechnen. Dazu müßten wir diesen Ausdruck auf die Form

$$A_0 \sin(\omega t + \psi) = A_0 \sin \psi \, \cos \omega t + A_0 \cos \psi \, \sin \omega t$$

bringen, d.h. wir müßten eine Phasenverschiebung ψ und eine Amplitude A_0 finden, so daß

$$\omega L I_0 = A_0 \sin \psi \quad \text{und} \quad R I_0 = A_0 \cos \psi$$

ist. Wegen der Beziehung $\sin^2 \psi + \cos^2 \psi = 1$ ist natürlich auch hier

$$A_0 = \sqrt{R^2 + \omega^2 L^2}\, I_0 ,$$

die Rechnung ist aber erheblich aufwendiger.

Auf Grund dieser vielen Vorteile hat sich die komplexe Zeigerrechnung allgemein durchgesetzt; ihre einfache Handhabbarkeit wiegt den Nachteil des zunächst etwas unanschaulichen Ansatzes mehr als auf, und im übrigen gewöhnt man sich nach etwas praktischer Übung auch sehr schnell daran.

Auch in der Wellenoptik erweist es sich oft als große Vereinfachung und Arbeitsersparnis, wenn man mit komplexen Schwingungen rechnet. Wir werden daher in diesem Kapitel nicht nur reelle, sondern auch komplexe Funktionen betrachten. Da wir viel differenzieren und integrieren müssen, wollen wir uns zunächst überlegen, was diese Operationen im Komplexen bedeuten und welche Gesetze dafür gelten.

Wer immer noch Probleme darin sieht, mit komplexen Größen zu rechnen, die keinerlei physikalische Realität haben, sollte sich daran erinnern, daß auch die reellen Zahlen eine mathematische Konstruktion ohne Entsprechung in der Realität sind. Die Erfahrung in Naturwissenschaften, Ingenieurwissenschaften, Wirtschafts- und Sozialwissenschaften haben gezeigt, daß die reellen Zahlen dort außerordentlich nützlich sein können, einen logischen Grund dafür gibt es aber nicht. Genauso verhält es sich mit vielen anderen mathematischen Theorien, insbesondere auch hier beim komplexen Formalismus zur Beschreibung elektrischer Größen.

Der Physik-Nobelpreisträger EUGENE WIGNER (1902–1995) bezeichnete dieses Phänomen im Titel einer seiner Arbeiten als „The unreasobable effectiveness of mathematics in the natural sciences“ *(Communications in pure and applied mathematics* **13** (1960)*, 1–14; zahlreiche Nachdrucke im WWW)*. Wirklich zum Tragen kommt diese schwer erklärbare Nützlichkeit der Mathematik allerdings nur in den Händen eines Anwenders, der sowohl ihre Möglichkeiten als auch ihre Grenzen für seinen jeweiligen Problembereich kennt.

Der Hauptinhalt des folgenden Paragraphen ist ein Steilkurs über Anwendungen des sogenannten Residuenkalküls auf reelle Integrale.

Ein Integral wie $\int_{-\infty}^{\infty} \frac{dx}{x^2+1}$ läßt sich auch mit elementaren Methoden einfach berechnen: Wie man entweder weiß oder in einer Formelsammlung nachschlägt, hat der Integrand den Arcustangens als Stammfunktion, und da der Tangens im Intervall $[-\frac{\pi}{2}, \pi/2]$ von $-\infty$ nach ∞ ansteigt, ist der Wert des uneigentlichen Integrals $\frac{\pi}{2} - (-\frac{\pi}{2}) = \pi$.

Auf ähnliche Weise könnte man zumindest im Prinzip alle Integrale der Form $\int_{-\infty}^{\infty} R(x)\,dx$ ausrechnen, bei denen der Integrand eine *rationale Funktion* (d.h. ein Quotient zweier Polynome) ist – vorausgesetzt, das uneigentliche Integral konvergiert und man kann den Nenner in höchstens quadratische Faktoren zerlegen. Dann kann man $R(x)$ als Summe von Partialbrüchen schreiben (s. **[HM1],** *Kap. 2,* §*3h4*), und für jeden dieser Partialbrüche läßt sich eine Stammfunktion angeben.

Die Konvergenz des Integrals ist ein Problem, um das kein Weg herumführt; was nicht existiert, kann man auch nicht ausrechnen.

Das zweite Problem, die Zerlegung des Nenners in höchstens quadratische Funktionen, ist prinzipiell lösbar, kann allerdings sehr schwierig sein: Man denke nur etwa an Nenner wie $x^7 + x + 2$ oder ähnliches. Zum Glück kennt die Computeralgebra eine Reihe von Algorithmen, mit denen man eine Stammfunktion von $R(x)$ bestimmen kann, *ohne* solche Faktoren zu kennen; die ersten solchen Algorithmen stammen bereits aus dem neunzehnten Jahrhundert, lange bevor es die ersten Computer gab. Selbst diese Algorithmen erfordern jedoch algebraische Techniken, für die wir in dieser Vorlesung keine Zeit haben.

Zum Glück können wir aber Integrale wie $\int_{-\infty}^{\infty} R(x)\,dx$ oft auch berechnen, *ohne* die Stammfunktion von $R(x)$ zu bestimmen. Voraussetzung dazu ist, daß das Integral konvergiert, daß wir die Nullstellen des Nenners kennen, und daß keine dieser Nullstellen reell ist. Integrale, bei denen diese drei Voraussetzungen erfüllt sind, kommen in der Elektrotechnik häufig vor; es lohnt sich also, diese Methode kennenzulernen.

Sie beruht darauf, daß Differentiation und Integration im Komplexen deutlich andere Eigenschaften haben als im Reellen.

b) Die Ableitung einer komplexen Funktion

Definition: Eine Funktion $f\colon D \to \mathbb{C}$ auf der offenen Menge $D \subseteq \mathbb{C}$ heißt *komplex differenzierbar* im Punkt $z \in D$, wenn der Grenzwert

$$f'(z) = \lim_{h\to 0} \frac{f(z+h) - f(z)}{h}$$

existiert und stetig ist. f heißt *komplex differenzierbar in* D, wenn f in jedem Punkt $z \in D$ komplex differenzierbar ist.

Der große Unterschied zum reellen Differentialquotienten liegt darin, daß h hier eine *komplexe* Zahl ist, die sich nicht nur von rechts und links, sondern aus allen Richtungen und auf jedem beliebigen Weg (für den $z + h$ noch in D liegt) dem Nullpunkt nähern kann. Um zu sehen, was das bedeutet, betrachten wir zunächst die beiden einfachsten Fälle, daß sich h ganz auf der reellen *bzw.* der imaginären Achse bewegt.

Konkret sei $z = x + iy$ und

$$f(z) = u(x,y) + iv(x,y)$$

mit zwei Funktionen $u, v\colon D \to \mathbb{R}$. (Hier wird D also als Teilmenge von $\mathbb{R}^2$ aufgefaßt.)

Für *reelles* h ist dann

$$\frac{f(z+h) - f(z)}{h} = \frac{u(x+h,y) - u(x,y)}{h} + i\frac{v(x+h,y) - v(x,y)}{h};$$

der Grenzwert für h gegen Null, so er existiert, ist damit gleich

$$\frac{\partial u}{\partial x}(x,y) + i\frac{\partial v}{\partial x}(x,y)\,.$$

Für rein *imaginäres* $h = ik$ ist entsprechend

$$\begin{aligned}\frac{f(z+h) - f(z)}{h} &= \frac{u(x,y+k) - u(x,y)}{ik} + i\frac{v(x,y+k) - v(x,y)}{ik} \\ &= \frac{v(x,y+k) - v(x,y)}{k} - i\frac{u(x,y+k) - u(x,y)}{k};\end{aligned}$$

der Grenzwert für h (oder k) gegen Null ist also, so er existiert, gleich

$$\frac{\partial v}{\partial y}(x,y) - i\frac{\partial u}{\partial y}(x,y)\,.$$

Falls f komplex differenzierbar ist, müssen beide Grenzwerte übereinstimmen, d.h.

$$\frac{\partial u}{\partial x}(x,y) = \frac{\partial v}{\partial y}(x,y) \quad \text{und} \quad \frac{\partial v}{\partial x}(x,y) = -\frac{\partial u}{\partial y}(x,y)$$

oder kurz $u_x = v_y$ und $v_x = -u_y$.

Diese Gleichungen heißen CAUCHY-RIEMANNsche Differentialgleichungen; sie sind eine notwendige Bedingung dafür, daß eine Funktion komplex differenzierbar ist.

Die Existenz einer Ableitung ist somit im Komplexen eine sehr viel einschneidendere Bedingung als im Reellen. Dafür liefert sie aber auch deutlich mehr!

Erinnern wie uns an die Analysis I:

- Wenn eine reelle Funktion differenzierbar ist, muß sie nicht stetig differenzierbar sein: Ein Gegenbeispiel ist etwa die Funktion

  $$f(x) = \begin{cases} x^2 \cos\frac{1}{x} & \text{für } x \neq 0 \\ 0 & \text{für } x = 0 \end{cases}.$$

  Für $x \neq 0$ ist ihre Ableitung nach Produkt- und Kettenregel gleich $\sin\frac{1}{x} + 2x\cos\frac{1}{x}$; für $x = 0$ ist

  $$f'(0) = \lim_{x\to 0}\frac{f(x)-f(0)}{x} = \lim_{x\to 0} x\cos\frac{1}{x} = 0\,,$$

  so daß $f'(x)$ für jedes $x \in \mathbb{R}$ existiert. Allerdings ist $f'(x)$ an der Stelle $x = 0$ nicht stetig, denn da $\sin\frac{1}{x}$ in jeder Umgebung der Null jeden Wert zwischen -1 und 1 annimmt, existiert $\lim_{x\to 0} f'(x)$ nicht, ist also insbesondere nicht gleich $f'(0)$, wie das bei einer stetigen Funktion der Fall sein müßte.
- Wenn eine Funktion n-mal stetig differenzierbar ist, muß sie nicht $(n+1)$-mal differenzierbar sein: Die Funktion

  $$f_n(x) = x^n\,|x| = \begin{cases} x^{n+1} & \text{für } x \geq 0 \\ -x^{n+1} & \text{für } x \leq 0 \end{cases}$$

  ist für $n = 0$ die Betragsfunktion und somit im Nullpunkt nicht differenzierbar. Für $n \geq 1$ ist f_n stetig differenzierbar mit Ableitung $f_n'(x) = (n+1)f_{n-1}(x)$, und damit ist klar, daß f_n zwar n-mal, nicht aber $(n+1)$-mal differenzierbar ist.
- Wenn eine Funktion beliebig oft stetig differenzierbar ist, heißt das noch immer nicht, daß sie auch nur lokal durch ihre TAYLOR-Reihe dargestellt wird. Die Funktion

  $$f(x) = \begin{cases} e^{-1/x^2} & \text{für } x \neq 0 \\ 0 & \text{für } x = 0 \end{cases}$$

  etwa ist beliebig oft stetig differenzierbar, und alle Ableitungen verschwinden bei $x = 0$, denn x^{-1/x^2} geht schneller gegen null als eine rationale Funktion gegen unendlich gehen kann. Damit stellt die TAYLOR-Reihe um $x = 0$ die Nullfunktion dar, nicht aber $f(x)$.

Im Komplexen kann all dies nicht passieren: Von den drei betrachteten Beispielfunktionen sind die erste und die dritte im Komplexen nicht einmal stetig im Nullpunkt, denn wenn man rein imaginäre Argumente einsetzt geht ihr Betrag gegen unendlich, und die Funktionen f_n erfüllen für kein $z \in \mathbb{C}$ die CAUCHY-RIEMANNschen Differentialgleichungen. Allgemein gilt:

Satz: Für eine offene Menge $D \subseteq \mathbb{C}$ und eine Funktion $f\colon D \to \mathbb{C}$ sind folgende Aussagen äquivalent:

a) f erfüllt die CAUCHY-RIEMANNschen Differentialgleichungen für alle $z \in D$.
b) f ist in jedem Punkt $z \in D$ differenzierbar.
c) f ist in jedem Punkt $z \in D$ stetig differenzierbar.
d) f ist in jedem Punkt $z \in D$ beliebig oft stetig differenzierbar.
e) f hat zu jedem Punkt $z \in D$ eine Umgebung, in der es durch seine TAYLOR-Reihe um z dargestellt wird.

Die Funktionen, die diese Bedingungen erfüllen, bezeichnet man als *holomorph*. Um nachzuweisen, daß eine Funktion holomorph ist, genügt es offensichtlich, *eine* der fünf Bedingungen *a)* bis *e)* nachzuweisen; die anderen sind dann automatisch erfüllt.

Der *Beweis* dieses Satzes steht (meist verteilt auf verschiedene Paragraphen) in jedem Lehrbuch der Funktionentheorie; abgesehen von der Implikation *b)* $\Rightarrow$ *c)* ist er auch im Skriptum von 2005 zu finden. Der Schritt von *a)* nach *b)* ist eine einfache Folgerung aus der Matrixdarstellung der komplexen Zahlen (*s.* **[HM1]**, *Kap. I,* §3*c)*). Für die restlichen Schritte auf dem Weg von *b)* nach *e)* werden komplexe Integrale benötigt.

c) Integration im Komplexen

Am einfachsten sind komplexe Integrale zu verstehen, wenn eine komplexe Funktion über ein reelles Intervall integriert werden soll: Wir schreiben $f\colon D \to C$ als $f(z) = u(z)+iv(z)$ mit Funktionen $u, v\colon D \to \mathbb{R}$ und definieren dann für ein Intervall $[a, b] \subset D$

$$\int_a^b f(t)\,dt = \int_a^b u(t)\,dt + i\int_a^b v(t)\,dt\,.$$

Rechts stehen, da t nur über $[a, b]$ variiert, rein reelle Integrale, von denen wir wissen, was sie bedeuten; damit ist auch die linke Seite definiert.

Falls der Integrand $f(z)$ eine Stammfunktion hat, d.h. wenn es eine holomorphe Funktion $F(z)$ gibt, deren Ableitung $f(z)$ ist, gilt genau wie im Reellen

$$\int_a^b f(z)\,dz = F(b) - F(a)\,,$$

denn natürlich ist die Ableitung des Realteils von $F(z)$ der Realteil von $f(z)$ und entsprechend auch für den Imaginärteil.

Man rechnet leicht nach, daß alle Potenzfunktionen $f(z) = z^n$ die CAUCHY-RIEMANNschen Differentialgleichungen erfüllen und somit holomorph sind mit der üblichen Ableitung $f'(z) = nz^{n-1}$. Da Ableiten natürlich auch im Komplexen eine lineare Operation ist, folgt damit, daß alle Polynome holomorph sind und wie gewohnt abgeleitet werden können.

Für die Exponentialfunktion $f(z) = e^{\lambda z}$ mit $\lambda \in \mathbb{C}$ zeigt eine kurze Rechnung mit der EULERschen Formel

$$e^{x+iy} = e^x(\cos y + i \sin y)\,,$$

daß $f'(z) = \lambda e^{\lambda z}$ ist, wie vom Reellen her gewohnt. Auf Grund der EULERschen Formeln

$$\cos z = \frac{e^{iz} + e^{-iz}}{2} \quad \text{und} \quad \sin z = \frac{e^{iz} - e^{-iz}}{2i}\,,$$

die wir (auch) im Komplexen als *Definition* von Sinus und Kosinus betrachten können, haben damit auch die trigonometrischen Funktionen ihre gewohnten Ableitungen, und natürlich gelten auch Produktregel, Quotientenregel und Kettenregel, deren Beweise wörtlich aus der Analysis I übernommen werden können – nur daß der in der Ableitungsdefinition gegen null gehende Parameter jetzt komplex statt reell ist.

Für alle Funktionen, die aus Polynomen, Exponentialfunktionen und trigonometrischen Funktionen zusammengesetzt sind, können wir somit Ableitungen nach genau denselben Formeln wie im Reellen berechnen, und damit können wir natürlich auch Stammfunktionen, da wo es möglich ist, mittels der gewohnten Tricks verschaffen.

Dies sind auch die einzigen komplexen Integrale die wir für den Rest dieses Semesters wirklich brauchen; lediglich für den Rest dieses Paragraphen brauchen wir, um die Idee hinter dem Residuenkalkül zur Berechnung reeller Integrale zu verstehen, auch allgemeinere komplexe Integrale.

Genau wie im letzten Semester bei den Kurvenintegralen haben wir auch jetzt bei den komplexen Integralen das Problem, daß es zwischen zwei Punkten a, b nicht mehr nur wie im Reellen nur einen Weg gibt: In der komplexen Zahlenebene gibt es viele Kurven, die in einem vorgegebenen Punkt beginnen und in einem anderen enden, und wie wir bei den reellen Kurvenintegralen gesehen haben, führen verschiedene Integrationswege im allgemeinen zu verschiedenen Werten des Integrals.

Glücklicherweise ist die Situation im Komplexen sehr viel überschaubarer als bei reellen Kurvenintegralen: Für auf ganz $\mathbb{C}$ holomorphe Integranden hängt das Integral nicht vom Weg ab, sondern nur von dessen Anfangs- und Endpunkt.

Genauer gilt folgende Aussage:

Cauchyscher Integralsatz: $f\colon D \to \mathbb{C}$ sei eine holomorphe Funktion, und die geschlossene Kurve γ sei die Randkurve einer offenen Teilmenge $G \subset D$, deren Abschluß $\overline{G}$ ganz in D liege. Dann ist

$$\int_\gamma f(z)\,dz = 0\,.$$

d) Der Residuensatz

Der CAUCHYsche Integralsatz nützt uns nichts für rationale Funktionen als Integranden, denn die sind ja (gekürzte Dartstellung vorausgesetzt) bei den Nullstellen des Nenners nicht einmal definiert, geschweige denn holomorph. Allerdings handelt es sich hier um relativ harmlose Definitionslücken, denn es sind lauter isolierte Punkte (die Nullstellen des Nenners), und wenn wir für so eine Nullstelle z_0 die Funktion mit $(z-z_0)^n$ multiplizieren, wobei n die Nullstellenordnung ist, wird zumindest das Produkt im Punkt z_0 wohldefiniert. Funktionen, bei denen nichts schlimmeres passiert, bezeichnet man als *meromorph:*

Definition: Eine Funktion f auf einer offenen Teilmenge $D \subseteq \mathbb{C}$ heißt *meromorph,* wenn gilt:

1. Es gibt eine Teilmenge $M \subseteq D$ ohne Häufungspunkte derart, daß f auf $D \setminus M$ durch eine holomorphe Funktion $D \setminus M \to \mathbb{C}$ gegeben ist.
2. Zu jedem Punkt $w \in M$ gibt es eine natürliche Zahl $n \in \mathbb{N}_0$, so daß $g(z) = (z-w)^n f(z)$ in einer Umgebung von w holomorph ist.

Für jeden Punkt $w \in M$ gibt es ein kleinstes $n \in \mathbb{N}_0$,s so daß $(z-w)^n g(z)$ in einer Umgebung von w holomorph ist. Im Falle $n = 0$ ist f selbst dort holomorph und wir können w aus M herausnehmen; andernfalls bezeichnen wir w als eine *Polstelle* von f und n als deren *Ordnung.* Bei einer rationalen Funktion, die als gekürzter Quotient zweier Polynome dargestellt ist, stimmt sie offensichtlich mit der Nullstellenordnung des Nennerpolynoms in w überein.

In den Punkten, die nicht in M liegen, ist eine meromorphe Funktion sogar holomorph, kann also um jeden Punkt durch eine TAYLOR-Reihe

dargestellt werden. In einer Polstelle $z_0 \in M$ gilt das natürlich nicht, aber es gibt nach Definition ein $n \in \mathbb{N}$, so daß $g(z) = (z-z_0)^n f(z)$ um $z = z_0$ holomorph ist und somit in eine TAYLOR-Reihe entwickelt werden kann:

$$g(z) = b_0 + b_1(z-z_0) + b_2(z-z_0)^2 + \cdots\,.$$

Dividiert man dies durch $(z-z_0)^n$, so erhält man für eine Umgebung von z_0 *mit Ausnahme des Punkts* $z = z_0$ *selbst* die Reihe

$$f(z) = \frac{g(z)}{(z-z_0)^n} = \frac{b_0}{(z-z_0)^n} + \frac{b_1}{(z-z_0)^{n-1}} + \ldots + \frac{b_{n-1}}{z-z_0} + b_n + b_{n+1}(z-z_0) + b_{n+2}(z-z_0)^2 + \cdots\,,$$

was wir auch in der Form

$$f(z) = a_{-n}(z-z_0)^{-n} + a_{-(n-1)}(z-z_0)^{-(n-1)} + \cdots + a_{-1}(z-z_0)^{-1} + a_0 + a_1(z-z_0) + a_2(z-z_0)^2 + \cdots \quad \text{mit} \quad a_j = b_{j-n}$$

schreiben können. Diese Darstellung heißt die LAURENT-Reihe von f um z_0.

PIERRE ALPHONSE LAURENT (1813–1854) war Kommandant eines Ingenieurkorps der französischen Armee und leitete unter anderem den Ausbau des Hafens von Le Havre. Seine Arbeit über die LAURENT-Reihen reichte er etwas zu spät für den großen Preis der Akademie von 1842 ein, so daß sie trotz CAUCHYs Fürsprache nicht berücksichtigt wurde. Ansonsten schrieb er anscheinend nur noch zwei weitere Arbeiten, die erst von seiner Witwe bei der Akademie eingereicht wurden. Die eine erschien 1863, die andere nie.

Eine LAURENT-Reihe unterscheidet sich somit nur dadurch von einer TAYLOR-Reihe, daß sie auch endlich viele Summanden mit negativen Exponenten haben darf. Diese treten genau dann auf, wenn die Funktion f im Punkt z_0 einen Pol hat. Wählt man n minimal, ist also $a_{-n} \neq 0$, so handelt es sich dabei offenbar genau um die Polordnung: Multipliziert man nämlich die LAURENT-Reihe mit $(z-z_0)^n$, so verschwinden alle negativen Potenzen; sie wird also zur TAYLOR-Reihe einer holomorphen Funktion. Multipliziert man dagegen mit einer kleineren Potenz von $(z-z_0)$, so steht nach a_{-n} weiterhin eine negative Potenz, das Produkt wird also für $z = z_0$ weiterhin unendlich.

Die Summe

$$H(z) = \sum_{k=-n}^{-1} a_k (z - z_0)^k = \sum_{\ell=1}^{n} \frac{a_{-\ell}}{(z - z_0)^\ell}$$

der Terme mit negativen Potenzen wird als *Hauptteil* der meromorphen Funktion $f(z)$ im Punkt z_0 bezeichnet; offensichtlich ist die Differenz $f(z) - H(z)$ in einer Umgebung von z_0 holomorph, da sie dort durch eine TAYLOR-Reihe dargestellt werden kann.

Angenommen, wir wollen eine meromorphe Funktion f integrieren über die Randkurve γ eines Gebiets G. Wir wollen annehmen, daß f auf einer offenen Menge $D \subseteq \mathbb{C}$ definiert sei, die G und γ enthält, und daß G beschränkt sei. Außerdem liege keine Polstelle von f auf der Randkurve γ.

Im Innern von G darf f Polstellen haben, und damit ist der CAUCHYsche Integralsatz nicht anwendbar.

Da G beschränkt ist und die Menge der Polstelle keine Häufungspunkte haben darf, wissen wir aber, daß es im Innern höchstens endlich viele Polstellen gibt; diese seien bei $z_1, \ldots, z_r$, und ihre Hauptteile seien $H_1(z), \ldots, H_r(z)$. Dann ist

$$f(z) = H_1(z) + \cdots + H_r(z) + g(z)$$

mit einer holomorphen Funktion $g(z)$; nach dem CAUCHYschen Integralsatz ist

$$\int_\gamma f(z)\,dz = \int_\gamma H_1(z)\,dz + \cdots + \int_\gamma H_r(z)\,dz\,,$$

hängt also nur von den Hauptteilen ab.

Tatsächlich können wir noch viel mehr sagen: Genau wie im Reellen hat $(z - z_i)^n$ die Stammfunktion $\dfrac{(z - z_i)^{n+1}}{n+1}$, und zwar für jedes *ganzzahlige* n mit der offensichtliche Ausnahme $n = -1$, für die $n + 1 = 0$ ist. Für $n < 1$ ist diese Stammfunktion für alle $z \neq z_i$ definiert, insbesondere also entlang des gesamten Integrationswegs γ, wo nach Voraussetzung keine Polstellen liegen, und damit können wir das Integral mittels dieser

Stammfunktion ausrechnen. Da Anfangs- und Endpunkt der geschlossenen Kurve γ übereinstimmen, ist

$$\int_\gamma a_{-n}(z - z_i)^{-n}\,dz = 0 \quad \text{für alle } n \neq 1\,.$$

Das Integral über $H_i(z)$ längs der geschlossenen Kurve γ hängt also nur ab vom Term mit $(z - z_0)^{-1}$.

Von daher ist klar, daß der Koeffizient a_{-1} eine besondere Rolle spielt und einen eigenen Namen verdient:

Definition: Für eine meromorphe Funktion f auf $D \subseteq \mathbb{C}$ mit LAURENT-Reihe

$$\sum_{k=-n}^{\infty} a_k (z - z_0)^k$$

um $z = z_0$ heißt der Koeffizient a_{-1} von $(z - z_0)^{-1}$ das *Residuum* von f im Punkt z_0; in Zeichen

$$a_{-1} = \operatorname{Res}_{z_0} f\,.$$

Somit ist

$$\int_\gamma f(z)\,dz = \operatorname{Res}_{z_1} f \cdot \int_\gamma \frac{dz}{z - z_1} + \cdots + \operatorname{Res}_{z_r} f \cdot \int_\gamma \frac{dz}{z - z_r}\,.$$

Um auch noch die verbleibenden Integrale auf der rechten Seite auszurechnen, beginnen wir mit dem Integral über $f(z) = 1/z$ längs eines Kreises um den Nullpunkt.

Im Reellen ist der Logarithmus die Stammfunktion von f, und das gilt im wesentlichen auch im Komplexen; nur haben wir bislang im Komplexen noch gar keinen Logarithmus definiert.

Im Reellen ist die Sache einfach: Die Funktion $x \mapsto e^x$ ist auf ganz $\mathbb{R}$ definiert und ist monoton steigend; ihr Bildbereich sind die positiven reellen Zahlen. Wegen der Monotonie gibt es eine eindeutig definierte Umkehrfunktion $\mathbb{R}_{>0} \to \mathbb{R}$, den natürlichen Logarithmus.

Im Komplexen ist die Funktion $z \mapsto e^z$ ebenfalls auf ganz $\mathbb{C}$ definiert, wahlweise über ihre TAYLOR-Reihe um Null oder die EULERsche Formel

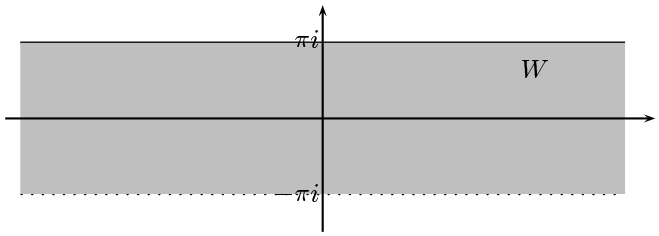
$$e^{x+iy} = e^x(\cos y + i \sin y)\,.$$

Wie letztere zeigt, besteht ihr Bildbereich aus allen komplexen Zahlen außer der Null. Die Formel zeigt aber auch, daß die Exponentialfunktion im Komplexen nicht einmal injektiv ist: Da Sinus und Kosinus Periode 2π haben, hat e^z die Periode $2\pi i$, d.h. $e^{z+2k\pi i} = e^z$ für alle $k \in \mathbb{Z}$. Eine Umkehrfunktion gibt es also nur, wenn wir den Definitionsbereich so einschränken, daß er keine zwei Punkte z, w enthält, die sich um ein (von Null verschiedenes) ganzzahliges Vielfaches von $2\pi i$ unterscheiden.

Ein solcher Bereich ist der Streifen

$$W = \left\{ z = x + iy \in \mathbb{C} \;\middle|\; x \in \mathbb{R} \text{ und } -\pi < y \leq \pi \right\};$$

es gibt also eine Umkehrfunktion $\mathbb{C} \setminus \{0\} \to W$.



Definition: Der Hauptwert $\operatorname{Ln} z$ des natürlichen Logarithmus der komplexen Zahl $z \in \mathbb{C} \setminus \{0\}$ ist die eindeutig bestimmte komplexe Zahl $w \in W$ mit $e^w = z$.

Der so definierte Logarithmus ist aber leider nicht holomorph auf ganz $\mathbb{C} \setminus \{0\}$, denn für eine negative reelle Zahl x ist $\operatorname{Ln} x = \ln|x| + \pi i$, für benachbarte Zahlen der Form $x - i\varepsilon$ mit kleinem $\varepsilon > 0$ aber liegt $\operatorname{Ln}(x - i\varepsilon)$ in der Nähe von $\ln|x| - \pi i$. Der Hauptwert des natürlichen Logarithmus ist also unstetig auf der negativen reellen Achse; überquert man diese, springt er um $2\pi i$.

Wir wollen das Integral längs einer Kreislinie um den Nullpunkt berechnen; dabei müssen wir natürlich die negative reelle Achse überqueren. Wenn wir, wie wir es auch in der HM1 meistens gemacht haben, im Gegenuhrzeigersinn integrieren, haben wir für einen Kreis mit Radius r auf der positiven reellen Achse $\ln r$ als Wert für den Logarithmus; gehen wir dann nach oben, wird ein positiver Imaginärteil addiert, bis wir beim Punkt $-r$ auf der negativen reellen Achse den Logarithmushauptwert $\ln r + \pi i$ erreicht haben. Wenn wir dann weiter den Kreis entlang gehen, sind wir bei Logarithmushauptwerten nahe $\ln r - \pi i$, deren Imaginärteil immer kleiner wird, bis wir auf der positiven reellen Achse bei $\ln r$ zurück sind.

Es ist klar, daß die unstetige Funktion $\operatorname{Ln} z$ nicht entlang der gesamten Kreislinie eine Stammfunktion von $1/z$ ist; wir können Integrale über $1/z$ nur entlang solcher Integrationswege durch den Hauptwert des Logarithmus ausdrücken, entlang derer der Hauptwert holomorph und damit insbesondere stetig ist, d.h. entlang von Wegen, die keinen Schnittpunkt mit der negativen reellen Achse (einschließlich Null) haben.

Ein solcher Integrationsweg ist beispielsweise für jedes $\varepsilon \in (0,\ \pi)$

$$\gamma_\varepsilon\colon \begin{cases} [-\pi + \varepsilon,\ \pi - \varepsilon] \to \mathbb{C} \\ t \mapsto re^{it} \end{cases},$$

d.h. der Kreis mit Radius r um den Nullpunkt, aus dem ein kleiner Bogen um den Schnittpunkt mit der negativen reellen Achse entfernt wurde. Hier können wir das Integral über eine Stammfunktion berechnen als

$$\int_{\gamma_\varepsilon} \frac{dz}{z} = \operatorname{Ln}\left(re^{i(\pi-\varepsilon)}\right) - \operatorname{Ln}\left(re^{i(-\pi+\varepsilon)}\right)$$

$$= \ln r + i(\pi - \varepsilon) - \ln r - i(-\pi + \varepsilon) = 2\pi i - 2\varepsilon i\,.$$

Lassen wir ε gegen Null gehen, konvergiert dies gegen $2\pi i$, das Integral über $1/z$ längs einer Kreislinie um den Nullpunkt, die im Gegenuhrzeigersinn durchlaufen wird, ist also $2\pi i$.

Es ist nun nicht schwierig, sich zu veranschaulichen (und sogar formal zu

beweisen), daß für *jede* geschlossene Randkurve γ eines einfach zusammenhängenden Gebiets, das den Nullpunkt als inneren Punkt enthält,

$$\int_\gamma \frac{dz}{z} = 2\pi i$$

ist, und da $\mathrm{iLn}(z - z + 0)$ Stammfunktion von $1/(z - z_0)$ ist, folgt auch, daß auch für jeden anderen Punkt z_0 im Innern des Gebiets gilt

$$\int_\gamma \frac{dz}{z - z_0} = 2\pi i\,.$$

Zusammen mit den obigen Argumenten führt dies zum

Residuensatz: Die Funktion f sei meromorph in $D \subseteq \mathbb{C}$ und γ sei eine ganz in D liegende Kurve, die Rand eines beschränkten Gebiets G sei und auf der f keine Pole habe. Dann hat f in G nur endlich viele Polstellen $z_1, \ldots, z_r$, und

$$\int_\gamma f(z)\,dz = 2\pi i \sum_{k=1}^{r} \mathrm{Res}_{z_k} f\,.$$

e) Berechnung der Residuen

Die Nützlichkeit dieses Satzes steht und fällt damit, daß wir die auf der rechten Seite auftretenden Residuen gut berechnen können.

Das Residuum einer meromorphen Funktion f an einer Stelle z_0 ist der Koeffizient von $(z - z_0)^{-1}$ in der LAURENT-Enwicklung von f und als solcher zumindest im Prinzip berechenbar. Für die Funktion

$$f(z) = \frac{\sin z}{z^4} = \frac{1}{z^4}\Big(z - \frac{z^3}{6} + \frac{z^5}{120} - \cdots\Big) = z^{-3} - \frac{z^{-1}}{6} + \frac{z}{120} - \cdots$$

etwa ist $\mathrm{Res}_0 f = \frac{1}{6}$. Für die rationalen Funktionen, die wir als Hauptanwendung im Auge haben, ist diese Vorgehensweise aber im allgemeinen recht aufwendig. Hier ist oft eine teilweise Partialbruchzerlegung günstiger, aber im Falle eines Pols erster Ordnung geht alles noch viel einfacher:

In diesem Fall hat die LAURENT-Entwicklung die Form

$$f(z) = \frac{a_{-1}}{z - z_0} + a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \cdots\,,$$

also ist

$$\mathrm{Res}_{z_0} f = a_{-1} = \lim_{z \to z_0} (z - z_0) f(z)\,.$$

Dies funktioniert natürlich nur für Pole erster Ordnung, denn für Pole höherer Ordnung divergiert der rechtsstehende Grenzwert gegen unendlich.

f) Anwendung auf reelle Integrale

Auf den ersten Blick erstaunlich, gerade für Anwendungen in der Elektrotechnik aber wichtig ist die Tatsache, daß sich auch eine ganze Reihe von bestimmten Integrale im Reellen am einfachsten über den Residuensatz berechnen lassen. Betrachten wir zum Beispiel das Integral

$$\int_{-\infty}^{\infty} \frac{dx}{x^4 + 1}\,.$$

Natürlich können wir via Partialbruchzerlegung eine Stammfunktion des Integranden finden, allerdings müssen wir dafür doch einiges arbeiten, und das Ergebnis

$$F(x) = \frac{\sqrt{2}}{8} \ln \frac{x^2 + \sqrt{2}\,x + 1}{x^2 - \sqrt{2}\,x + 1} + \frac{\sqrt{2}}{4} \arctan(\sqrt{2}\,x + 1) + \frac{\sqrt{2}}{4} \arctan(\sqrt{2}\,x - 1)$$

ist alles andere als angenehm.

Um auch dieses Integral über den Residuenkalkül ausrechnen zu können, setzen wir den Integranden fort zu einer komplexen Funktion

$$f(z) = \frac{1}{z^4 + 1}\,;$$

diese ist holomorph in allen Punkten $z \in \mathbb{C}$, in denen der Nenner $z^4 + 1$ nicht verschwindet.

Nach der dritten binomischen Formel ist $(z^4+1)(z^4-1) = (z^8-1)$, also

$$z^4 + 1 = \frac{z^8 - 1}{z^4 - 1}\,.$$

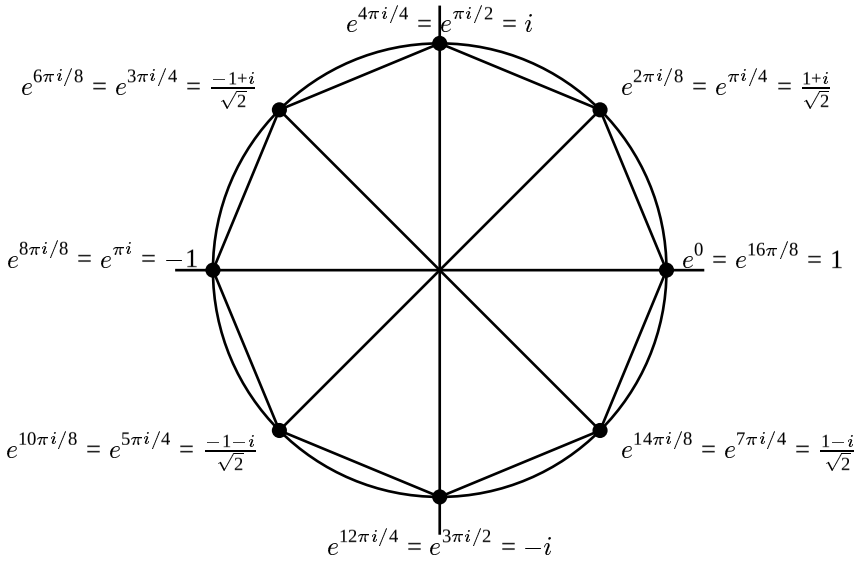
Die Nullstellen des Polynoms $z^n - 1$ sind jene komplexen Zahlen, deren n-te Potenz gleich Eins ist; man bezeichnet sie als die n-ten Einheitswurzeln. Da ein Polynom vom Grad n über einem Körper höchstens n Nullstellen haben kann, kann es höchstens n von ihnen geben; da für jede natürliche Zahl k

$$\left(e^{2k\pi i/n}\right)^n = e^{n\cdot 2k\pi i/n} = e^{2k\pi i} = 1$$

ist, gibt es genau die n Einheitswurzeln

$$1 = e^0, \quad e^{2k\pi i/n}, \quad e^{4k\pi i/n}, \quad \ldots, \quad e^{(n-1)\cdot 2\pi i/n}.$$

Auf dem Einheitskreis sind sie die Eckpunkte eines regelmäßigen n-Ecks, was die folgende Zeichnung für den Fall $n = 8$ illustriert:



Ist m ein Teiler von n, so ist jede m-te Einheitswurzel erst recht eine m-ten Einheitswurzeln; wir bezeichnen eine n-te Einheitswurzel als *primitiv*, wenn es keinen echten Teiler m von n gibt, für den sie bereits m-te Einheitswurzel ist.

Eine achte Einheitswurzel ist offenbar genau dann primitiv, wenn sie nicht gleichzeitig vierte Einheitswurzel ist; die Nullstellen von $z^4 + 1$ sind also genau die primitiven achten Einheitswurzeln

$$e^{\pi i/4}, \quad e^{3\pi i/4}, \quad e^{5\pi i/4} \quad \text{und} \quad e^{7\pi i/4}.$$

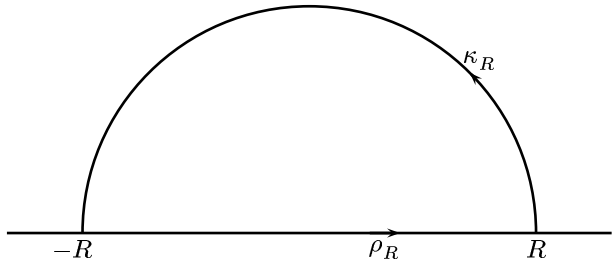
Wir betrachten nun für $R > 1$ einen Integrationsweg γ_R, der zusammengesetzt ist aus dem eigentlich interessierenden reellen Integrationsweg

$$\rho_R\colon \begin{cases} [-R, R] \to \mathbb{C} \\ \qquad t \mapsto t \end{cases}$$

von $-R$ bis R und einem Halbkreis

$$\kappa_R\colon \begin{cases} [0, \pi] \to \mathbb{C} \\ \qquad t \mapsto Re^{it} \end{cases}$$

in der oberen Halbebenen von $\mathbb{C}$, der von R im Gegenuhrzeigersinn nach $-R$ führt.



Beides zusammen bildet eine geschlossene Kurve, und Polstellen des Integranden gibt es nur bei den primitiven achten Einheitswurzeln, von denen allerdings nur $e^{\pi i/4}$ und $e^{3\pi i/4}$ im Halbkreis liegen. Nach dem

Residuensatz ist daher für $R > 1$

$$\int\limits_{\gamma_R} \frac{dz}{z^4+1} = 2\pi i\big(\mathrm{Res}_{e^{\pi i/4}} f + \mathrm{Res}_{e^{3\pi i/4}} f\big)\,.$$

Die Residuen lassen sich wie oben bestimmen, beispielsweise ist

$$\begin{aligned}
\mathrm{Res}_{e^{\pi i/4}} f &= \lim_{z\to e^{\pi i/4}} \frac{z-e^{\pi i/4}}{z^4+1} \\
&= \lim_{z\to e^{\pi i/4}} \frac{z-e^{\pi i/4}}{(z-e^{\pi i/4})(z+e^{\pi i/4})(z-e^{3\pi i/4})(z+e^{3\pi i/4})} \\
&= \frac{1}{2e^{\pi i/4}(e^{\pi i/4}-e^{3\pi i/4})(e^{\pi i/4}+e^{3\pi i/4})} \\
&= \frac{1}{2e^{\pi i/4}(e^{\pi i/2}-2e^{3\pi i/2})} \\
&= \frac{1}{2e^{\pi i/4}\big(i-(-i)\big)} = \frac{e^{-\pi i/4}}{4i} = \frac{\frac12(\sqrt2-\sqrt2\,i)}{4i} \\
&= \frac{\sqrt2}{8}(-1-i)\,,
\end{aligned}$$

und genauso könnte man auch $\mathrm{Res}\, e^{\pi i/4} f = \frac{\sqrt2}{8}(1-i)$ berechnen. Eine Alternative wäre die Regel von DE L'HOSPITAL: Danach ist

$$\begin{aligned}
\mathrm{Res}_{e^{\pi i/4}} f &= \lim_{z\to e^{\pi i/4}} \frac{z-e^{\pi i/4}}{z^4+1} \\
&= \lim_{z\to e^{\pi i/4}} \frac{1}{4z^3} = \frac{1}{4e^3\pi i/4} = \frac14 e^{-3\pi i/4} \qquad \text{und} \\
\mathrm{Res}_{e^{3\pi i/4}} f &= \lim_{z\to e^{3\pi i/4}} \frac{z-e^{3\pi i/4}}{z^4+1} \\
&= \lim_{z\to e^{3\pi i/4}} \frac{1}{4z^3} = \frac{1}{4e^{9\pi i/4}} = \frac14 e^{-9\pi i/4} = \frac14 e^{-\pi i/4}\,,
\end{aligned}$$

was zumindestest für dieses f geringfügig einfache war als die obige Rechnung. Beide zusammen ergeben

$$\mathrm{Res}_{e^{\pi i/4}} f + \mathrm{Res}_{e^{3\pi i/4}} f = \frac{e^{-3\pi i/4}+e^{-\pi i/4}}{4}\,.$$

Um dies weiter auszurechnen kann man entweder, in diesem Fall sehr einfach, die Polarkoordinatendarstellungen $e^{-3\pi i/4}$ und $e^{-\pi i/4}$ verwenden, oder aber man versucht, die Summe über die EULERschen Formeln als trigonometrische Funktion zu interpretieren:

$$\begin{aligned}
\frac{e^{-3\pi i/4}+e^{-\pi i/4}}{4} &= e^{-\pi i/2}\cdot\frac{e^{-\pi i/4}+e^{\pi i/4}}{4} = \frac{e^{-\pi i/2}}{2}\cdot\cos\frac{\pi}{4} \\
&= -\frac{i}{2}\cdot\frac{\sqrt2}{2} = -\frac{i\sqrt2}{4}
\end{aligned}$$

Damit ist also

$$\int\limits_{\gamma_R} \frac{dz}{z^4+1} = 2\pi i\big(\mathrm{Res}\, e^{\pi i/4} f + \mathrm{Res}_{e^{3\pi i/4}} f\big) = \pi\frac{\sqrt2}{2}\,.$$

Was uns wirklich interessiert ist allerdings nicht das Integral über γ_R, sondern das über ρ_R. Wir können es aus dem über γ_R berechnen, wenn wir das Integral über den Halbkreisbogen κ_R kennen. Dieses Integral kennen wir zwar nicht, aber wir können uns leicht überlegen, daß es für $R\to\infty$ gegen Null geht:

$$\int\limits_{\kappa_R} \frac{dz}{z^4+1} = \int\limits_0^\pi \frac{iRe^{it}}{R^4e^{4it}+1}\,dt\,,$$

und da der Integrand rechts für $R\to\infty$ überall gegen Null geht, gilt dasselbe auch für das Integral über das endliche Intervall $[0,\pi]$. Somit ist

$$\begin{aligned}
\int\limits_{-\infty}^{\infty} \frac{dz}{z^4+1} &= \lim_{R\to\infty}\int\limits_{-R}^{R} \frac{dz}{z^4+1} \\
&= \lim_{R\to\infty}\int\limits_{\gamma_R} \frac{dz}{z^4+1} - \lim_{R\to\infty}\int\limits_{\kappa_R} \frac{dz}{z^4+1} = \frac{\sqrt2}{2}\pi\,.
\end{aligned}$$

Ähnlich kann man bei anderen Integralen von $-\infty$ nach ∞ vorgehen, vorausgesetzt der Integrand hat keine Polstellen auf der reellen Achse, die Residuen der Polstellen in der oberen Halbebene sind bekannt und für

das Integral über den Halbkreisbogen κ_R kann zumindest der Grenzwert für $R \to \infty$ bestimmt werden.

Für uns interessant ist vor allem der folgende Fall:

Satz: $f(x) = P(x)/Q(x)$ sei Quotient zweier Polynome $P(x)$ und $Q(x)$ mit $\deg P \leq \deg Q - 2$, und $Q(x)$ habe keine reelle Nullstelle. Dann ist

$$\int_{-\infty}^{\infty} f(x)\,dx = 2\pi i \sum_{k=1}^{r} \operatorname{Res}_{z_i} f\,,$$

wobei $z_1, \ldots, z_r$ die komplexen Nullstellen von Q mit positivem Imaginärteil sind.

Beweis: $Q(z)$ als Polynom kann höchstens $\deg Q$ Nullstellen haben, insbesondere also nur endlich viele. Unter diesen seien z_1 bis z_r diejenigen mit positivem Imaginärteil, und R sei eine reelle Zahl, die größer sei als die Beträge aller z_i. Dazu betrachten wir, wie oben im Beispiel, einen Integrationsweg γ_R, der zusammengesetzt ist aus dem eigentlich interessierenden reellen Integrationsweg

$$\rho_R\colon \begin{cases} [-R, R] \to \mathbb{C} \\ \qquad t \mapsto t \end{cases}$$

von $-R$ bis R und einem Halbkreis

$$\kappa_R\colon \begin{cases} [0, \pi] \to \mathbb{C} \\ \qquad t \mapsto Re^{it} \end{cases}$$

in der oberen Halbebenen von $\mathbb{C}$, der von R im Gegenuhrzeigersinn nach $-R$ führt. Nach dem Residuensatz ist dann

$$\int_{\gamma_R} f(z)\,dz = 2\pi i \sum_{k=1}^{r} \operatorname{Res}_{z_i} f\,.$$

Andererseits ist

$$\int_{\gamma_R} f(z)\,dz = \int_{\rho_r} f(z)\,dz + \int_{\kappa_R} f(z)\,dz = \int_{-R}^{R} f(x)\,dx + \int_{\kappa_R} f(z)\,dz\,,$$

also

$$\int_{-R}^{R} f(x)\,dx = 2\pi i \sum_{k=1}^{r} \operatorname{Res}_{z_i} f - \int_{\kappa_R} f(z)\,dz\,.$$

Dabei ist

$$\int_{\kappa_R} f(z)\,dz = \int_{\kappa_R} \frac{P(z)}{Q(z)}\,dz = \int_{0}^{\pi} \frac{P(Re^{it}) \cdot iRe^{it}}{Q(Re^{it})}\,dt\,,$$

wobei der Zähler rechts trotz des zusätzlichen Faltors R immer noch kleineren Grad hat als der Nenner. Somit geht der Integrand unabhängig von t für $R \to \infty$ gegen null. Da das Integrationsintervall von 0 bis 2π trotz wachsendem R immer dasselbe bleibt, geht dann auch das Integral gegen null.

Für $R \to \infty$ wird die obige Formel daher zu

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\,dx &= \lim_{R\to\infty} \int_{-R}^{R} f(x)\,dx \\ &= 2\pi i \sum_{k=1}^{r} \operatorname{Res}_{z_i} f - \lim_{R\to\infty} \int_{\kappa_R} f(z)\,dz \\ &= 2\pi i \sum_{k=1}^{r} \operatorname{Res}_{z_i} f\,, \end{aligned}$$

wie behauptet. ∎

Die Nützlichkeit des Residuenkalküls für reelle Integrale ist nicht auf rationale Integranden beschränkt; beispielsweise läßt sich damit (auf Umwegen) auch ein Integral wie

$$\int_{-\infty}^{\infty} \frac{\sin x}{x}\,dx = \pi$$

berechnen, wobei hier die Stammfunktion des Integranden nicht einmal elementar angebbar ist. Da sie trotzdem beispielsweise in der Elektrotechnik eine wichtige Rolle spielt, definiert man sie als den sogenannten *Integralsinus*

$$\mathrm{Si}(x) = \int_0^x \frac{\sin t}{t}\, dt\,;$$

da $\dfrac{\sin t}{t}$ eine gerade Funktion ist, zeigt die obige Formel somit, daß $\mathrm{Si}(x)$ für $x \to \infty$ gegen $\dfrac{\pi}{2}$ konvergiert.

§ 2: Reelle und komplexe Fourier-Reihen

Wir beginnen mit einem einfachen und anschaulichen Beispiel für den Aufbau einer komplizierten Funktion aus reinen Schwingungen; Ziel des Paragraphen wird sein, eine (fast) *beliebige* periodische Funktion möglichst exakt als Überlagerung solcher Schwingungen darzustellen.

a) Die schwingende Saite

Ein Ton werde erzeugt durch eine schwingende Saite. Wir wollen der Einfachheit halber annehmen, daß diese ausschließlich senkrecht zu ihrer Ruhelage schwingt und daß ihre Schwingung auf eine feste Ebene begrenzt ist; die Physiker bezeichnen dies als eine transversale linear polarisierte Schwingung. Zumindest in erster Näherung kann man, bei nicht zu extremer Auslenkung der Saiten, einige Musikinstrumente so beschreiben.

Der Zustand der Saite zu einem *festen Zeitpunkt* wird beschrieben durch eine Funktion der Längenkoordinate, die wir wie üblich mit x bezeichnen wollen. Der Wert dieser Funktion an jeder Stelle x ist aber, da die Saite schwingt, auch eine Funktion der Zeit. Wir haben also insgesamt eine Funktion $f(x,t)$ sowohl der Längenkoordinate als auch der Zeit, die angibt, wie weit der Punkt mit Längenkoordinate x zum Zeitpunkt t von seiner Ruhelage entfernt ist. Falls wir annehmen, daß die Schwingung in der (x,y)-Ebene stattfindet, ist $f(x,t)$ also die y-Koordinate des Punktes

mit Längenkoordinate x zum Zeitpunkt t. Da wir nur transversale linear polarisierte Schwingungen betrachten, hat dieser Punkt die Koordinaten

$$\big(x, f(x,t)\big)\,;$$

falls wir auch longitudinale Schwingungen zugelassen hätten, würde auch die x-Koordinate von der Zeit abhängen, und falls wir uns nicht auf linear polarisierte Schwingungen festgelegt hätten, gäbe es noch eine z-Koordinate.

Die Saite ist an beiden Enden fest eingespannt; wir wählen die Koordinaten auf der x-Achse so, daß diese Enden den Werten $x = 0$ und $x = L$ entsprechen, wobei $L \in \mathbb{R}$ die Länge der Saite bezeichnet. Da die Enden nicht schwingen können, muß notwendigerweise

$$f(0,t) = 0 \quad \text{und} \quad f(L,t) = 0$$

sein; nur für $0 < x < L$ kann $f(x,t)$ wirklich von t abhängen.

Wie könnte f aussehen? In ihrer Ruhelage ist die Saite eine Strecke; die einfachste Form einer Schwingung könnte darin bestehen, daß diese Strecke durch einen Teil einer Sinuslinie ersetzt wird. Da die Funktion an den Stellen 0 und L verschwinden muß und der Sinus bei allen ganzzahligen Vielfachen von π verschwindet, kommen daher Funktionen der Art

$$f(x,t) = A(t) \cdot \sin\left(\frac{k\pi}{L}\, x\right)$$

in Frage, wobei $A(t)$ irgendeine Funktion der Zeit ist und k eine natürliche Zahl. Abbildung zwei zeigt die entsprechenden Funktionen für $k = 1$ bis 4 und $A(t) \equiv 1$.

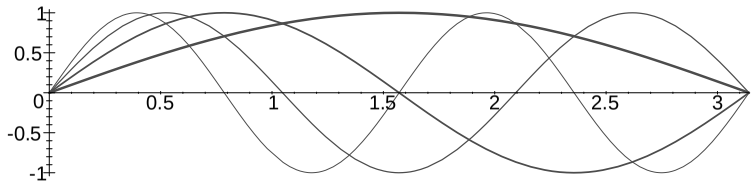


Abb. 2: Eine Schwingung mit Oberschwingungen

Bei einem realen Musikinstrument werden diese Oberschwingungen nicht alle dieselbe Amplitude haben; aus Kapitel I, §1*f)* etwa ist das Beispiel der g-Saite einer Geige bekannt, das hier noch einmal in Abbildung drei dargestellt ist: Die gestrichelte Kurve ist die Grundschwingung mit Amplitude eins, die fett eingezeichnete Kurve die Gesamtschwingung, und die sonstigen Kurven sind die reinen Teilschwingungen mit ihren jeweiligen Amplituden. (Wer selbst solche Kurven konstruieren und auch die dazugehörigen Töne hören möchte, findet ein Java-Applet unter `http://www.gac.edu/~huber/fourier/`)

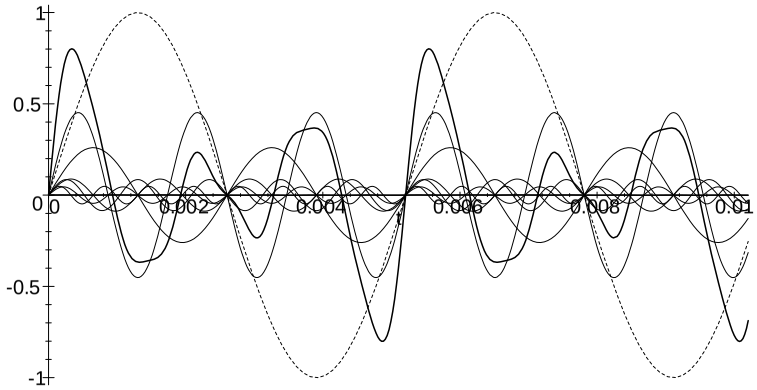


Abb. 3: Ton der g-Saite einer Geige und seine Komponenten

Offensichtlich spielt die Grundschwingung kaum eine Rolle: Wie man sowohl hier als auch genauer an der Darstellung der Größenverhältnisse der Koeffizienten in Abbildung vier sieht, sind die Schwingungen mit doppelter und dreifacher Grundfrequenz am stärksten ausgeprägt, d.h. also die Oktave und vor allem die darüberliegende Quinte.

Über die zeitabhängige Auslenkungsfunktion $A(t)$ wurde bislang noch nichts gesagt; da wir periodische Schwingungen erwarten, liegt es nahe, auch hier einen Ansatz mit trigonometrischen Funktionen zu machen. Wenn wir die Zeitachse so festlegen, daß sich die Saite zum Zeitpunkt $t = 0$ in Ruhelage befindet, ist der Sinus die geeignete Funktion; wir

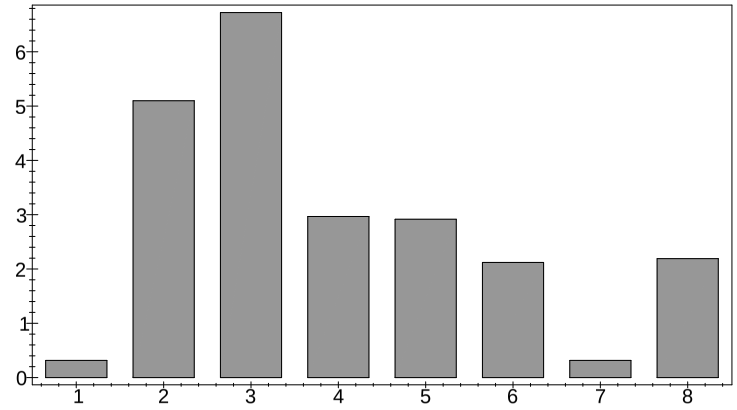


Abb. 4: Koeffizienten von Grund- und Oberschwingungen

versuchen es daher für eine reine Schwingung mit einem Ansatz der Form

$$A(t) = a \cdot \sin \omega t \,,$$

wobei a und ω reelle Konstanten sind, von denen wir zumindest ω als positiv voraussetzen können. Insgesamt erwarten wir somit im einfachsten Fall Funktionen der Form

$$f(x,t) = a \cdot \sin \omega t \cdot \sin\left(\frac{k\pi}{L}\, x\right) \,.$$

ω und k hängen natürlich voneinander ab: Wie jedermann aus Physik- und Musikunterricht weiß, führt ein doppelt so großer Wert von k zu einer doppelt so hohen Kreisfrequenz ω.

b) Die Differentialgleichung der schwingenden Saite

Wer sich allerdings kurz überlegt, *warum* dem so ist, wird wohl in den meisten Fällen nur auf den „Grund“ kommen, daß dies eben allgemein bekannt sei. Tatsächlich stecken dahinter einige nicht ganz triviale Überlegungen, die man für die Zwecke dieser Vorlesung zwar nicht unbedingt kennen muß, die aber für etwaige Interessenten trotzdem im Kleindruck beigefügt sind:

Da wir alles so einfach wie möglich halten wollen, gehen wir aus von einer Saite mit konstantem Querschnitt und konstanter Massendichte; letztere können wir dann beschreiben durch die Masse pro Längeneinheit, die für konkrete Saiten gemessen wird in Gramm pro Zentimeter oder Milligramm pro Zentimeter. Wir bezeichnen diese (lineare) Massendichte mit σ.

Die zweite wesentliche physikalische Größe für eine schwingende Saite ist deren *Spannung.* Auch hier beschränken wir uns wieder auf das einfachste physikalische Modell, in dem das HOOKEsche Gesetz gilt: Wir betrachten die Saite als eine elastische Feder, die eine natürliche Länge L_0 hat. Da sie aber in ein Musikinstrument eingespannt ist, wurde sie auf eine Länge $L > L_0$ gedehnt; nach dem HOOKEschen Gesetz wirkt somit eine Rückstellkraft $\lambda L/L_0$, die proportional ist zur Überdehnung L/L_0 mit der Federkonstanten λ als Proportionalitätsfaktor.

In der Ruhelage ist diese Rückstellkraft bedeutungslos: Da die Saite an beiden Enden fest eingespannt ist, kann sie ihre Länge nicht verringern. Anders wird es, wenn die Saite aus der Ruhelage entfernt wird: Dann hat die Federkraft in allen Punkten, an denen die (Tangente der) Saite nicht parallel zur x-Achse ist, auch eine Kraftkomponente in y-Richtung.

Die Lage der ausgelenkten Saite zu einem festen Zeitpunkt t wird beschrieben durch die Funktion

$$g(x) = f(x,t)\,,$$

die die y-Koordinate des Punkts x angibt.

Betrachten wir das Saitenstück zwischen $x = x_1$ und $x = x_2$. Im Punkt x_1 habe die Tangente den Winkel α gegenüber der Horizontalen, im Punkt x_2 sei dieser Winkel β. Falls x_1 und x_2 einigermaßen nahe beieinander liegen, können wir die Saite zwischen x_1 und x_2 in erster Näherung als eine Gerade betrachten. Diese Gerade sei um den Winkel γ gegenüber der Horizontalen geneigt; dann hat das Stück zwischen $x = x_1$ und $x = x_2$ die Länge

$$\frac{x_2 - x_1}{\cos\gamma}\,,$$

denn der Kosinus eines Winkels im rechtwinkligen Dreieck ist gleich Ankathete durch Hypothenuse. Gegenüber ihrer Ruhelage ist die ausgelenkte Saite daher noch um einen weiteren (lokalen) Faktor $1/\cos\gamma$ gestreckt.

Die Rückstellkraft *in Richtung der ausgelenkten Saite* ist daher gleich

$$\lambda \cdot \frac{L}{L_0} \cdot \frac{1}{\cos\gamma}\,,$$

und die Komponente in y-Richtung ist im Punkt x_1 gleich

$$\lambda \cdot \frac{L}{L_0} \cdot \frac{1}{\cos\gamma} \cdot \sin\alpha$$

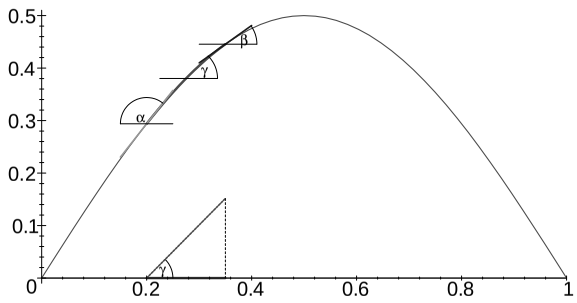


Abb. 5: Eine ausgelenkte Saite

und im Punkt x_2 entsprechend gleich

$$-\lambda \cdot \frac{L}{L_0} \cdot \frac{1}{\cos\gamma} \cdot \sin\beta\,.$$

Da α und β bei gängigen Musikinstrumenten ziemlich klein sind, machen wir kein großen Fehler, wenn wir die Näherungsformeln

$$\sin x \approx x \approx \tan x$$

für im Bogenmaß gemessene Winkel x benutzen; demnach können wir also den Sinus in obigen Formeln ohne großen Genauigkeitsverlust durch einen Tangens ersetzen.

Der Tangens des Steigungswinkels der Tangenten des Graphs einer Funktion $y = g(x)$ gegenüber der Horizontalen ist gleich der Ableitung $g'(x)$, die Rückstellkräfte im Anfangs- und im Endpunkt des betrachteten Kurvenstücks sind also näherungsweise gleich

$$-\lambda \cdot \frac{L}{L_0} \cdot \frac{1}{\cos\gamma} \cdot g'(x_1) \quad \text{und} \quad \lambda \cdot \frac{L}{L_0} \cdot \frac{1}{\cos\gamma} \cdot g'(x_2)\,.$$

Bei einem hinreichen kleinen Saitenstück ist die resultierende Rückstellkraft gleich der Summe dieser beiden Komponenten, also

$$\lambda \cdot \frac{L}{L_0} \cdot \frac{1}{\cos\gamma} \cdot g'(x_2) - \lambda \cdot \frac{L}{L_0} \cdot \frac{1}{\cos\gamma} \cdot g'(x_1) = \lambda \cdot \frac{L}{L_0} \cdot \frac{1}{\cos\gamma} \cdot \Big(g'(x_2) - g'(x_1)\Big)\,.$$

Da wir von sehr kleinen Winkeln γ ausgehen, liegt $\cos\gamma$ sehr nahe bei eins und kann, da wir hier ohnehin nur näherungsweise argumentieren, gleich eins gesetzt werden; die Rückstellkraft ist also näherungsweise gleich

$$\lambda \cdot \frac{L}{L_0} \cdot \Big(g'(x_2) - g'(x_1)\Big)\,.$$

Diese Kraft bedingt nach dem zweiten NEWTONschen Gesetz eine Bewegung der Massenpunkte auf der Saite. Ein solcher Massenpunkt mit x-Koordinate x_0 hat eine zeitabhängige Auslenkung

$$h(t) = f(x_0, t)\,,$$

und für die Kraft, die dies bewirkt, gilt nach dem zweiten NEWTONschen Gesetz

$$\text{Kraft} = \text{Masse} \times \text{Beschleunigung} .$$

Bei einem hinreichend kleinen Saitenstück können wir die Masse näherungsweise gleich der Masse,

$$m = \sigma \cdot (x_2 - x_1)$$

des gesamten Stück setzen. Die Beschleunigung ist gleich der zweiten Ableitung $\ddot{h}(t)$ der Auslenkung, also gilt insgesamt

$$\lambda \cdot \frac{L}{L_0} \cdot \Big(g'(x_2) - g'(x_1) \Big) = \sigma \cdot (x_2 - x_1) \cdot \ddot{h}(t)$$

oder

$$\frac{g'(x_2) - g'(x_1)}{x_2 - x_1} = \frac{\sigma L_0}{\lambda L} \ddot{h}(t) .$$

Lassen wir nun x_2 und x_1 simultan gegen einen dazwischenliegenden Punkt x gehen, konvergiert die linke Seite gegen $g''(x)$, wir bekommen also die Gleichung

$$g''(x) = \frac{\sigma L_0}{\lambda L} \ddot{h}(t) .$$

Damit sind wir fast fertig; wir müssen uns nur noch klarmachen, daß die beiden Funktionen $g(x)$ und $h(t)$ spezielle Werte der Funktion $f(x,t)$ berechnen: Für einen oben festgehaltenen (aber nicht weiter spezifizierten) Zeitpunkt t ist $g(x) = f(x,t)$, und für einen ebenfalls festgehaltenen (aber nicht weiter spezifizierten) Punkt x auf der Saite ist $h(t) = f(x,t)$. Daher ist

$$g''(x) = f_{xx}(x,t) \quad \text{und} \quad \ddot{h}(t) = f_{tt}(x,t) ,$$

und die Differentialgleichung der schwingenden Saite wird zu

$$f_{xx}(x,t) = \frac{\sigma L_0}{\lambda L} f_{tt}(x,t)$$

oder, wie man meist schreibt,

$$f_{tt}(x,t) = \frac{\lambda L}{\sigma L_0} f_{xx}(x,t) .$$

Da es uns auf exakte Zahlenwerte nicht ankommt, wählen wir noch eine Abkürzung für den Bruch; da er positiv ist, können wir ihn als Quadrat schreiben und definieren

$$c^2 \underset{\text{def}}{=} \frac{\lambda L}{\sigma L_0} .$$

Mit dieser neuen Bezeichnung wird die Differentialgleichung der schwingenden Saite zu

$$f_{tt}(x,t) = c^2 f_{xx}(x,t) .$$

Sie allein legt $f(x,t)$ bei weitem noch nicht eindeutig fest: Sind φ und ψ *irgendwelche* zweifach stetig differenzierbare Funktionen einer Veränderlichen, so überzeugt man sich leicht (Kettenregel), daß

$$f(x,t) = \varphi(x - ct) + \psi(x + ct)$$

eine Lösung dieser Gleichung ist, die sogenannte D'ALEMBERTsche Lösung. Sie zeigt, daß man die Konstante c interpretieren kann als Schallgeschwindigkeit *innerhalb der Saite;* die beiden Terme beschreiben Erregungen, die sich gegenläufig auf der Saite fortbewegen.

Wir waren oben ausgegangen von speziellen sinusförmigen Lösungen der Form

$$f(x,t) = A(t) \cdot \sin\left(\frac{k\pi}{L} x\right) = a \cdot \sin \omega t \cdot \sin\left(\frac{k\pi}{L} x\right)$$

und müssen nun sehen, für welche Parameterwerte dies Lösungen sind.

Die partiellen Ableitungen von f sind

$$f_t(x,t) = a\omega \cdot \cos \omega t \cdot \sin\left(\frac{k\pi}{L} x\right)$$

$$f_{tt}(x,t) = -a\omega^2 \cdot \sin \omega t \cdot \sin\left(\frac{k\pi}{L} x\right) = -\omega^2 f(x,t)$$

$$f_x(x,t) = a \cdot \sin \omega t \cdot \left(\frac{k\pi}{L}\right) \cos\left(\frac{k\pi}{L}\right)$$

$$f_{xx}(x,t) = -a \cdot \sin \omega t \cdot \left(\frac{k\pi}{L}\right)^2 \sin\left(\frac{k\pi}{L} x\right) = -\left(\frac{k\pi}{L}\right)^2 f(x,t) ,$$

also ist

$$f_{tt}(x,t) = \left(\frac{k\pi}{\omega L}\right)^2 f_{xx}(x,t) .$$

Die Differentialgleichung ist somit genau dann erfüllt, wenn

$$c = \frac{k\pi}{\omega L} \quad \text{oder} \quad \omega = k \cdot \frac{\pi}{cL} .$$

Wie diese Rechnung zeigt, wächst zumindest für die hier betrachteten einfachen Schwingungen die Frequenz in der Tat linear mit k, die Frequenzen der Obertöne sind also ganzzahlige Vielfache der Grundfrequenz.

c) Orthogonalitätsrelationen

Wie eingangs erwähnt, wollen wir in diesem Paragraphen (fast) beliebige peridodische Funktionen durch Linearkombinationen von reinen Schwingungen beschreiben; bevor wir damit beginnen, müssen wir uns zunächst überlegen, welche Funktionen genau wir betrachten wollen.

Wir dürfen uns auf keinen Fall nur auf stetige Funktionen beschränken; Rechteckimpulse beispielsweise spielen eine sehr wichtige Rolle in der

Informationstechnik. Andererseits wollen wir aber nicht soweit gehen, auch Funktionen wie

$$f(t) = \begin{cases} \sin t & \text{für rationale } t \\ \cos t & \text{für irrationale } t \end{cases}$$

zu betrachten, wir müssen also einen Kompromiss finden.

Definition: Eine Funktion $f: \mathbb{R} \to \mathbb{R}$ oder $f: \mathbb{R} \to \mathbb{C}$ heißt *stückweise stetig,* wenn die Menge aller Unstetigkeitsstellen von f keine Häufungspunkte hat und wenn für jede Unstetigkeitsstelle t die links- und rechtsseitigen Grenzwerte

$$\lim_{s \to t^-} f(t) \quad \text{und} \quad \lim_{s \to t^+} f(t)$$

existieren. (Wir verlangen nicht, daß sie gleich sind: Das gilt nur, wenn f an der Stelle t stetig ist.)

Damit sind also Rechteckimpulse und allgemeiner alle Funktionen, die bis auf isolierte Sprungstellen stetig sind, stückweise stetig.

Auch periodische Funktionen sollten wir vorsichtshalber zumindest einmal formal definieren:

Definition: Eine Funktion $f: \mathbb{R} \to \mathbb{R}$ oder $f: \mathbb{R} \to \mathbb{C}$ heißt *periodisch zur Periode T,* wenn für alle $t \in \mathbb{R}$ gilt:

$$f(t + T) = f(t)\,.$$

Ist f peridiodisch zur Periode T, so offensichtlich auch zur Periode $2T$ oder $-5T$ *usw.*; falls es einen kleinsten positive Wert T gibt, zu dem f periodisch ist, bezeichnen wir diesen als *die* Periode von f. In diesem Sinne haben also $\sin t$ und $\cos t$ *die* Periode 2π, wohingegen konstante Funktionen für jedes $T \in \mathbb{R}$ periodisch zur Periode T sind, so daß wir hier nicht von *der* Periode reden können.

Eine periodische Funktion ist eindeutig bestimmt durch ihre Werte in irgendeinem abgeschlossenen Intervall J der Länge T, denn für jedes $t \in \mathbb{R}$ gibt es ein $k \in \mathbb{Z}$, so daß $t - kT \in J$, und wegen der Periodizität muß $f(t) = f(t - kT)$ sein. Ein solches Intervall der Länge T bezeichnen wir kurz als ein *Periodenintervall.*

Da jede unendliche Menge in einem abgeschlossenen Intervall $[a, b]$ einen Häufungspunkt hat, kann eine stückweise stetige Funktion in jedem solchen Intervall höchstens endlich viele Unstetigkeitsstellen haben; insbesondere gibt es also bei einer stückweise stetigen periodischen Funktion in jedem Periodenintervall höchstens endlich viele Unstetigkeitsstellen, und durch diese sind *alle* Unstetigkeiten der Funktion festgelegt.

Wir betrachten im folgenden für jede reelle Zahl $T > 0$ die beiden Mengen

$$\mathrm{L}_T(\mathbb{R}, \mathbb{R}) = \left\{ f: \mathbb{R} \to \mathbb{R} \;\middle|\; \begin{array}{l} f \text{ stückweise stetig und} \\ \text{periodisch zur Periode } T \end{array} \right\}$$

und

$$\mathrm{L}_T(\mathbb{R}, \mathbb{C}) = \left\{ f: \mathbb{R} \to \mathbb{C} \;\middle|\; \begin{array}{l} f \text{ stückweise stetig und} \\ \text{periodisch zur Periode } T \end{array} \right\}.$$

Da Linearkombinationen periodischer Funktionen zur selben Periode T wieder periodisch mit T sind und die Nullfunktion periodisch ist zu jeder Periode, ist $\mathrm{L}_T(\mathbb{R}, \mathbb{R})$ ein $\mathbb{R}$-Vektorraum und $\mathrm{L}_T(\mathbb{R}, \mathbb{C})$ ein $\mathbb{C}$-Vektorraum.

Da auch das Produkt zweier stückweise stetiger Funktionen stückweise stetig ist, existiert für $f, g \in \mathrm{L}_T(\mathbb{R}, \mathbb{R})$ das Integral

$$(f, g) \underset{\text{def}}{=} \frac{1}{T} \int_0^T f(t) g(t)\, dt\,,$$

und das so definierte Produkt hat *fast* alle Eigenschaften eines Skalarprodukts: Symmetrie und Bilinearität sind klar, und da Quadrate in $\mathbb{R}$ stets nichtnegativ sind, ist auch für alle $f \in \mathrm{L}_T(\mathbb{R}, \mathbb{R})$

$$(f, f) = \frac{1}{T} \int_0^T f(t)^2\, dt \geq 0\,.$$

Leider kann aber für eine nur stückweise stetige Funktion $(f, f) = 0$ sein, ohne daß f gleich der Nullfunktion wäre, beispielsweise für

$$f: \mathbb{R} \to \mathbb{R}; \quad t \mapsto \begin{cases} 1 & \text{falls } t = kT \text{ mit } k \in \mathbb{Z} \\ 0 & \text{sonst} \end{cases}.$$

Damit ist $L_T(\mathbb{R},\mathbb{R})$ kein EUKLIDischer Vektorraum; das gerade eingeführte Produkt wird uns aber trotzdem im folgenden sehr nützlich sein.

Entsprechend definieren wie auf $L_T(\mathbb{R},\mathbb{C})$ ein Produkt durch

$$(f,g) \underset{\text{def}}{=} \frac{1}{T}\int_0^T f(t)\overline{g(t)}\,dt\,;$$

es hat alle Eigenschaften eines HERMITEschen Skalarprodukts *außer* der positiven Definitheit.

Wenn wir wirklich EUKLIDische oder HERMITEsche Vektorräume wollen, können wir uns auf die Unterräume

$$L_T^0(\mathbb{R},\mathbb{R}) = \left\{ f \in L_T(\mathbb{R},\mathbb{R}) \;\middle|\; f(t) = \frac{1}{2}\left(\lim_{s\to t^+} f(s) + \lim_{s\to t^-} f(s)\right)\right\}$$

und

$$L_T^0(\mathbb{R},\mathbb{C}) = \left\{ f \in L_T(\mathbb{R},\mathbb{C}) \;\middle|\; f(t) = \frac{1}{2}\left(\lim_{s\to t^+} f(s) + \lim_{s\to t^-} f(s)\right)\right\}$$

beschränken, für deren Elemente in jeder Sprungstelle t der Funktionswert gleich dem arithmetischen Mittel aus links- und rechtsseitigem Grenzwert ist; sie bilden offensichtlich einen Untervektorraum, und jedes Element, das von der Nullfunktion verschieden ist, hat auf einem abgeschlossenen Intervall Funktionswerte mit positivem Betrag, so daß sein Skalarprodukt mit sich selbst nicht verschwinden kann.

Für praktische Anwendungen ist es gleichgültig, ob man mit L oder L^0 arbeiten will, denn jeder Funktion aus L kann man eine Funktion aus L^0 zuordnen, die sich höchstens in den Sprungstellen von dieser unterscheidet; der Unterschied zwischen den beiden Funktionen ist somit nicht meßbar.

Fundamental für das weitere ist die folgende Orthogonalitätseigenschaft:

Lemma: Mit $\omega = \dfrac{2\pi}{T}$ und $k,\ell \in \mathbb{Z}$ ist

$$(e^{ik\omega t},\, e^{i\ell\omega t}) = \begin{cases} 0 & \text{falls } k \neq \ell \\ 1 & \text{falls } k = \ell \end{cases}.$$

Beweis:

$$T\cdot(e^{ik\omega t},\, e^{i\ell\omega t}) = \int_0^T e^{ik\omega t}\overline{e^{i\ell\omega t}}\,dt = \int_0^T e^{ik\omega t}e^{-i\ell\omega t}\,dt$$

$$= \int_0^T e^{i(k-\ell)\omega t}\,dt\,.$$

Für $k = \ell$ integrieren wir hier die Konstante eins über ein Intervall der Länge T, das Integral ist also T. Für $k \neq \ell$ hat der Integrand die Stammfunktion

$$\frac{e^{i(k-\ell)\omega t}}{i(k-\ell)}\,,$$

die wegen der Beziehung $\omega T = 2\pi$ periodisch ist mit Periode T; das Integral verschwindet also. Division durch T liefert die Behauptung. ∎

Zerlegen wir die komplexe Exponentialfunktion in Real- und Imaginärteil, erhalten wir die etwas umständlicheren entsprechenden Beziehungen für trigonometrische Funktionen. Da Kosinus eine gerade und Sinus eine ungerade Funktion ist, sind negative k und ℓ uninteressant; wir begnügen uns daher mit

Lemma: Mit $\omega = \dfrac{2\pi}{T}$ und $k,\ell \in \mathbb{N}_0$ ist

$$(\cos k\omega t,\, \cos \ell\omega t) = \begin{cases} 0 & \text{falls } k \neq \ell \\ 1/2 & \text{falls } k = \ell \neq 0 \\ 1 & \text{falls } k = \ell = 0 \end{cases},$$

$$(\sin k\omega t,\, \sin \ell\omega t) = \begin{cases} 0 & \text{falls } k \neq \ell \\ 1/2 & \text{falls } k = \ell \neq 0 \\ 0 & \text{falls } k = \ell = 0 \end{cases} \quad \text{und}$$

$$(\cos k\omega t,\, \sin \ell\omega t) = 0\,.$$

Beweis: Wir verwenden das gerade bewiesene Lemma; danach ist mit

dem KRONECKER-δ ausgedrückt

$$\begin{aligned}\delta_{k\ell} &= (e^{ik\omega t},\, e^{i\ell\omega t}) = (\cos k\omega t + i\sin k\omega t,\ \cos \ell\omega t + i\sin \ell\omega t)\\ &= (\cos k\omega t,\, \cos \ell\omega t) + (\sin k\omega t,\, \sin \ell\omega t)\\ &\quad + i(\sin k\omega t,\, \cos \ell\omega t) - i(\cos k\omega t,\, \sin \ell\omega t)\,,\end{aligned}$$

also ist

$$(\cos k\omega t,\, \cos \ell\omega t) + (\sin k\omega t,\, \sin \ell\omega t) = \delta_{k\ell}$$

und

$$(\sin k\omega t,\, \cos \ell\omega t) - (\cos k\omega t,\, \sin \ell\omega t) = 0\,.$$

Diese Gleichungen gelten auch, wenn wir ℓ durch $-\ell$ ersetzen; sie werden dann zu

$$(\cos k\omega t,\, \cos \ell\omega t) - (\sin k\omega t,\, \sin \ell\omega t) = \delta_{k,-\ell}$$

und

$$(\sin k\omega t,\, \cos \ell\omega t) + (\cos k\omega t,\, \sin \ell\omega t) = 0\,.$$

Addiert *bzw.* subtrahiert man jeweils zwei der sich nur im Vorzeichen unterscheidenden Gleichungen, folgt, daß für $k, \ell \geq 0$

$$2(\cos k\omega t,\, \cos \ell\omega t) = \delta_{k\ell} + \delta_{k,-\ell} = \begin{cases} 0 & \text{für } k \neq \ell \\ 1 & \text{für } k = \ell \neq 0 \\ 2 & \text{für } k = \ell = 0 \end{cases}$$

ist und

$$2(\sin k\omega t,\, \sin \ell\omega t) = \delta_{k\ell} + \delta_{k,-\ell} = \begin{cases} 0 & \text{für } k \neq \ell \\ 1 & \text{für } k = \ell \neq 0 \\ 0 & \text{für } k = \ell = 0 \end{cases};$$

außerdem ist

$$(\sin k\omega t,\, \cos \ell\omega t) = (\cos k\omega t,\, \sin \ell\omega t) = 0\,.$$

Damit ist alles bewiesen. ∎

Ein Leser, der noch nicht davon überzeugt ist, daß komplexe Zahlen und Funktionen auch im Reellen nützlich sind, sollte versuchen, dies rein reell zu beweisen: er muß also zeigen, daß

$$\int_0^T \cos k\omega t\ \cos \ell\omega t\, dt = \begin{cases} 0 & \text{falls } k \neq \ell \\ T/2 & \text{falls } k = \ell \neq 0 \\ T & \text{falls } k = \ell = 0 \end{cases}$$

und

$$\int_0^T \sin k\omega t\ \sin \ell\omega t\, dt = \begin{cases} 0 & \text{falls } k \neq \ell \\ T/2 & \text{falls } k = \ell \neq 0 \\ 0 & \text{falls } k = \ell = 0 \end{cases}$$

ist, sowie

$$\int_0^T \cos k\omega t\ \sin \ell\omega t\, dt = 0\,.$$

Eine ganze Reihe dieser Integrationen sind trivial, und *alle* sind rein reell durchführbar. Dennoch spart der Umweg übers Komplexe viel Zeit.

d) Harmonische Analyse trigonometrischer Polynome

Die Funktionen $e^{k\cdot i\omega t}$ bilden natürlich keine Basis von $\mathrm{L}_T(\mathbb{R},\, \mathbb{C})$, genauso wenig wie die Funktionen $\cos k\omega t$ und $\sin \ell\omega t$ eine Basis von $\mathrm{L}_T(\mathbb{R}, \mathbb{R})$ bilden: Basisdarstellungen sind schließlich stets *endliche* Linearkombinationen, und eine endliche Linearkombination von trigonometrischen oder Exponentialfunktionen ist insbesondere stetig.

Trotzdem ist es ganz nützlich, zur Demonstration der weiteren Vorgehensweise zunächst die Untervektorräume zu betrachten, die von diesen Funktionen erzeugt werden:

Definition: *a)* Der Vektorraum $P_T(\mathbb{C})$ aller komplexer trigonometrischer Polynome der Periode T ist der von den Funktionen $e^{k\cdot i\omega t}$ mit $k \in \mathbb{Z}$ aufgespannte Untervektorraum von $\mathrm{L}^2_T(\mathbb{R}, \mathbb{C}\,)$.
b) Der Vektorraum $P_T(\mathbb{R})$ aller reeller trigonometrischer Polynome der Periode T ist der von den Funktionen $\cos k\omega t$ für $k \in \mathbb{N}_0$ und den Funktionen $\sin \ell\omega t$ mit $\ell \in \mathbb{N}$ aufgespannte Untervektorraum von $\mathrm{L}^2_T(\mathbb{R}, \mathbb{R}\,)$.

Die gerade bewiesenen Orthogonalitätsrelationen können wir dann auch so formulieren, daß die Funktionen $e^{k\cdot i\omega t}$ mit $k \in \mathbb{Z}$ eine Orthonormalbasis von $P_T(\mathbb{C})$ bilden, während die Funktionen $1, \sqrt{2}\cos k\omega t$ und $\sqrt{2}\sin \ell\omega t$ mit $k, \ell \in \mathbb{N}$ eine Orthonormalbasis von $P_T(\mathbb{R})$ bilden.

Zumindest für Funktionen aus $P_T(\mathbb{C})$ und $P_T(\mathbb{R})$ ist damit klar, wie man sie in reine Schwingungen zerlegen kann: Ist allgemein V ein

EUKLIDischer oder HERMITEscher Vektorraum und $\mathcal{B}$ eine Orthonormalbasis von V, so läßt sich ein beliebiger Vektor $\vec{v} \in V$ gemäß

$$\vec{v} = \sum_{\vec{b}\in\mathcal{B}} (\vec{v},\, \vec{b})\, \vec{b}$$

als (endliche) Linearkombination der Basisvektoren ausdrücken.

Für $f \in P_T(\mathbb{C})$ ist somit

$$f(t) = \sum_{k\in\mathbb{Z}} c_k e^{ki\omega t} \quad \text{mit} \quad c_k = \left(f(t),\, e^{ki\omega t}\right) = \frac{1}{T}\int_0^T f(t) e^{-ki\omega t}\, dt\,,$$

und für $f \in P_T(\mathbb{R})$ ist

$$f(t) = c_0 + \sum_{k=1}^{\infty} a_k \cos k\omega t + \sum_{\ell=1}^{\infty} b_\ell \sin \ell\omega t$$

mit

$$c_0 = \left(f(t), 1\right) = \frac{1}{T}\int_0^T f(t)\, dt$$

$$a_k = \sqrt{2}\cdot\left(f(t), \sqrt{2}\cos k\omega t\right) = \frac{2}{T}\int_0^T f(t)\cos k\omega t\, dt$$

$$b_\ell = \sqrt{2}\cdot\left(f(t), \sqrt{2}\sin \ell\omega t\right) = \frac{2}{T}\int_0^T f(t)\sin \ell\omega t\, dt\,.$$

Die Summen in diesen Formeln sind natürlich nur formal unendlich; da ein trigonometisches Polynom nach Definition *endliche* Linearkombination der Basisfunktionen ist, können in jeder dieser Summen höchstens endlich viele Summanden von Null verschieden sein.

Da die Formeln für reelle trigonometrische Polynome deutlich unangenehmer sind als die für komplexe, lohnt es sich oft, auch für reelle Funktionen den Umweg über das Komplexe zu gehen. Das ist immer

möglich, denn auf Grund der EULERschen Beziehungen ist jedes reelle trigonometrische Polynom gleichzeitig ein komplexes:

$$\begin{aligned}
&a_0 + \sum_{k=1}^{N} a_k \cos k\omega t + \sum_{\ell=1}^{M} b_\ell \sin \ell\omega t \\
&= a_0 + \sum_{k=1}^{N} a_k \frac{e^{k\cdot i\omega t} + e^{-k\cdot i\omega t}}{2} + \sum_{\ell=1}^{M} b_\ell \frac{e^{\ell\cdot i\omega t} e^{-\ell\cdot i\omega t}}{2i} \\
&= a_0 + \sum_{k=1}^{N} \frac{a_k}{2} e^{k\cdot i\omega t} + \sum_{k=1}^{N} \frac{a_k}{2} e^{-k\cdot i\omega t} - i\sum_{\ell=1}^{M} \frac{b_\ell}{2} e^{\ell\cdot i\omega t} + i\sum_{\ell=1}^{M} \frac{b_\ell}{2} e^{-\ell\cdot i\omega t} \\
&= a_0 + \sum_{k=1}^{N} \frac{a_k - ib_k}{2} e^{k\cdot i\omega t} + \sum_{k=1}^{N} \frac{a_k + ib_k}{2} e^{-k\cdot i\omega t}\,.
\end{aligned}$$

Schreibt man dies in der üblichen Weise als komplexes trigonometrisches Polynom $\sum c_k e^{k\cdot i\omega t}$, ist also

$$c_k = \begin{cases} \frac{1}{2}(a_k - ib_k) & \text{für } k > 0 \\ a_0 & \text{für } k = 0 \\ \frac{1}{2}(a_{-k} + ib_{-k}) & \text{für } k < 0 \end{cases}\,.$$

Insbesondere sind c_k und c_{-k} für alle k komplex konjugiert zueinander; c_0 ist reel und somit zu sich selbst konjugiert. Aus obigen Formeln folgt auch, daß umgekehrt

$$a_k = 2\,\mathfrak{Re}\, c_k \quad \text{und} \quad b_\ell = -2\,\mathfrak{Im}\, c_\ell$$

ist; man kann also leicht zwischen reeller und komplexer Darstellung umrechnen.

Damit ist auch klar, daß $P_T(\mathbb{R}) = P_T(\mathbb{C}) \cap \mathrm{L}_T(\mathbb{R}, \mathbb{R})$ ist; die reellen trigonometrischen Polynome sind also genau jene komplexe trigonometrische Polynome, die nur reelle Werte annehmen

Gefühlsmäßig würde man trigonometrische Polynome wohl nicht so definieren wie in diesem Abschnitt, sondern als Polynome in $\sin\omega t$ und $\cos\omega t$. Als kleine Anwendung obiger Überlegung folgt, daß dies in der Tat trigonometrische Polynome im Sinne der hießigen Definition sind, denn wegen der EULERschen Formel ist klar, daß es komplexe

trigonometrische Polynome sind, und natürlich nehmen sie nur reelle Werte an.

e) Harmonische Analyse periodischer Funktionen

Die Bedingung, daß $f(t)$ als Summe endlich vieler reiner Schwingungen gegeben sein soll, schränkt die Brauchbarkeit obiger Resultate leider erheblich ein: Ein periodischer Rechteckimpuls etwa läßt sich so nicht behandeln.

Wir können aber jedes beliebige Element von $\mathrm{L}^2_T(\mathbb{R}, \mathbb{C})$ die Skalarprodukte $c_k = (f,\ e^{ik\omega t})$ berechnen und hoffen, daß sie für eine harmonische Analyse von f nützlich sind; wir definieren

Definition: Die FOURIER-Transformierte einer Funktion $f \in \mathrm{L}^2_T(\mathbb{R}, \mathbb{C})$ ist die Funktion

$$\widehat{f}\colon \begin{cases} \mathbb{Z} \to \mathbb{C} \\ k \mapsto (f,\ e^{k\cdot i\omega t}) = \dfrac{1}{T}\displaystyle\int_0^T f(t)e^{-k\cdot i\omega t}\,dt\,. \end{cases}$$

(Man beachte, daß diese FOURIER-Transformierte einer *periodischen* Funktion nur auf $\mathbb{Z}$ definiert ist: Periodische Funktionen haben kein kontinuierliches Frequenzspektrum, sondern nur Oberschwingungen zu ganzzahligen Vielfachen der Grundfrequenz).

JEAN BAPTISTE JOSEPH FOURIER (1768–1830) begann zunächst mit einer Ausbildung zum Priester, beendete diese jedoch nicht, sondern wurde stattdessen Mathematiklehrer. 1793 trat er dem lokalen Revolutionskomittee bei, 1798 begleitete er Napoleon auf dessen Ägyptenfeldzug. Nach dem Rückzug aus Ägypten ernannte ihn dieser zum Präfekten von Isère; dort in Grenoble begann er mit seinen Arbeiten über Wärmeleitung, aus denen die FOURIER-Reihen hervorgingen. Nach Napoleons endgültiger Vertreibung wurde FOURIER 1817 in die Akademie der Wissenschaften gewählt; 1822 wurde er Sekretär der mathematischen Sektion.

Als *komplexe* FOURIER-*Reihe* von f bezeichnen wir die zunächst nur formale unendliche Summe

$$\sum_{k=-\infty}^{\infty} \widehat{f}(k)e^{-k\cdot i\omega t}\,,$$

als *reelle* FOURIER-*Reihe* von $f \in \mathrm{L}_T(\mathbb{R},\ \mathbb{R})$ entsprechend

$$c_0 + \sum_{k=1}^{\infty} a_k \cos k\omega t + \sum_{\ell=1}^{\infty} b_\ell \sin \ell\omega t$$

mit c_0, a_k und b_ℓ wie im vorigen Abschnitt.

Natürlich ist im Augenblick weder klar, ob diese Summen überhaupt existieren, d.h. also, ob die angegebenen Reihen für alle (oder zumindest fast alle) $t \in \mathbb{R}$ konvergieren, noch ist klar, ob sie dort, wo sie konvergieren, gegen den Funktionswert $f(t)$ konvergieren.

§ 3: Erste Beispiele von Fourier-Reihen

Bevor wir uns solchen allgemeinen Fragen zuwenden, wollen wir zunächst anhand einiger Beispiele sehen, was wir realistischerweise erwarten können.

a) Rechenregeln

Als erstes wollen wir uns überlegen, wie wir bei der Berechnung von FOURIER-Koeffizienten überflüssigen Rechenaufwand vermeiden können.

Das größte Potential für Vereinfachungen bieten *Symmetrien* der Funktion. Die beiden wichtigsten Symmetrien sind die Eigenschaften, *gerade* oder *ungerade* zu sein: Eine Funktion $f\colon \mathbb{R} \to \mathbb{R}$ ist gerade, wenn $f(-t) = f(t)$ ist für *alle* $t \in \mathbb{R}$; sie ist ungerade, wenn $f(-t) = -f(t)$ ist für *alle* $t \in \mathbb{R}$. Auch Symmetrien bezüglich anderer Punkte als $t = 0$ lassen sich gelegentlich erfolgreich ausnützen.

A priori läßt sich keine Symmetrie bezüglich $t = 0$ für die Berechnung der hier interessierenden bestimmten Integrale mit Grenzen 0 und T

ausnutzen; da wir es hier aber mit periodischen Funktionen zu tun haben, sind wir nicht an diese Integrationsgrenzen gebunden:

Lemma: Ist die Funktion g periodisch mit Periode T, so ist für jedes $\tau \in \mathbb{R}$

$$\int_0^T g(t)\,dt = \int_\tau^{\tau+T} g(t)\,dt\,.$$

Beweis: Wir können τ schreiben als

$$\tau = kT + \tau_0 \quad \text{mit} \quad 0 \le \tau_0 < T \quad \text{und} \quad k \in \mathbb{Z}\,.$$

Wegen der Periodizität von f ist

$$\int_\tau^{\tau+T} g(t)\,dt = \int_{\tau_0}^{\tau_0+T} g(t)\,dt\,;$$

es reicht also, den Fall $0 \le \tau < T$ zu betrachten. Hierfür ist

$$\begin{aligned}\int_\tau^{\tau+T} g(t)\,dt &= \int_\tau^T g(t)\,dt + \int_T^{\tau+T} g(t)\,dt = \int_\tau^T g(t)\,dt + \int_0^\tau g(t)\,dt \\ &= \int_0^\tau g(t)\,dt + \int_\tau^T g(t)\,dt = \int_0^T g(t)\,dt\,.\end{aligned}$$

■

Speziell für $\tau = -T/2$ ist also auch

$$a_0 = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\,dt \quad \text{und} \quad a_k = \frac{2}{T}\int_{-T/2}^{T/2} f(t)\cos k\omega t\,dt$$

für alle $k \in \mathbb{N}$, und

$$b_\ell = \frac{2}{T}\int_{-T/2}^{T/2} f(t)\sin \ell\omega t\,dt \quad \text{für alle } \ell \in \mathbb{N}\,.$$

Ist nun f eine ungerade Funktion, so sind auch alle Funktionen $f(t)\cos k\omega t$ ungerade, d.h.

$$a_0 = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\,dt = 0 \quad \text{und} \quad a_k = \int_{-T/2}^{T/2} f(t)\cos k\omega t\,dt = 0$$

für alle $k \in \mathbb{N}$. Die Funktion $f(t)\sin \ell\omega t$ ist Produkt zweier ungerader Funktionen und somit gerade; dies liefert die Beziehung

$$b_\ell = \frac{2}{T}\int_{-T/2}^{T/2} f(t)\sin k\omega t\,dt = \frac{4}{T}\int_0^{T/2} f(t)\sin k\omega t\,dt\,,$$

die je nach der speziellen Form von f entweder nützlich ist oder auch nicht. Auf jeden Fall gibt es aber bei einer ungeraden Funktion in der FOURIER-Reihe keine Kosinusterme (einschließlich des konstanten Terms zu $\cos 0 = 1$); nur Sinusterme können von Null verschiedene Koeffizienten haben.

Für eine gerade Funktion f ist $f(t)\cdot\sin \ell\omega t$ als Produkt einer geraden und einer ungeraden Funktion ungerade, d.h.

$$b_\ell = \frac{2}{T}\int_{-T/2}^{T/2} f(t)\sin \ell\omega t\,dt = 0$$

für alle ℓ. Somit sind keine Sinusterme möglich; nur Kosinusterme (einschließlich des konstanten Terms) können auftreten. Weiter ist

$$\int_{-T/2}^{T/2} f(t)\cos k\omega t\,dt = 2\int_0^{T/2} f(t)\cos k\omega t\,dt\,,$$

was wiederum in Abhängigkeit von der speziellen Gestalt von f entweder nützlich ist oder auch nicht.

Nach diesen Vorbereitungen kommen wir nun endgültig zu konkreten Beispielen; das erste davon ist gerade in der Digitaltechnik von großer Bedeutung:

b) Periodische Rechteckimpulse

Hier betrachten wir die Funktion

$$f(t) = \begin{cases} h & \text{für } 0 \le t < \frac{T}{2} \\ -h & \text{für } \frac{T}{2} \le t < T \end{cases} \quad \text{mit} \quad f(t{+}T) = f(t) \quad \text{für alle} \quad t \in \mathbb{R}\,;$$

offensichtlich ist $f \in \mathrm{L}_T(\mathbb{R}, \mathbb{R})$.

Außerdem ist f eine ungerade Funktion, d.h. es gibt nur Sinusterme. Für diese ist

$$b_\ell = \frac{2}{T}\int_0^T f(t)\sin \ell\omega t\,dt = \frac{2}{T}\left(\int_0^{T/2} h\sin \ell\omega t\,dt - \int_{T/2}^{T} h\sin \ell\omega t\,dt\right),$$

aber wir können uns die Auswertung des zweiten Integrals sparen, wenn wir uns daran erinnern, daß für eine ungerade Funktion b_ℓ auch berechnet werden kann als $\frac{4}{T}$ mal dem Integral von Null bis zu rhalben Periode. Somit ist

$$b_\ell = \frac{4}{T}\int_0^{T/2} h\sin \ell\omega t\,dt = \frac{4h}{T}\left(-\frac{\cos \ell\omega\frac{T}{2} - 1}{\ell\omega}\right) = \frac{4h}{T}\left(\frac{(-1)^{\ell+1}+1}{\ell\omega}\right),$$

denn wegen $\omega T = 2\pi$, ist $\omega T/2 = \pi$ und $\cos \ell\pi = (-1)^\ell$. Somit ist

$$b_\ell = \begin{cases} 0 & \text{für gerade } \ell \\ \frac{4h}{T}\cdot\frac{2}{\ell\omega} = \frac{4h}{\pi\ell} & \text{für ungerade } \ell \end{cases}$$

und

$$S_f(t) = \frac{4h}{\pi}\sum_{\ell=1}^{\infty}\frac{\sin(2\ell-1)\omega t}{(2\ell-1)}\,.$$

Wir sollten nicht zu optimistisch sein und erwarten, daß diese FOURIER-Reihe in jedem Punkt t gegen $f(t)$ konvergiert: Wir hätten einen Rechteckimpuls mit Periode T im Intervall $[0, T)$ beispielsweise auch durch

$$g(t) = \begin{cases} h & \text{für } 0 \le t \le \frac{T}{2} \\ -h & \text{für } \frac{T}{2} < t < T \end{cases}$$

definieren können. $f(t)$ und $g(t)$ unterscheiden sich im Intervall $[0, T]$ nur an der Stelle $t = \frac{T}{2}$, was bei der Berechnung der Integrale für die FOURIER-Koeffizienten keine Rolle spielt. Die beiden Funktionen haben daher dieselbe FOURIER-Reihe, und diese kann, selbst wenn sie konvergiert, an der Stelle $t = \frac{T}{2}$ nicht sowohl gegen $f(\frac{T}{2}) = -h$ und $g(\frac{T}{2}) = h$ konvergieren. (Tatsächlich konvergiert sie, da $\omega\frac{T}{2} = \pi$ ist und der Sinus bei allen Vielfachen von π verschwindet, gegen Null, d.h. den Mittelwert der beiden Funktionswerte.)

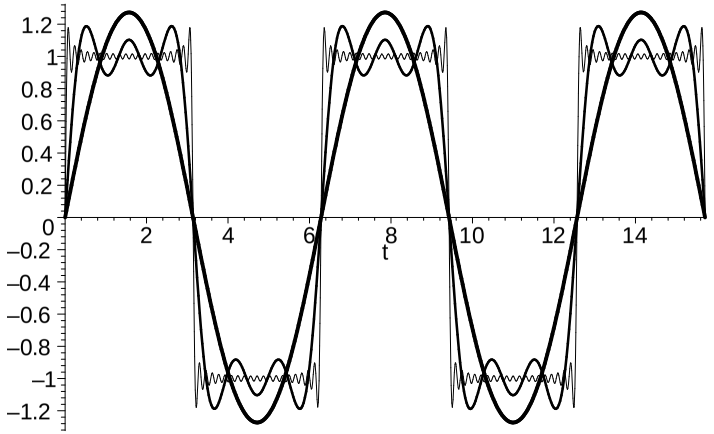


Abb. 6: FOURIER-Polynome für Rechteckimpulse

Experimentell konvergiert die berechnete Reihe abgesehen von den Sprungstellen anscheinend recht gut: Abbildung sechs zeigt die Teilsummen mit oberen Grenzen 1, 3 und 20, die die Funktion f offensichtlich immer besser annähern. Gerade für die größeren Werte ist dieses Bild natürlich etwas gestört durch numerische Fehler und *alias*-Effekte der Rastergraphik.

Keine solche Störung sind allerdings die Überschwingungen an den Unstetigkeitsstellen von f: Dieses sogenannte GIBBS-Phänomen ist eine mathematisch unvermeidbare Eigenschaft von FOURIER-Reihen stückweise stetiger Funktionen, mit der wir uns in Kürze näher beschäftigen werden.

Im Augenblick sei nur kurz auf eine Anwendung dieser Überschwingungen hingewiesen: Die Pixel auf einem Computerbildschirm werden durch Rechteckimpulse geschaltet, wobei aus physikalischen Gründen Oberschwingungen hoher Frequenz bei der Übertragung so stark gedämpft werden, daß für alle praktischen Zwecke nur so etwas wie eine endliche Teilsumme der FOURIER-Reihe übertragen wird.

Das tatsächliche Signal hat somit eher die Gestalt einer der Kurven aus Abbildung sechs als die eines (physikalisch nicht zu realisierenden) „echten“ Rechteckimpulses.

Die in der Abbildung zu sehenden höherfrequenten Anteile lassen sich aber problemlos mit einem geeignet eingestellten Funkempfänger auffangen und können dann zur Rekonstruktion des Bildschirminhalts verwendet werden.

Zumindest bei sensitiven Anwendungen muß ein Computer daher so abgeschirmt sein, daß von dieser Strahlungen nichts aus dem Gehäuse dringt. Bei einem Standardgehäuse hat man hier nicht die geringste Chance; Computer im Hochsicherheitsbereich brauchen ihre eigenen Spezialgehäuse.

c) Sägezahnimpulse

Hier betrachten wir die Funktion

$$f(t) = \frac{T}{4} - \frac{t}{2} \quad \text{für} \quad 0 < t < T \quad \text{und} \quad f(0) = 0\,,$$

periodisch fortgesetzt mit Periode T auf ganz $\mathbb{R}$.

Dies ist eine ungerade Funktion, denn für $0 < t < T$ ist

$$f(-t) = f(-t+T) = \frac{T}{4} - \frac{(-t+T)}{2} = \frac{t}{2} - \frac{T}{4} = -f(t)\,,$$

und $f(0) = 0$, wie es sich für eine ungerade Funktion gehört. Die FOURIER-Reihe von f enthält daher nur Sinusterme.

Zu deren Berechnung setzen wir wie üblich $\omega = \frac{2\pi}{T}$ und erhalten den

Koeffizienten von $\sin \ell\omega t$ als

$$\begin{aligned} b_\ell &= \frac{2}{T}\int_0^T f(t)\sin \ell\omega t\,dt = \frac{2}{T}\int_0^T \left(\frac{T}{4} - \frac{t}{2}\right)\sin \ell\omega t\,dt \\ &= \frac{2}{T}\cdot\frac{T}{4}\int_0^T \sin \ell\omega t\,dt - \frac{2}{T}\cdot\frac{1}{2}\int_0^T t\sin \ell\omega t\,dt \\ &= -\frac{1}{T}\int_0^T t\sin \ell\omega t\,dt\,, \end{aligned}$$

da das Integral einer Sinusfunktion über eine oder mehrere volle Perioden verschwindet. Zur weiteren Rechnung wenden wir die Methode der partiellen Integration an:

$$\int u(t)\cdot\dot v(t)\,dt = u(t)\cdot v(t) - \int \dot u(t)\cdot v(t)\,dt$$

ergibt für $u(t) = t$ und $\dot v(t) = \sin \ell\omega t$ mit $v = -\frac{1}{\ell\omega}\cos \ell\omega t$ die Beziehung

$$\begin{aligned} \int t\sin \ell\omega t\,dt &= -t\frac{\cos \ell\omega t}{\ell\omega} + \frac{1}{\ell\omega}\int \cos \ell\omega t\,dt \\ &= -t\frac{\cos \ell\omega t}{\ell\omega} + \frac{1}{\ell^2\omega^2}\sin \ell\omega t + C\,. \end{aligned}$$

Somit ist

$$\begin{aligned} b_\ell &= -\frac{1}{T}\left(\frac{-T\cdot\cos \ell\omega T + 0\cdot\cos 0}{\ell\omega} + \frac{\sin \ell\omega T - \sin 0}{\ell^2\omega^2}\right) \\ &= \frac{1}{\ell\omega}\cos(\ell\cdot 2\pi) = \frac{1}{\ell\omega} \end{aligned}$$

und

$$S_f(t) = \sum_{\ell=1}^{\infty}\frac{\sin \ell\omega t}{\ell\omega}\,.$$

Wieder haben wir keine Ahnung, ob und gegebenenfalls wohin diese Reihe konvergiert – außer bei den ganzzahligen Vielfachen von $T/2$,

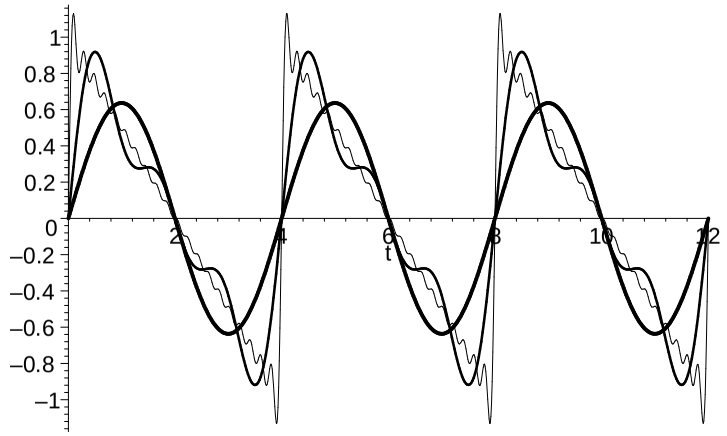


Abb. 7: FOURIERpolynome für die Sägezahnschwingung

denn dort verschwinden alle Sinusfunktionen in den Zählern, so daß die Summe gleich null ist.

Abbildung 7 zeigt die Teilsummen mit 1, 3 und 20 Summanden für $T = 4$; anscheinend nähern diese die Funktion recht gut an, allerdings gibt es wieder Überschwingungen an den Sprungstellen, denn für $T = 4$ haben wir einen Sägezahn, der zwischen +1 und -1 hin- und herpendelt.

d) Der Sinus hyperbolicus

Als letztes Beispiel berechnen wir die FOURIER-Reihe von

$f(t) = \sinh t$ für $-\pi < t \leq \pi$, periodisch fortgesetzt mit Periode 2π .

Die Koeffizienten der komplexen FOURIER-Reihe sind

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-k \cdot i\omega t}\, dt \,.$$

Man darf nun aber keineswegs den Fehler machen, daraus zu folgern,

daß

$$\text{???} \qquad c_k = \frac{1}{2\pi} \int_0^{2\pi} \sinh t\, e^{-k \cdot i\omega t}\, dt \qquad \text{???}$$

sei, denn $f(t)$ stimmt *nur* im Intervall $(-\pi, \pi]$ mit $\sinh t$ überein; für $\pi < t \leq 2\pi$ ist $f(t) = \sinh(t - 2\pi)$. Falls wir die mit Fragezeichen versehene Formel benutzen, berechnen wir tatsächlich die FOURIER-Reihe von

$g(t) = \sinh t$ für $0 < t \leq 2\pi$, periodisch fortgesetzt mit Periode 2π ,

und das ist, wie die Abbildungen acht und neun zeigen, eine völlig andere Funktion: f ist eine ungerade Funktion mit einem Wertebereich, der durch die beiden Extrema $\pm \sinh \pi \approx \pm 11{,}54873936$ begrenzt ist, g dagegen eine Funktion mit Werten zwischen null und $\sinh 2\pi \approx 267{,}7448943$, die weder gerade noch ungerade ist.

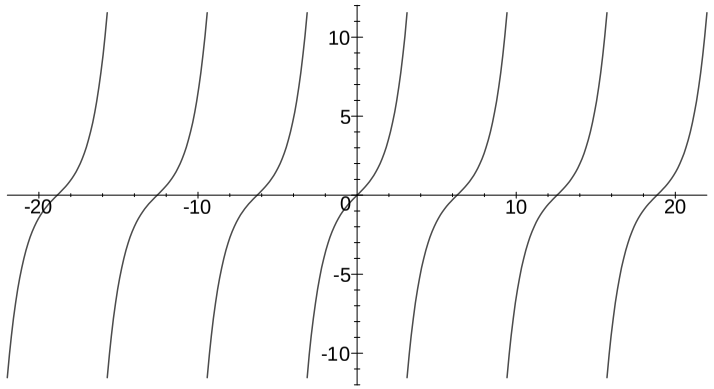


Abb. 8: Die Funktion $f(t)$

Wie groß der Unterschied zwischen den beiden Funktionen wirklich ist, sieht man am besten, wenn man sie wie in Abbildung zehn in einem

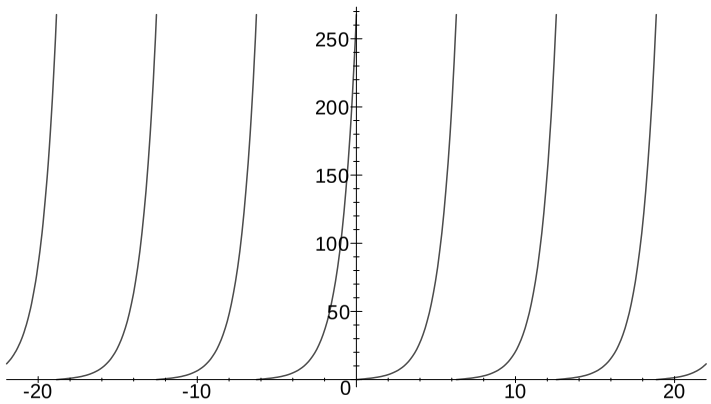


Abb. 9: Die Funktion $g(t)$

gemeinsamen Koordinatensystem abbildet: Die fett gezeichneten Kurvenstücke sind beiden Funktionen gemeinsam, und dort, wo f und g nicht übereinstimmen, ist f durch eine ausgezogene, g durch eine gestrichelte Kurve dargestellt.

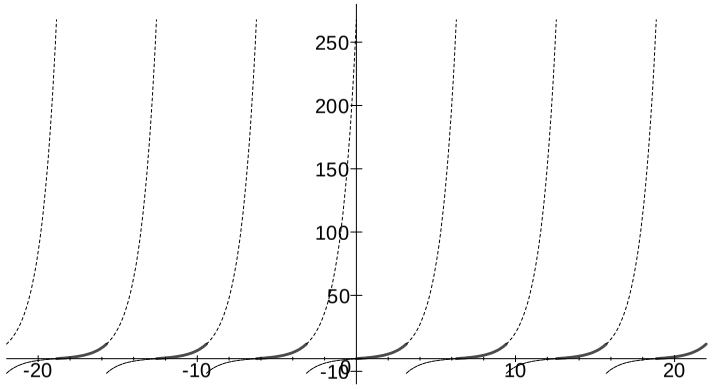


Abb. 10: $f(t)$ und $g(t)$ im gleichen Koordinatensystem

Wenn wir mit einer Integration von 0 bis 2π arbeiten wollen, müssen

wir also das Integral in zwei Teilintegrale aufteilen:

$$\int_0^{2\pi} f(t)e^{-k\cdot it}\,dt = \int_0^{\pi} \sinh t\, e^{-k\cdot it}\,dt + \int_{\pi}^{2\pi} \sinh(t-2\pi)\, e^{-k\cdot it}\,dt$$

Zum Glück wissen wir aber aus §3*a)*, daß wir bei einer periodischen Form über jedes beliebige Periodenintervall integrieren dürfen, ohne etwas am Ergebnis zu verändern: Das wurde dort zwar nur für reelle Integrale gezeigt, aber da ein komplexes Integral auf zwei reelle zurückgeführt werden kann, gilt es auch dafür. Also ist auch

$$\begin{aligned}
c_k &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)e^{-k\cdot it}\,dt = \frac{1}{2\pi}\int_{-\pi}^{\pi} \sinh t\, e^{-k\cdot it}\,dt \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{e^t - e^{-t}}{2}\, e^{-k\cdot it}\,dt = \frac{1}{4\pi}\int_{-\pi}^{\pi} \left(e^{t(1-ki)} - e^{-t(1+ki)}\right)\,dt \\
&= \frac{1}{4\pi}\left(\frac{e^{t(1-ki)}}{1-ki}\,\Big|_{-\pi}^{\pi} - \frac{e^{-t(1+ki)}}{-(1+ki)}\,\Big|_{-\pi}^{\pi}\right) \\
&= \frac{1}{4\pi}\left(\frac{e^{\pi}e^{-ki\pi} - e^{-\pi}e^{ki\pi}}{1-ki} + \frac{e^{-\pi}e^{-ki\pi} - e^{\pi}e^{ki\pi}}{1+ki}\right) \\
&= \frac{(-1)^k}{4\pi}(e^{\pi} - e^{-\pi})\left(\frac{1}{1-ki} - \frac{1}{1+ki}\right) \\
&= \frac{(-1)^k \sinh\pi}{2\pi}\,\frac{(1+ki)-(1-ki)}{1+k^2} = \frac{\sinh\pi}{\pi}\cdot\frac{(-1)^k\cdot ik}{k^2+1}\,.
\end{aligned}$$

Die komplexe FOURIER-Reihe ist somit

$$S_f(t) = i\frac{\sinh\pi}{\pi}\sum_{k=-\infty}^{\infty}\frac{(-1)^k k}{k^2+1}e^{ikt}\,.$$

Da der Koeffizient von e^{ikt} eine ungerade Funktion von k ist, fallen beim Einsetzen von $e^{ikt} = \cos kt + i\sin kt$ die Kosinusterme weg, während

sich die Sinusterme zu k und zu $-k$ gegenseitig verdoppeln; wir erhalten also die reelle Form

$$S_f(t) = i\frac{\sinh\pi}{\pi}\sum_{k=-\infty}^{\infty}\frac{(-1)^k k}{k^2+1} i\sin kt = -2\frac{\sinh\pi}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k k}{k^2+1}\sin kt\,,$$

die gleichzeitig die reelle FOURIER-Reihe von f ist.

Bei der direkten Berechnung über die Koeffizientenformeln für die reelle Reihe wären die Vorüberlegungen aus §3*a)* ebenfalls nützlich gewesen: Da f eine ungerade Funktion ist, treten nur Sinusterme auf, und deren Koeffizienten sind

$$b_\ell = \frac{1}{\pi}\int_{-\pi}^{\pi}\sinh t\,\sin\ell t\,dt\,.$$

Der Integrand hier ist in Exponentialform gleich

$$\frac{e^t-e^{-t}}{2}\cdot\frac{e^{i\ell t}-e^{-i\ell t}}{2i} = \frac{e^{t(1+i\ell)}}{4i} - \frac{e^{t(1-i\ell)}}{4i} - \frac{e^{-t(1-i\ell)}}{4i} + \frac{e^{-t(1+i\ell)}}{4i}\,,$$

und die Stammfunktion des Summanden

$$\frac{e^{\pm t(1\pm i\ell)}}{4i} \quad\text{ist}\quad \frac{e^{\pm t(1\pm i\ell)}}{\pm 4i(1\pm i\ell)}\,.$$

Die Stammfunktion des Integranden ist daher

$$\frac{e^{t(1+i\ell)}}{4i(1+i\ell)} - \frac{e^{t(1-i\ell)}}{4i(1-i\ell)} - \frac{e^{-t(1-i\ell)}}{-4i(1-i\ell)} + \frac{e^{-t(1+i\ell)}}{-4i(1+i\ell)}$$
$$= \frac{e^{t(1+i\ell)}-e^{-t(1+i\ell)}}{4i(1+i\ell)} - \frac{e^{t(1-i\ell)}-e^{-t(1-i\ell)}}{4i(1-i\ell)}\,,$$

d.h.
$$b_\ell = \frac{e^{t(1+i\ell)}-e^{-t(1+i\ell)}}{4\pi i(1+i\ell)}\Bigg|_{-\pi}^{\pi} - \frac{e^{t(1-i\ell)}-e^{-t(1-i\ell)}}{4i(1-i\ell)}\Bigg|_{-\pi}^{\pi}$$
$$= \frac{\left(e^{\pi(1+i\ell)}-e^{-\pi(1+i\ell)}\right)-\left(e^{-\pi(1+i\ell)}-e^{\pi(1+i\ell)}\right)}{4\pi i(1+i\ell)}$$
$$-\frac{\left(e^{\pi(1-i\ell)}-e^{-\pi(1-i\ell)}\right)-\left(e^{-\pi(1-i\ell)}-e^{\pi(1-i\ell)}\right)}{4\pi i(1-i\ell)}$$

$$= \frac{(-1)^\ell\left(e^\pi-e^{-\pi}-e^{-\pi}+e^\pi\right)}{4\pi i(1+i\ell)} - \frac{(-1)^\ell\left(e^\pi-e^{-\pi}-e^{-\pi}+e^\pi\right)}{4\pi i(1-i\ell)}$$
$$= (-1)^\ell\left(\frac{\sinh\pi}{\pi i(1+i\ell)} - \frac{\sinh\pi}{\pi i(1-i\ell)}\right)$$
$$= (-1)^\ell\frac{\sinh\pi}{\pi i}\left(\frac{1}{1+i\ell}-\frac{1}{1-i\ell}\right)$$
$$= (-1)^\ell\frac{\sinh\pi}{\pi i}\frac{(1-i\ell)-(1+i\ell)}{1+\ell^2} = (-1)^\ell\frac{\sinh\pi}{\pi}\frac{-2\ell}{\ell^2+1}\,.$$

Das sind, abgesehen von der anderen Bezeichnung für den Index, genau die oben berechneten Koeffizienten.

Wir können das Integral auch ganz ohne komplexe Zahlen ausrechnen: Zweimalige Anwendung der Regel für partielle Integration liefert

$$b_\ell = \frac{1}{\pi}\int_{-\pi}^{\pi}\sinh t\,\sin\ell t\,dt = \frac{1}{\pi}\cosh t\sin\ell t\,\Big|_{-\pi}^{\pi} - \frac{\ell}{\pi}\int_{-\pi}^{\pi}\cosh t\cos\ell t\,dt$$
$$= -\frac{\ell}{\pi}\left(\sinh t\cos\ell t\Big|_{-\pi}^{\pi} + \ell\int_{-\pi}^{\pi}\sinh t\sin\ell t\,dt\right)$$
$$= -\frac{\ell}{\pi}\left((-1)^\ell\cdot 2\sinh\pi + \ell\pi b_\ell\right) = -\frac{\ell}{\pi}(-1)^\ell\cdot 2\sinh\pi - \ell^2 b_\ell\,.$$

Somit ist

$$(1+\ell^2)\,b_\ell = -\frac{2\sinh\pi}{\pi}(-1)^\ell\ell \quad\text{und}\quad b_\ell = -\frac{2\sinh\pi}{\pi}\frac{(-1)^\ell\ell}{\ell^2+1}\,.$$

Damit haben wir die FOURIER-Reihe von f auf drei verschiedene Weisen berechnet; das Ergebnis war natürlich in allen drei Fällen dasselbe, der Weg dorthin aber recht verschieden. Es hängt sowohl vom Problem als auch von persönlichen Vorlieben ab, welchen Rechengang man vorzieht; gerade bei Funktionen, bei denen die FOURIER-Reihe sowohl Sinus- als auch Kosinusterme enthält, wird aber oft der Weg über die komplexe FOURIER-Reihe am schnellsten sein, da man dann nur ein Integral berechnen muß.

e) Konvergenz der berechneten Reihen

Als nächstes wollen, zunächst für Sägezahnschwingungen, die Konvergenz der FOURIER-Reihe untersuchen. Für $t = 0$ und damit auch für alle Vielfachen von T sind alle Summanden null, die Reihe konvergiert also gegen null.

Für t aus dem offenen Intervall $(0, T)$ können wir folgendermaßen vorgehen: Die Summanden $\dfrac{\sin \ell\omega t}{\ell\omega}$ sind Stammfunktionen der Funktionen $\cos \ell\omega t$; also ist

$$\sum_{\ell=1}^{N} \frac{\sin \ell\omega t}{\ell\omega} \quad \text{Stammfunktion von} \quad \sum_{\ell=1}^{N} \cos \ell\omega t\,.$$

Auch die Funktion f läßt sich im Intervall $(0, T)$ als Stammfunktion schreiben: Dort ist

$$f(t) = \frac{T}{4} - \frac{t}{2} = \int_{T/2}^{t} \left(\frac{-1}{2}\right) d\tau\,,$$

und da auch

$$\int_{T/2}^{t} \cos \ell\omega\tau \, d\tau = \frac{\sin \ell\omega t - \sin \ell\omega T/2}{\ell\omega} = \frac{\sin \ell\omega t - \sin \ell\pi}{\ell\omega} = \frac{\sin \ell\omega t}{\ell\omega}$$

ist, erhalten wir die Differenz zwischen der N-ten Teilsumme und $f(t)$ als Integral:

$$\begin{aligned}\sum_{\ell=1}^{N} \frac{\sin \ell\omega t}{\ell\omega} - f(t) &= \int_{T/2}^{t} \left(\sum_{\ell=1}^{N} \cos \ell\omega\tau - \left(\frac{-1}{2}\right)\right) d\tau \\ &= \int_{T/2}^{t} \left(\frac{1}{2} + \sum_{\ell=1}^{N} \cos \ell\omega\tau\right) d\tau\,.\end{aligned}$$

Diesen Integranden können wir über die komplexe Darstellung des Kosinus ausrechnen:

$$\begin{aligned}\frac{1}{2} + \sum_{\ell=1}^{N} \cos \ell\omega\tau &= \frac{1}{2} + \sum_{\ell=1}^{N} \frac{e^{\ell\cdot i\omega\tau} + e^{-\ell\cdot i\omega\tau}}{2} = \frac{1}{2} \sum_{\ell=-N}^{N} e^{\ell\cdot i\omega\tau} \\ &= \frac{1}{2} e^{-N\cdot i\omega\tau} \sum_{\ell=0}^{2N} e^{\ell\cdot i\omega\tau}\end{aligned}$$

ist im wesentlichen eine geometrische Reihe, und die läßt sich bekanntlich leicht ausrechnen: Da

$$(1-q)\sum_{j=0}^{r} q^j = \sum_{j=0}^{r} q^j - \sum_{j=1}^{r+1} q^j = 1 - q^{r+1}$$

ist, folgt für $q \neq 1$ die Formel

$$\sum_{j=0}^{r} q^j = \frac{1 - q^{r+1}}{1-q}\,.$$

In unserem Fall ist $q = e^{i\omega\tau}$ und somit

$$\sum_{\ell=0}^{2N} e^{\ell\cdot i\omega\tau} = \frac{1 - e^{(2N+1)\cdot i\omega\tau}}{1 - e^{i\omega\tau}}\,.$$

Also ist

$$\begin{aligned}\frac{1}{2} + \sum_{\ell=1}^{N} \cos \ell\omega\tau &= \frac{1}{2} e^{-N\cdot i\omega\tau} \frac{1 - e^{(2N+1)\cdot i\omega\tau}}{1 - e^{i\omega\tau}} \\ &= \frac{1}{2} \frac{e^{-N\cdot i\omega\tau} - e^{(N+1)\cdot i\omega\tau}}{1 - e^{i\omega\tau}}\,.\end{aligned}$$

Erweiterung des Bruchs mit $e^{\frac{1}{2}\cdot i\omega\tau}$ führt auf die symmetrischere Form

$$\frac{1}{2} \frac{e^{-(N+\frac{1}{2})\cdot i\omega\tau} - e^{(N+\frac{1}{2})\cdot i\omega\tau}}{e^{-\frac{1}{2}\cdot i\omega\tau} - e^{\frac{1}{2}\cdot i\omega\tau}} = \frac{\sin\left(N+\frac{1}{2}\right)\omega\tau}{2\sin\frac{\omega\tau}{2}}\,.$$

Damit ist

$$\sum_{\ell=1}^{N} \frac{\sin \ell\omega t}{\ell\omega} - f(t) = \int_{T/2}^{t} \frac{\sin\left(N+\frac{1}{2}\right)\omega\tau}{2\sin\frac{\omega\tau}{2}}\, d\tau\,.$$

Die FOURIER-Reihe konvergiert genau dann im Punkt t gegen $f(t)$, wenn dieses Integral für $N \to \infty$ gegen null geht.

Die Suche nach einer Stammfunktion sieht ziemlich hoffnungslos aus; trotzdem hilft partielle Integration zu einem besseren Verständnis des Integrals. Wir wenden die Regel an mit

$$u(\tau) = \frac{1}{2\sin\frac{\omega\tau}{2}} \quad \text{und} \quad \dot{v}(\tau) = \sin\left(N+\frac{1}{2}\right)\omega\tau\,,$$

d.h.

$$v(\tau) = -\frac{\cos\left(N+\frac{1}{2}\right)\omega\tau}{\left(N+\frac{1}{2}\right)\omega}\,;$$

das Integral wird zu

$$-\frac{\cos\left(N+\frac{1}{2}\right)\omega t}{(2N+1)\omega\sin\frac{\omega t}{2}} + \int_{T/2}^{t} \frac{\cos\left(N+\frac{1}{2}\right)\omega\tau}{\left(N+\frac{1}{2}\right)\omega}\,\frac{d}{d\tau}\left(\frac{1}{2\sin\frac{\omega\tau}{2}}\right)d\tau\,,$$

denn an der unteren Grenze ist

$$\cos\left(N+\frac{1}{2}\right)\omega\frac{T}{2} = \cos\left(N+\frac{1}{2}\right)\pi = 0\,.$$

Auf das noch verbleibende Integral wenden wir den Mittelwertsatz der Integralrechnung in seiner allgemeinen Form an:

Für eine im Intervall $[a,\,b]$ stetige Funktion v und eine in $[a,\,b]$ integrierbare Funktion w gibt es einen Wert $\zeta \in [a,\,b]$, so daß gilt

$$\int_a^b v(\tau)w(\tau)\,d\tau = v(\zeta)\int_a^b w(\tau)\,d\tau\,.$$

Für alle, die den Satz nicht in dieser Form kennen, sei der *Beweis* kurz nachgetragen: Als stetige Funktion nimmt v im Intervall $[a,\,b]$ sowohl seinen Maximalwert $v_{\max}$ als auch seinen Minimalwert $v_{\min}$ an. Der Wert des linksstehenden Integrals liegt dann zwischen

$$v_{\min}\int_a^b w(\tau)\,d\tau \quad \text{und} \quad v_{\max}\int_a^b w(\tau)\,d\tau\,,$$

es gibt also einen Wert $v_0 \in [v_{\min},\,v_{\max}]$, so daß

$$\int_a^b v(\tau)w(\tau)\,d\tau = v_0\int_a^b w(\tau)\,d\tau$$

ist. Nach dem Zwischenwertsatz nimmt v als stetige Funktion diesen Wert v_0 irgendwo an, es gibt also ein $\zeta \in [a,\,b]$, so daß $v(\zeta) = v_0$ ist. Damit ist der Satz bewiesen. ∎

Hier setzen wir

$$v(\tau) = \frac{\cos\left(N+\frac{1}{2}\right)\omega\tau}{\left(N+\frac{1}{2}\right)\omega} \quad \text{und} \quad w(\tau) = \frac{d}{d\tau}\left(\frac{1}{2\sin\frac{\omega\tau}{2}}\right)\,;$$

wir erhalten

$$\begin{aligned}
&\int_{T/2}^{t} \frac{\cos\left(N+\frac{1}{2}\right)\omega\tau}{\left(N+\frac{1}{2}\right)\omega}\,\frac{d}{d\tau}\left(\frac{1}{2\sin\frac{\omega\tau}{2}}\right)d\tau\\
&= \frac{\cos\left(N+\frac{1}{2}\right)\omega\zeta}{(2N+1)\omega}\int_{T/2}^{t}\frac{d}{d\tau}\left(\frac{1}{\sin\frac{\omega\tau}{2}}\right)d\tau\\
&= \frac{\cos\left(N+\frac{1}{2}\right)\omega\zeta}{(2N+1)\omega}\left(\frac{1}{\sin\frac{\omega t}{2}}-1\right)\,,
\end{aligned}$$

denn $\sin\dfrac{\omega T}{4} = \sin\dfrac{\pi}{2} = 1$. Also ist

$$\begin{aligned}
&\sum_{\ell=1}^{N}\frac{\sin\ell\omega t}{\ell\omega} - f(t)\\
&= -\frac{\cos\left(N+\frac{1}{2}\right)\omega t}{(2N+1)\omega\sin\frac{\omega t}{2}} + \frac{\cos\left(N+\frac{1}{2}\right)\omega\zeta}{(2N+1)\omega}\left(\frac{1}{\sin\frac{\omega t}{2}}-1\right)\,.
\end{aligned}$$

Für $0 < t < T$ ist $1/\sin\frac{\omega t}{2} \geq 1$, also

$$0 \leq \frac{1}{\sin\frac{\omega t}{2}} - 1 < \frac{1}{\sin\frac{\omega t}{2}}\,,$$

und da der Kosinus nur Werte zwischen -1 und $+1$ annimmt, folgt, daß

$$\left| \sum_{\ell=1}^{N} \frac{\sin \ell\omega t}{\ell\omega} - f(t) \right| \leq \frac{2}{(2N+1)\omega \sin \frac{\omega t}{2}}$$

für alle t mit $0 < t < T$.

Für $N \to \infty$ geht die rechte Seite gegen null, also ist

$$S_f(t) = \lim_{N\to\infty} \sum_{\ell=1}^{N} \frac{\sin \ell\omega t}{\ell\omega} = f(t)$$

für alle t mit $0 < t < T$. Für $t = 0$ stehen links und rechts Nullen, so daß die Gleichung auch dort gilt, und da beide Seiten periodisch sind mit Periode T, gilt sie tatsächlich für alle $t \in \mathbb{R}$.

Auf abgeschlossenen Teilintervallen von $(0, T)$ ist die Konvergenz sogar gleichmäßig, denn im Intervall $[\varepsilon, T - \varepsilon]$ ist $\sin \frac{\omega t}{2} \geq \sin \frac{\omega \varepsilon}{2}$, d.h.

$$\left| \sum_{\ell=1}^{N} \frac{\sin \ell\omega t}{\ell\omega} - f(t) \right| \leq \frac{2}{(2N+1)\omega \sin \frac{\omega \varepsilon}{2}}$$

für alle $t \in [\varepsilon, T - \varepsilon]$.

Mit diesem Resultat können wir nun auch die Konvergenz der FOURIER-Reihe für Rechteckimpulse genauer untersuchen:

Für $0 < t < \frac{T}{2}$ liegt auch $\frac{T}{2} - t$ im Intervall $(0,\ T)$, d.h. mit der gerade betrachteten Funktion f für Sägezahnimpulse ist

$$f(t) + f\left(\frac{T}{2} - t\right) = \frac{T}{4} - \frac{t}{2} + \frac{T}{4} - \frac{T}{4} + \frac{t}{2} = \frac{T}{4}\,.$$

Für $\frac{T}{2} < t < T$ liegt $\frac{T}{2} - t$ im Intervall $(-T,\ 0)$, d.h. $\frac{3T}{2} - t$ liegt in $(0,\ T)$ und

$$f(t) + f\left(\frac{T}{2} - t\right) = f(t) + f\left(\frac{3T}{2} - t\right) = \frac{T}{4} - \frac{t}{2} + \frac{T}{4} - \frac{3T}{4} + \frac{t}{2} = -\frac{T}{4}\,.$$

Für $t = 0$ sowie auch für $t = \frac{T}{2}$ ist $f(t) + f(\frac{T}{2} - t) = 0$, insgesamt ist also

$$f(t) + f\left(\frac{T}{2} - t\right) = \begin{cases} \frac{T}{4} & \text{für } 0 < t < \frac{T}{2} \\ -\frac{T}{4} & \text{für } \frac{T}{2} < t < T \\ 0 & \text{für } t = 0,\ \frac{T}{2} \end{cases},$$

periodisch fortgesetzt mit Periode T. Somit beschreibt $f(t) + f(\frac{T}{2} - t)$ einen Rechteckimpuls.

Da $S_f(t) = f(t)$ für alle $t \in \mathbb{R}$ ist

$$f(t) + f\left(\frac{T}{2} - t\right) = S_f(t) + S_f\left(\frac{T}{2} - t\right)$$

und

$$\begin{aligned} S_f\left(\frac{T}{2} - t\right) &= \sum_{\ell=1}^{\infty} \frac{\sin \ell\omega \left(\frac{T}{2} - t\right)}{\ell\omega} = \sum_{\ell=1}^{\infty} \frac{\sin\left(\frac{\ell\omega T}{2} - \ell\omega t\right)}{\ell\omega} \\ &= \sum_{\ell=1}^{\infty} \frac{\sin(\ell\pi - \ell\omega t)}{\ell\omega}\,. \end{aligned}$$

Für gerades ℓ ist

$$\sin(\ell\pi - x) = \sin(-x) = -\sin x$$

und für ungerades ℓ ist

$$\sin(\ell\pi - x) = \sin(\pi - x) = -\sin(x - \pi) = \sin x\,,$$

denn bei Verschiebung um π wird der $\sin x$ zu $-\sin x$. Damit folgt

$$S_f\left(\frac{T}{2} - t\right) = \sum_{\ell=1}^{\infty} (-1)^{\ell+1} \frac{\sin \ell\omega t}{\ell\omega}$$

und

$$S_f(t) + S_f\left(\frac{T}{2} - t\right) = \sum_{\ell=1}^{\infty} \frac{\sin \ell\omega t}{\ell\omega} + \sum_{\ell=1}^{\infty} (-1)^{\ell+1} \frac{\sin \ell\omega t}{\ell\omega}\,.$$

Für endliche Teilsummen heben sich bei dieser Addition einfach die Terme mit geraden Indizes weg, während die mit ungeradem Index

verdoppelt werden, d.h.

$$\sum_{\ell=1}^{N} \frac{\sin \ell\omega t}{\ell\omega} + \sum_{\ell=1}^{N} (-1)^{\ell+1} \frac{\sin \ell\omega t}{\ell\omega} = 2 \sum_{\ell=1}^{N} \frac{\sin(2\ell-1)\omega t}{(2\ell-1)\omega} .$$

Durch Grenzübergang $N \to \infty$ folgt

$$S_f(t) + S_f\left(\frac{T}{2} - t\right) = 2 \sum_{\ell=1}^{\infty} \frac{\sin(2\ell-1)\omega t}{(2\ell-1)\omega} ,$$

d.h.

$$2 \sum_{\ell=1}^{\infty} \frac{\sin(2\ell-1)\omega t}{(2\ell-1)\omega} = \begin{cases} \frac{T}{4} & \text{für } 0 < t < \frac{T}{2} \\ -\frac{T}{4} & \text{für } \frac{T}{2} < t < T \\ 0 & \text{für } t = 0, \frac{T}{2} \end{cases} .$$

Multiplikation beider Seiten mit $4h/T$ führt wegen

$$\frac{8h}{T\omega} = \frac{8h}{2\pi} = \frac{4h}{\pi}$$

zur Formel

$$\frac{4h}{\pi} \sum_{\ell=1}^{\infty} \frac{\sin(2\ell-1)\omega t}{(2\ell-1)} = \begin{cases} h & \text{für } 0 < t < \frac{T}{2} \\ -h & \text{für } \frac{T}{2} < t < T \\ 0 & \text{für } t = 0, \frac{T}{2} \end{cases} .$$

Damit ist also auch die Konvergenz der FOURIER-Reihe der Rechteckschwingung geklärt. Als kleine Anwendung können wir den Wert $t = \frac{T}{4}$ einsetzen; für diesen ist

$$\sin(2\ell-1)\omega \frac{T}{4} = \sin(2\ell-1)\frac{\pi}{2} = (-1)^{\ell+1} ,$$

d.h.

$$\frac{4h}{\pi} \sum_{\ell=1}^{\infty} \frac{(-1)^{\ell+1}}{(2\ell-1)} = h$$

und somit

$$\sum_{\ell=1}^{\infty} \frac{(-1)^{\ell+1}}{(2\ell-1)} = 1 - \frac{1}{3} + \frac{1}{5} - \cdots = \frac{\pi}{4} .$$

f) Das Gibbs-Phänomen

Wie in den Abbildungen sechs und sieben zu sehen ist, zeigen zumindest die dort dargestellten FOURIER-Polynome Überschwingungen an den Sprungstellen. Wir wollen uns davon überzeugen, daß diese auch bei FOURIER-Polynome mit beliebig vielen Summanden nicht verschwinden.

Beginnen wir mit den Rechteckschwingungen! Wir versuchen zunächst, die Summe

$$S_N(t) = \frac{4h}{\pi} \sum_{\ell=1}^{N} \frac{\sin(2\ell-1)\omega t}{(2\ell-1)}$$

in etwas kompakterer Form darzustellen. Wegen

$$\frac{\sin(2\ell-1)\omega t}{(2\ell-1)\omega} = \int_0^t \cos(2\ell-1)\omega\tau \, d\tau$$

ist

$$\frac{S_N(t)}{\omega} = \frac{4h}{\pi} \sum_{\ell=1}^{N} \int_0^t \cos(2\ell-1)\omega\tau \, d\tau = \frac{4h}{\pi} \int_0^t \sum_{\ell=1}^{N} \cos(2\ell-1)\omega\tau \, d\tau ,$$

und diesen letzten Integranden können wir über seine komplexe Darstellung ausrechnen. Um den Nenner zwei zu eliminieren, berechnen wir den zweifachen Wert

$$2 \sum_{\ell=1}^{N} \cos(2\ell-1)\omega\tau = \sum_{\ell=1}^{N} (e^{(2\ell-1)\cdot i\omega\tau} + e^{-(2\ell-1)\cdot i\omega\tau})$$

$$= \sum_{\ell=-(N-1)}^{N} e^{(2\ell-1)\cdot i\omega\tau} = e^{(-2N+1)\cdot i\omega\tau} \sum_{\ell=0}^{2N-1} e^{2\ell\cdot i\omega\tau} .$$

Letztere Summe ist eine geometrische Reihe mit Quotient $e^{2i\omega\tau}$; nach der Summenformel hat sie den Wert

$$\frac{1 - e^{4N\cdot i\omega\tau}}{1 - e^{2i\omega\tau}} .$$

Die gesuchte Summe ist also

$$e^{-2N\cdot i\omega\tau}\frac{1-e^{4N\cdot i\omega\tau}}{1-e^{2i\omega\tau}} = \frac{e^{-2N\cdot i\omega\tau}}{e^{-i\omega\tau}}\frac{1-e^{4N\cdot i\omega\tau}}{1-e^{2i\omega\tau}}$$
$$= \frac{e^{-2N\cdot i\omega\tau}-e^{2N\cdot i\omega\tau}}{e^{-i\omega\tau}-e^{i\omega\tau}} = \frac{\sin 2N\omega\tau}{\sin\omega\tau}$$

und

$$S_N = \frac{2h\omega}{\pi}\int_0^t \frac{\sin 2N\omega\tau}{\sin\omega\tau}d\tau\,.$$

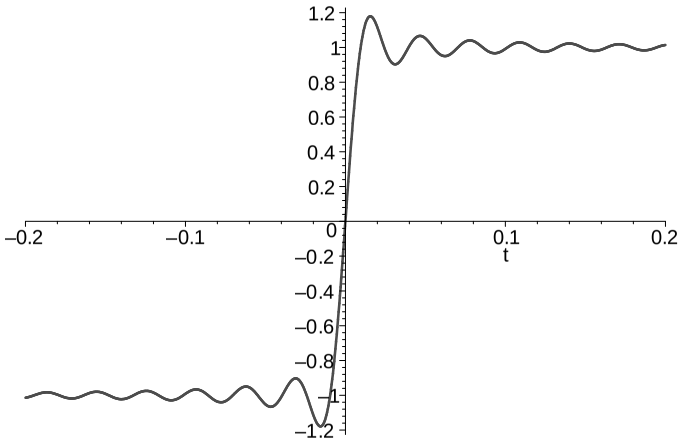


Abb. 11: Das GIBBSphänomen in Großaufnahme

Als nächstes suchen wir nach dem Betrag der Überschwingung. Abbildung elf zeigt diese in Großaufnahme für $S_{100}(t)$; wir wollen das Maximum unmittelbar nach der Sprungstelle bestimmen und das Integral dort auswerten.

Im Maximum verschwindet die Ableitung des Integrals, also der Integrand

$$\frac{\sin 2N\omega\tau}{\sin\omega\tau}\,.$$

Bei der ersten positiven Nullstelle t_N ist $2N\omega t_N = \pi$ die erste positive Nullstelle des Sinus, d.h.

$$t_N = \frac{\pi}{2N\omega}\,.$$

Insbesondere rückt t_0 mit wachsendem N immer näher zur Null; je größer N wird, desto enger lokalisiert wird die Überschwingung.

Mit der Substitution $u = 2N\omega\tau$ wird

$$S_N(t_N) = \frac{2h\omega}{\pi}\int_0^{t_N}\frac{\sin 2N\omega\tau}{\sin\omega\tau}\,d\tau = \frac{2h}{\pi}\int_0^{\pi}\frac{\sin u}{2N\sin\frac{u}{2N}}\,du\,.$$

Für große Werte von N ist das Argument des Sinus im Nenner des Integranden sehr klein; wir machen also keinen großen Fehler, wenn wir den Sinus durch sein Argument annähern, und für $N\to\infty$ geht der Fehler gegen Null. Somit ist

$$\lim_{N\to\infty} S_N(t_N) = \lim_{N\to\infty}\frac{2h}{\pi}\int_0^{\pi}\frac{\sin u}{2N\cdot\frac{u}{2N}}\,du = \frac{2h}{\pi}\int_0^{\pi}\frac{\sin u}{u}\,du\,.$$

Die Stammfunktion von $\frac{\sin u}{u}$ ist nicht in geschlossener Form durch trigonometrische Funktionen, Logarithmen, Exponentialfunktionen und ähnliches ausdrückbar, sie ist aber sehr wichtig und hat daher einen eigenen Namen:

Definition: Die Funktion

$$\mathrm{Si}(t) = \int_0^t \frac{\sin u}{u}\,du$$

heißt *Integralsinus* oder *sinus integralis.*

Der Integralsinus existiert für beliebige reelle Argumente t, denn die Nullstelle des Integranden bei $u = 0$ ist harmlos, da

$$\lim_{u\to 0}\frac{\sin u}{u} = 1$$

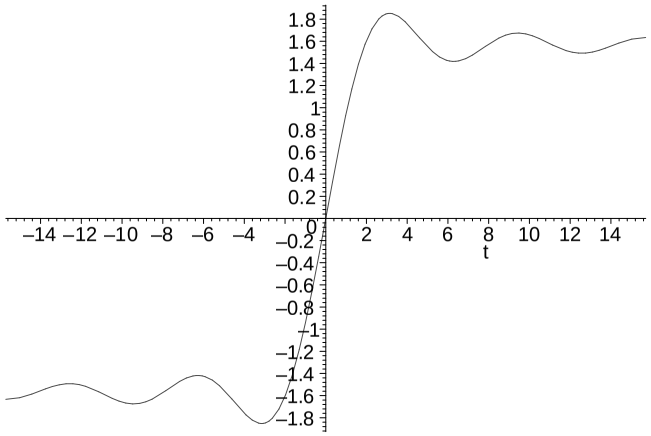


Abb. 12: Der Integralsinus

ist. Wie Abbildung zwölf zeigt, konvergiert er für $t \to \pm\infty$ relativ schnell gegen einen konstanten Wert. Wie am Ende von §1 erwähnt wurde, kann man mit Hilfe des Residuensatzes zeigen, daß

$$\int_{-\infty}^{\infty} \frac{\sin z}{z}\, dz = \pi$$

ist; im Skriptum von 2005 ist diese Rechnung auch nachzulesen. Da der Integrand gerade ist, ist das Integral von Null bis unendlich die Hälfte davon, also $\pi/2$.

Der uns interessierende Wert des Integralsinus an der Stelle π läßt sich nicht in einfacher Weise durch bekannte Konstanten ausdrücken und muß daher numerisch berechnet werden; man erhält

$$\mathrm{Si}(\pi) \approx 1{,}8519305198 \quad \text{und} \quad \frac{2h}{\pi}\,\mathrm{Si}(\pi) \approx 1{,}1789797445 \cdot h\,.$$

Die FOURIER-Polynome überschwingen also auch für $N \to \infty$ den Funktionswert noch um knapp 18%.

Ersetzen wir die Funktion f durch $f+c$ für irgendeine Konstante c, ändert sich nur der konstante Term der FOURIER-Reihe; das GIBBS-Phänomen

bleibt auch in seiner Größe unverändert. Verändert wird dagegen der Funktionswert; es ist daher besser, die Überschwingung auf die unverändert gebliebene Sprunghöhe zu beziehen. Wir sollten daher besser sagen, daß die Überschwingung knapp neun Prozent der Sprunghöhe ausmacht.

Auch beim Sägezahn können wir die Überschwingung explizit berechnen: Aus Abschnitt *e)* wissen wir bereits, daß hier

$$\sum_{\ell=1}^{N} \frac{\sin \ell\omega t}{\ell\omega} - f(t) = \int_{T/2}^{t} \frac{\sin\left(N+\frac12\right)\omega\tau}{2\sin\frac{\omega\tau}{2}}\, d\tau$$

ist. Auch die Ableitung dieser Funktion ist wieder der Integrand; für ihre erste positive Nullstelle t_N ist

$$\left(N+\frac12\right)\omega t_N = \pi\,, \quad \text{d.h.} \quad t_N = \frac{\pi}{\left(N+\frac12\right)\omega} = \frac{T}{2N+1}\,.$$

Mit der Substitution $u = \left(N+\frac12\right)\omega\tau$ wird

$$\int_{T/2}^{t_N} \frac{\sin\left(N+\frac12\right)\omega\tau}{2\sin\frac{\omega\tau}{2}}\, d\tau = \int_{(N+\frac12)\,\pi}^{\pi} \frac{\sin u}{(2N+1)\omega\sin\frac{u}{2N+1}}\, du\,.$$

Wieder können wir den Sinus im Nenner für große N durch sein Argument ersetzen und erhalten somit als Limes für $N \to \infty$

$$\lim_{N\to\infty} \int_{(N+\frac12)\,\pi}^{\pi} \frac{\sin u}{\omega u}\, du = \frac{1}{\omega}\left(\mathrm{Si}(\pi) - \lim_{t\to\infty}\mathrm{Si}(t)\right) = \frac{\mathrm{Si}(\pi) - \frac{\pi}{2}}{\omega}$$

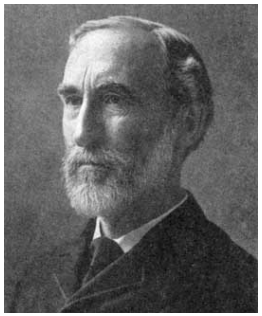
$$= \frac{T}{2\pi}\,\mathrm{Si}(\pi) - \frac{T}{4}\,.$$

Dabei ist $T/4$ der Wert, der überschwungen wird, der Maximalwert des FOURIER-Polynoms geht also für $N \to \infty$ gegen $\frac{T}{2\pi}\,\mathrm{Si}(\pi)$. Dividiert man dies durch $\frac{T}{4}$, ergibt sich wieder der Quotient

$$\frac{T}{2\pi}\,\mathrm{Si}(\pi)\cdot\frac{4}{T} = \frac{2}{\pi}\,\mathrm{Si}(\pi)\,,$$

auch hier gibt es also wieder eine Überschwingung um knapp 18% des Funktionswerts oder knapp neun Prozent der Sprunghöhe.

Dies ist kein Zufall: Wie wir im nächsten Paragraphen sehen werden, tritt an *jeder* Unstetigkeitsstelle einer Funktion das GIBBS-Phänomen ein, wonach die FOURIER-Polynome den Sprung um knapp neun Prozent überschwingen.



JOSIAH WILLARD GIBBS (1839–1903) promovierte 1863 an der amerikanischen Yale-Universität mit einer Arbeit über Zahnradgetriebe; die erste amerikanische Dissertation auf dem Gebiet des Ingenieurwesens. Danach unterrichtete er in Yale Latein und Naturphilosophie, bis er 1866 nach Europa fuhr, wo er 1868/1869 in Heidelberg bei KIRCHHOFF und HELMHOLTZ studierte. 1871 wurde er in Yale Professor für mathematische Physik, 1873 publizierte er seine erste Arbeit, die sich, wie viele spätere, mit Thermodynamik befaßte. Sehr einflußreich waren auch seine Arbeiten zur elektromagnetischen Theorie des Lichts und zur Vektoranalysis.

g) Die Besselsche Ungleichung

In einem EUKLIDischen oder HERMITEschen Vektorraum V mit (HERMITEschem) Skalarprodukt $(\cdot,\cdot)$ gilt für jede Orthonormalbasis $\mathcal{B}$, daß für zwei Vektoren

$$\vec{v} = \sum_{\vec{b}\in\mathcal{B}} \lambda_{\vec{b}}\vec{b} \quad \text{und} \quad \vec{w} = \sum_{\vec{b}\in\mathcal{B}} \mu_{\vec{b}}\vec{b}$$

das Skalarprodukt berechnet werden kann als

$$(\vec{v},\vec{w}) = \sum_{\vec{b}\in\mathcal{B}} \lambda_{\vec{b}}\overline{\mu_{\vec{b}}}\,,$$

wobei wegen der Basiseigenschaft von $\mathcal{B}$ natürlich wieder alle Summen endlich sind, auch wenn die Basis $\mathcal{B}$ unendlich sein sollte. Insbesondere ist

$$(\vec{v},\vec{v}) = \sum_{\vec{b}\in\mathcal{B}} \lambda_{\vec{b}}\overline{\lambda_{\vec{b}}} = \sum_{\vec{b}\in\mathcal{B}} |\lambda_{\vec{b}}|^2 \ .$$

Für trigonometrische Polynome

$$f(t) = \sum_{k\in\mathbb{Z}} c_k e^{k\cdot i\omega t} \in \mathrm{P}_T(\mathbb{R},\,\mathbb{C})$$

ist daher $(f,\,f) = \sum_{k\in\mathbb{Z}} |c_k|^2$, wobei auch hier wieder, da es sich um ein trigonometrisches *Polynom* handelt, in beiden unendlichen Summen nur endlich viele Summanden ungleich null sind.

$\mathrm{L}_T(\mathbb{R},\,\mathbb{R})$ ist zwar kein EUKLIDischer Vektorraum und $\mathrm{L}_T(\mathbb{R},\,\mathbb{C})$ kein HERMITEscher, aber wir hoffen doch, daß sich die trigonometrischen Funktionen *bzw.* komplexen Exponentialfunktionen wenigstens so ähnlich verhalten wie eine Orthonormalbasis; vielleicht sollte auch gelten, daß für eine Funktion $f \in \mathrm{L}_T(\mathbb{R},\,\mathbb{C})$ mit FOURIER-Reihe

$$f(t) = \sum_{k\in\mathbb{Z}} c_k e^{k\cdot i\omega t} \quad \text{gilt} \quad (f,\,f) = \sum_{k\in\mathbb{Z}} |c_k|^2 \ ,$$

obwohl *hier* wirklich unendliche Summen stehen können.

Bei der Untersuchung der Konvergenz von FOURIER-Reihen wird diese Frage eine wesentliche Rolle spielen; als ersten Einstieg dazu beweisen wir die BESSELsche Ungleichung:

Lemma: Für die FOURIER-Koeffizienten $c_k = \frac{1}{T}\int_0^T f(t)e^{-k\cdot i\omega t}\,dt$ einer Funktion $f \in \mathrm{L}_T(\mathbb{R},\,\mathbb{C})$ ist

$$\sum_{k=-\infty}^{\infty} c_k\overline{c_k} = \sum_{k=-\infty}^{\infty} |c_k|^2 \le \frac{1}{T}\int_0^T f(t)\overline{f(t)}\,dt = \frac{1}{T}\int_0^T |f(t)|^2\,dt\,;$$

insbesondere konvergiert die linke Summe also.

Beweis: Die rechte Seite der Ungleichung ist gerade das HERMITEsche Produkt $(f,\,f)$, und auch die FOURIER-Koeffizienten lassen sich als Produkte

$$c_k = (f,\,e^{k\cdot i\omega t})$$

schreiben. Wir wollen noch einige weitere Produkte ausrechnen.

Zunächst definieren wir die Teilsumme

$$S_N(t) \underset{\text{def}}{=} \sum_{k=-N}^{N} c_k e^{k\cdot i\omega t}\,.$$

Dies ist ein trigonometrisches Polynom, und wir wissen daher schon aus §2*a)*, daß

$$(S_N,\, S_N) = \sum_{k=-N}^{N} |c_k|^2$$

ist.

Als nächstes betrachten wir das Produkt

$$(f - S_N,\, f - S_N) = (f,\, f) - (S_N,\, f) - (f,\, S_N) + (S_N,\, S_N)\,.$$

Auch wenn das HERMITEsche Produkt auf $\mathrm{L}_T(\mathbb{R},\, \mathbb{C})$ nicht positiv definit ist, wissen wir doch, daß die linke (und damit auch die rechte) Seite zumindest nichtnegativ ist.

Wegen der Linearität des HERMITEschen Produkts im ersten Argument ist weiter

$$\begin{aligned}(S_N,\, f) &= \left(\sum_{k=-N}^{N} c_k e^{k\cdot i\omega t},\, f\right) = \sum_{k=-N}^{N} c_k \left(e^{k\cdot i\omega t},\, f\right)\\ &= \sum_{k=-N}^{N} c_k \overline{(f,\, e^{k\cdot i\omega t})} = \sum_{k=-N}^{N} c_k \overline{c_k} = \sum_{k=-N}^{N} |c_k|^2\\ &= (S_N,\, S_N)\,,\end{aligned}$$

und da dies eine reelle Zahl ist, folgt auch

$$(f,\, S_N) = \overline{(S_N,\, f)} = (S_N,\, f) = (S_N,\, S_N)\,.$$

Fassen wir alles zusammen, ist also

$$0 \le (f - S_N,\, f - S_N) = (f,\, f) - (S_N,\, S_N)$$

und damit

$$(S_N,\, S_N) \le (f, f) \quad \text{für alle} \quad n \in \mathbb{N}\,.$$

Ausgeschrieben wird das zu

$$\sum_{k=-N}^{N} |c_k|^2 \le \frac{1}{T}\int_0^T |f(t)|^2\, dt \quad \text{für alle} \quad n \in \mathbb{N}\,,$$

und der Grenzübergang $N \to \infty$ führt zur gewünschten Ungleichung

$$\sum_{k=-\infty}^{\infty} |c_k|^2 \le \frac{1}{T}\int_0^T |f(t)|^2\, dt\,,$$

wie behauptet. ∎

FRIEDRICH WILHELM BESSEL (1784–1846) verließ die Schule schon im Alter von 14 Jahren und wurde Lehrling eines Handelshauses. Dessen Überseehandel veranlaßte ihn zur Beschäftigung mit Geographie, Spanisch und Englisch und schließlich (für die Navigation) mit Astronomie. Nachdem er 1804 die Bahn des HALLEYschen Kometen berechnet hatte, bekam er 1806 eine Stelle als Astronom eines privaten Observatoriums. Seine Beiträge zur Mathematik entstanden aus seinen astronomischen Berechnungen. 1809 wurde er zum Direktor des Königsberger Observatoriums und zum Professor der Astronomie ernannt.

Aufgrund der Beziehungen zwischen reellen und komplexen FOURIER-Reihen (für die genau dieselben Formeln gelten, die wir in Abschnitt *d)* für trigonometrische Polynome hergeleitet haben) folgt, daß für die reelle FOURIER-Reihe einer reellwertigen Funktion gilt

$$|a_0|^2 + \frac{1}{2}\sum_{k=1}^{\infty}\left(|a_k|^2 + |b_k|^2\right) \le \frac{1}{T}\int_0^T f(t)^2\, dt\,.$$

Insbesondere konvergieren also auch die beiden Reihen

$$\sum_{k=0}^{\infty} a_k^2 \quad \text{und} \quad \sum_{\ell=1}^{\infty} b_\ell^2\,.$$

h) Harmonische Analyse als lineare Abbildung

Die BESSELsche Ungleichung zeigt, daß nicht alle Folgen reeller oder komplexer Zahlen als FOURIER-Koeffizienten eine stückweise stetigen Funktion auftreten können: Zumindest muß die Summe ihrer Betragsquadrate konvergieren, d.h. die $c\colon \mathbb{Z} \to \mathbb{C}$ mit $c(k) = c_k$, die jedem $k \in \mathbb{Z}$ den FOURIER-Koeffizienten c_k zuordnet, liegt im Vektorraum

$$\ell^2(\mathbb{Z},\, \mathbb{C}) \underset{\text{def}}{=} \left\{ c\colon \mathbb{Z} \to \mathbb{C} \;\middle|\; \sum_{k=-\infty}^{\infty} |c(k)|^2 < \infty \right\}.$$

Um den Zusammenhang zwischen Funktionen und Koeffizientenfolgen besser zu verstehen, definieren wir für $f \in \mathrm{L}_T(\mathbb{R},\, \mathbb{C})$ die Funktion

$$\widehat{f}\colon \begin{cases} \mathbb{Z} \to \mathbb{C} \\ k \mapsto \dfrac{1}{T} \displaystyle\int_0^T f(t) e^{-k \cdot i\omega t}\, dt \end{cases};$$

$\widehat{f}(k)$ ist also gerade der k-te FOURIER-Koeffizient von f.

Lemma: Die Zuordnung $f \mapsto \widehat{f}$ definiert eine lineare Abbildung von $\mathrm{L}_T(\mathbb{R},\, \mathbb{C})$ nach $\ell^2(\mathbb{Z},\, \mathbb{C})$.

Der *Beweis* ist trivial wegen der Linearität der Integration. ∎

Als nächstes definieren wir auf $\ell^2(\mathbb{Z},\, \mathbb{C})$ ein HERMITEsches Skalarprodukt durch die Vorschrift

$$(c,\, d) \underset{\text{def}}{=} \sum_{k=-\infty}^{\infty} c_k \overline{d_k}\,.$$

Wir müssen zeigen, daß dies erstens wohldefiniert ist, daß die Summe also überhaupt konvergiert, und daß es zweitens alle Forderungen an ein HERMITEsches Skalarprodukt erfüllt.

Da für zwei komplexe Zahlen z und w ist $\big(|z| - |w|\big)^2 \geq 0$, also

$$|zw| = |z\overline{w}| \leq \frac{1}{2}\big(|z|^2 + |w|^2\big)\,.$$

Für zwei Funktionen $c, d \in \ell^2(\mathbb{Z},\, \mathbb{C})$ ist daher für jede natürliche Zahl N

$$\left| \sum_{k=-N}^{N} c(k)\overline{d(k)} \right| \leq \sum_{k=-N}^{N} \left| c(k)\overline{d(k)} \right| \leq \frac{1}{2} \left(\sum_{k=-N}^{N} |c(k)|^2 + \sum_{k=-N}^{N} |d(k)|^2 \right).$$

Nach Definition von $\ell^2(\mathbb{Z},\, \mathbb{C})$ konvergiert die rechte Seite für $N \to \infty$, also auch die linke.

Das Nachrechnen der Forderungen an ein HERMITEsches Skalarprodukt ist nun einfach: Abgesehen von der Tatsache, daß die Summen nicht mehr endlich sind, geht alles ganz genauso wie beim HERMITEschen Standardskalarprodukt auf $\mathbb{C}^n$.

Mit diesem Skalarprodukt ausgedrückt bekommt die BESSELsche Ungleichung die kompakte Form

$$\left(\widehat{f},\, \widehat{f}\right) \leq (f,\, f) \quad \text{für alle} \quad f \in \mathrm{L}_T(\mathbb{R},\, \mathbb{C})\,.$$

§ 4: Periodische Faltungen

Abgesehen von den beiden Beispielen der Rechteckschwingung und des Sägezahns wissen wir bislang noch von keiner FOURIER-Reihe, ob und gegebenenfalls wohin sie konvergiert. In diesem Paragraphen soll dies zumindest für stückweise differenzierbare Funktionen geklärt werden. Als zentral wird sich dabei eine neue Konstruktion herausstellen, die *Faltung zweier Funktionen.*

a) Faltungen

Definition: Für $f, g \in \mathrm{L}_T(\mathbb{R},\, \mathbb{C})$ bezeichnen wir die Funktion

$$f \star g(t) = \frac{1}{T} \int_0^T f(t - \tau) g(\tau)\, d\tau$$

als (periodische) Faltung von f und g.

Anschaulich kann man sich die Faltung als eine Art Mittelung von f vorstellen mit einer Gewichtsfunktion g. Im zweidimensionalen (nichtperiodischen) Analogon kann man beispielsweise ein optisch defokusiertes Bild so beschreiben: Bei einer perfekten optischen Abbildung einer Ebene hängt jeder Bildpunkt von genau einem Punkt der Ebenen ab; ist das System aber defokusiert, so kommen auch noch Einflüsse der Nachbarpunkte dazu, die umso größer sind, je näher die Punkte beieinanderliegen. Eine Helligkeitsverteilung $f(s,t)$ wird dann abgebildet auf

$$F(s,t) = \iint_{\mathbb{R}^2} e^{\left(s-\sigma)^2+(t-\tau)^2\right)/2a} f(\sigma,\tau)\, d\sigma\, d\tau\,,$$

wobei der Parameter a umso größer ist, je stärker das Bild defokusiert ist. Für kleines a ist der Effekt also eher ein Weichzeichnen als eine echte Unschärfe, und dieser Glättungseffekt ist ein allgemeines Charakteristikum von Faltungen:

Lemma: Für $f, g \in \mathrm{L}_T(\mathbb{R},\ \mathbb{C})$ ist $f \star g$ eine stetige Funktion.

(Man beachte, daß f und g beide nur als stückweise stetig vorausgesetzt sind!)

Beweis: Als stückweise stetige periodische Funktion ist g insbesondere beschränkt: Für jedes offene Intervall $(t_j,\ t_{j+1})$, in dem f stetig ist, müssen nach Definition der stückweisen Stetigkeit der rechtsseitige Grenzwert $\lim\limits_{\substack{t\to t_j\\ t>t_j}} g(t)$ und der linksseitige Grenzwert $\lim\limits_{\substack{t\to t_{j+1}\\ t<t_{j+1}}} g(t)$ existieren; damit kann die Einschränkung von g auf das offene Intervall $(t_j,\ t_{j+1})$ fortgesetzt werden zu einer stetigen Funktion auf dem abgeschlossenen Intervall $[t_j,\ t_{j+1}]$ (die an den Intervallenden natürlich nicht mit g übereinstimmen muß). Damit hat der Betrag dieser Funktion ein endliches Maximum M_j, das auch eine Schranke für g im offenen Intervall $(t_j,\ t_{j+1})$ ist. Nimmt man nun als M das Maximum aller M_j sowie auch der Beträge $\left|g(t_j)\right|$ der Funktionswerte an den potentiellen Sprungstellen, so ist $|g(t)| \le M$ für alle $t \in [0,\ T]$ und damit auch für alle $t \in \mathbb{R}$.

Seien nun t_1 und t_2 Punkte aus $\mathbb{R}$; dann ist

$$\begin{aligned} |f\star g(t_1) - f\star g(t_2)| &\le \frac{1}{T}\int_0^T |f(t_1-\tau) - f(t_2-\tau)|\,|g(\tau)|\, d\tau \\ &\le \frac{M}{T}\int_0^T |f(t_1-\tau) - f(t_2-\tau)|\, d\tau\,. \end{aligned}$$

Das noch verbliebene Integral mißt die Fläche zwischen den Graphen von $f(t)$ und $f(t+t_2-t_1)$ über eine Periode von f; wegen der stückweisen Stetigkeit von f geht diese gegen null für $t_2 \to t_1$. ∎

b) Die Fourier-Reihe einer Faltung

Die Nützlichkeit von Faltungen für FOURIER-Reihen ergibt sich aus folgender Formel:

Satz: Für $f, g \in \mathrm{L}_T(\mathbb{R}, \mathbb{C})$ ist

$$\widehat{f\star g}(k) = \widehat{f}(k)\cdot\widehat{g}(k) \quad \text{für alle} \quad k \in \mathbb{Z}\,;$$

die komplexen FOURIER-Koeffizienten von $f \star g$ sind also gerade die Produkte der komplexen FOURIER-Koeffizienten von f und von g.

Der *Beweis* erfolgt durch Nachrechnen: Der k-te FOURIER-Koeffizient

c_k von $f \star g$ ist

$$c_k = \frac{1}{T}\int_0^T f\star g(t)e^{-k\cdot i\omega t}\,dt = \frac{1}{T^2}\int_0^T\left(\int_0^T f(t-\tau)g(\tau)\,d\tau\right)e^{-k\cdot i\omega t}\,dt$$
$$= \frac{1}{T^2}\iint\limits_{\substack{0\le t\le T\\ 0\le\tau\le T}} f(t-\tau)g(\tau)e^{-k\cdot i\omega t}\,d\tau\,dt$$
$$= \frac{1}{T^2}\iint\limits_{\substack{0\le t\le T\\ 0\le\tau\le T}} f(t-\tau)e^{-k\cdot i\omega(t-\tau)}\cdot g(\tau)e^{-k\cdot i\omega\tau}\,dt\,d\tau$$
$$= \frac{1}{T}\int_0^T\left(\frac{1}{T}\int_0^T f(t-\tau)e^{-k\cdot i\omega(t-\tau)}\,dt\right)g(\tau)e^{-k\cdot i\omega\tau}\,d\tau\,.$$

Der Inhalt der letzten Klammer kann mit Hilfe der Substitution $u = t-\tau$ im Integral berechnet werden:

$$\frac{1}{T}\int_0^T f(t-\tau)e^{-k\cdot i\omega(t-\tau)}\,dt = \frac{1}{T}\int_{-\tau}^{T-\tau} f(u)e^{-k\cdot i\omega u}\,du = \widehat{f}(k)\,,$$

denn wie wir uns schon überlegt haben, kommt es bei einer periodischen Funktion nicht darauf an, über welches Intervall der Länge T wir integrieren. Somit ist

$$c_k = \frac{1}{T}\int_0^T \widehat{f}(k)\cdot g(\tau)e^{-k\cdot i\omega\tau}\,d\tau = \widehat{f}(k)\cdot\widehat{g}(k)\,,$$

wie behauptet. ∎

Sind $S_f(t) = \sum_{k=-\infty}^{\infty} c_k e^{k\cdot i\omega t}$ und $S_g(t) = \sum_{k=-\infty}^{\infty} d_k e^{k\cdot i\omega t}$ die FOURIER-Reihen von f und g, ist die FOURIER-Reihe von $f\star g$ somit

$$S_{f\star g}(t) = \sum_{k=-\infty}^{\infty} c_k d_k e^{k\cdot i\omega t}\,.$$

Nach der BESSELschen Ungleichung konvergieren die Summen

$$\sum_{k=-\infty}^{\infty}|c_k|^2 \quad\text{und}\quad \sum_{k=-\infty}^{\infty}|d_k|^2\;;$$

außerdem ist für jedes k

$$|c_k d_k| \le \frac{1}{2}\left(|c_k|^2 + |d_k|^2\right);$$

also konvergiert auch

$$\sum_{k=-\infty}^{\infty}|c_k d_k| = \sum_{k=-\infty}^{\infty}\left|c_k d_k e^{k\cdot i\omega t}\right|\,.$$

(Man beachte, daß $\left|e^{k\cdot i\omega t}\right| = 1$ ist für jede reelle Zahl t.) Damit haben wir gezeigt, daß die FOURIER-Reihe von $f\star g$ absolut und gleichmäßig konvergiert.

Das hat eine wichtige Konsequenz:

Lemma: Konvergiert die FOURIER-Reihe einer Funktion h gleichmäßig gegen eine Funktion S_h, so ist S_h stetig und hat dieselben FOURIER-Koeffizienten wie h.

Beweis: Die FOURIER-Reihe von h sei

$$S_h(t) = \sum_{k=-\infty}^{\infty} c_k e^{k\cdot i\omega t}\,.$$

Da alle Summanden $c_k e^{k\cdot i\omega t}$ stetige Funktionen sind, ist wegen der gleichmäßigen Konvergenz der Reihe auch die Summe eine stetige Funktion; deren k-ter FOURIER-Koeffizient ist

$$\widehat{S_h}(k) = \left(S_h,\, e^{k\cdot i\omega t}\right) = \lim_{N\to\infty}\sum_{\ell=-N}^{N} c_\ell\left(e^{\ell\cdot i\omega t}, e^{k\cdot i\omega t}\right) = c_k\,,$$

wie behauptet. ∎

Insgesamt haben wir damit bewiesen

Satz: Für $f, g \in \mathrm{L}_T(\mathbb{R}, \mathbb{C})$ konvergiert die FOURIER-Reihe von $f \star g$ absolut und gleichmäßig gegen eine stetige Funktion. Diese hat dieselben FOURIER-Koeffizienten wie $f \star g$. ∎

Damit wissen wir zwar immer noch nicht, *wohin* die FOURIER-Reihe von $f \star g$ konvergiert, aber wir wissen immerhin, *daß* sie für Funktionen, die als Faltungen darstellbar sind, konvergiert, und wir wissen auch, daß für die Differenz zwischen Ausgangs- und Grenzfunktion sämtliche FOURIER-Koeffizienten verschwinden. Wir müssen daher einerseits Funktionen mit verschwindenden FOURIER-Koeffizienten genauer untersuchen und andererseits versuchen, eine möglichst große Klasse von Funktionen als Faltungen darzustellen, auf die wir den gerade bewiesenen Satz anwenden können.

c) Faltung mit einem Sägezahn

Als erstes konkretes Beispiel (von dem sich zeigen wird, daß es zumindest einen Teil der zweiten Aufgabe lösen wird) betrachten wir für eine beliebige Funktion $f \in \mathrm{L}_T(\mathbb{R}, \mathbb{C})$ die Faltung mit dem Sägezahn

$$s(t) = \frac{T}{2} - t \quad \text{für} \quad 0 < t < T \quad \text{und} \quad s(0) = 0\,,$$

periodisch fortgesetzt mit Periode T; es handelt sich hier um das Zweifache der in §2*c)* betrachteten Funktion.

Für $t, \tau \in [0,\, T)$ liegt $t - \tau$ genau dann wieder in $[0,\, T)$, wenn $\tau \le t$ ist; andernfalls liegt $t - \tau$ im Intervall $(-T,\, 0)$, wo

$$s(t) = s(t+T) = \frac{T}{2} - (t+T) = -\frac{T}{2} - t$$

ist. Somit ist für $t, \tau \in (0,\, T)$

$$s(t-\tau) = \begin{cases} \frac{T}{2} - t + \tau & \text{für } \tau < t \\ 0 & \text{für } \tau = t \\ -\frac{T}{2} - t + \tau & \text{für } \tau > t \end{cases}$$

und $s \star f(t)$ ist gleich

$$\begin{aligned}
&\frac{1}{T}\int_0^T s(t-\tau) f(\tau)\,d\tau \\
&= \frac{1}{T}\int_0^t \left(\frac{T}{2} - (t-\tau)\right) f(\tau)\,d\tau + \frac{1}{T}\int_t^T \left(-\frac{T}{2} - (t-\tau)\right) f(\tau)\,d\tau \\
&= \frac{1}{T}\left(\frac{T}{2}\int_0^t f(\tau)\,d\tau - \frac{T}{2}\int_t^T f(\tau)\,d\tau + \int_0^T (\tau - t) f(\tau)\,d\tau\right) \\
&= \frac{1}{2}\left(\int_0^t f(\tau)\,d\tau - \int_t^T f(\tau)\,d\tau\right) + \frac{1}{T}\int_0^T (\tau - t) f(\tau)\,d\tau\,.
\end{aligned}$$

Ist $F(t)$ eine Stammfunktion von $f(t)$, so ist der erste Summand gleich

$$\frac{1}{2}\big(F(t) - F(0) - F(T) + F(t)\big) = F(t) - \frac{F(0) + F(T)}{2}\,.$$

Auch das letzte Integral läßt sich durch partielle Integration weiter ausrechnen zu

$$\begin{aligned}
\int_0^T (\tau - t) f(\tau)\,d\tau &= (\tau - t) F(\tau)\,\Big|_0^T - \int_0^T F(\tau)\,d\tau \\
&= (T - t) F(T) + t F(0) - \int_0^T F(\tau)\,d\tau \\
&= t\big(F(0) - F(T)\big) + T F(T) - \int_0^T F(\tau)\,d\tau\,.
\end{aligned}$$

Insgesamt ist also

$$s \star f(t) = F(t) + \frac{t}{T}\big(F(0) - F(T)\big) - \frac{F(0) + F(T)}{2} + F(T) - \frac{1}{T}\int_0^T F(\tau)\,d\tau\,.$$

Abgesehen von dem Term $\frac{t}{T}\big(F(0) - F(T)\big)$ ist das eine Stammfunktion von $f(t)$, denn die drei hinteren Terme sind Konstanten, die nicht von t abhängen. Ist also insbesondere $F(T) = F(0)$, so ist $s \star f$ eine Stammfunktion von f. Dies wollen wir ausnutzen, um differenzierbare Funktionen sowie eine leichte Verallgemeinerung davon als Faltungen auszudrücken und so die Konvergenz ihrer FOURIER-Reihen zu beweisen.

d) Fourier-Reihen stetiger stückweise differenzierbarer Funktionen

Definition: Eine stückweise stetige Funktion heißt stückweise differenzierbar, wenn es nur isolierte Punkte gibt, in denen f nicht stetig differenzierbar ist, und wenn auch in diesen Ausnahmepunkten sowohl der linksseitige als auch der rechtsseitige Grenzwert von $\dot{f}(t)$ existieren.

Hier interessieren wir uns für periodische Funktionen; für diese bedeutet die Definition, daß es pro Periodenintervall höchstens endliche viele Punkte geben darf, in denen die Ableitung nicht definiert ist, aber auch dort muß sie einen linksseitigen und einen rechtsseitigen Grenzwert haben.

Einfache Beispiele stückweise differenzierbarer Funktionen sind die Rechteckschwingungen aus §2*b)*, die überall außer in den Sprungstellen stetig differenzierbar sind und Ableitung null haben; in den Sprungstellen verschwindet daher auch sowohl der linksseitige als auch der rechtsseitige Grenzwert. Die Funktion ist aber trotzdem nicht differenzierbar in den Sprungstellen, da sie dort nicht einmal stetig ist. (Wäre f differenzierbar, so würde die Ableitung identisch verschwinden, die Funktion müßte also nach dem üblichen Argument über den Mittelwertsatz der Differentialrechnung konstant sein.)

Genauso ist beim Sägezahn aus §2*c)* die Ableitung überall außer in den Sprungstellen gleich -1; in den Sprungstellen ist die Funktion nicht differenzierbar, aber beide Grenzwerte der Ableitung sind gleich -1.

Weiteres Beispiel einer stetigen stückweise differenzierbaren Funktion ist etwa

$$f(t) = |t| \quad \text{für} \quad |t| \le 1\,,$$

periodisch fortgesetzt mit Periode zwei. Für alle $t \notin \mathbb{Z}$ ist f differenzierbar; falls die größte ganze Zahl kleiner t gerade ist, ist die Ableitung +1, ansonsten -1. Bei einer geraden ganzen Zahl ist der linksseitige Grenzwert der Ableitung -1 und der rechtsseitige +1, bei einer ungeraden ist es umgekehrt.

Für eine stückweise differenzierbare Funktion können wir nicht wirklich von der abgeleiteten Funktion reden, da diese nicht in jedem Punkt existieren muß. Wir können aber eine Funktion $\varphi(t)$ definieren, die überall dort mit $\dot{f}(t)$ übereinstimmt, wo $\dot{f}(t)$ existiert; in den übrigen Punkten setzen wir $\varphi(t)$ auf irgendeinen beliebigen Wert, zum Beispiel auf null.

Die Funktion

$$\widetilde{f}(t) = \int_0^t \varphi(\tau)\, d\tau\,.$$

ist stetig und stückweise differenzierbar, und sie hängt nicht ab vom Wert von φ in den Ausnahmepunkten; ihre Ableitung stimmt dort, wo sie definiert ist, mit φ überein. Falls f stetig ist, unterscheidet sich $\widetilde{f}$ daher nur um eine Konstante von f, und auch f ist eine Stammfunktion von φ.

Setzen wir nun noch zusätzlich voraus, daß $f \in \mathrm{L}_T(\mathbb{R}, \mathbb{C})$ periodisch ist, folgt aus der Rechnung im vorigen Abschnitt, daß sich $f(t)$ nur um eine Konstante vom Faltungsprodukt $s \star \varphi$ unterscheidet, denn

$$f(T) - f(0) = 0$$

für eine Funktion mit Periode T.

Aus Abschnitt *b)* wissen wir, daß die FOURIER-Reihe von $s \star \varphi$ wie auch von jeder anderen Faltung absolut und gleichmäßig konvergiert; da die Addition einer Konstanten hieran nichts ändert folgt also zusammen mit den übrigen Resultaten aus Abschnitt *b)*

Satz: Ist $f \in \mathrm{L}_T(\mathbb{R}, \mathbb{C})$ stetig und stückweise differenzierbar, konvergiert die FOURIER-Reihe von f absolut und gleichmäßig gegen eine stetige Funktion S_f, die dieselben FOURIER-Koeffizienten hat wie f. ∎

e) Der Eindeutigkeitssatz

Auch wenn wir nun wissen, daß die FOURIER-Reihe zumindest für stetige stückweise differenzierbare Funktionen konvergiert, wissen wir noch nicht, wohin sie konvergiert. Diese Frage soll in diesem Abschnitt geklärt werden.

Beginnen wir mit dem einfachsten Fall einer Funktion, deren sämtliche FOURIER-Koeffizienten verschwinden, und die außerdem noch im Vektorraum $L_T^0(\mathbb{R}, \mathbb{C})$ liegt, d.h. in jedem Punkt $t \in \mathbb{R}$ ist $f(t)$ der Mittelwert aus dem rechtsseitigen Grenzwert $f(t^+)$ und dem linklsseitigen Grenzwert $f(t^-)$.

Satz: Sind für $f \in L_T^0(\mathbb{R}, \mathbb{C})$ alle FOURIER-Koeffizienten null, ist auch $f(t) \equiv 0$.

Beweis: Auch hier arbeiten wir wieder mit einem Faltungsintegral, und zwar wollen wir versuchen, die Funktion f als Faltung von sich selbst mit einer geeigneten Funktion g auszudrücken; wir suchen also nach einer Funktion g, so daß $f \star g = f$ ist. Falls g diese Eigenschaft für *beliebige* Funktionen f haben soll, die durch ihre FOURIER-Reihe dargestellt werden, müssen dann alle FOURIER-Koeffizienten von g gleich eins sein, denn bei einer Faltung zweier Funktionen multiplizieren sich die FOURIER-Koeffizienten. g hat also die FOURIER-Reihe

$$S_g(t) = \sum_{k=-\infty}^{\infty} e^{k\cdot i\omega t} \quad \text{mit} \quad \omega = \frac{2\pi}{T},$$

was offensichtlich unmöglich ist: Nach der BESSELschen Ungleichung müßte sonst nämlich

$$\sum_{k=-\infty}^{\infty} |c_k|^2 = \sum_{k=-\infty}^{\infty} 1$$

konvergieren, was natürlich nicht der Fall ist.

Wir können aber eine kleine Modifikation dieser Reihe, betrachten, nämlich

$$I_r(t) = \sum_{k=-\infty}^{\infty} r^{|k|} e^{k\cdot i\omega t}.$$

Hier sorgt der Exponent $|k|$ im Falle $|r| < 1$ für eine starke Dämpfung der Koeffizienten mit großem Index, so daß es zumindest mit der BESSELschen Ungleichung keine Probleme mehr gibt: Die entsprechende Summe ist zusammengesetzt aus zwei konvergenten geometrischen Reihen. Falls auch $I_r(t)$ selbst konvergiert, können wir damit rechnen und hoffen, daß wir irgendwann einmal den Grenzübergang $r \to 1$ machen können, was dann *ungefähr* der Faltung mit der nicht existenten Funktion g entspricht.

Wir berechnen $I_r(t)$ über zwei geometrische Reihen:

Beschränken wir uns zunächst auf positive Indizes, so ist nach der Summenformel

$$\sum_{k=0}^{\infty} r^{|k|} e^{k\cdot i\omega t} = \frac{1}{1 - re^{i\omega t}} = \frac{1 - re^{-i\omega t}}{(1 - re^{i\omega t})(1 - re^{-i\omega t})}$$
$$= \frac{1 - r\cos\omega t + ir\sin\omega t}{1 + r^2 - 2r\cos\omega t}.$$

Ersetzt man in einem der Summanden k durch $-k$, so ändert sich nichts am Koeffizienten $r^{|k|}$ und auch nichts am Realteil von $e^{k\cdot i\varphi}$; der Imaginärteil allerdings ändert sein Vorzeichen. Somit ist

$$\sum_{k=-\infty}^{0} r^{|k|} e^{k\cdot i\omega t} = \frac{1 - r\cos\omega t - ir\sin\omega t}{1 + r^2 - 2r\cos\omega t}$$

der konjugiert komplexe Wert zu obiger Summe. Der Summand eins für $k = 0$ wurde in beiden Summen berücksichtigt, tritt aber in der Gesamtsumme nur einmal auf; also ist

$$I_r(t) = \frac{1 - r\cos\omega t + ir\sin\omega t}{1 + r^2 - 2r\cos\omega t} + \frac{1 - r\cos\omega t - ir\sin\omega t}{1 + r^2 - 2r\cos\omega t} - 1$$
$$= \frac{2 - 2r\cos\omega t}{1 + r^2 - 2r\cos\omega t} - 1 = \frac{2 - 2r\cos\omega t - (1 + r^2 - 2r\cos\omega t)}{1 + r^2 - 2r\cos\omega t}$$
$$= \frac{1 - r^2}{1 + r^2 - 2r\cos\omega t}.$$

Diese Funktion wollen wir uns für verschiedene Werte von r etwas genauer ansehen.

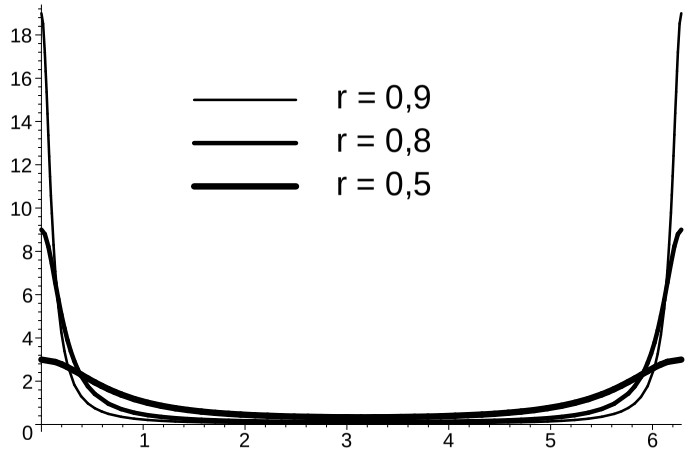


Abb. 13: $I_r(t) = \frac{1-r^2}{1+r^2-2r\cos t}$

Abbildung 13 zeigt, daß sich I_r im offenen Intervall $(0,\, T)$ für $r \to 1$ immer stärker an die t-Achse annähert, wohingegen die Funktionswerte an den Intervallenden immer stärker ansteigen. In der Tat können wir für $0 < t < T$ beim Grenzübergang $r \to 1$ einfach $r = 1$ setzen und erhalten

$$I_1(t) = \frac{1-1^2}{2-2\cos\omega t} = 0\,.$$

Für $t = 0$ oder T verschwindet dagegen auch der Nenner, und

$$\lim_{r\to 1} I_r(0) = \lim_{r\to 1} I_r(T) = \lim_{r\to 1} \frac{1-r^2}{1+r^2-2}$$
$$= \lim_{r\to 1} \frac{(1+r)(1-r)}{(1-r)^2} = \lim_{r\to 1} \frac{(1+r)}{(1-r)} = \infty\,.$$

Die Fläche unter der Kurve I_r im Intervall $[0,\, T]$ ist

$$\int_0^T \frac{1-r^2}{1+r^2-2r\cos\omega t}\,dt\,,$$

ein nicht sehr angenehm aussehendes Integral.

Die Mathematik stellt allerdings seit über hundert Jahren Algorithmen zur Verfügung, mit denen sich nicht nur entscheiden läßt, ob Funktionen wie I_r eine elementar ausdrückbare Stammfunktion haben, sondern auch berechnen, wie diese Stammfunktion dann aussieht. In den gängigen Computeralgebrasystemen sind diese Algorithmen zumindest teilweise implementiert, und wenn es auch zu weit führen würde, hier zu erklären, wie man eine Stammfunktion des Integranden

$$I_r(t) = \frac{1-r^2}{1+r^2-2r\cos\omega t}$$

findet, läßt sich doch das Ergebnis

$$F_r(t) = \int I_r(t)\,dt = -\frac{2}{\omega}\arctan\left(\frac{r+1}{r-1}\,\tan\frac{\omega t}{2}\right) + C$$

leicht verifizieren: Nach der Kettenregel ist zunächst

$$\frac{d}{dt}\arctan\left(a\tan\frac{\omega t}{2}\right) = \frac{1}{1+a^2\tan^2\frac{\omega t}{2}}\cdot\frac{a\omega}{2}\cdot\frac{1}{\cos^2\frac{\omega t}{2}}$$
$$= \frac{1}{2}\cdot\frac{a\omega}{\cos^2\frac{\omega t}{2}+a^2\sin^2\frac{\omega t}{2}} = \frac{1}{2}\cdot\frac{a\omega}{1+(a^2-1)\sin^2\frac{\omega t}{2}}\,.$$

Mit der Beziehung

$$\sin^2\frac{\omega t}{2} = \frac{1-\cos\omega t}{2}$$

wird das zu

$$\frac{1}{2}\,\frac{a\omega}{1+\frac{a^2-1}{2}-\frac{a^2-1}{2}\cos\omega t} = \frac{a\omega}{(a^2+1)-(a^2-1)\cos\omega t}\,.$$

Für $a = \dfrac{r+1}{r-1}$ ist

$$a^2+1 = 2\cdot\frac{r^2+1}{(r-1)^2} \quad\text{und}\quad a^2-1 = \frac{4r}{(r-1)^2}\,,$$

also wird der Ausdruck zu

$$\frac{\dfrac{r+1}{r-1}\cdot\omega}{2\cdot\dfrac{r^2+1}{(r-1)^2}-\dfrac{4r}{(r-1)^2}\cos\omega t}=\frac{(r^2-1)\omega}{2(r^2+1)-4r\cos\omega t}\,.$$

Um die Ableitung von F_r zu berechnen, müssen wir das noch mit $-2/\omega$ multiplizieren, was genau den Integranden ergibt. Somit ist F_r eine Stammfunktion von I_r.

Daß es mit dieser Stammfunktion ein Problem gibt, sieht man spätestens dann, wenn man naiv einsetzt und auf

$$\int_0^T I_r(t)\,dt = F_r(T) - F_r(0) = 0 - 0 = 0$$

kommt, denn aus geometrischen Gründen ist völlig klar, daß das Integral für $r < 1$ positiv sein muß.

Das Problem ist natürlich die Singularität des Tangens im Punkt $\frac{\pi}{2}$: Für $t = \frac{T}{2}$ ist F_r nicht definiert, da der Tangens dort gegen plus oder minus unendlich geht – je nachdem, von welcher Seite wir kommen. Damit ist F_r keine auf dem ganzen Integrationsintervall definierte Stammfunktion, und das Integral kann nicht einfach durch Einsetzen der oberen und der unteren Grenze berechnet werden.

Der linksseitige und der rechtsseitige Grenzwert von F_r existieren allerdings auch für $t = \frac{T}{2}$:

Für $0 < t < \frac{T}{2}$ ist $\tan\frac{\omega t}{2}$ positiv und geht gegen $+\infty$ für $t \to \frac{T}{2}$. Da der Faktor vor dem Tangens für alle $r \in (-1,\,1)$ negativ ist, folgt

$$\lim_{\substack{t\to T/2\\ t<T/2}} F_r(t) = -\frac{2}{\omega}\lim_{u\to-\infty}\arctan u = -\frac{2}{\omega}\frac{-\pi}{2} = \frac{T}{2}\,.$$

Für $\frac{T}{2} < t < T$ dagegen ist $\tan\frac{\omega t}{2}$ negativ und geht gegen $-\infty$ für $t \to \frac{T}{2}$. Somit ist

$$\lim_{\substack{t\to T/2\\ t>T/2}} F_r(t) = -\frac{2}{\omega}\lim_{u\to\infty}\arctan u = -\frac{2}{\omega}\cdot\frac{\pi}{2} = -\frac{T}{2}\,.$$

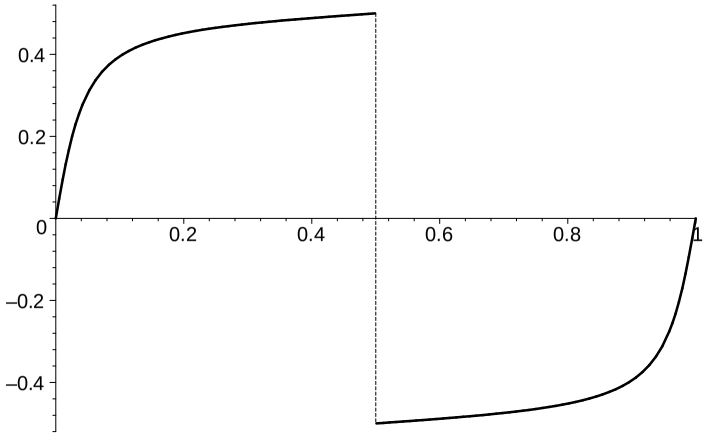


Abb. 14: $F_r(t)$ für $r = 0{,}8$ und $T = 2$

Abbildung 14 zeigt die Funktion F_r mit ihrer Sprungstelle in der Intervallmitte.

Zur Berechnung des Integrals von I_r über eine Periode spalten wir es auf in zwei Teilintegrale über Intervalle, in denen wir eine Stammfunktion haben, und erhalten

$$\begin{aligned}\int_0^T I_r(t)\,dt &= \int_0^{\frac{T}{2}} I_r(t)\,dt + \int_{\frac{T}{2}}^T I_r(t)\,dt\\ &= \lim_{\substack{t\to T/2\\ t<T/2}} F_r(t) - F_r(0) + F_r(T) - \lim_{\substack{t\to T/2\\ t>T/2}} F_r(t)\\ &= \frac{T}{2} - 0 + 0 + \frac{T}{2} = T\,.\end{aligned}$$

Somit ist

$$\frac{1}{T}\int_0^T I_r(t)\,dt = 1 \quad \text{für alle} \quad r \in (-1,\,1)$$

und

$$\frac{1}{T}\int_0^{\frac{T}{2}} I_r(t)\,dt = \frac{1}{T}\int_{\frac{T}{2}}^{T} I_r(t)\,dt = \frac{1}{2}\,.$$

Lassen wir nun r gegen eins gehen, ändert sich natürlich nichts an diesen Formeln, aber die beiden Integrale konzentrieren sich immer mehr auf die Umgebung von $t = 0$ *bzw.* $t = T$. Damit ist anschaulich ziemlich klar, daß für eine Funktion $f \in \mathrm{L}_T(\mathbb{R},\,\mathbb{C})$ gilt

$$\begin{aligned}\lim_{r\to 1} I_r \star f(t) &= \lim_{r\to 1}\frac{1}{T}\int_0^T f(t-\tau)I_r(\tau)\,d\tau\\ &= \lim_{r\to 1}\frac{1}{T}\left(\int_0^{\frac{T}{2}} f(t-\tau)I_r(\tau)\,d\tau + \int_{\frac{T}{2}}^{T} f(t-\tau)I_r(\tau)\,d\tau\right)\\ &= \frac{1}{2}\big(f(t^-)+f(t^+)\big)\,,\end{aligned}$$

denn für $r \to 1$ geht $I_r(t)$ gegen null für alle t im Integrationsbereich außer $t = 0$ und $t = T$. Auf die genauen Abschätzungen zum Beweis dieser Aussage sei verzichtet.

Stattdessen wenden wir die Aussage an auf die Funktion f aus dem Satz; dafür besagt sie, daß

$$\lim_{r\to 1} I_r \star f(t) = f(t) \quad \text{für alle} \quad t \in \mathbb{R}\,.$$

Für $|r| < 1$ ist

$$I_r(t) = \sum_{k=-\infty}^{\infty} r^{|k|} e^{k\cdot i\omega t}$$

absolut und gleichmäßig konvergent; deshalb kann die Summation mit Integration vertauscht werden und wir erhalten für $I_r \star f$ auch den

Ausdruck

$$\begin{aligned}I_r \star f(t) &= \frac{1}{T}\int_0^T \sum_{k=-\infty}^{\infty} r^{|k|} e^{k\cdot i\omega(t-\tau)} f(\tau)\,d\tau\\ &= \sum_{k=-\infty}^{\infty} r^{|k|} e^{k\cdot i\omega t}\cdot\frac{1}{T}\int_0^T f(\tau)e^{-k\cdot i\omega\tau}\,d\tau\\ &= \sum_{k=-\infty}^{\infty} r^{|k|} e^{k\cdot i\omega t}\cdot\widehat{f}(k)\,.\end{aligned}$$

Da nach Voraussetzung alle FOURIER-Koeffizienten von f verschwinden, ist also $I_r \star f(t) = 0$ für alle r vom Betrag kleiner eins. Damit ist aber auch

$$f(t) = \lim_{r\to 1} I_r \star f(t) = 0 \quad \text{für alle} \quad t \in \mathbb{R}\,,$$

wie behauptet. ∎

Das war ein langer Beweis für eine nicht sonderlich aufregende Aussage; der Satz hat jedoch weitreichende Konsequenzen:

Eindeutigkeitssatz: *a)* Haben zwei Funktionen $f, g \in \mathrm{L}^0_T(\mathbb{R},\,\mathbb{C})$ dieselben FOURIER-Koeffizienten, so sind sie gleich.
b) Haben zwei Funktionen $f, g \in \mathrm{L}_T(\mathbb{R},\,\mathbb{C})$ dieselben FOURIER-Koeffizienten, so unterscheiden sie sich höchstens an Unstetigkeitsstellen.

Beweis: a) ist klar, denn dann erfüllt $f - g$ die Voraussetzungen des gerade bewiesenen Satzes.

Um *b)* auf *a)* zurückzuführen, definieren zwei neue Funktion

$$\widetilde{f}(t) = \frac{1}{2}\big(f(t^+)+f(t^-)\big) \quad \text{und} \quad \widetilde{g}(t) = \frac{1}{2}\big(g(t^+)+g(t^-)\big)\,;$$

aus $\mathrm{L}^0_T(\mathbb{R},\,\mathbb{C})$. Diese unterscheiden sich von f *bzw.* g höchstens in deren Unstetigkeitsstellen; insbesondere haben f und $\widetilde{f}$ sowie g und $\widetilde{g}$ also dieselben FOURIER-Koeffizienten. Nach *a)* ist daher $\widetilde{f} = \widetilde{g}$, und die

ursprünglichen Funktionen f und g unterscheiden sich davon höchstens in ihren jeweiligen Unstetigkeitsstellen. ∎

Speziell können wir diesen Satz anwenden auf eine stetige stückweise differenzierbare Funktion f: Aus dem vorigen Abschnitt wissen wir, daß deren FOURIER-Reihe absolut und gleichmäßig gegen eine stetige Funktion S_f konvergiert, die dieselben FOURIER-Koeffizienten wie f hat. Aus dem Eindeutigkeitssatz folgt also

Satz: Die FOURIER-Reihe einer stetigen stückweise differenzierbaren Funktion $f \in \mathrm{L}_T(\mathbb{R}, \mathbb{C})$ konvergiert absolut und gleichmäßig gegen f. ∎

Kombinieren wir dies mit den speziellen Beispielen aus §3, so erhalten wir den folgenden

Hauptsatz: $f \in \mathrm{L}_T(\mathbb{R}, \mathbb{C})$ sei stückweise stetig differenzierbar.
a) Ist f in einem abgeschlossenen Intervall $[a, b]$ stetig, so konvergiert die FOURIER-Reihe dort gleichmäßig gegen f.
b) In jedem Punkt $t \in \mathbb{R}$ konvergiert die FOURIER-Reihe gegen

$$\frac{1}{2}\big(f(t^+) + f(t^-)\big)\,.$$

c) In jeder Sprungstelle von f tritt das GIBBS-Phänomen auf, d.h. die Teilsummen der FOURIER-Reihe überschwingen die Funktion um einen Betrag, der asymptotisch gleich der Sprunghöhe mal einem Faktor

$$\frac{1}{2}\left(\frac{2}{\pi}\,\mathrm{Si}(\pi) - 1\right) \approx 0.089489872$$

ist.

Beweis: Für stetiges f ist *a)* klar nach dem vorigen Satz und sowohl *b)* als auch *c)* sind auch klar, da es keine Sprungstellen gibt, so daß der Wert in *b)* immer gleich $f(t)$ ist.

Für unstetiges f seien $t_1, \ldots, t_r$ die Unstetigkeitsstellen im Intervall $[0, T)$; die Sprunghöhen dort seien

$$a_i \underset{\text{def}}{=} f(t_i^-) - f(t_i^+)\,.$$

Mit der aus §3*c)* bekannten Sägezahnschwingung

$$s(t) = \begin{cases} \frac{T}{4} - \frac{t}{2} & \text{für } 0 < t < T \\ 0 & \text{für } t = 0 \end{cases}\,,$$

periodisch fortgesetzt mit Periode T, ist dann

$$s_i(t) \underset{\text{def}}{=} \frac{2a_i}{T} s(t - t_i)$$

eine weitere stückweise differenzierbare Funktion, die ebenfalls Sprunghöhe a_i an der Stelle t_i hat. Also ist auch

$$\widetilde{f}(t) = f(t) - \sum_{j=1}^{r} s_j(t)$$

eine stückweise differenzierbare Funktion, die nun aber *keine* Sprungstellen mehr hat; $\widetilde{f}$ ist also stetig und erfüllt daher alle Behauptungen des Satzes.

Für die Funktion s haben wir die drei Behauptungen des Satzes in §3*e)* und *f)* explizit nachgerechnet; da sie unter Verschiebung und Reskalierung invariant sind, gelten sie auch für die Funktionen s_i. Damit gelten sie aber auch für $f(t) = \widetilde{f}(t) + \sum_{j=1}^{r} s_j(t)$. ∎

Da praktisch alle Funktionen, deren FOURIER-Reihen man in technischen Anwendungen betrachtet, stückweise differenzierbar sind, wollen wir es für die punktweise Konvergenz bei diesem Satz bewenden lassen; für sonstige stückweise stetige Funktionen wollen wir uns im nächsten Abschnitt mit einer schwächeren Konvergenzaussage begnügen.

f) Der Satz von Parseval

Im Zusammenhang mit der BESSELschen Ungleichung haben wir bereits die HERMITEschen Produkte in $\mathrm{L}_T(\mathbb{R}, \mathbb{C})$ und $\ell^2(\mathbb{Z}, \mathbb{C})$ miteinander verglichen; jetzt wollen wir sehen, daß sie dieselben Werte liefern. Auch hierbei arbeiten wir mit Faltungen; wesentliches Hilfsmittel ist der folgende, implizit schon im vorigen Abschnitt verwendete

Satz: Für $f, g \in \mathrm{L}_T(\mathbb{R}, \mathbb{C})$ konvergiert die FOURIER-Reihe von $f \star g$ überall gleichmäßig gegen $f \star g$; sind $c_k = \widehat{f}(k)$ und $d_k = \widehat{g}(k)$ die FOURIER-Koeffizienten von f und g, ist also für jedes $t \in \mathbb{R}$

$$f \star g(t) = \sum_{k=-\infty}^{\infty} c_k d_k e^{-k \cdot i\omega t} .$$

∎

Aus diesem Satz lassen sich sehr einfach Eigenschaften der Faltung ableiten, z.B. gilt

Lemma: *a)* Für $f, g \in \mathrm{L}_T(\mathbb{R}, \mathbb{C})$ ist $f \star g = g \star f$
b) Für $f, g, h \in \mathrm{L}_T(\mathbb{R}, \mathbb{C})$ ist $(f \star g) \star h = f \star (g \star h)$

Beweis: Betrachtet man die FOURIER-Reihen, werden die Behauptungen einfach zum Kommutativ- und Assoziativgesetz der Multiplikation komplexer Zahlen.

∎

(Man könnte das Lemma natürlich auch direkt durch Integration beweisen.)

Wir wollen den obigen Satz verwenden, um folgende Formel zu beweisen:

Satz von Parseval: Für $f, g \in \mathrm{L}_T(\mathbb{R}, \mathbb{C})$ ist

$$(f,\ g) = \left(\widehat{f},\ \widehat{g}\right),$$

d.h. für die FOURIER-Koeffizienten c_k von f und d_k von g ist

$$\sum_{k=-\infty}^{\infty} c_k \overline{d_k} = \frac{1}{T} \int_0^T f(t) \overline{g(t)}\, dt .$$

Zum *Beweis* brauchen wir eine Funktion, deren FOURIER-Koeffizienten die Zahlen $\overline{d_k}$ sind. Komplexe Konjugation der FOURIER-Reihe zu g führt zu

$$\overline{\sum_{k=-\infty}^{\infty} d_k e^{k \cdot i\omega t}} = \sum_{k=-\infty}^{\infty} \overline{d_k e^{k \cdot i\omega t}} = \sum_{k=-\infty}^{\infty} \overline{d_k} e^{-k \cdot i\omega t} .$$

Ersetzen wir hierin noch t durch $-t$, erhalten wir

$$\sum_{k=-\infty}^{\infty} \overline{d_k} e^{k \cdot i\omega t}$$

als FOURIER-Reihe von $\widetilde{g}(t) = \overline{g(-t)}$.

Somit ist $f \star \widetilde{g}$ eine Funktion mit FOURIER-Reihe

$$\sum_{k=-\infty}^{\infty} c_k \overline{d_k} e^{k \cdot i\omega t} .$$

Da die FOURIER-Reihe einer Faltung stets gleichmäßig konvergiert und dieselben FOURIER-Koeffizienten hat wie die Faltung selbst, folgt aus dem Eindeutigkeitssatz des vorigen Abschnitts, daß diese Reihe in jedem Punkt t gegen $f \star \widetilde{g}(t)$ konvergiert. Speziell für $t = 0$ ist daher einerseits

$$f \star \widetilde{g}(0) = \sum_{k=-\infty}^{\infty} c_k \overline{d_k}$$

und andererseits

$$f \star \widetilde{g}(0) = \widetilde{g} \star f(0) = \frac{1}{T} \int_0^T \widetilde{g}(-\tau) f(\tau)\, d\tau = \frac{1}{T} \int_0^T f(\tau) \overline{g(\tau)}\, d\tau .$$

Damit ist der Satz bewiesen.

∎

Korollar: Sind $c_k = \widehat{f}(k)$ die FOURIER-Koeffizienten von f, so ist

$$\sum_{k=-\infty}^{\infty} |c_k|^2 = \frac{1}{T} \int_0^T |f(t)|^2\, dt .$$

Gelegentlich wird auch dieses Korollar als Satz von PARSEVAL bezeichnet.

Der französische Mathematiker MARC-ANTOINE PARSEVAL DES CHÊNES (1755–1836) publizierte nur fünf mathematische Arbeiten; die 1799 veröffentlichte zweite davon enthält den hier betrachteten Satz.
Als überzeugter Royalist kam PARSEVAL während der französischen Revolution 1792 ins

Gefängnis; später mußte er aus Frankreich fliehen, weil ihn NAPOLEON wegen regimekritischer Gedichte verhaften lassen wollte.

Obiges Korollar liefert oft interessante spezielle Werte unendlicher Reihen: Für den Sägezahn mit Periode 2π etwa ist im Intervall $(0, 2\pi)$

$$s(t) = \frac{\pi - t}{2} = \sum_{k=1}^{\infty} \frac{\sin kt}{k} = \sum_{k=1}^{\infty} \frac{e^{ikt} - e^{-ikt}}{2ki} = \sum_{k=-\infty}^{\infty} c_k e^{ikt}$$

mit

$$c_k = \begin{cases} -\frac{i}{2k} & \text{für } k > 0 \\ 0 & \text{für } k = 0 \\ \frac{i}{2k} & \text{für } k < 0 \end{cases} .$$

Somit ist

$$\sum_{k=-\infty}^{\infty} |c_k|^2 = 2 \sum_{k=1}^{\infty} \left(\frac{1}{2k}\right)^2 = \frac{1}{2} \sum_{k=1}^{\infty} \frac{1}{k^2} .$$

Außerdem ist

$$\frac{1}{2\pi} \int_0^{2\pi} \left(\frac{\pi - t}{2}\right)^2 dt = \frac{1}{8\pi} \int_{-\pi}^{\pi} t^2 \, dt = \frac{1}{8\pi} \cdot 2 \cdot \frac{\pi^3}{3} = \frac{\pi^2}{12} .$$

Somit ist

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} .$$

Der Satz von PARSEVAL liefert auch schnell die im vorigen Abschnitt angekündigte schwächere Konvergenzaussage für beliebige stückweise stetige periodische Funktionen: Mit dem HERMITEschen Produkt in $L_T(\mathbb{R}, \mathbb{C})$ können wir in der üblichen Weise eine Norm

$$\|f\|_{2,T} = \sqrt{(f, f)} = \sqrt{\frac{1}{T} \int_0^T |f(t)|^2 \, dt}$$

definieren und sagen, eine Folge $f_1, f_2, \ldots$ konvergiere in der L^2-Norm gegen f, wenn

$$\lim_{n \to \infty} \|f_n - f\|_{2,T} = 0$$

ist. Dann gilt:

Satz: Für jedes $f \in L_T(\mathbb{R}, \mathbb{C})$ konvergiert die FOURIER-Reihe in der L^2-Norm gegen f.

Beweis: Wie wir beim Beweis der BESSELschen Ungleichung gesehen haben, gilt mit $c_k = \widehat{f}(k)$ für die Teilsumme

$$S_N = \sum_{k=-N}^{N} c_k e^{k \cdot i\omega t} ,$$

daß

$$\|f - S_N\|_{2,T}^2 = (f - S_N, \, f - S_N) = (f, \, f) - \sum_{k=-N}^{N} |c_k|^2 .$$

Nach dem Satz von PARSEVAL geht die rechts stehende Differenz für $N \to \infty$ gegen null, also auch $\|f - S_N\|_{2,T}$. ∎

g) Hilbert-Räume

In §2 hatten wir das Problem der harmonischen Analyse verglichen mit dem Problem, einen Vektor $\vec{v} \in \mathbb{R}^n$ bezüglich einer Orthonormalbasis darzustellen; die einfachste Lösung besteht bekanntlich darin, die Skalarprodukte von $\vec{v}$ mit den Basisvektoren $\vec{b}_i$ zu bilden; dann ist

$$\vec{v} = \sum_{k=1}^{n} (\vec{v} \cdot \vec{b}_k) \, \vec{b}_k .$$

Genauso haben wir nun bei der Berechnung einer komplexen FOURIER-Reihe einer Funktion $f \in L_T(\mathbb{R}, \mathbb{C})$ die Skalarprodukte $c_k = (f, e^{k \cdot i\omega t})$ gebildet für alle ganzen Zahlen $k \in \mathbb{Z}$ und gesehen, daß zumindest für stückweise differenzierbare Funktionen aus $L_T^0(\mathbb{R}, \mathbb{R})$ *bzw.* $L_T^0(\mathbb{R}, \mathbb{C})$

$$f(t) = \sum_{k=-\infty}^{\infty} c_k \, e^{k \cdot i\omega t}$$

ist, eine sehr ähnliche Situation also.

Es gibt allerdings einen entscheidenden Unterschied: Im Gegensatz zur endlichen Summe oben haben wir hier eine Summe, in der im allgemeinen unendlich viele der Summanden von Null verschieden sind. Eine solche unendliche Summe kann nur sinnvoll definiert werden, wenn wir einen Konvergenzbegriff haben, wie dies etwa in $\mathbb{R}^n$ oder $\mathbb{C}^n$ der Fall ist; über Körpern wie $\mathbb{Q}$ oder auch über endlichen Körpern hätten unendliche Summen überhaupt keine sinnvolle Interpretation.

In der linearen Algebra verlangt man daher aus gutem Grund von einer Basis, daß sich alle Vektoren als *endliche* Linearkombinationen von Basisvektoren darstellen lassen, auch wenn die Basis selbst unendlich sein kann.

Aus diesem Grund bilden die Funktionen $e^{k\cdot i\omega t}$ mit $k \in \mathbb{Z}$ keine Basis von $\mathrm{L}_T(\mathbb{R},\, \mathbb{C})$, und sie sind auch keine Basis des Untervektorraums aus allen stückweise differenzierbaren Funktionen mit der Mittelwerteigenschaft: Sie bilden nur eine Basis des sehr viel kleineren Untervektorraums der (komplexen) trigonometrischen Polynome. Die in §3*b)* betrachtete Rechteckschwingung f mit Periode T gehört bereits nicht mehr zu diesem Untervektorraum und ist in der Tat linear unabhängig von den Funktionen $e^{k\cdot i\omega t}$: Andernfalls gäbe es nämlich eine nichttriviale endliche Linearkombination

$$\lambda_0 f(t) + \lambda_1 e^{k_1\cdot i\omega t} + \cdots + \lambda_r e^{k_r\cdot i\omega t} = 0\,,$$

in der λ_0 nicht verschwinden darf, da sonst die $e^{k\cdot i\omega t}$ linear abhängig wären.

Also wäre

$$f(t) = -\frac{\lambda_1}{\lambda_0} e^{k_1\cdot i\omega t} + \cdots - \frac{\lambda_r}{\lambda_0} e^{k_r\cdot i\omega t}\,,$$

$f(t)$ als endliche Linearkombination stetiger und differenzierbarer Funktionen selbst stetig und differenzierbar, was natürlich absurd ist.

Wenn wir die reinen Schwingungen zu einer Basis von $\mathrm{L}_T(\mathbb{R},\, \mathbb{C})$ ergänzen wollen, müssen wir also beispielsweise den Rechteckimpuls als neues Basiselement dazunehmen und, wie man sich leicht überlegt, auch noch den Sägezahn. Tatsächlich muß man noch eine ganze Reihe

anderer Funktionen mit hinzunehmen, und bis heute hat es niemand geschafft, eine Basis von $\mathrm{L}_T(\mathbb{R},\, \mathbb{C})$ hinzuschreiben.

Der letzte Satz sollte zumindest diejenigen nicht verwundern, die sich aus dem letzten Semester noch an die dortige Diskussion im Zusammenhang mit der Existenz von Basen erinnern: Wir hatten dies nur für endlichdimensionale Vektorräume bewiesen, da es im unendlichdimensionalen Fall logische Schwierigkeiten gibt: Man muß das absolut nichtkonstruktive Auswahlaxiom benutzen, das zu allem Überfluß auch noch von den restlichen Axiomen der Mengenlehre unabhängig ist, so daß zumindest im Prinzip auch eine Mathematik ohne Auswahlaxiom möglich ist. Dort ist dann nicht beweisbar, daß jeder Vektorraum eine Basis hat – was freilich nicht bedeutet, daß nun beweisbar wäre, daß irgendein Vektorraum *keine* Basis hätte.

Von daher ist zumindest für alle praktischen Zwecke das System der Funktionen $e^{k\cdot i\omega t}$ die beste Annäherung an eine Basis, die man bekommen kann.

Sie ist in vielfacher anderer Hinsicht eine Verallgemeinerung der Standardbasis des $\mathbb{R}^n$ oder $\mathbb{C}^n$: Beispielsweise besagt der Satz von PARSEVAL, daß wir den Abstand (bezüglich der L_2-Norm) zwischen zwei Funktionen, die beide bezüglich dieser Basis dargestellt sind, genauso ausrechnen können, wie wir das vom EUKLIDischen oder HERMITEschen Abstand her gewohnt sind – nur daß wir es jetzt eben mit einer unendlichen Summe von Quadraten zu tun haben.

Da eine solche Situation in Funktionenräumen alles andere als selten vorkommt und FOURIER-Reihen beileibe nicht die einzigen unendlichen Summen ihrer Art mit praktischer Bedeutung sind, hat die Mathematik hierfür einen Begriffsapparat nebst zugehörigen Techniken entwickelt, die wir hier in dieser Vorlesung zwar nicht unbedingt brauchen, die aber in manchen Gebieten beispielsweise der Signalverarbeitung oder der optischen Übertragungstechnik eine wichtige Rolle spielen, die Theorie der HILBERTräume:

Definition: Ein EUKLIDischer oder HERMITEscher Vektorraum V heißt HILBERT-Raum, wenn jede CAUCHY-Folge aus V gegen ein Element von V konvergiert.

DAVID HILBERT (1862–1943) wurde in Königsberg geboren, wo er auch zur Schule und zur Universität ging. Er promovierte dort 1885 mit einem Thema aus der Invariantentheorie, habilitierte sich 1886 und bekam 1893 einen Lehrstuhl. 1895 wechselte er an das damalige Zentrum der deutschen wie auch internationalen Mathematik, die Universität Göttingen, wo er bis zu seiner Emeritierung im Jahre 1930 lehrte. Seine Arbeiten umfassen ein riesiges Spektrum aus unter anderem Invariantentheorie, Zahlentheorie, Geometrie, Funktionalanalysis, Logik und Grundlagen der Mathematik sowie auch zur Relativitätstheorie. Er gilt als einer der Väter der modernen Algebra.

Offensichtlich ist jeder EUKLIDische oder HERMITEsche Vektorraum endlicher Dimension ein HILBERT-Raum, denn wir können den Raum vermöge irgendeiner Orthonormalbasis mit $\mathbb{R}^n$ *bzw.* $\mathbb{C}^n$ identifizieren und das CAUCHYsche Konvergenzkriterium komponentenweise anwenden. Von den unendlichdimensionalen Vektorräumen $\mathrm{L}_T(\mathbb{R},\,\mathbb{C})$ und $\mathrm{L}_T(\mathbb{R},\,\mathbb{R})$, die uns in Augenblick interessieren, ist leider keiner ein HILBERT-Raum, denn wie wir schon gesehen haben ist weder $\mathrm{L}_T(\mathbb{R},\,\mathbb{R})$ ein EUKLIDischer noch $\mathrm{L}_T(\mathbb{R},\,\mathbb{C})$ ein HERMITEscher Vektorraum, da es stückweise stetige Funktionen gibt, deren Norm $\sqrt{(f,\,f)}$ verschwindet, ohne daß f die Nullfunktion wäre.

Wenn wir uns auf die Unterräume $\mathrm{L}^0_T(\mathbb{R},\,\mathbb{C})$ und $\mathrm{L}^0_T(\mathbb{R},\,\mathbb{R})$ beschränken, haben wir zwar Skalarprodukte, aber die Vollständigkeit ist alles andere als klar.

Dieses Problem wollen wir (wie in den Anwendungen üblich) weitgehend ignorieren; wir bezeichnen einfach für *jeden* $\mathbb{R}$- oder $\mathbb{C}$-Vektorraum V mit einem Produkt, das bis auf die Forderung

$$(f,\,f) = 0 \Longrightarrow f \equiv 0$$

EUKLIDsch *bzw.* HERMITEsch ist, eine Teilmenge $H \subseteq V$ als HILBERT-*Raumbasis*, wenn es für jedes Element $f \in V$ eine Folge $(h_i)_{i\in\mathbb{N}}$ von Elementen aus H und eine Folge $(\lambda_i)_{i\in\mathbb{N}}$ gibt, so daß

$$\delta_N \underset{\text{def}}{=} \sum_{i=1}^{N} \lambda_i h_i - f$$

für $N \to \infty$ die Eigenschaft hat, daß $\lim\limits_{N\to\infty} (\delta_N,\,\delta_N) = 0$ ist.

In diesem Sinne ist das System der Funktionen $e^{k\cdot i\omega t}$ nach der Diskussion im vorigen Abschnitt eine HILBERT-Raumbasis von $\mathrm{L}_T(\mathbb{R},\,\mathbb{C})$ (auch wenn wir das nur für den Untervektorraum der stückweise differenzierbaren Funktionen beweisen haben), und damit ist auch das System der Funktionen $1, \sin k\omega t, \cos \ell\omega t$ eine HILBERT-Raumbasis von $\mathrm{L}_T(\mathbb{R},\,\mathbb{R})$.

Solche HILBERT-Raumbasen sind in dieser Allgemeinheit leider noch nicht sonderlich nützlich für praktische Anwendungen: Will man einen Vektor $\vec{v} \in \mathbb{R}^n$ als Linearkombination von *irgendeiner* Basis $\mathcal{B} \subset \mathbb{R}^n$ darstellen, muß man ein lineares Gleichungssystem mit n Gleichungen in n Unbekannten lösen. Für endliches n ist das für große n zwar nicht mehr sehr angenehm, aber doch grundsätzlich möglich und per Computer auch noch für n in der Größenordnung von hundert Tausend durchaus praktikabel.

Wird die Dimension allerdings unendlich, so läßt sich ein System aus unendlich vielen Gleichungen in unendlich vielen Variablen nur in sehr speziellen Fällen wirklich lösen; einer davon ist der, den wir bei der Berechnung des FOURIER-Reihe ausgenutzt haben: Im Falle einer abzählbaren orthogonalen HILBERT-Raumbasis $\left\{\vec{b}_i \;\middle|\; i \in \mathbb{N}\right\}$ ist

$$\vec{v} = \sum_{i=1}^{\infty} \frac{\vec{v}\cdot\vec{b}_i}{\vec{b}_i\cdot\vec{b}_i}$$

leicht berechenbar; noch einfacher wird es, wenn wir von einer orthonormalen HILBERT-Raumbasis ausgehen, da dann alle Nenner eins sind.

Solche orthogonale *bzw.* orthonormale HILBERT-Raumbasen bezeichnet man als *vollständige Orthogonalsysteme* bzw. *vollständige Orthonormalsysteme.*

h) Die Poisson-Formel

Der Konvergenzbeweis für FOURIER-Reihen war sehr abstrakt; in diesem Anschnitt wollen wir sehen, daß die Methoden, die wir dabei kennengelernt haben, auch nützlich sein können bei der Lösung eines praktischen Problems:

Bei bildgebenden Verfahren der Medizintechnik, bei Werkstoffuntersuchungen, Wärmeleitungsproblemen und vielen anderen Anwendungen hat man es oft mit folgendem *Randwertproblem* zu tun: Man kennt eine Funktion am Rand einer Fläche oder eines Volumens und möchte sie auch im (physikalisch oft unzugänglichen) Innern berechnen.

Sofern man keine einschränkenden Annahmen über die Funktion macht, ist dieses Problem natürlich weit von einer eindeutigen Lösbarkeit entfernt; in vielen interessanten Fällen ist es allerdings eindeutig lösbar.

Wir wollen hier nur ein ganz einfaches Beispiel betrachten: eine auf der Kreislinie bekannte Funktion, die ins Kreisinnere hinein fortgesetzt werden soll. Dabei wollen wir verlangen, daß die Funktion überall der *Kontinuitätsgleichung* $\Delta u = 0$ genügt, wie das beispielsweise für elektrische Potentiale in Abwesenheit von Ladungen der Fall ist. Solche Funktionen haben wir bereits in §1 betrachtet und dort als *harmonische* Funktionen bezeichnet; wir wissen, daß sie gerade die Realteile holomorpher Funktionen sind.

Aus §1*d)* wissen wir, daß holomorphe Funktionen nach der CAUCHYschen Integralformel im Innern eines einfach zusammenhängenden Gebiets durch ihre Werte auf der Randkurve bestimmt sind; hier wollen mit Hilfe von FOURIER-Reihen eine entsprechende Formel für harmonische Funktionen finden.

In Polarkoordinaten ausgedrückt ist der LAPLACE-Operator nach **[HM1]**, Kap. 2, §2*g1)* gleich

$$\Delta u = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\varphi\varphi} ;$$

für $u(r,\varphi) = r^k \cos k\varphi$ und $u(r,\varphi) = r^k \sin k\varphi$ rechnet man leicht nach, daß $\Delta u = 0$ ist. Damit ist auch

$$\Delta(r^n e^{\pm k\cdot i\varphi}) = 0$$

und allgemeiner

$$\Delta u(r,\varphi) = 0 \quad \text{für} \quad u(r,\varphi) = \sum_{k=-\infty}^{\infty} c_k r^{|k|} e^{k\cdot i\omega t} ,$$

falls die Summe konvergiert. Für ein Randwertproblem ist die Funktion brauchbar, wenn zusätzlich auch noch

$$u_0(\varphi) \underset{\text{def}}{=} \lim_{r\to 1} u(r,\varphi)$$

für alle Winkel φ existiert.

Ein spezielles Beispiel einer konvergenten Summe ist die Funktion I_r aus Abschnitt *e);* wenn wir als Argument anstelle von ωt die Winkelvariable φ einsetzen, ist für $|r| < 1$

$$I_r(\varphi) = \sum_{k=-\infty}^{\infty} r^{|k|} e^{k\cdot i\varphi} = \frac{1-r^2}{1+r^2-2r\cos\varphi} .$$

Da $I_r \star u_0$ für stetige Funktionen im Limes $r \to 1$ gerade u_0 ist, ist für eine vorgegebene Funktion u_0

$$u(r,\varphi) = I_r \star u_0(\varphi) = \frac{1}{2\pi}\int_0^{2\pi} \frac{(1-r^2)u_0(\psi)}{1+r^2-2r\cos(\varphi-\psi)}\,d\psi$$

eine Fortsetzung ins Innere mit $\Delta u(r,\varphi) = 0$. Das ist die POISSONsche Integralformel zur Lösung dieses einfachen Randwertproblems.

SIMÉON DENIS POISSON (1781–1840) studierte zunächst Medizin, dann ab 1798 Mathematik an der *Ecole Polytechnique* bei LAPLACE und LAGRANGE. 1802 bekam er eine Stelle als Astronom am *Bureau des Longitudes*, 1809 wurde er Professor für reine Mathematik an der neugegründeten *Faculté des Sciences*. Er arbeitete hauptsächlich über bestimmte Integrale und FOURIER-Theorie, schrieb aber auch ein wichtiges Lehrbuch der Wahrscheinlichkeitstheorie (in dem die POISSON-Verteilung erstmalig auftauchte) und Arbeiten über Mechanik, Astronomie, Elektrizität und Magnetismus.

§5: Fourier- und Laplace-Transformationen

In den vorigen Paragraphen haben wir periodische Funktionen mittels ihrer FOURIER-Reihen als Überlagerungen reiner Schwingungen dargestellt. Diese Zerlegung einer Funktion in Sinus- und Kosinusschwingungen verschiedener Frequenzen ist nicht nur für periodische Funktionen

nützlich; angesichts der Tatsache, daß das Verhalten vieler elektronischer Bauteile von der Frequenz abhängt, würde man gerne *jede* Funktion entsprechend zerlegen. Es ist allerdings klar, daß FOURIER-Reihen, wie wir sie bislang kennen, dazu nicht geeignet sind: Da dort alle beteiligten Frequenzen Vielfache eine festen Grundfrequenz sind, muß auch die Summe zumindest die der Grundfrequenz entsprechende Periode haben.

Daher brauchen wir für nichtperiodische Funktionen im Allgemeinen ein kontinuierliches Frequenzspektrum; dieses liefert uns für hinreichend gutartige Funktionen die FOURIER-Transformation. Die LAPLACE-Transformation ist eine Variante davon, die zwar inhaltlich etwas schwerer zu interpretieren ist als die FOURIER-Transformation, die dafür aber für größere Funktionenklassen existiert. Außerdem gibt es zur LAPLACE-Transformation sehr viel ausführlichere Tabellen als zur FOURIER-Transformation

a) Fourier-Reihen und Fourier-Integrale

Zur Konstruktion der FOURIER-Transformation gehen wir aus von FOURIER-Reihen:

Für eine beliebige reelle Funktion $f\colon \mathbb{R} \to \mathbb{R}$ wählen wir dazu zunächst eine (große) Periode T und betrachten die Funktion $f_T\colon \mathbb{R} \to \mathbb{R}$, die auf dem Intervall $(-T/2,\, T/2]$ mit f übereinstimmt und dann periodisch mit Periode T auf ganz $\mathbb{R}$ fortgesetzt wird. Mit $\omega = 2\pi/T$ ist die FOURIER-Reihe von f_T gleich

$$\sum_{k=-\infty}^{\infty} c_k e^{ki\omega t} \quad \text{mit} \quad c_k = \widehat{f_T}(k) = \frac{1}{T} \int_{-T/2}^{T/2} f(t) e^{-ki\omega t}\, dt\,.$$

Um f selbst darzustellen, müssen wir T gegen unendlich gehen lassen; um das Verhalten von c_k bei Veränderung von T kontrollieren zu können, definieren wir dazu eine Funktion $C(\nu)$ als

$$C(\nu) \underset{\text{def}}{=} \int_{-T/2}^{T/2} f(t) e^{-i\nu t}\, dt\,.$$

Mit dieser Definition ist

$$c_k = \frac{1}{T} C(k\omega) = \frac{\omega}{2\pi} C(k\omega)\,,$$

und die FOURIER-Reihe von f_T läßt sich schreiben als

$$\frac{1}{2\pi} \sum_{k=-\infty}^{\infty} C(k\omega) e^{i(k\omega)t}\, \omega\,.$$

Wäre dies eine endliche Summe, etwa

$$\frac{1}{2\pi} \sum_{k=-N}^{N} C(k\omega) e^{i(k\omega)t}\, \omega\,,$$

so könnten wir sie auffassen als RIEMANN-Summe für

$$\int_{-N\omega}^{(N+1)\omega} \frac{1}{2\pi} C(\nu) e^{i\nu t}\, d\nu$$

bei einer äquidistanten Unterteilung mit Intervallbreite ω. Falls also ω gegen Null geht (und damit $T = 2\pi/\omega$ gegen unendlich) und gleichzeitig N gegen unendlich, konvergiert die FOURIER-Reihe gegen das Integral

$$\int_{-\infty}^{\infty} \frac{1}{2\pi} C(\nu) e^{i\nu t}\, d\nu = \frac{1}{2\pi} \int_{-\infty}^{\infty} C(\nu) e^{i\nu t}\, d\nu\,,$$

sofern dieses existiert. Im Idealfall sollte also gelten

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} C(\nu) e^{i\nu t}\, d\nu \quad \text{mit} \quad C(\nu) \underset{\text{def}}{=} \int_{-\infty}^{\infty} f(t) e^{-i\nu t}\, dt\,.$$

Um dies genauer zu untersuchen, geben wir diesen Konstruktionen Namen:

Definition: Für $f\colon \mathbb{R} \to \mathbb{C}$ bezeichnen wir die Funktion

$$\widehat{f}\colon \begin{cases} \mathbb{R} \to \mathbb{C} \\ \omega \mapsto \displaystyle\int_{-\infty}^{\infty} f(t)\, e^{-i\omega t}\, dt \end{cases},$$

so sie existiert, als FOURIER-Transformierte von f.

Damit sollte dann also gelten

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\omega) e^{i\omega t}\, d\omega\,,$$

und diese Konstruktion, die f aus $\widehat{f}$ rekonstruiert, heißt *inverse* FOURIER-*Transformation:*

Definition: Für $g\colon \mathbb{R} \to \mathbb{C}$ bezeichnen wir die Funktion

$$\check{g}\colon \begin{cases} \mathbb{R} \to \mathbb{C} \\ t \mapsto \dfrac{1}{2\pi} \displaystyle\int_{-\infty}^{\infty} g(\omega)\, e^{i\omega t}\, d\omega \end{cases}$$

als inverse FOURIER-Transformierte von g.

Offensichtlich ist

$$\check{f}(\omega) = \frac{1}{2\pi} \widehat{f}(-\omega) \quad \text{und} \quad \widehat{g}(t) = 2\pi \check{g}(-t)\,.$$

Je nach Buch oder Vorlesung werden die Vorfaktoren gelegentlich auch anders gewählt, beispielsweise stand in der HM II bis 1998 der Faktor $1/2\pi$ vor der FOURIER-Transformation selbst statt vor ihrer inversen. Die jetzt gewählte Definition paßt besser zu der aus der hiesigen Elektrotechnik; dort wird die FOURIER-Transformation als

$$F(j\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t}\, dt$$

definiert, wobei j,wie in der Elektrotechnik üblich, für die in der Mathematik und Physik mit i bezeichnete imaginäre Einheit $\sqrt{-1}$ steht. Demnach ist also $F(j\omega) = \widehat{f}(\omega)$.

Einige Autoren bevorzugen es auch, aus Symmetriegründen bei beiden Transformationen einen Vorfaktor $1/\sqrt{2\pi}$ zu verwenden, so daß je nach Buch durchaus sehr verschiedene Dinge gemeint sein können, wenn

von „der“ FOURIER-Transformation und ihrer Umkehrung die Rede ist. In allen Fällen sind die Faktoren aber so aufeinander abgestimmt, daß für hinreichend gutartige Funktionen die Beziehungen

$$\check{\widehat{f}}(t) = f(t) \quad \text{und} \quad \widehat{\check{f}}(t) = f(t)$$

gelten.

b) Die Laplace-Transformation

Die Existenz der FOURIER-Transformierten, d.h. die Konvergenz des uneigentlichen Integrals aus der Definition, sowie auch die obigen Formeln für $\check{\widehat{f}}$ und $\widehat{\check{f}}$ sind leider alles andere als selbstverständlich: Für $f(t) = 1$ oder auch $f(t) = e^{i\omega t}$ oder $f(t) = t^n$ und in vielen weiteren Fällen kann das Integral für $\widehat{f}(\omega)$ schon aus ganz trivialen Gründen nicht existieren.

Offensichtlich hat das FOURIER-Integral

$$\int_{-\infty}^{\infty} f(t) e^{-i\omega t}\, dt$$

Konvergenzprobleme sowohl an der oberen als auch an der unteren Grenze. Bei vielen Anwendung interessieren Funktionen vor allem für positive Werte von t (die „Zukunft“), während negative Werte (die „Vergangenheit“) vernachlässigt werden können. Um daher eine gegebene Funktion f so abzuändern, daß das FOURIER-Integral an der unteren Grenze keine Konvergenzprobleme mehr hat, setzen wir sie für $t < 0$ einfach auf null.

Für positive t dürfen wir nicht so radikal vorgehen; schließlich soll das Ergebnis noch etwas mit der Funktion f zu tun haben. Deshalb dämpfen wir hier die Funktion nur durch eine Exponentialfunktion. Insgesamt betrachten wir also anstelle von $f(t)$ die Funktion

$$g_r(t) = \begin{cases} 0 & \text{für } t < 0 \\ f(t) e^{-rt} & \text{für } t > 0 \end{cases}.$$

Den Funktionswert an der Stelle 0 legen wir so fest, daß die Funktion dort rechtsseitig stetig ist, d.h.

$$g(0) = f(0^+) = \lim_{t \to 0^+} f(t)\,.$$

Die FOURIER-Transformierte dieser Funktion g_r bezeichnen wir, wenn sie existiert, als LAPLACE-Transformierte von f an der Stelle $s = r + i\omega$, in Zeichen

$$\mathcal{L}\{f(t)\}(s) \underset{\text{def}}{=} \int_0^\infty f(t)e^{-st}\,dt\,.$$

Für gängige Funktionen f ist $\mathcal{L}\{f(t)\}(s)$ ist den meisten Formelsammlungen zu finden; es gibt auch umfangreiche Tabellenwerke, die ausschließlich der LAPLACE-Transformation gewidmet sind. Im allgemeinen wird sie nur für hinreichend große Werte von $r = \mathfrak{Re}\, s$ existieren.

Die inverse LAPLACE-Transformation läßt sich aus der inversen FOURIER-Transformation ableiten: Wegen $\mathcal{L}\{f(t)\}(r+i\omega) = \widehat{g}_r(t)$ sollte $g_r(t)$ die inverse FOURIER-Transformierte von $\mathcal{L}\{f(t)\}(r + i\omega)$ sein; für $t > 0$ sollte daher

$$f(t) = e^{rt}g_r(t) = \frac{e^{rt}}{2\pi}\int_{-\infty}^{\infty} \mathcal{L}\{f(t)\}(r + i\omega)e^{i\omega t}\,d\omega$$

sein. Für $t < 0$ können wir natürlich keine entsprechende Formel erwarten, da die Funktionswerte von f auf der negativen Achse bei der Berechnung der LAPLACE-Transformation ignoriert werden.

c) Erste Beispiele

Als erstes Beispiel betrachten wir die Funktion $f(t) = \sin\omega t$. Ihr FOURIER-Integral

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} \sin\omega t\, e^{-i\omega t}\,dt$$

ist ein unendliches Integral über eine periodische Funktion, existiert also nicht. Auch ein CAUCHYscher Hauptwert kann nicht existieren, denn wenn auch $\sin\omega t$ eine ungerade Funktion ist, ist der Integrand als ganzes weder gerade noch ungerade, da $e^{-i\omega t} = \cos\omega t - i\sin\omega t$ gerade Realteil, aber ungeraden Imaginärteil hat.

Das LAPLACE-Integral

$$\mathcal{L}\{\sin\omega t\}(s) = \int_0^\infty \sin\omega t\, e^{-st}\,dt$$

existiert für rein imaginäres $s = i\omega$ aus dem gleichen Grund nicht, und für ein s mit negativem Realteil kann es natürlich schon gar nicht existieren. Ist aber $\mathfrak{Re}\, s > 0$, so liefert die Regel für partielle Integration

$$\mathcal{L}\{\sin\omega t\}(s) = \int_0^\infty \sin\omega t\, e^{-st}\,dt = \sin\omega t\frac{e^{-st}}{-s}\,\Big|_0^\infty - \int_0^\infty \omega\cos\omega t\frac{e^{-st}}{-s}\,dt$$

etwas Verwertbares: e^{-st} geht dann nämlich für $t \to \infty$ gegen null, und an der unteren Grenze $t = 0$ verschwindet $\sin\omega t$, so daß der erste Summand rechts insgesamt verschwindet. Der Integral ganz hinten ist bis auf den Faktor $-\omega/s$ die LAPLACE-Transformierte des Kosinus, d.h.

$$\mathcal{L}\{\sin\omega t\}(s) = \frac{\omega}{s}\mathcal{L}\{\cos\omega t\}(s)\,.$$

Auf das LAPLACE-Integral für den Kosinus wenden wir wieder die Regel der partiellen Integration an:

$$\mathcal{L}\{\cos\omega t\}(s) = \int_0^\infty \cos\omega t\, e^{-st}\,dt = \cos\omega t\frac{e^{-st}}{-s}\,\Big|_0^\infty + \int_0^\infty \omega\sin\omega t\frac{e^{-st}}{-s}\,dt\,.$$

Hier bekommen wir an der unteren Grenze des ersten Terms rechts den Wert eins für den Kosinus, und an der oberen Grenze geht natürlich wieder der Exponentialfaktor gegen null, so daß

$$\mathcal{L}\{\cos\omega t\}(s) = \frac{1}{s} + \frac{\omega}{s}\mathcal{L}\{\sin\omega t\}(s)$$

ist. Insgesamt ist

$$\mathcal{L}\{\sin\omega t\}(s) = \frac{\omega}{s}\mathcal{L}\{\cos\omega t\}(s) = \frac{\omega}{s^2} - \frac{\omega^2}{s^2}\mathcal{L}\{\sin\omega t\}(s)$$

oder

$$\left(1 + \frac{\omega^2}{s^2}\right)\mathcal{L}\{\sin\omega t\}(s) = \frac{\omega}{s^2}\,.$$

Multiplikation mit s^2 macht daraus

$$(s^2+\omega^2)\mathcal{L}\{\sin\omega t\}(s) = \omega \quad \text{und} \quad \mathcal{L}\{\sin\omega t\}(s) = \frac{\omega}{s^2+\omega^2}$$

für $\mathfrak{Re}\, s > 0$. Damit kennen wir auch

$$\mathcal{L}\{\cos\omega t\}(s) = \frac{s}{\omega}\mathcal{L}\{\sin\omega t\}(s) = \frac{s}{s^2+\omega^2}\,.$$

Auch bei den Potenzfunktionen $t \mapsto t^n$ für $n \in \mathbb{N}_0$ gibt es offensichtlich keine Chance, daß das FOURIER-Integral

$$\int_{-\infty}^{\infty} t^n e^{-i\omega t}\,dt = \int_{-\infty}^{\infty} t^n \cos\omega t\,dt - i\int_{-\infty}^{\infty} t^n \sin\omega t\,dt$$

existiert; selbst der CAUCHYsche Hauptwert existiert nicht, denn wenn der Integrand des Realteils ungerade ist, ist der des Imaginärteils gerade und umgekehrt.

Die LAPLACE-Transformation verlangt die Berechnung von

$$\mathcal{L}\{t^n\}(s) = \int_0^{\infty} t^n e^{-st}\,dt\,.$$

Dieses Integral könnten wir induktiv durch partielle Integration berechnen; falls wir allerdings für n auch nicht ganzzahlige Werte einsetzen wollen, läßt es sich nicht mehr durch die uns bislang bekannten Funktionen ausdrücken.

Es läßt sich jedoch leicht zurückführen auf die sogenannte EULERsche *Gammafunktion*, die für positive Werte von x (oder allgemeiner für komplexe x mit positivem Realteil) definiert ist als

$$\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}\,dt\,.$$

Dies ist ein uneigentliches Integral: Es ist natürlich immer uneigentlich an der oberen Grenze, und für $0 < x < 1$ zusätzlich auch noch an der unteren. (Für $\mathfrak{Re}\, x \le 0$ divergiert das Integral.)

Die untere Grenze ist harmlos, denn für $0 < x < 1$ ist

$$e^{-t}t^{x-1} \le t^{x-1}\,,$$

und für $0 < x < 1$ hat die Stammfunktion $\frac{t^x}{x}$ der rechten Seite einfach den Wert Null für $t = 0$.

Auch die obere Grenze ist unproblematisch: Da die Exponentialfunktion stärker wächst als jede Potenz, ist für hinreichend große Werte von t

$$e^t \ge t^{r+x-1} \iff e^{-t}t^{x-1} \le \frac{K}{t^r}\,.$$

Dies gilt insbesondere für $r = 2$, und da $\int_1^{\infty} \frac{dr}{r^2} = 1$ konvergiert, gilt dasselbe für $\Gamma(x)$ an seiner oberen Grenze.

Die wichtigste Eigenschaft der Γ-Funktion folgt durch partielle Integration: Für $x > 0$ ist

$$\Gamma(x) = \int_0^{\infty} e^{-t}t^{x-1}\,dt = e^{-t}\frac{t^x}{x}\Big|_0^{\infty} + \frac{1}{x}\int_0^{\infty} e^{-t}t^x\,dt = \frac{\Gamma(x+1)}{x}$$

oder

$$\Gamma(x+1) = x\Gamma(x)\,.$$

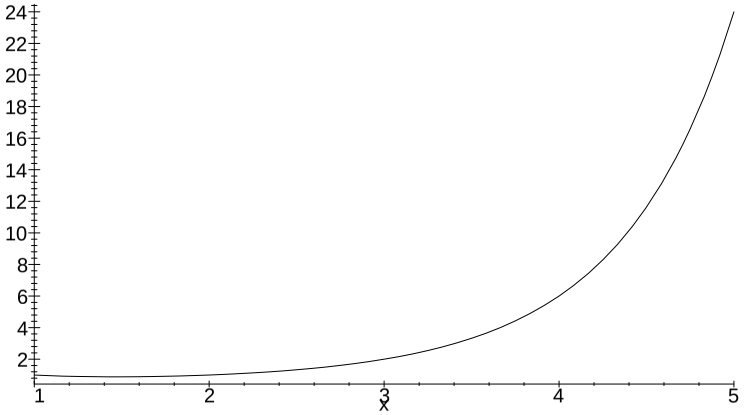


Abb. 15: Die Γ-Funktion

Aus dem elementar berechenbaren Wert

$$\Gamma(1) = \int_0^\infty e^{-t}\,dt = -e^{-t}\,\Big|_0^\infty = 1$$

ergibt sich somit, daß für alle natürliche Zahlen n gilt $\Gamma(n) = (n-1)!$; die Γ-Funktion ist also nicht als eine Art stetig gemachter Fakultätsfunktion.

GAUSS definierte auf andere Weise eine stetige Funktion $\Pi(x)$, für die $\Pi(n) = n!$ ist, aber wie sich bald herausstellte, ist $\Pi(x) = \Gamma(x+1)$, so daß nur eine der beiden Funktionen wirklich gebraucht wird. Nach einigen Modewechseln im letzten Jahrhundert entscheidet man sich heute meist für $\Gamma(x)$: Diese Funktionswerte sind in Tafelwerken tabelliert, und numerische Verfahren für ihre Berechnung stehen in den einschlägigen Unterprogrammbibliotheken und Computeralgebrasystemen zur Verfügung. Tatsächlich läßt sich $\Gamma(x)$ mit Hilfe komplexer Integrale sogar fortsetzen zu einer auf ganz $\mathbb{C}$ meromorphen Funktion mit der Eigenschaft $\Gamma(z+1) = z\Gamma(z)$; die einzigen Pole sind die durch diese Eigenschaft erzwungenen bei Null und den negativen ganzen Zahlen.

Mit dieser Funktion können wir die LAPLACE-Transformierte von $f(t) = t^n$ sogar für jede *reelle* Zahl $n > -1$ leicht ausdrücken: Mit der Substitution $u = st$ wird

$$\mathcal{L}\{t^n\}(s) = \int_0^\infty t^n e^{-st}\,dt = \int_0^\infty \frac{u^n}{s^n} e^{-u}\,\frac{du}{s} = \frac{1}{s^{n+1}} \int_0^\infty u^n e^{-u}\,du = \frac{\Gamma(n+1)}{s^{n+1}}\,.$$

Für nichtnegative ganze Zahlen $n \in \mathbb{N}_0$ vereinfacht sich dies zu

$$\mathcal{L}\{t^n\}(s) = \frac{n!}{s^{n+1}}\,.$$

Insbesondere ist die LAPLACE-Transformierte einer konstanten Funktion $f(t) = a$ gleich a/s. Genau dieselbe Transformierte hat natürlich auch die Sprungfunktion

$$f(t) = \begin{cases} a & \text{für } t \geq 0 \\ 0 & \text{für } t < 0 \end{cases},$$

denn auf Werte an negativen Stellen kommt es bei der LAPLACE-Transformation nicht an.

Als nächstes wollen wir negative Potenzen t^{-n} betrachten. Deren LAPLACE-Transformation ist gegeben durch

$$\mathcal{L}\{t^{-n}\}(s) = \int_0^\infty \frac{e^{-st}}{t^n}\,dt\,,$$

und dieses sowohl an der oberen als auch an der unteren Grenze uneigentliche Integral existiert leider nicht: Für reelles $s > 0$ etwa ist für jedes $a > 0$ die Funktion e^{-as}/t^n überall im Intervall $(0,\, a]$ kleiner oder gleich dem Integranden; da ihre Stammfunktion $e^{-as}/(1-n)t^{n-1}$ für $n > 1$ und $e^{-as} \ln t$ für $n = 1$ für $t \to 0$ gegen unendlich geht, existiert das Integral

$$\int_0^a \frac{e^{-as}}{t^n}\,dt$$

für kein $a > 0$, und damit existiert erst recht das obige LAPLACE-Integral nicht.

Als Kuriosität am Rande sei erwähnt, daß dafür aber (nur) der CAUCHYsche Hauptwert des entsprechenden FOURIER-Integrals

$$\int_{-\infty}^\infty \frac{e^{-i\omega t}}{t^n}\,dt$$

existiert: Für $n = 1$ haben wir in §3*f)* im Zusammenhang mit dem Integralsinus nachgerechnet, daß

$$\int_{-\infty}^\infty \frac{e^{i\omega t}}{t}\,dt = \pi i \quad \text{für alle } \omega > 0\,.$$

Ersetzen wir hier ω durch $-\omega$, wird der Integrand komplex konjugiert, also auch der CAUCHYsche Hauptwert des Integrals, und damit ist

$$\int_{-\infty}^\infty \frac{e^{i\omega t}}{t}\,dt = -\pi i \quad \text{für alle } \omega < 0\,.$$

Für $\omega = 0$ haben wir das Integral über $1/t$, das natürlich ebenfalls nicht existiert, das aber den CAUCHYschen Hauptwert null hat, da der Integrand ungerade ist. Der CAUCHYsche Hauptwert des FOURIER-Integrals ist also

$$\int_{-\infty}^\infty \frac{e^{-i\omega t}}{t}\,dt = \begin{cases} -\pi i & \text{für } \omega > 0 \\ 0 & \text{für } \omega = 0 \\ \pi i & \text{für } \omega < 0 \end{cases}.$$

Für $n > 1$ kann man genau wie in §3*f)* argumentieren und erhält (mit den dortigen Bezeichnungen) die Beziehung

$$\int_{-\infty}^{\infty} \frac{e^{i\omega t}}{t^n}\,dt = \lim_{\delta\to 0} \int_{\alpha_\delta} \frac{e^{i\omega z}}{z^n}\,dz$$

für den CAUCHYschen Hauptwert. Reihenentwicklung der Exponentialfunktion führt auf

$$\int_{\alpha_\delta} \frac{e^{i\omega z}}{z^n} = \int_{\alpha_\delta} \sum_{k=0}^{\infty} \frac{(i\omega)^k z^{k-n}}{k!}\,dz = \sum_{k=0}^{\infty} \frac{(i\omega)^k}{k!} \int_{\alpha_\delta} z^{k-n}\,dz\,.$$

Für $k = n-1$ ist das rechtsstehende Integral

$$\int_{\alpha_\delta} z^{-1}\,dz = \operatorname{Ln}(\delta) - \operatorname{Ln}(-\delta) = \operatorname{Ln}(-1) = -\pi i$$

unabhängig von δ; im Falle $k \neq n-1$ verschwindet

$$\int_{\alpha_\delta} z^{k-n}\,dz = \frac{\delta^{k-n+1} - (-\delta)^{k-n+1}}{k}$$

für $k \equiv n-1 \bmod 2$ und ist gleich $2\delta^k/k$ sonst. Da die geometrische Reihe $2\sum_{k=1}^{\infty} \delta^k$ eine konvergente Majorante der Summe aller solcher Terme ist und für $\delta \to 0$ gegen null geht, folgt

$$\int_{-\infty}^{\infty} \frac{e^{i\omega t}}{t^n}\,dt = \lim_{\delta\to 0} \int_{\alpha_\delta} \frac{e^{i\omega z}}{z^n}\,dz = \frac{(i\omega)^{n-1}}{(n-1)!} \cdot \pi i \quad \text{für } \omega > 0\,.$$

Für $\omega < 0$ wird wieder der Integrand komplex konjugiert, also auch das Ergebnis; im Faktor $(i\omega)^{n-1}$ sorgt ω selbst für die komplexe Konjugation, so daß rechts nur πi konjugiert werden muß, d.h. der CAUCHYsche Hauptwert des FOURIER-Integrals ist

$$\int_{-\infty}^{\infty} \frac{e^{-i\omega t}}{t^n}\,dt = \begin{cases} -\frac{(i\omega)^{n-1}}{(n-1)!} \cdot \pi & \text{für } \omega > 0 \\ \frac{(i\omega)^{n-1}}{(n-1)!} \cdot \pi & \text{für } \omega > 0 \end{cases}.$$

Für $\omega = 0$ haben wir das Integral über $1/t^n$, daß für ungerades n den CAUCHYschen Hauptwert null hat und für gerades n gegen unendlich divergiert.

Als letztes Beispiel wollen wir eine der wichtigsten Funktionen der Elektrotechnik betrachten, den *Rechteckimpuls*. Wir beschränken uns hier auf einen zum Nullpunkt symmetrischen Impuls der Form

$$f(t) = \begin{cases} a & \text{für } -b \le t \le b \\ 0 & \text{sonst} \end{cases}.$$

Hier ist

$$\begin{aligned} \widehat{f}(\omega) &= \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,dt = \int_{-b}^{b} ae^{-i\omega t}\,dt = \frac{a}{-i\omega} e^{-i\omega t}\Big|_{-b}^{b} \\ &= \frac{a}{\omega} \frac{e^{i\omega b} - e^{-i\omega b}}{i} = \frac{2a\sin\omega b}{\omega}\,. \end{aligned}$$

Mit der in der Elektrotechnik sehr wichtigen Funktion

$$\operatorname{sinc} t = \frac{\sin t}{t}$$

läßt sich dies auch schreiben als

$$\widehat{f}(\omega) = 2ab \operatorname{sinc} \omega b\,.$$

(Anstelle von $\operatorname{sinc} t$ schreiben manche Autoren auch $\operatorname{si} t$, man darf die Funktion aber nicht mit Ihrer Stammfunktion, dem Integralsinus $\operatorname{Si} t$, verwechseln.)

Die LAPLACE-Transformierte dieses Rechteckimpulses ist

$$\mathcal{L}\{f(t)\}(s) = \int_0^{\infty} f(t)e^{-st}\,dt = \int_0^{b} ae^{-st}\,dt = \frac{a}{s}\left(1 - e^{-sb}\right),$$

und das ist gleichzeitig auch die LAPLACE-Transformierte der Rechteckimpulse

$$g(t) = \begin{cases} a & \text{für } 0 \le t \le b \\ 0 & \text{sonst} \end{cases} \quad \text{und} \quad h(t) = \begin{cases} a & \text{für } t \le b \\ 0 & \text{sonst} \end{cases}.$$

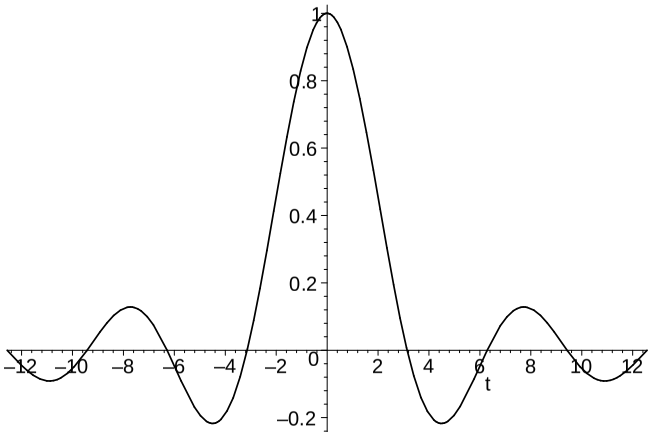


Abb. 16: Die Funktion $\operatorname{sinc} t = \frac{\sin t}{t}$

Dagegen existiert die FOURIER-Transformierte $\widehat{h}(\omega)$ nicht einmal, und auch

$$\widehat{g}(\omega) = \int_{-\infty}^{\infty} g(t)e^{-i\omega t}\,dt = \int_{0}^{b} ae^{-i\omega t}\,dt = \frac{a}{-i\omega}e^{-i\omega t}\Big|_0^b$$
$$= \frac{ia}{\omega}(e^{-i\omega b} - 1)$$

ist deutlich verschieden von $\widehat{f}(\omega)$.

d) Erste Rechenregeln

Die gerade betrachteten Beispiele sind zwar mit die wichtigsten Funktionen, die in den gängigen Anwendungen auftauchen; allerdings findet man sie dort selten als *reine* Sinus- oder Kosinusschwingungen oder als *reine* Potenzen; häufiger sind Linearkombinationen dieser Funktionen, eventuell noch verbunden mit Phasenverschiebungen und anderen Transformationen des Arguments. In diesem Abschnitt sollen die wichtigsten Rechenregeln zusammengestellt werden, die in solchen Situationen gebraucht werden.

Die fundamentalste und einfachste Regel ist

Lemma: Sowohl die FOURIER- als auch die LAPLACE-Transformation sind lineare Operationen, d.h. für zwei Funktionen f, g und zwei komplexe Zahlen λ, μ gilt

$$\widehat{\lambda f + \mu g}(\omega) = \lambda\widehat{f}(\omega) + \mu\widehat{g}(\omega)$$

und

$$\mathcal{L}\{\lambda f + \mu g\}(s) = \lambda\mathcal{L}\{f\}(s) + \mu\mathcal{L}\{g\}(s)\,,$$

sofern jeweils beide Seiten existieren. ∎

Damit ist etwa

$$\mathcal{L}\{a\cos\omega t + b\sin\omega t\}(s) = \frac{as + b\omega}{s^2 + \omega^2}\,,$$

und entsprechend läßt sich auch die LAPLACE-Transformation jedes trigonometrischen Polynoms berechnen.

Für ein Polynom $f(t) = a_n t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0$ ist entsprechend

$$\mathcal{L}\{f(t)\}(s) = \frac{n!a_n}{s^{n+1}} + \frac{(n-1)!a_{n-1}}{s^n} + \ldots + \frac{a_1}{s^2} + \frac{a_0}{s}$$
$$= \frac{n!a_n + (n-1)!a_{n-1}s + \cdots + a_1 s^{n-1} + a_0 s^n}{s^{n+1}}\,.$$

Ebenfalls sehr einfach läßt sich der Effekt von Verschiebungen sowohl im Zeit- als auch im Frequenzbereich ausdrücken: Die FOURIER-Transformierte von $g(t) = f(t+c)$ berechnet sich mittels der Substitution $u = t + c$ als

$$\widehat{g}(\omega) = \int_{-\infty}^{\infty} f(t+c)e^{-i\omega t}\,dt = \int_{-\infty}^{\infty} f(u)e^{-i\omega(u-c)}\,du$$
$$= e^{i\omega c}\int_{-\infty}^{\infty} f(u)e^{-i\omega u}\,du = e^{i\omega c}\widehat{f}(\omega)$$

und die LAPLACE-Transformierte als

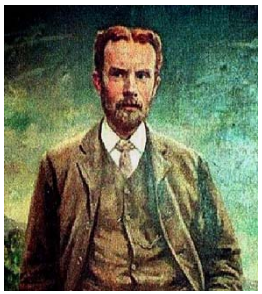
$$\mathcal{L}\{g(t)\}(s) = \int_0^\infty f(t+c)e^{-st}\,dt = \int_c^\infty f(u)e^{-s(u-c)}\,du$$
$$= e^{sc}\int_c^\infty f(u)e^{-su}\,du\,,$$

eine Formel, die wegen der geänderten unteren Integrationsgrenze nicht sonderlich nützlich aussieht. Allerdings war hier auch von vornherein nicht viel zu erwarten, denn schon zur Definition der LAPLACE-Transformation müssen wir schließlich einen festen Zeitpunkt als Nullpunkt der Zeitskala auszeichnen, und niemand sollte sich wundern, daß Transformationen, die diese Auszeichnung nicht respektieren, bei der LAPLACE-Transformation zu Problemen führen. Man kann freilich das hintere Integral zumindest für $c > 0$ mit Gewalt als LAPLACE-Transformation auffassen, indem man den Integranden mit einer geeigneten Sprungfunktion multipliziert:

Die HEAVISIDE-Funktion ist definiert durch

$$H(t) = \begin{cases} 1 & \text{für } t > 0 \\ 0 & \text{für } t < 0 \end{cases}.$$

Ein Funktionswert an der Stelle Null wird üblicherweise nicht definiert, die die Funktion praktisch immer nur in Integralen auftritt, wo es auf diesen einen Wert ohnehin nicht ankommt.



OLIVER HEAVISIDE (1850–1925) wurde in den Londoner Slums geboren und hatte nur wenig formale Ausbildung; an akademischer Wissenschaft hatte er zeitlebens kein Interesse. Unter dem Einfluß seines Onkels CHARLES WHEATSTONE (nach dem die gleichnamige Brücke benannt ist) wandte er sich zur Telegraphie und arbeitete als Telegraphist zunächst in Dänemark, dann in Newcastle. In diesem Zusammenhang beschäftigte er sich mit Elektrizitätslehre; unter anderem brachte er die MAXWELLschen Originalgleichungen (20 Gleichungen in 20 Unkenannten) auf die heute übliche Form. Weiter untersuchte er die Bedingungen für die störungsfreie Übertragung eines Signals, sagte die für weltweiten Kurzwellenfunk wesentliche Reflektionseigenschaft der Ionosphäre (HEAVISIDE-Schicht) voraus und entwickelte einen Operatorkalkül zur Übersetzung von Differentialgleichungen in algebraische Gleichungen.

Mit Hilfe dieser Funktion können wir obige Gleichung umschreiben als

$$\mathcal{L}\{f(t+c)\}(s) = e^{sc}\mathcal{L}\{f(t)H(t-c)\}(s)\,.$$

Im Frequenzbereich gibt es keine solchen Probleme; hier ist einfach

$$\widehat{f}(\omega+c) = \int_{-\infty}^\infty f(t)e^{-i(\omega+c)t}\,dt = \widehat{e^{-ict}f}(\omega)$$

und

$$\mathcal{L}\{f(t)\}(s+c) = \int_0^\infty f(t)e^{-(s+c)t}\,dt = \mathcal{L}\{e^{-ct}f(t)\}(s)\,.$$

Diese Formeln sind auch rückwärts gelesen sehr nützlich: Wollen wir etwa die LAPLACE-Transformierte einer gedämpften Schwingung

$$f(t) = e^{-\lambda t}(a\cos\omega t + b\sin\omega t)$$

berechnen, so ist

$$\mathcal{L}\{f(t)\}(s) = \mathcal{L}\{a\cos\omega t + b\sin\omega t\}(s+\lambda) = \frac{a(s+\lambda)+b\omega}{(s+\lambda)^2+\omega^2}\,.$$

Auch die LAPLACE-Transformierte der Exponentialfunktion selbst läßt sich so ausrechnen: Für $\mathfrak{Re}\,s > \lambda$ ist

$$\mathcal{L}\{e^{\lambda t}\}(s) = \mathcal{L}\{1\}(s-\lambda) = \frac{1}{s-\lambda}\,,$$

denn auf $1 = t^0$ können wir die Formel für t-Potenzen anwenden.

Die FOURIER-Transformierte einer Exponentialfunktion existiert natürlich nicht, da das FOURIER-Integral immer an mindestens einer der beiden Grenzen divergiert.

Schließlich können wir noch ohne großen Aufwand den Effekt einer Streckung im Zeit- oder Frequenzbereich ausrechnen: Für $g(t) = f(ct)$

zeigt die Substitution $u = ct$, daß gilt

$$\widehat{g}(\omega) = \int_{-\infty}^{\infty} f(ct)e^{-i\omega t}\,dt = \int_{-\infty}^{\infty} f(u)e^{-i\frac{\omega}{c}u}\,\frac{du}{c} = \frac{1}{c}\widehat{f}\left(\frac{\omega}{c}\right)$$

und

$$\mathcal{L}\{g(t)\}(s) = \int_{0}^{\infty} f(ct)e^{-st}\,dt = \int_{0}^{\infty} f(u)e^{-\frac{s}{c}u}\,\frac{du}{c} = \frac{1}{c}\mathcal{L}\{f(t)\}\left(\frac{s}{c}\right).$$

Das Verhalten von FOURIER- und LAPLACE-Transformation im Zusammenhang mit Ableitungen, Produkten und Faltungen werden wir in den nächsten beiden Paragraphen ausführlich untersuchen.

§6: Ableitungen und Differentialgleichungen

Als erstes wollen wir die Transformationen von Ableitungen und die Ableitungen von Transformationen betrachten; wie sich zeigen wird, führt dies zu einer Methode, mit der sich Anfangswertprobleme für lineare Differentialgleichungen oft recht bequem lösen lassen. Es lohnt sich daher, zuerst noch etwas in Vorbereitungen zu investieren, um das notwendige Werkzeug bereitszustellen.

a) Ableitungen unter dem Integralzeichen

Lemma: *a)* Die Funktion $h\colon [a,b]\times[c,d]\to\mathbb{R}$ sei stetig. Dann ist auch

$$H\colon \begin{cases} [a,b]\to\mathbb{R} \\ \omega\mapsto \displaystyle\int_c^d h(\omega,t)\,dt \end{cases}$$

stetig.
b) Ist h zusätzlich r-mal stetig partiell nach der ersten Variablen ω differenzierbar, so ist

$$\frac{d^rH}{d\omega^r}(\omega) = \int_c^d \frac{\partial^r h}{\partial\omega^r}(\omega,t)\,dt\,.$$

Beweis: a) Da $[a,b]$ und $[c,d]$ abgeschlossene Intervalle sind, ist h nicht nur stetig, sondern sogar gleichmäßig stetig. Es gibt also zu jedem $\varepsilon>0$ ein $\delta>0$, so daß für jedes $t\in[c,d]$ gilt

$$|h(\omega_1,t)-h(\omega_2,t)|<\varepsilon \quad\text{falls}\quad |\omega_1-\omega_2|<\delta\,.$$

Für solche ω_1 und ω_2 ist dann

$$|H(\omega_1)-H(\omega_2)| \le \int_c^d |h(\omega_1,t)-h(\omega_2,t)|\,dt < \varepsilon(d-c)\,.$$

Da $d-c$ eine Konstante ist, läßt sich dies durch Wahl von $\varepsilon=\eta/(d-c)$ unter jedes vorgegebene $\eta>0$ drücken.

b) Für $\Delta\neq 0$ ist

$$\frac{H(\omega+\Delta)-H(\omega)}{\Delta} = \int_c^d \frac{h(\omega+\Delta,t)-h(\omega,t)}{\Delta}\,dt\,,$$

und der rechtsstehende Integrand ist nach dem Mittelwertsatz der Differentialrechnung gleich $\frac{\partial h}{\partial\omega}(\xi,t)$ für ein ξ zwischen ω und $\omega+\Delta$. Für $\Delta\to 0$ geht dies gegen $\frac{\partial h}{\partial\omega}(\omega,t)$, und da die partielle Ableitung als stetig vorausgesetzt wurde, gilt wegen *a)*, daß

$$H'(\omega) = \lim_{\xi\to\omega}\int_c^d \frac{\partial h}{\partial\omega}(\xi,t)\,dt = \int_c^d \frac{\partial h}{\partial\omega}(\omega,t)\,dt$$

ist, wie behauptet. Für $r>1$ folgt die Behauptung induktiv. ∎

Als erste Anwendung folgt ein Satz über die Vertauschung der Integrationsreihenfolge, der für die Inversion der FOURIER-Transformation fundamental sein wird:

Satz von Fubini: Für eine stetige Funktion $h\colon[a,b]\times[c,d]\to\mathbb{R}$ ist

$$\int_a^b\left(\int_c^d h(\omega,t)\,dt\right)d\omega = \int_c^d\left(\int_a^b h(\omega,t)\,d\omega\right)dt\,.$$

Beweis: Das folgt entweder aus der zweidimensionalen Integrationstheorie in **[HM1]**, Kap. II, §6*b)*, da beide Seiten das Integral

$$\iint\limits_{[a,b]\times[c,d]} h(\omega,t)\,d\omega\,dt$$

berechnen, wobei das Rechteck $[a,b]\times[c,d]$ als Normalbereich einmal vom Typ I und einmal vom Typ II aufgefaßt wird. Es folgt aber auch leicht aus dem gerade bewiesenen Lemma:

Für $a \le \omega \le b$ sei

$$H(\omega) = \int\limits_c^d \left(\int\limits_a^\omega h(\xi,t)\,d\xi \right) dt\,.$$

Nach der zweiten Aussage des gerade bewiesenen Lemmas ist dann

$$H'(\omega) = \int\limits_c^d h(\omega,t)\,dt\,,$$

also ist die linke Seite der zu beweisenden Gleichung

$$\int\limits_a^b \left(\int\limits_c^d h(\omega,t)\,dt \right) d\omega = \int\limits_a^b H'(\omega)\,d\omega = H(b) - H(a) = H(b)\,,$$

und das ist nach Konstruktion gleich der rechten Seite. ∎

Der italienische Mathematiker GUIDO FUBINI (1879–1943) arbeitete zunächst auf dem Gebiet der Differentialgeometrie, interessierte sich dann aber immer mehr für analytische Themen wie Differentialgleichungen und Funktionen mehrerer komplexer Veränderlicher. 1901 wurde er Professor in Catania auf Sizilien, später in Genua und ab 1908 in Turin, wo er blieb, bis er 1939 trotz seiner angegriffenen Gesundheit wegen des italienischen Faschismus nach USA emigrierte und ans Institute for Advanced Study in Princeton wechselte. Der hier zitierte Satz ist zwar sein bekanntestes, aber ganz sicher nicht sein bedeutendstes Ergebnis.

Wir interessieren uns im Augenblick nicht für Integrale über endliche Intervalle, sondern für Integrale über die gesamte reelle Gerade; bevor wir den gerade bewiesenen Satz auf FOURIER-Integrale anwenden können, müssen wir also noch den Grenzübergang $a, c \to -\infty$ und $b, d \to +\infty$ durchführen. Nach dem WEIERSTRASSschen Konvergenzkriterium gibt es hier keine Probleme, *falls* die betroffenen Integrale absolut konvergent sind. Die für uns interessante Version des Satzes von FUBINI ist also

Satz: Die stetige Funktion $h\colon \mathbb{R}^2 \to \mathbb{R}$ sei so, daß die uneigentlichen Integrale

$$\int\limits_{-\infty}^{\infty} \left(\int\limits_{-\infty}^{\infty} |h(\omega,t)|\,dt \right) d\omega \quad \text{und} \quad \int\limits_{-\infty}^{\infty} \left(\int\limits_{-\infty}^{\infty} |h(\omega,t)|\,d\omega \right) dt$$

beide konvergieren. Dann ist

$$\int\limits_{-\infty}^{\infty} \left(\int\limits_{-\infty}^{\infty} h(\omega,t)\,dt \right) d\omega = \int\limits_{-\infty}^{\infty} \left(\int\limits_{-\infty}^{\infty} h(\omega,t)\,d\omega \right) dt\,.$$

∎

b) Transformationen und Ableitungen

Die Aussagen des vorigen Abschnitts führen geradewegs auf Eigenschaften der FOURIER-Transformation; als erstes erhalten wir

Lemma: Ist die Funktion $f\colon \mathbb{R} \to \mathbb{C}$ mindestens r mal stetig differenzierbar und existieren die FOURIER-Transformationen von $f, t^r f$ und $f^{(r)}$, so ist

$$\frac{d^r}{d\omega^r}\widehat{f}(\omega) = (-i)^r \widehat{t^r f}(\omega) \quad \text{oder} \quad \widehat{t^r f}(\omega) = i^r \frac{d^r}{d\omega^r}\widehat{f}(\omega)$$

und

$$\omega^r \widehat{f}(\omega) = (-i)^r \widehat{f^{(r)}}(\omega) \quad \text{oder} \quad \widehat{f^{(r)}}(\omega) = (i\omega)^r \widehat{f}(\omega)\,.$$

Beweis: Nach dem Lemma im vorigen Abschnitt ist

$$\frac{d^r \widehat{f}(\omega)}{d\omega^r} = \int_{-\infty}^{\infty} \frac{d^r}{d\omega^r}\left(f(t)e^{-i\omega t}\right) dt$$
$$= \int_{-\infty}^{\infty} (-it)^r f(t)e^{-i\omega t}\, dt = (-i)^r \widehat{t^r f}(\omega)\,,$$

womit die erste Aussage bewiesen wäre.

Für die zweite begnügen wir uns der Einfachheit halber mit dem Fall $r = 1$, aus dem die allgemeine Aussage per Induktion folgt. Für $r = 1$ ist

$$\omega \widehat{f}(\omega) = \int_{-\infty}^{\infty} f(t) \cdot \omega e^{-i\omega t}\, dt\,,$$

und dieses Integral läßt sich durch partielle Integration weiter umformen. Dazu nehmen wir $f(t)$ als den ersten Faktor und $\omega e^{-i\omega t}$ als den zweiten; letzterer hat die Stammfunktion

$$\frac{\omega e^{-i\omega t}}{-i\omega} = ie^{-i\omega t}\,,$$

und wir erhalten

$$\int_{-\infty}^{\infty} f(t) \cdot \omega e^{-i\omega t}\, dt = \quad f(t) \cdot ie^{-i\omega t}\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \dot{f}(t) \cdot ie^{-i\omega t}\, dt\,.$$

Da die FOURIER-Transformierte von f existiert, geht $f(t)$ für $t \to \pm\infty$ gegen null; der erste Summand verschwindet also, und übrig bleibt

$$\omega \widehat{f}(\omega) = -i \int_{-\infty}^{\infty} \dot{f}(t)e^{-i\omega t}\, dt = -i \cdot \widehat{\dot{f}}(\omega)\,.$$

Damit ist das Lemma bewiesen. ∎

Für die LAPLACE-Transformation gelten ähnliche Regeln: Falls die Funktion f mindestens r mal stetig differenzierbar ist und die LAPLACE-Transformierten ihrer Ableitungen existieren, ist nach der Regel über

partielle Integration

$$\mathcal{L}\{\dot{f}(t)\}(s) = \int_0^{\infty} \dot{f}(t)e^{-st}\, dt = f(t)e^{-st}\Big|_0^{\infty} + s\int_0^{\infty} f(t)e^{-st}\, dt$$
$$= -f(0) + s\mathcal{L}\{f(t)\}(s)$$

und damit induktiv

$$\mathcal{L}\{f^{(r)}(t)\}(s) = s^r\mathcal{L}\{f(t)\}(s) - s^{r-1}f(0) - s^{r-2}\dot{f}(0) - \cdots - f^{(r-1)}(0)\,.$$

Dies ist etwas komplizierter als bei der FOURIER-Transformation, wo wir keine Funktionswerte an der Stelle null berücksichtigen mußten, aber für die Anwendung auf Differentialgleichungen ist das meist ein *Vorteil:*

In der Praxis hat man es fast immer mit sogenannten *Anfangswertproblemen* zu tun, d.h. man kennt den Zustand eines Systems (beschreiben durch eine Funktion $f(t)$ der Zeit) zu einem gewissen Zeitpunkt t_0, den wir der Einfachheit halber als null annehmen wollen, und man kennt Naturgesetze für die weitere Entwicklung des Systems. Letztere haben meist die Form von Differentialgleichungen; ein Anfangswertproblem besteht darin, daß man anhand der Differentialgleichung und der bekannten Funktionswerte zum Zeitpunkt t_0 die weitere Entwicklung des Systems berechnen will, d.h. die Funktion f. Wir werden uns im nächsten Kapitel ausführlich mit Differentialgleichungen beschäftigen; hier beschränken wir uns auf den in der Elektrotechnik sehr häufigen Fall einer linearen Differentialgleichung mit konstanten Koeffizienten, d.h. wir suchen eine Funktion $y(t)$, für die gilt

$$y^{(n)}(t) + a_{n-1}(t)y^{(n-1)}(t) + \cdots + a_0(t)y(t) = b(t)$$

und

$$y(0) = y_0, \quad \dot{y}(0) = y_1, \quad \ldots, \quad y^{n-1}(0) = y_{n-1}\,.$$

Anwendung der LAPLACE-Transformation macht daraus die algebrai-

sche Gleichung

$$\begin{aligned} & s^n\mathcal{L}\{y(t)\}(s) - s^{n-1}y_0 - s^{n-2}y_1 - \cdots - y_{(n-1)} \\ + a_{n-1}&s^{n-1}\mathcal{L}\{y(t)\}(s) - s^{n-2}y_0 - s^{n-3}y_1 - \cdots - y_{(n-2)} \\ & \vdots \\ & + a_1 s\mathcal{L}\{y(t)\}(s) - y_0 \\ & + a_0\mathcal{L}\{y(t)\}(s) \qquad = \qquad \mathcal{L}\{b(t)\}(s) \end{aligned}$$

für $\mathcal{L}\{y(t)\}(s)$. Mit dem *charakteristischen Polynom*

$$f(s) = s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0$$

der Differentialgleichung und den Abkürzungen

$$c_{n-1} = y_0, \quad c_{n-2} = y_0 + y_1, \quad \ldots \quad c_0 = y_0 + y_1 + \cdots + y_{n-1}$$

läßt sich dies auch kürzer schreiben als

$$f(s)\cdot\mathcal{L}\{y(t)\}(s) - c_{n-1}s^{n-1} - c_{n-2}s^{n-2} - \cdots - c_1 s - c_0 = \mathcal{L}\{b(t)\}(s)$$

oder

$$\mathcal{L}\{y(t)\}(s) = \frac{\mathcal{L}\{b(t)\}(s) + c_{n-1}s^{n-1} + c_{n-2}s^{n-2} + \cdots + c_1 s + c_0}{f(s)}\,.$$

Falls sowohl die gesuchte Lösungsfunktion als auch die rechte Seite $b(t)$ LAPLACE-transformierbar sind, können wir also die LAPLACE-Transformierte $\mathcal{L}\{y(t)\}(s)$ der Lösungsfunktion leicht berechnen. Rücktransformation (meist anhand einer Tabelle) führt dann auf die Lösung $y(t)$ des Anfangswertproblems – sofern wir wissen, daß eine Funktion $y(t)$ zumindest für $t > 0$ durch $\mathcal{L}\{y(t)\}(s)$ eindeutig bestimmt ist.

Letzteres ist (abgesehen von kleineren Besonderheiten, die für praktische Anwendungen kaum eine Rolle spielen) in der Tat der Fall, allerdings wird der entsprechende Beweis – genau wie im Fall der FOURIER-Reihen – recht viel Arbeit kosten. Um zu sehen, daß sich diese Arbeit auch lohnt, wollen wir daher zunächst einige Anwendungen betrachten.

Bevor wir damit beginnen, sollten wir uns noch die Ableitung der LAPLACE-Transformierten einer Funktion anschauen: Wegen der Vertauschbarkeit der Integration über t und der Ableitung nach s erhalten

wir ohne jede Mühe die Formel

$$\begin{aligned} \frac{d}{ds}\mathcal{L}\{f(t)\}(s) &= \frac{d}{ds}\int_0^\infty f(t)e^{-st}\,dt = \int_0^\infty \frac{d}{ds}\left(f(t)e^{-st}\right)dt \\ &= -\int_0^\infty tf(t)e^{-st}\,dt = -\mathcal{L}\{tf(t)\}(s)\,. \end{aligned}$$

Durch Induktion folgt für beliebiges $r \in \mathbb{N}$ die Formel

$$\frac{d^r}{ds^r}\mathcal{L}\{f(t)\}(s) = (-1)^r\mathcal{L}\{t^r f(t)\}(s)\,,$$

die bis auf den Vorfaktor $(-1)^r$ genauso aussieht wie die entsprechende Formel

$$\frac{d^r}{d\omega^r}\widehat{f}(\omega) = i^r\widehat{t^r f}(\omega)$$

für die FOURIER-Transformation.

c) Ungedämpfte Schwingungen

Nach diesen Vorbereitungen können wir daran gehen, einige physikalisch interessante Anfangswertprobleme zu lösen.

Als erstes Beispiel betrachten wir die extrem einfache Gleichung für eine Masse an einer Feder, die sich reibungsfrei in x-Richtung bewegen kann:

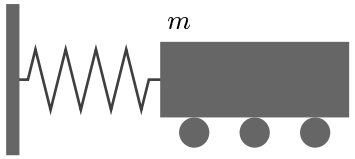


Abb. 17: Eine schwingende Masse

Nach dem HOOKEschen Gesetz wirkt auf diese Masse eine Rückstellkraft $\lambda x(t)$, die proportional ist zur Auslenkung $x(t)$ aus der Ruhelage; nach dem zweiten NEWTONschen Gesetz ist diese Kraft (eine zeitlich konstante Masse m vorausgesetzt) gleich $m\ddot{x}(t)$. Insgesamt ist also

$$m\ddot{x}(t) + \lambda x(t) = 0 \quad \text{oder} \quad \ddot{x}(t) + \frac{\lambda}{m} x(t) = 0\,.$$

Anwendung der LAPLACE-Transformation macht daraus

$$s^2\mathcal{L}\{x(t)\}(s) - s\cdot x(0) - \dot{x}(0) + \frac{\lambda}{m}\mathcal{L}\{x(t)\}(s) = 0$$

oder

$$\mathcal{L}\{x(t)\}(s) = \frac{s\cdot x(0) + \dot{x}(0)}{s^2 + \frac{\lambda}{m}}\,.$$

Die schwingende Masse m soll natürlich positiv sein, und auch λ ist größer als null, da $\lambda x(t)$ die *Rückstellkraft* ist. Also ist

$$\mathcal{L}\{x(t)\}(s) = \frac{x(0)\cdot s}{s^2+\omega^2} + \frac{\dot{x}(0)}{s^2+\omega^2} \quad \text{mit} \quad \omega = \sqrt{\frac{\lambda}{m}}\,.$$

Hier erkennen wir die gerade berechneten LAPLACE-Transformierten

$$\mathcal{L}\{\cos\omega t\}(s) = \frac{s}{s^2+\omega^2} \quad \text{und} \quad \mathcal{L}\{\sin\omega t\}(s) = \frac{\omega}{s^2+\omega^2}$$

und folgern, daß $x(t)$, falls LAPLACE-transformierbar, die Form

$$x(t) = x(0)\cos\omega t + \frac{\dot{x}(0)}{\omega}\sin\omega t$$

haben muß mit $\omega = \sqrt{\lambda/m}$. Die Masse schwingt also ungedämpft mit Frequenz $\sqrt{\lambda/m}$.

d) Gedämpfte Schwingungen

Ungedämpfte Schwingungen wir im letzten Abschnitt wird man in der Realität eher selten beobachten: In den meisten Fällen führen Reibungseffekte schließlich zum Abklingen der Schwingung. Die Reibungskraft wird in den einfachsten Modellen als proportional zur Geschwindigkeit angesetzt, d.h. die linke Seite der Differentialgleichung wird durch ein

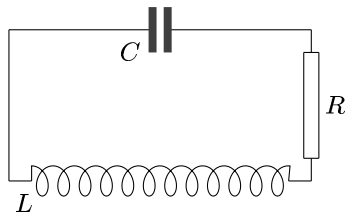


Abb. 18: Ein elektrischer Schwingkreis

konstantes Vielfaches von $\dot{x}(t)$ ergänzt. Dieselbe Art von Differentialgleichung erhalten wir auch, wenn wir eine Spule, einen Kondensator und einen Widerstand wie in Abbildung 18 hintereinanderschalten:

Damit hier ein Strom fließt, nehmen wir an, daß der Kondensator zur Zeitpunkt $t = 0$ eine Ladung Q_0 enthalte; die Ladung zum Zeitpunkt t sei $Q(t)$. Dann beträgt der Spannungsabfall am Kondensator

$$U_1(t) = \frac{Q(t)}{C}\,,$$

der an der Spule ist nach der LENZschen Regel gleich

$$U_2(t) = L\dot{I}(t) = L\ddot{Q}(t)\,,$$

wobei $I(t) = \dot{Q}(t)$ die Stromstärke bezeichnet, und am Widerstand haben wir natürlich nach dem OHMschen Gesetz

$$U_3(t) = RI(t) = R\dot{Q}(t)\,.$$

Diese drei Spannungen müssen sich zu Null addieren, d.h.

$$L\ddot{Q}(t) + R\dot{Q}(t) + \frac{Q(t)}{C} = 0 \quad \text{oder} \quad \ddot{Q}(t) + \frac{R}{L}\dot{Q}(t) + \frac{Q(t)}{LC} = 0\,.$$

Um bei der Lösung dieser Gleichung keine komplizierten Konstanten mitschleppen zu müssen, schreiben wir die Gleichung bis auf weiteres in der Form

$$\ddot{Q}(t) + \rho\dot{Q}(t) + \sigma Q(t) = 0 \quad \text{mit} \quad \rho = \frac{R}{L} \text{ und } \sigma = \frac{1}{LC}\,.$$

Außerdem schreiben wir $y(t)$ anstelle von $Q(t)$, um es einerseits mit gewohnten Variablen zu tun zu haben und andererseits, weil wir diesen

Typ von Gleichungen noch auf viele andere Probleme anwenden können, bei denen die gesuchte Funktion nicht als Ladung interpretiert werden kann. Wir interessieren uns für das Anfangswertproblem

$$\ddot{y}(t) + \rho\dot{y}(t) + \sigma y(t) = 0 \quad \text{mit} \quad y(0) = y_0 \quad \text{und} \quad \dot{y}(0) = y_1 .$$

Um zu sehen, wie sich die Lösungen solcher Gleichungen verhalten können, betrachten wir einige konkrete Beispiele. Beginnen wir mit dem Anfangswertproblem

$$\ddot{y}(t) + 8\dot{y}(t) + 25y(t) = 0 \quad \text{mit} \quad y(0) = 1,\ \dot{y}(0) = 2 .$$

Für die LAPLACE-Transformierte $Y(s) = \mathcal{L}\{y(t)\}(s)$ gilt dann

$$s^2Y(s) - s - 2 + 8\big(sY(s) - 1\big) + 25Y(s) = (s^2 + 8s + 25)Y(s) - s - 10 = 0,$$

also ist $Y(s) = \dfrac{s + 10}{s^2 + 8s + 25}$.

Wenn wir von der (bislang noch nicht bewiesenen) Annahme ausgehen, daß die gesuchte Funktion $y(t)$ durch ihre LAPLACE-Transformierte eindeutig bestimmt ist, müssen wir nun eine Funktion $y(t)$ finden, deren LAPLACE-Transformierte gleich $Y(s)$ ist. Unter den wenigen Beispielen, die wir bislang kennen, haben nur die Transformationen von Sinus und Kosinus quadratische Nenner, allerdings sind diese von der Art $s^2 + \omega^2$. Um einen linearen Term zu bekommen, müssen wir s durch $s + \lambda$ ersetzen; dies entspricht, wie wir gesehen haben, der Multiplikation mit einer Exponentialfunktion $e^{-\lambda t}$:

$$\mathcal{L}\{e^{-\lambda t}\cos\omega t\}(s) = \frac{s + \lambda}{(s + \lambda)^2 + \omega^2} \quad \text{und}$$
$$\mathcal{L}\{e^{-\lambda t}\sin\omega t\}(s) = \frac{\omega}{(s + \lambda)^2 + \omega^2} .$$

Wir müssen daher versuchen, den Nenner auf die Form $(s + \lambda)^2 + \omega^2$ zu bringen und den Zähler dann als Linearkombination von $s + \lambda$ und ω zu schreiben. Dies leistet einer der ältesten Tricks der Algebra, die schon

seit über 2000 Jahre bekannte quadratische Ergänzung:

$$\begin{aligned} Y(s) &= \frac{s + 10}{s^2 + 8s + 25} = \frac{s + 10}{(s + 4)^2 + 9} = \frac{s + 4}{(s + 4)^2 + 3^2} + 2 \cdot \frac{3}{(s + 4)^2 + 3^2} \\ &= \mathcal{L}\{e^{-4t}\cos 3t\}(s) + 2\mathcal{L}\{e^{-4t}\sin 3t\}(s) \\ &= \mathcal{L}\{e^{-4t}(\cos 3t + 2\sin 3t)\}(s) . \end{aligned}$$

Wenn wir, wie auch bei allen folgenden Beispielen, davon ausgehen, daß eine Funktion durch ihre LAPLACE-Transformation zumindest für alle positiven Werte von t eindeutig bestimmt ist, kennen wir also die gesuchte Funktion

$$y(t) = e^{-4t}(\cos 3t + 2\sin 3t) .$$

Sie beschreibt eine gedämpfte Schwingung der Art, wie sie in Abbildung 19 zu sehen ist. (Die Funktion $y(t)$ geht bezogen auf ihre Periode zu schnell gegen Null um ein interessantes Bild zu geben.)

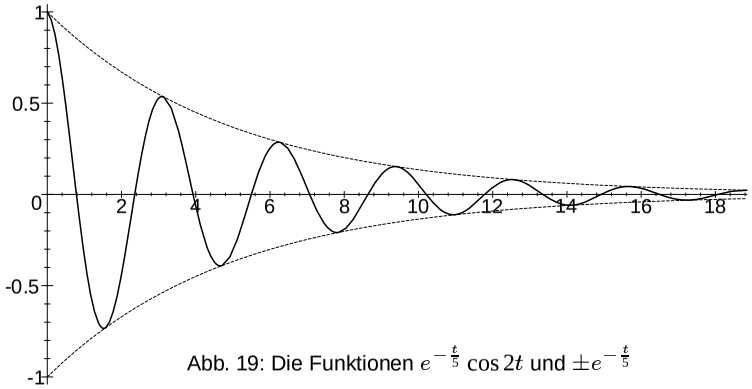


Abb. 19: Die Funktionen $e^{-\frac{t}{5}}\cos 2t$ und $\pm e^{-\frac{t}{5}}$

Als nächstes Beispiel betrachten wir das Anfangswertproblem

$$\ddot{y}(t) + 8\dot{y}(t) + 15y(t) = 0 \quad \text{mit} \quad y(0) = 1,\ \dot{y}(0) = 3 .$$

LAPLACE-Transformation macht daraus

$$s^2Y(s) - s - 3 + 8\big(sY(s) - 1\big) + 15Y(s) = (s^2 + 8s + 15)Y(s) - s - 11 = 0,$$

wobei wir wieder, wie auch in allen folgenden Beispielen, zur Abkürzung

$$Y(s) = \mathcal{L}\{y(t)\}(s)$$

setzen. Auflösen nach $Y(s)$ führt auf

$$Y(s) = \frac{s+11}{s^2+8s+15} = \frac{s+11}{(s+4)^2-1}\,.$$

Wegen des Minuszeichens im Nenner können wir dies nicht als LAPLACE-Transformation einer gedämpften Schwingung schreiben. Dafür sagt uns dieses Minuszeichen, daß der Nenner zwei *reelle* Nullstellen hat, nämlich -4 ± 1, also -3 und -5. Damit ist der Nenner auch gleich $(s+3)(s+5)$; via Partialbruchzerlegung können wir $Y(s)$ daher als Summe zweier rationaler Funktionen mit Nenner $s+3$ *bzw.* $s+5$ schreiben. Der Ansatz

$$\frac{a}{s+3} + \frac{b}{s+5} = \frac{(a+b)s+5a+3b}{(s+3)(s+5)} = \frac{s+11}{(s+3)(s+5)}$$

führt auf das lineare Gleichungssystem

$$a+b=1 \quad \text{und} \quad 5a+3b=11\,.$$

Subtraktion von fünfmal der ersten Gleichung von der zweiten ergibt $-2b=6$, also ist $b=-3$ und $a=4$. Damit ist

$$Y(s) = \frac{4}{s+3} - \frac{3}{s+5}\,.$$

Aus dem vorigen Paragraphen wissen wir, daß $1/s$ die LAPLACE-Transformation der Konstanten Eins ist, also ist $1/(s+\lambda)$ die von $e^{-\lambda t}$. Somit ist $Y(s)$ die LAPLACE-Transformierte von $4e^{-3t} - 3e^{-5t}$, und dies ist auch die gesuchte Lösungsfunktion. In diesem Beispiel geht $y(t)$ also exponentiell gegen Null.

Als letztes Beispiel zu diesem Typ von Gleichungen wollen wir noch das Anfangswertproblem

$$\ddot{y}(t) + 8\dot{y}(t) + 16 = 0 \quad \text{mit} \quad y(0)=1,\ \dot{y}(0)=2$$

betrachten. Hier ist

$$s^2Y(s) - s - 2 + 8\big(sY(s)-1\big) + 16Y(s) = (s^2+8s+16)Y(s) - s - 10 = 0\,,$$

also

$$Y(s) = \frac{s+10}{s^2+8s+16} = \frac{s+10}{(s+4)^2} = \frac{s+4}{(s+4)^2} + \frac{6}{(s+4)^2} = \frac{1}{s+4} + \frac{6}{(s+4)^2}\,.$$

Da $1/s^2$ die LAPLACE-Transformierte von t ist, entspricht der zweite Summand der Funktion $6te^{-4t}$, d.h. $Y(s)$ ist die LAPLACE-Transformierte von

$$y(t) = e^{-4t} + 6te^{-4t} = (1+6t)e^{-4t}\,.$$

Hier haben wir also ein Produkt einer Exponentialfunktion mit einer linearen Funktion. Wie Abbildung 20 zeigt, dominiert in solchen Fällen für kleine t die lineare Funktion, aber langfristig sorgt natürlich der Dämpfungsfaktor $e^{-\lambda t}$ dafür, daß sich die Funktion asymptotisch der Null annähert.

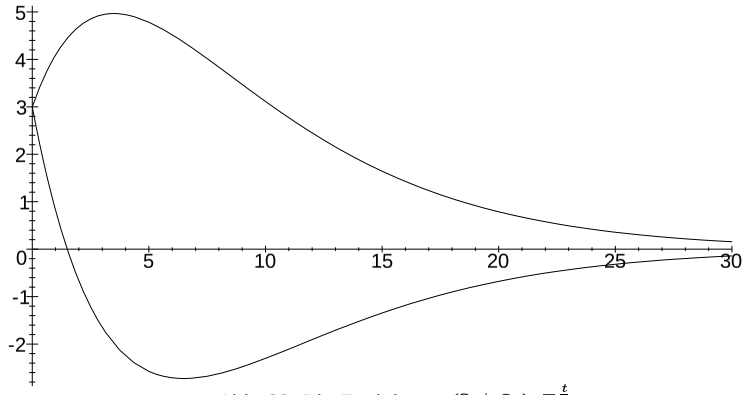


Abb. 20: Die Funktionen $(3 \pm 2t)e^{-\frac{t}{5}}$

Kehren wir zurück zum allgemeinen Fall, dem Anfangswertproblem

$$\ddot{y}(t) + \rho\dot{y}(t) + \sigma y(t) = 0 \quad \text{mit} \quad y(0)=y_0 \quad \text{und} \quad \dot{y}(0)=y_1\,.$$

Anwendung der LAPLACE-Transformation ergibt

$$s^2\mathcal{L}\{y(t)\}(s) - sy_0 - y_1 + \rho\big(s\mathcal{L}\{y(t)\}(s) - y_0\big) + \sigma\mathcal{L}\{y(t)\}(s) = 0$$

und damit

$$\mathcal{L}\{y(t)\}(s) = \frac{sy_0+y_1+\rho y_0}{s^2+\rho s+\sigma} = \frac{sy_0+y_1+\rho y_0}{\left(s+\frac{\rho}{2}\right)^2 + \sigma - \frac{\rho^2}{4}}\,.$$

Falls $\sigma > \rho^2/4$ können wir $\omega = \sqrt{\sigma - \frac{\rho^2}{4}}$ setzen und haben dann

$$\mathcal{L}\{y(t)\}(s) = \frac{sy_0 + y_1 + \rho y_0}{(s+\rho/2)^2 + \omega^2} = \frac{y_0(s+\rho/2)}{(s+\rho/2)^2+\omega^2} + \frac{y_1 + y_0\rho/2}{(s+\rho/2)^2+\omega^2}\,.$$

In diesen beiden Summanden erkennen wir (bis auf konstante Faktoren) die LAPLACE-Transformierten von $e^{-\rho t/2}\cos\omega t$ und $e^{-\rho t/2}\sin\omega t$, d.h.

$$\begin{aligned} y(t) &= y_0 e^{-\rho t/2}\cos\omega t + \frac{y_1 + y_0\rho/2}{\omega} e^{-\rho t/2}\sin\omega t \\ &= e^{-\rho t/2}\left(y_0\cos\omega t + \frac{y_1 + y_0\rho/2}{\omega}\sin\omega t\right)\,. \end{aligned}$$

Der Kondensator entlädt sich also, wie es physikalisch zu erwarten war, aber der zeitliche Verlauf ist gegeben durch eine gedämpfte Schwingung. Die Dämpfung wird mit wachsendem $\rho = R/L$ immer stärker, d.h. je größer der Widerstand und je kleiner die Induktivität ist, desto schneller geht die Lösungsfunktion gegen Null.

Falls σ kleiner ist als $\rho^2/4$, können wir $\omega = \sqrt{\frac{\rho^2}{4} - \sigma}$ setzen und haben

$$\mathcal{L}\{y(t)\}(s) = \frac{sy_0 + y_1 + \rho y_0}{(s+\rho/2)^2 - \omega^2} = \frac{sy_0 + y_1 + \rho y_0}{(s+\rho/2+\omega)(s+\rho/2-\omega)}\,.$$

Wie im obigen Beispiel ist hier eine Partialbruchzerlegung fällig; wegen

$$\frac{1}{(s+\rho/2)-\omega} - \frac{1}{(s+\rho/2)+\omega} = \frac{2\omega}{(s+\rho/2)^2-\omega^2}$$

ergibt sich die LAPLACE-Transformierte zu

$$\begin{aligned} \mathcal{L}\{y(t)\}(s) &= \frac{sy_0 + y_1 + \rho y_0}{(s+\rho/2)^2 - \omega^2} \\ &= \frac{1}{2\omega}\left(\frac{sy_0 + y_1 + \rho y_0}{(s+\rho/2)-\omega} - \frac{sy_0 + y_1 + \rho y_0}{(s+\rho/2)+\omega}\right) \\ &= \frac{1}{2\omega}\left(\frac{y_0(s+\rho/2-\omega) + y_0\rho/2 + y_1 + y_0\omega}{s+\rho/2-\omega}\right. \\ &\qquad \left. - \frac{y_0(s+\rho/2+\omega) + y_0\rho/2 + y_1 - y_0\omega}{s+\rho/2+\omega}\right) \\ &= \frac{1}{2\omega}\left(y_0 + \frac{y_0\rho/2 + y_1 + y_0\omega}{s+\rho/2-\omega} - y_0 - \frac{y_0\rho/2 + y_1 - y_0\omega}{s+\rho/2+\omega}\right) \\ &= \frac{y_0\rho/2 + y_1 + y_0\omega}{2\omega}\cdot\frac{1}{s+\rho/2-\omega} \\ &\qquad - \frac{y_0\rho/2 + y_1 - y_0\omega}{2\omega}\cdot\frac{1}{s+\rho/2+\omega}\,. \end{aligned}$$

Diese Summanden können wir nun als LAPLACE-Transformierte identifizieren und erhalten

$$\begin{aligned} \mathcal{L}\{y(t)\}(s) &= \frac{y_0\rho/2 + y_1 + y_0\omega}{2\omega}\cdot\mathcal{L}\{e^{-(\rho/2-\omega)t}\}(s) \\ &\qquad - \frac{y_0\rho/2 + y_1 - y_0\omega}{2\omega}\cdot\mathcal{L}\{e^{-(\rho/2+\omega)t}\}(s)\,. \end{aligned}$$

Die Lösung ist also

$$y(t) = \frac{y_0\rho/2 + y_1 + y_0\omega}{2\omega}\cdot e^{-(\rho/2-\omega)t} - \frac{y_0\rho/2 + y_1 - y_0\omega}{2\omega}\cdot e^{-(\rho/2+\omega)t}\,.$$

Wegen der Positivität von σ ist

$$\omega = \sqrt{\frac{\rho^2}{4} - \sigma} < \sqrt{\frac{\rho^2}{4}} = \frac{\rho}{2}\,;$$

daher sind dies zwei Exponentialfunktionen, die für $t \to \infty$ gegen null gehen. Damit entlädt sich der Kondensator in diesem Fall ohne Schwingungen gemäß einer Summe zweier abfallender Exponentialfunktionen.

Die Bedingung $\sigma < \frac{\rho^2}{4}$ übersetzt sich in

$$\frac{1}{LC} < \frac{R^2}{L^2} \quad \text{oder} \quad R > \sqrt{\frac{L}{C}}\,;$$

wenn der Widerstand zu groß ist, dämpft er also so stark, daß es keine Schwingungskomponente mehr gibt.

Bleibt noch der Fall $\sigma = \rho^2/4$. Hier ist

$$\mathcal{L}\{y(t)\}(s) = \frac{sy_0 + y_1 + \rho y_0}{(s+\rho/2)^2} = \frac{y_0(s+\rho/2)}{(s+\rho/2)^2} + \frac{y_1 + y_0\rho/2}{(s+\rho/2)^2}$$
$$= \frac{y_0}{s+\rho/2} + \frac{y_1 + y_0\rho/2}{(s+\rho/2)^2}\,.$$

Da $1/s$ die LAPLACE-Transformierte der Eins ist und $1/s^2$ die der Identität, folgt

$$y(t) = \left(y_0 + \left(y_1 + y_0\frac{\rho}{2}\right)t\right)\, e^{-\frac{\rho}{2}t}$$

Produkt einer linearen Funktion und einer abfallenden Eyponentialfunktion.

e) Erzwungene Schwingungen

Im Stromkreis aus dem letzten Abschnitt floß nur deshalb ein Strom, weil der Kondensator aus irgendeinem Grund bereits aufgeladen war; üblicher wäre, daß ein Strom fließt, weil der Stromkreis eine Stromquelle enthält. Wir ergänzen deshalb den Stromkreis aus Abbildung 18 durch eine Wechselstromquelle mit Kreisfrequenz ω_0. Die Differentialgleichung wird damit zu

$$L\ddot{Q}(t) + R\dot{Q}(t) + \frac{Q(t)}{C} = a\cos\omega_0 t + b\sin\omega_0 t\,,$$

in abstrakt-mathematischer Schreibweise geht es also um Differentialgleichungen der Form

$$\ddot{y}(t) + \rho\dot{y}(t) + \sigma y(t) = a\cos\omega_0 t + b\sin\omega_0 t\,.$$

Betrachten wir auch hierzu wieder zunächst einige Beispiele, etwa das Anfangswertproblem

$$\ddot{y}(t) + 8\dot{y}(t) + 25y(t) = 40\cos t + 40\sin t \quad \text{mit} \quad y(0) = 2,\ \dot{y}(0) = 4\,.$$

Die LAPLACE-Transformation macht daraus

$$s^2Y(s) - 2s - 4 + 8\big(sY(s) - 2\big) + 25Y(s)$$
$$= (s^2 + 8s + 25)Y(s) - 2s - 20 = \frac{40s+40}{s^2+1}\,,$$

und wir erhalten

$$Y(s) = \frac{2s+20}{s^2+8s+25} + \frac{40s+40}{(s^2+1)(s^2+8s+25)}\,.$$

Den ersten Summanden kennen wir im wesentlichen bereits aus dem letzten Abschnitt; beim zweiten hilft offensichtlich nur eine Partialbruchzerlegung: Wir setzen an

$$\frac{40s+40}{(s^2+1)(s^2+8s+25)} = \frac{as+b}{s^2+1} + \frac{cs+d}{s^2+8s+25}$$
$$= \frac{(as+b)(s^2+8s+25) + (cs+d)(s^2+1)}{(s^2+1)(s^2+8s+25)}$$
$$= \frac{(a+c)s^3 + (8a+b+d)s^2 + (25a+8b+c)s + 25b + d}{(s^2+1)(s^2+8s+25)}\,,$$

d.h. $a+c=0$. $\quad 8a+b+d=0,\quad 25a+8b+c=40 \quad$ und $\quad 25b+d=40.$

Damit ist $c = -a$ und $d = 40 - 25b$; setzen wir das ein in die beiden mittleren Gleichungen, folgt

$$8a + b + 40 - 25b = 8a - 24b + 40 = 0 \quad \text{und} \quad 24a + 8b = 40\,.$$

Subtrahiert man dreimal die erste Gleichung von der zweiten, folgt, daß $80b = 160$ ist, also $b = 2$ und $a = 1$. Damit kennen wir auch $c = -1$ und $d = -10$ und

$$Y(s) = \frac{2s+20}{s^2+8s+25} + \frac{s+2}{s^2+1} - \frac{s+10}{s^2+8s+25}$$
$$= \frac{s+2}{s^2+1} + \frac{s+10}{s^2+8s+25}$$
$$= \frac{s}{s^2+1} + \frac{2}{s^2+1} + \frac{s+4}{(s+4)^2+3^2} + 2\cdot\frac{3}{(s+4)^2+3^2}$$
$$= \mathcal{L}\{\cos t + 2\sin t + e^{-4t}\big(\cos 3t + 2\sin 3t\big)\}(s)\,.$$

Wir erhalten somit die Lösungsfunktion

$$y(t) = \cos t + 2\sin t + e^{-4t}\left(\cos 3t + 2\sin 3t\right).$$

Sie ist Summe aus einer gedämpften Schwingung, wie wir sie ohne die rechte Seite hätten, und einer reinen Schwingung der anregenden Frequenz, die sich langfristig durchsetzt. Im Vergleich zur rechten Seite hat sie jedoch sowohl eine andere Amplitude als auch eine andere Phase. Abbildung 21 zeigt, wieder mit besser zum Zeichnen geeigneten Parametern, eine solche Summe.

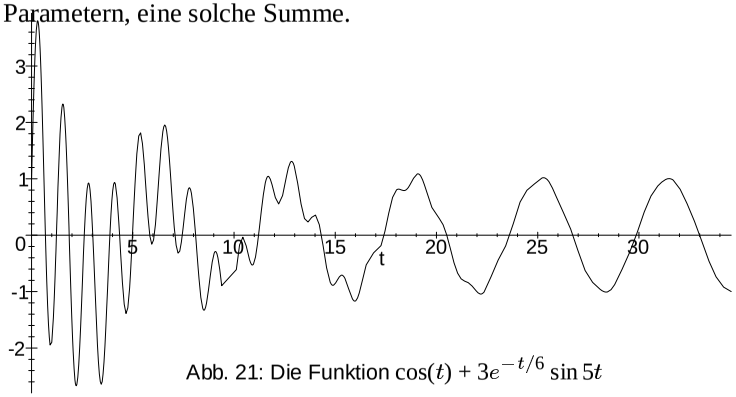


Abb. 21: Die Funktion $\cos(t) + 3e^{-t/6}\sin 5t$

Als zweites Beispiel betrachten wir

$$\ddot{y}(t) + 8\dot{y}(t) + 15y(t) = 30\cos t + 20\sin t \quad \text{mit} \quad y(0) = 2,\ \dot{y}(0) = 1\,.$$

Hier führt die LAPLACE-Transformation auf

$$s^2Y(s) - 2s - 1 + 8\big(sY(s) - 2\big) + 15Y(s)$$
$$= (s^2 + 8s + 15)Y(s) - 2s - 17 = \frac{30s + 20}{s^2 + 1}$$

oder

$$Y(s) = \frac{2s + 17}{s^2 + 8s + 15} + \frac{30s + 20}{(s^2 + 1)(s^2 + 8s + 15)}\,.$$

Wieder ist für den zweiten Summanden eine Partialbruchzerlegung notwendig.

$$\frac{30s + 20}{(s^2 + 1)(s^2 + 8s + 15)} = \frac{as + b}{s^2 + 1} + \frac{cs + d}{(s^2 + 8s + 15}$$

$$= \frac{(as + b)(s^2 + 8s + 15) + (cs + d)(s^2 + 1)}{(s^2 + 1)(s^2 + 8s + 15)}$$
$$= (a + c)s^3 + (8a + b + d)s^2 + (15a + 8b + c)s + 15b + d\,.$$

Wir müssen also das lineare Gleichungssystem

$$a + c = 0, \quad 8a + b + d = 0, \quad 15a + 8b + c = 30 \quad \text{und} \quad 15b + d = 20$$

lösen. Die erste und die letzte Gleichung erlauben auch hier wieder, $c = -a$ und $d = 20 - 15b$ durch a und b auszudrücken; Einsetzen in die beiden mittleren Gleichungen führt auf

$$8a - 14b = -20 \quad \text{und} \quad 14a + 8b = 30\,,$$

was sich zu

$$4a - 7b = -10 \quad \text{und} \quad 7a + 4b = 15$$

kürzen läßt. Vier mal erste plus sieben mal zweite Gleichung ergibt $65a = 65$ oder $a = 1$, also ist $b = 2$, $c = -1$ und $d = -10$. Somit ist

$$Y(s) = \frac{2s + 17}{s^2 + 8s + 15} + \frac{s + 2}{s^2 + 1} - \frac{s + 10}{s^2 + 8s + 15}$$
$$= \frac{s + 2}{s^2 + 1} + \frac{s + 7}{(s + 4)^2 - 1}\,.$$

Auch hier ist für den zweiten Summanden eigentlich wieder eine Partialbruchzerlegung notwendig, allerdings läßt sie sich in diesem Falle auch umgehen: Wegen der Linearität der LAPLACE-Transformation ist nämlich

$$\mathcal{L}\{\cosh \lambda t\}(s) = \mathcal{L}\left\{\frac{e^{\lambda t} + e^{-\lambda t}}{2}\right\} = \frac{1}{2}\left(\frac{1}{s - \lambda} + \frac{1}{s + \lambda}\right) = \frac{s}{s^2 - \lambda^2}$$

und

$$\mathcal{L}\{\sinh \lambda t\}(s) = \mathcal{L}\left\{\frac{e^{\lambda t} - e^{-\lambda t}}{2}\right\} = \frac{1}{2}\left(\frac{1}{s - \lambda} - \frac{1}{s + \lambda}\right) = \frac{\lambda}{s^2 - \lambda^2}\,.$$

Kombinieren wir dies mit der Regel, daß Multiplikation mit einer Exponentialfunktion das Argument verschiebt, erhalten wir über die Zerlegung

$$Y(s) = \frac{s + 2}{s^2 + 1} + \frac{s + 4}{(s + 4)^2 - 1} + \frac{3}{(s + 4)^2 - 1}$$

als Funktion mit Laplace-Transformation $Y(s)$

$$\begin{aligned} y(t) &= \cos t + 2\sin t + e^{-4t}(\cosh t + 3\sinh t) \\ &= \cos t + 2\sin t + \frac{e^{-3t} + e^{-5t} + 3e^{-3t} - 3e^{-5t}}{2} \\ &= \cos t + 2\sin t + 2e^{-3t} - e^{-5t}\,. \end{aligned}$$

Auch hier setzt sich also die anregende Schwingung durch.

Als letztes Beispiel betrachten wir

$$\ddot{y}(t) + 25y(t) = \cos 5t \quad \text{mit} \quad y(0) = 1,\ \dot{y}(0) = 0\,.$$

LAPLACE-Transformation führt auf

$$s^2Y(s) - s + 25Y(s) = \frac{s}{s^2+25} \quad \text{oder} \quad Y(s) = \frac{s}{s^2+25} + \frac{s}{(s^2+25)^2}\,.$$

Der erste Summand ist einfach die LAPLACE-Transformierte von $\cos 5t$; der zweite ist uns bislang noch nicht begegnet.

Wenn wir daran denken, daß die Quotienregel der Differentiation das Quadrat des Nenners der abgeleiteten Funktion im Nenner hat, liegt es nahe, mit der Ableitung der LAPLACE-Transformierten des Sinus zu vergleichen:

$$\frac{d}{ds}\mathcal{L}\{\sin\omega t\}(s) = \frac{d}{ds}\frac{\omega}{s^2+\omega^2} = \frac{-2\omega s}{(s^2+\omega^2)^2} = -2\omega\cdot\frac{s}{(s^2+\omega^2)^2}\,.$$

Wie wir aus Abschnitt *b)* wissen, ist die Ableitung der LAPLACE-Transformierten einer Funktion gleich der LAPLACE-Transformierten der $-t$-fachen Funktion. In unserem Fall folgt

$$\mathcal{L}\{t\sin\omega t\}(s) = \frac{2\omega s}{(s^2+\omega^2)^2}$$

und speziell

$$\mathcal{L}\{t\sin 5t\}(s) = \frac{10s}{(s^2+25)^2}\,.$$

Somit ist hier die Lösung gleich

$$y(t) = \cos 5t + \frac{t}{10}\sin 5t\,.$$

Der zweite Term ist eine Schwingung mit linear ansteigender Amplitude; wie die obige Rechnung zeigt, kommt er daher, daß die Eigenfrequenz der linken Seite gleich der anregenden Frequenz auf der rechten Seite ist. Da unbegrenzt wachsende Amplituden nichts gutes bedeuten, redet man hier von der sogenannten *Resonanzkatastrophe.* Abbildung 22 zeigt ein Beispiel.

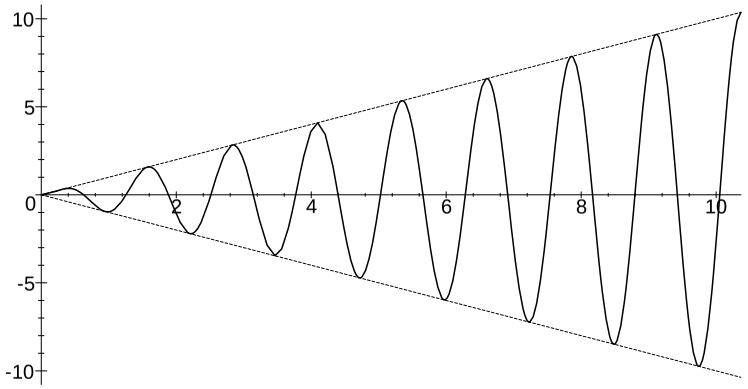


Abb. 22: Die Funktion $t\sin 5t$

Nach diesen Beispielen wollen wir auch das Anfangswertproblem dieses Abschnitts systematisch betrachten, allerdings sei die rechte Seite der Einfachheit halber als eine reine Kosinusfunktion der Form $A_0\cos\omega_0 t$ angenommen. Wir betrachten somit das Anfangswertproblem

$$\ddot{y}(t) + \rho\dot{y}(t) + \sigma y(t) = c\cos\omega_0 t \quad \text{mit} \quad y(0) = y_0,\ \dot{y}(0) = y_1\,.$$

Anwendung der LAPLACE-Transformation führt auf

$$(s^2 + \rho s + \sigma)\mathcal{L}\{y(t)\}(s) - sy_0 - y_1 - \rho y_0 = \frac{cs}{s^2+\omega_0^2}$$

und damit

$$\mathcal{L}\{y(t)\}(s) = \frac{sy_0 + y_1 + \rho y_0}{s^2+\rho s+\sigma} + \frac{cs}{(s^2+\omega_0^2)(s^2+\rho s+\sigma)}\,.$$

Die Umkehrung der LAPLACE-Transformation des ersten Summanden kennen wir: Das ist die Lösung des im vorigen Abschnitt betrachteten

Anfangswertproblems. Für den zweiten Summanden brauchen wir, wie im obigen Beispiel, eine Partialbruchzerlegung: Falls die beiden Faktoren des Nenners verschieden sind, können wir mit geeigneten Konstanten $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ schreiben

$$\frac{cs}{(s^2+\omega_0^2)(s^2+\rho s+\sigma)} = \frac{\alpha s+\beta}{s^2+\omega_0^2} + \frac{\gamma s+\delta}{s^2+\rho s+\sigma}\,.$$

Multiplikation mit dem Hauptnenner führt auf die Polynomgleichung

$$\begin{aligned} cs &= (\alpha s+\beta)(s^2+\rho s+\sigma) + (\gamma s+\delta)(s^2+\omega_0^2) \\ &= (\alpha+\gamma)s^3 + (\beta+\alpha\rho+\delta)s^2 + (\beta\rho+\alpha\sigma+\gamma\omega_0^2)s + \beta\sigma+\delta\omega_0^2\,, \end{aligned}$$

also auf das lineare Gleichungssystem

$$\alpha+\gamma = 0, \quad \alpha+\rho\beta+\delta = 0, \quad \sigma\alpha+\rho\beta+\omega_0^2\gamma = c \quad \text{und} \quad \sigma\beta+\omega_0^2\delta = 0\,.$$

Aus der ersten und der letzten Gleichung erhalten wir die Beziehungen

$$\gamma = -\alpha \quad \text{und} \quad \delta = -\frac{\sigma}{\omega_0^2}\beta\,;$$

damit bleiben nur noch zwei Gleichungen für die beiden Unbekannten α und β übrig:

$$\rho\alpha + \left(1-\frac{\sigma}{\omega_0^2}\right)\beta = 0 \quad \text{und} \quad (\sigma-\omega_0^2)\alpha+\rho\beta = c\,.$$

Falls ρ nicht verschwindet, führt die erste Gleichung zu

$$\alpha = \frac{\dfrac{\sigma}{\omega_0^2}-1}{\rho}\cdot\beta = \frac{\sigma-\omega_0^2}{\rho\omega_0^2}\cdot\beta\,,$$

und damit ist nach der zweiten Gleichung

$$\left(\frac{(\sigma-\omega_0^2)^2}{\rho\omega_0^2}+\rho\right)\beta = c$$

oder

$$\beta = \frac{c}{\frac{(\sigma-\omega_0^2)^2}{\rho\omega_0^2}+\rho} = \frac{c\rho\omega_0^2}{(\sigma-\omega_0^2)^2+\rho^2\omega_0^2}\,.$$

Damit sind auch α, γ und δ bekannt:

$$\alpha = -\gamma = \frac{\sigma-\omega_0^2}{\rho\omega_0^2}\cdot\beta = \frac{c(\sigma-\omega_0^2)}{(\sigma-\omega_0^2)^2+\rho^2\omega_0^2}$$

und

$$\delta = -\frac{\sigma}{\omega_0^2}\beta = \frac{-c\rho\sigma}{(\sigma-\omega_0^2)^2+\rho^2\omega_0^2}\,.$$

Bleibt noch der Fall $\rho = 0$ zu behandeln. Dann bleibt vom Gleichungssystem für α und β nur noch

$$\left(1-\frac{\sigma}{\omega_0^2}\right)\beta = 0 \quad \text{und} \quad (\sigma-\omega_0^2)\alpha = c$$

übrig. Ist $\sigma \neq \omega_0^2$, folgt, daß

$$\beta = \delta = 0 \quad \text{und} \quad \alpha = -\gamma = \frac{c}{\sigma-\omega_0^2}$$

sein muß, was offensichtlich genau die obigen Formeln im Spezialfall $\rho = 0$ sind.

Für $\sigma = \omega_0^2$ wird die zweite Gleichung zu $0\cdot\alpha = c \neq 0$ und damit unlösbar; das ist nicht weiter verwunderlich, denn das entspricht dem Fall, daß im obigen Ansatz zur Partialbruchzerlegung die beiden Nenner gleich sind, was natürlich nicht funktionieren kann.

In allen anderen Fällen kennen wir nun reelle Zahlen $\alpha, \beta, \gamma, \delta$, so daß

$$\frac{cs}{(s^2+\omega_0^2)(s^2+\rho s+\sigma)} = \frac{\alpha s+\beta}{s^2+\omega_0^2} + \frac{\gamma s+\delta}{s^2+\rho s+\sigma}$$

ist. Vom ersten Summanden wissen wir, daß

$$\mathcal{L}\left\{\alpha\cos\omega_0 t + \frac{\beta}{\omega_0}\sin\omega_0 t\right\} = \frac{\alpha s+\beta}{s^2+\omega_0^2}$$

ist; den zweiten Summanden müssen wir wie oben durch quadratische Ergänzung

$$s^2+\rho s+\sigma = \left(s-\frac{\rho}{2}\right)^2 + \sigma - \frac{\rho^2}{4}$$

umformen, und genau wie dort hängt es vom Vorzeichen von $\sigma - \frac{\rho^2}{4}$ ab, ob wir gedämpfte Schwingungen mit Frequenz $\omega = \sqrt{\sigma-\frac{\rho^2}{4}}$ oder

abfallende Eyponentialfunktionen erhalten. In jedem Fall ist die Lösung Linearkombination einer reinen Schwingung mit der erregenden Frequenz ω_0, im elektrischen Schwingkreis also der Frequenz der Wechselstromquelle, und einer Funktion, die für $t \to \infty$ gegen null geht. Langfristig setzt sich, wie in Abbildung 21 zeigt die erregende Frequenz durch.

Bleibt noch der zurückgestellte Fall, daß beide Nenner gleich sind, d.h. $\rho = 0$ und $\sigma = \omega_0^2$ ist. Dann müssen wir, wie im letzten der konkreten Beispiele dieses Abschnitts, eine Funktion finden, deren LAPLACE-Transformierte gleich

$$\frac{cs}{(s^2+\omega_0^2)^2}$$

ist. Dort haben wir gesehen, daß

$$\mathcal{L}\{t\sin\omega t\}(s) = \frac{2\omega s}{(s^2+\omega^2)^2}$$

ist, also folgt

$$\frac{cs}{(s^2+\omega_0^2)^2} = \mathcal{L}\left\{\frac{c}{2\omega_0}\cdot t\sin\omega_0 t\right\}(s)\,,$$

und wir haben auch diesen Fall gelöst: Er führt auf die bereits im letzten der konkreten Beispiele aufgetretene *Resonanzkatastrophe:* Die erregende Schwingung hat dieselbe Frequenz wie der Schwingkreis, und das führt, bei Abwesenheit einer jeglichen Dämpfung, zu einer katastrophalen Aufschaukelung. Auch bei Dämpfung ist Resonanz zu beobachten: Die oben berechneten Koeffizienten $\alpha, \beta, \gamma, \delta$, haben allesamt den Nenner

$$(\sigma-\omega_0^2)^2 + \rho^2\omega_0^2\,,$$

werden also umso größer, je näher σ bei ω_0^2 liegt, jedoch verhindert der Dämpfungsterm ρ, daß der Nenner je wirklich verschwindet. Bei kleinem ρ kann die Resonanz bei und um $\sigma = \omega_0^2$ allerdings in der Praxis trotzdem problematisch und in Extremfällen (z.B. bei Brücken) sogar katastrophal sein.

Mit den Formeln, die schon haben, könnten wir nun leicht die vollständigen Lösungen für jeden der behandelten Fälle hinschreiben, aber die

bisherige Diskussion zeigt, daß das doch zu sehr langen Formeln führen würde.Die LAPLACE-Transformation ist zwar sehr gut geeignet, um die Lösung eines *konkreten* Anfangswertproblems hinzuschreiben – dann sind $\alpha, \beta, \gamma, \delta$ keine komplizierten Ausdrücke, sondern einfach reelle Zahlen –, aber für abstraktere Überlegungen führt sie zu eher unübersichtlichen Ergebnissen. Wir werden daher im nächsten Kapitel alternative Methoden kennenlernen, die mehr über die Struktur der Lösungen von Differentialgleichungen aussagen.

§ 7: Die Fourier-Transformation auf dem Schwartz-Raum

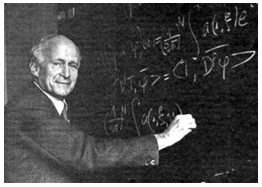
a) Der Schwartz-Raum der stark abfallenden Funktionen

Wie die Beispiele aus §5 zeigen, ist die Existenz von FOURIER- und LAPLACE-Integralen alles andere als sicher. In diesem Abschnitt wollen wir eine Klasse von Funktionen betrachten, für die es garantiert keine Probleme gibt, und wir wollen für diese Funktionen weitere Eigenschaften von FOURIER- und LAPLACE-Transformation herleiten. Im nächsten Paragraphen werden wir diese Ergebnisse verallgemeinern auf die Funktionen, die uns wirklich interessieren.

Definition: Eine Funktion $f\colon \mathbb{R} \to \mathbb{C}$ heißt *stark abfallend*, wenn sie beliebig oft stetig differenzierbar ist und die Funktionen

$$t \mapsto \left| t^r f^{(k)}(t) \right|$$

für alle $k, r \geq 0$ beschränkt sind. Die Menge aller stark abfallender Funktionen bezeichnen wir als SCHWARTZ-Raum $\mathcal{S}(\mathbb{R})$.



LAURENT SCHWARTZ (1915–2002) wurde in Paris geboren, studierte zunächst an der dortigen Ecole Normale Supérieure, dann an der Universität Straßburg. 1945 wurde er Professor in Nancy und entwickelte dort die mathematische Theorie der bislang nur von Physikern wie DIRAC und HEAVISIDE betrachteten Distributionen. Für diese Arbeiten wurde er 1950 mit der Fields Medal ausgezeichnet, dem bedeutendsten Preis in der Mathematik. Von 1953 bis zu seiner Emeritierung 1983 lehrte er in Paris, bis 1980 an der Ecole Polytechnique, dann an der Universität Paris VII.

Es ist klar, daß auch Summen und skalare Vielfache von stark abfallenden Funktionen stark abfallend sind; der SCHWARTZ-Raum $\mathcal{S}(\mathbb{R})$ ist daher ein $\mathbb{C}$-Vektorraum.

Beispiele: *a)* Die Funktion $f(t) = e^{-t^2}$ liegt in $\mathcal{S}(\mathbb{R})$: Sie ist beliebig oft stetig differenzierbar; ihre Ableitungen haben jeweils die Form $P(t)e^{-t^2}$ mit einem geeigneten Polynom P. Da die Exponentialfunktion schneller wächst als jedes Polynom, geht e^{-t^2} schneller gegen Null als ein Polynom gegen unendlich gehen kann, das Produkt geht also für $t \to \pm\infty$ gegen Null und ist daher auf ganz $\mathbb{R}$ beschränkt.

b) Sei

$$f(t) = \begin{cases} e^{\dfrac{-1}{(t-a)(b-t)}} & \text{falls } a < t < b \\ 0 & \text{sonst} \end{cases} .$$

Da diese Funktion außerhalb des Intervalls (a, b) verschwindet und im Innern stetig ist, ist sie natürlich beschränkt. Ihre Ableitungen sind Produkte aus rationalen Funktionen mit f selbst; da $f(t)$ für $t \to a$ oder $t \to b$ erheblich schneller gegen null geht als eine rationale Funktion gegen unendlich gehen kann, haben alle Ableitungen an den Intervallgrenzen den Wert null; die Funktion ist also beliebig oft stetig differenzierbar. Die Beschränktheitsbedingungen sind problemlos: Im kompakten Intervall $[a, b]$ ist jede stetige Funktion beschränkt, und außerhalb sind alle hier betrachteten Funktionen null.

Ein erster Hinweis darauf, daß wir in $\mathcal{S}(\mathbb{R})$ nur selten Probleme mit der Existenz von Integralen haben dürften, gibt das folgende

Lemma: *a)* Für eine Funktion $f \in \mathcal{S}(\mathbb{R})$ existieren

$$\int_{-\infty}^{\infty} f(t)\,dt \quad \text{und} \quad \int_{-\infty}^{\infty} f(t)\overline{f(t)}\,dt\,.$$

b) Für $f \in \mathcal{S}(\mathbb{R})$ und $\omega \in \mathbb{R}$ existiert das FOURIER-Integral

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,dt\,;$$

für $g \in \mathcal{S}(\mathbb{R})$ und $t \in \mathbb{R}$ existiert das inverse FOURIER-Integral

$$\check{g}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} g(\omega)e^{i\omega t}\,d\omega\,.$$

c) Die Abbildung

$$\begin{cases} \mathcal{S}(\mathbb{R}) \times \mathcal{S}(\mathbb{R}) \to \mathbb{C} \\ \qquad (f, g) \mapsto \displaystyle\int_{-\infty}^{\infty} f(t)\overline{g(t)}\,dt \end{cases}$$

macht $\mathcal{S}(\mathbb{R})$ zu einem HERMITEschen Vektorraum.

Beweis: a) Da sowohl $f(t)$ als auch $t^2 f(t)$ beschränkt sind, ist auch $(1 + t^2)f(t)$ beschränkt, es gibt also eine Konstante $C \in \mathbb{R}$, so daß

$$|f(t)| \le \frac{C}{1+t^2} \quad \text{für alle } t \in \mathbb{R}\,.$$

Da das uneigentliche Integral

$$\int_{-\infty}^{\infty} \frac{dt}{1+t^2} = \lim_{a,b\to\infty}\int_{-a}^{b} \frac{dt}{1+t^2} = \lim_{a,b\to\infty}\big(\arctan b - \arctan(-a)\big)$$
$$= \frac{\pi}{2} + \frac{\pi}{2} = \pi$$

konvergiert, ist es eine konvergente Majorante des Integrals über f, so daß nach dem Majorantenkriterium auch das letztere konvergiert. Damit ist auch *b)* bewiesen, d.h. die Konvergenz aller Integrale $\widehat{f}(\omega)$ und $\check{g}(t)$, denn da $e^{\pm i\omega t}$ den Betrag eins hat, ist auch für jedes $\omega \in \mathbb{R}$ *bzw.* $t \in \mathbb{R}$

$$|f(t)| = \left|f(t)\cdot e^{-i\omega t}\right| \le \frac{C}{1+t^2}$$

bzw.

$$|g(\omega)| = \left|g(\omega)\cdot e^{i\omega t}\right| \le \frac{C}{1+\omega^2}\,.$$

Genauso läßt sich auch das Integral über $f(t)\overline{f(t)}$ abschätzen, denn da $|tf(t)|$ beschränkt ist, ist auch $\left|t^2 f(t)\overline{f(t)}\right|$ und damit $(1 + t^2)f(t)\overline{f(t)}$ beschränkt. (Betragsstriche sind hier natürlich überflüssig.)

b) Wie wir gerade gesehen haben, konvergiert das rechtsstehende Integral im Spezialfall $f = g$. Für beliebiges $f, g \in \mathcal{S}(\mathbb{R})$ und beliebige reelle Zahlen $a \leq b$ gilt nach der CAUCHY-SCHWARZschen Ungleichung in der etwas allgemeineren Form aus **[HM1]**, *Kap. 1,* §6*c)*

$$\left| \int_a^b f(t)\overline{g(t)}\,dt \right| \leq \sqrt{\left| \int_a^b f(t)\overline{f(t)}\,dt \right|} \cdot \sqrt{\left| \int_a^b g(t)\overline{g(t)}\,dt \right|},$$

und somit konvergiert mit der rechten Seite auch die linke für $a \to -\infty$ und $b \to \infty$.

Die Eigenschaften eines HERMITEschen Skalarprodukts sind klar bis auf die Eigenschaft, daß nur die Nullfunktion Skalarprodukt null mit sich selbst haben darf, aber da wir es hier mit beliebig oft stetig differenzierbaren und damit insbesondere stetigen Funktionen zu tun haben, folgt dies genauso wie in **[HM1]**, *Kap. 1,* §6*a)* für das Skalarprodukt auf dem Vektorraum aller stetiger Funktionen $[0, 1] \to \mathbb{R}$. ∎

Da mit einer Funktion f auch alle deren Ableitungen sowie ihre Produkte mit Polynomen stark abfallend sind, gelten im übrigen auch die Formeln aus dem letzten Paragraphen über FOURIER-Transformationen und Ableitungen, ohne daß wir uns über die dort notwendigen, für Funktionen aus dem SCHWARTZ-Raum aber automatisch erfüllten Zusatzvoraussetzungen Gedanken machen müssen.

b) Die Fourier-Transformierte der Gauß-Funktion

Ein wesentliches Ziel dieses Paragraphen ist der Beweis, daß zumindest auf dem SCHWARTZ-Raum die inverse FOURIER-Transformation wirklich invers zur FOURIER-Transformation ist. Die Strategie ist folgende: Wir zeigen zunächst, daß dies für *eine* spezielle Funktion $f \in \mathcal{S}(\mathbb{R})$ gilt, und folgern daraus in einem zweiten Schritt, daß dies für *alle* $f \in \mathcal{S}(\mathbb{R})$ der Fall ist.

Für die eine spezielle Funktion aus $\mathcal{S}(\mathbb{R})$ haben wir nicht viel Auswahl: Wir kennen bislang im wesentlichen nur zwei Beispiele, nämlich $f(t) = e^{-t^2}$ und $f(t) = e^{-1/(t-a)(b-t)}$ auf (a, b) und null sonst. Da das

erste Beispiel etwas harmloser aussieht, nehmen wir dieses, und da es den Aufwand kaum vergrößert, später aber nützlich sein wird, verallgemeinern wir es leicht zu

$$f(t) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{\frac{-t^2}{2\sigma^2}} \quad \text{mit} \quad \sigma \in \mathbb{R}\,.$$

Diese Funktion heißt GAUSS-Funktion mit Varianz σ^2; ihr Graph wird auch als *Glockenkurve* bezeichnet. Abbildung 23 zeigt die Kurven für $\sigma = 1/2$ (gepunktet), $\sigma = 1$ (ausgezogen) und $\sigma = 2$ (gestrichelt); wie man sieht, wird die Kurve flacher für größere σ, wohingegen kleine σ zu einem schärfer ausgeprägten Maximum führen. Im Zusammenhang mit der Fehlerrechnung und Statistik werden uns am Ende des Semesters noch genauer mit dieser Funktion beschäftigen.

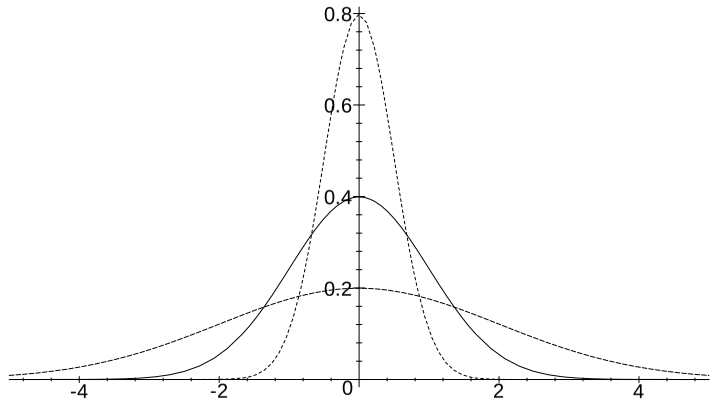


Abb. 23: Gaußkurven für $\sigma = \frac{1}{2}$, 1 und 2

Nach Definition ist

$$\widehat{f}(\omega) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \int_{-\infty}^{\infty} e^{\frac{-t^2}{2\sigma^2}} e^{-i\omega t}\,dt\,,$$

aber da schon die Stammfunktion von e^{-t^2} nicht elementar ausdrückbar ist, haben wir sicherlich wenig Chancen, dieses Integral über eine Stammfunktion zu berechnen.

Das Lemma aus dem vorigen Abschnitt erlaubt uns aber, Aussagen über die Ableitung von $\widehat{f}(\omega)$ machen:

$$\frac{d\widehat{f}}{d\omega}(\omega) = (-i)\cdot \widehat{tf}(\omega) = \frac{-i}{\sqrt{2\pi}\cdot\sigma}\int\limits_{-\infty}^{\infty} te^{-\frac{t^2}{2\sigma^2}} e^{-i\omega t}\,dt\,.$$

Der neue Integrand ist ziemlich ähnlich zur Ableitung des alten, denn

$$\frac{d}{dt}e^{-\frac{t^2}{2\sigma^2}-i\omega t} = -\left(\frac{t}{\sigma^2}+i\omega\right)e^{-\frac{t^2}{2\sigma^2}-i\omega t}$$

oder

$$\frac{d}{dt}\left(-\sigma^2 e^{-\frac{t^2}{2\sigma^2}-i\omega t}\right) = (t+i\omega\sigma^2)\,e^{-\frac{t^2}{2\sigma^2}-i\omega t}\,.$$

Die Funktion, die hier abgeleitet wird, geht für $t\to\pm\infty$ gegen null, d.h.

$$\int\limits_{-\infty}^{\infty}(t+i\omega\sigma^2)\,e^{-\frac{t^2}{2\sigma^2}-i\omega t}\,dt = -\sigma^2 e^{-\frac{t^2}{2\sigma^2}-i\omega t}\Big|_{-\infty}^{\infty} = 0\,,$$

und damit ist

$$\int\limits_{-\infty}^{\infty} te^{-\frac{t^2}{2\sigma^2}-i\omega t}\,dt = -i\omega\sigma^2\int\limits_{-\infty}^{\infty} e^{-\frac{t^2}{2\sigma^2}-i\omega t}\,dt\,.$$

Die Ableitung von $\widehat{f}(\omega)$ ist daher

$$\frac{d\widehat{f}}{d\omega}(\omega) = \frac{-i}{\sqrt{2\pi\sigma}}(-i\omega\sigma^2)\int\limits_{-\infty}^{\infty} e^{-\frac{t^2}{2\sigma^2}-i\omega t}\,dt = -\omega\sigma^2\cdot\widehat{f}(\omega)\,.$$

Somit ist $\widehat{f}(\omega)$ eine Lösung der Differentialgleichung

$$\frac{dg}{d\omega}(\omega) = -\omega\sigma^2\cdot g(\omega)\,.$$

Diese Differentialgleichung hat offensichtlich die Nullfunktion als eine ihrer Lösungen; falls sie auch eine Lösung $g(\omega)$ hat, die nicht für alle

Werte von ω verschwindet, können wir zumindest in der Umgebung solcher Werte durch $g(\omega)$ dividieren und erhalten

$$\frac{g'(\omega)}{g(\omega)} = -\omega\sigma^2\,.$$

Da die Ableitung der Logarithmusfunktion die Funktion $1/x$ ist, zeigt die Kettenregel, daß die linke Seite dieser Gleichung die Ableitung von $\ln g(\omega)$ ist. Durch Integration beider Seiten folgt

$$\ln g(\omega) = -\frac{\omega^2\sigma^2}{2} + C \Longrightarrow g(\omega) = e^C e^{-\frac{\sigma^2\omega^2}{2}}\,.$$

Somit ist $\widehat{f}(\omega)$ ein konstantes Vielfaches von $e^{-\frac{\sigma^2\omega^2}{2}}$, d.h.

$$\widehat{f}(\omega) = \widehat{f}(0)\cdot e^{-\frac{\sigma^2\omega^2}{2}}\,.$$

Damit ist uns die FOURIER-Transformierte von f bekannt bis auf die Konstante

$$\widehat{f}(0) = \frac{1}{\sqrt{2\pi\sigma}}\int\limits_{-\infty}^{\infty} e^{-\frac{t^2}{2\sigma^2}}\,dt\,.$$

In **[HM1]**, Kap. 2, §6*c)* hatten wir auf dem Umweg über ein zweidimensionales Integral

$$\int\limits_{-\infty}^{\infty} e^{-t^2}\,dt = \sqrt{\pi}$$

berechnet; über die Substitution $u = t/\sqrt{2}\sigma$ folgt daraus sofort

$$\widehat{f}(0) = \frac{1}{\sqrt{2\pi\sigma}}\int\limits_{-\infty}^{\infty} e^{-\frac{t^2}{2\sigma^2}}\,dt = 1\,.$$

Als Endergebnis erhalten wir somit

$$\widehat{f}(\omega) = e^{-\frac{\sigma^2\omega^2}{2}} = \frac{\sqrt{2\pi}}{\sigma}\cdot\frac{1}{\sqrt{2\pi}(1/\sigma)}e^{\frac{-\omega^2}{2\cdot(1/\sigma)^2}}\,,$$

wobei die kompliziertere zweite Form zeigt, daß es sich abgesehen vom Vorfaktor $\sqrt{2\pi}/\sigma$ wieder um eine GAUSS-Funktion handelt, allerdings mit Varianz $1/\sigma^2$.

Mit der Abkürzung

$$N_\sigma(t) \underset{\text{def}}{=} \frac{1}{\sqrt{2\pi}\sigma}\, e^{\frac{-t^2}{2\sigma^2}}$$

können wir kurz schreiben

$$\widehat{N_\sigma}(\omega) = \frac{\sqrt{2\pi}}{\sigma} N_{1/\sigma}(\omega)\,.$$

Damit kennen wir natürlich auch die inverse FOURIER-Transformierte einer GAUSS-Funktion, denn nach den allgemeinen Rechenregeln ist

$$\check{N}_\sigma(\omega) = \frac{1}{2\pi}\widehat{N_\sigma}(\omega) = \frac{1}{\sqrt{2\pi}\sigma} N_{\frac{1}{\sigma}}(\omega)\,.$$

Insbesondere können wir damit nachrechnen, daß die *inverse* FOURIER-Transformation zumindest in diesem Beispiel tatsächlich invers zur FOURIER-Transformation ist, d.h.

$$\check{\widehat{N}}_\sigma(t) = \frac{\sqrt{2\pi}}{\sigma}\check{N}_{\frac{1}{\sigma}}(t) = \frac{\sqrt{2\pi}}{\sigma}\cdot\frac{1}{\sqrt{2\pi}\frac{1}{\sigma}} N_\sigma(t) = N_\sigma(t)\,.$$

Genauso zeigt man, daß auch $\widehat{\check{N}}_{\sigma(t)} = N_\sigma(t)$ ist; die beiden Transformationen sind hier also in der Tat invers zueinander.

c) Die Umkehrung der Fourier-Transformation

Wie angekündigt, soll aus dem Beispiel des vorigen Abschnitts nun in einem zweiten Schritt gefolgert werden, daß dies nicht nur für die Funktionen N_σ gilt, sondern für *alle* Funktionen aus $\mathcal{S}(\mathbb{R})$, d.h.

Satz: Die FOURIER-Transformation und die inverse FOURIER-Transformation definieren zueinander inverse lineare Abbildungen

$$\begin{cases} \mathcal{S}(\mathbb{R}) \to \mathcal{S}(\mathbb{R}) \\ f \mapsto \widehat{f} \end{cases} \quad \text{und} \quad \begin{cases} \mathcal{S}(\mathbb{R}) \to \mathcal{S}(\mathbb{R}) \\ g \mapsto \check{g} \end{cases}.$$

Insbesondere sind also beide Abbildungen Isomorphismen, und für alle $f \in \mathcal{S}(\mathbb{R})$ ist

$$\check{\widehat{f}}(t) = f(t)\,.$$

Für das HERMITEsche Skalarprodukt auf $\mathcal{S}^2(\mathbb{R})$ gilt

$$\int_{-\infty}^{\infty} f(t)\overline{g(t)}\,dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{f}(\omega)\overline{\widehat{g}(\omega)}\,d\omega = 2\pi\int_{-\infty}^{\infty} \check{f}(\omega)\overline{\check{g}(\omega)}\,d\omega\,,$$

und damit insbesondere auch

$$\|f\|_2 = \frac{1}{\sqrt{2\pi}}\left\|\widehat{f}\right\|_2 = \sqrt{2\pi}\left\|\check{f}\right\|_2 \quad \text{mit} \quad \|f\|_2 = \sqrt{(f,\,f)}\,.$$

Beweis: Die Linearität ist, wie bei jedem Integral, klar; das Problem ist, ob $\widehat{f}$ und $\check{g}$ stark abfallend sind. Betrachten wir zunächst nur die Produkte $\omega^r\widehat{f}(\omega)$. Für diese ist

$$\left|\omega^r\widehat{f}(\omega)\right| = \left|(-i)^r\widehat{f^{(r)}}(\omega)\right| = \left|\int_{-\infty}^{\infty} f^{(r)}(t)e^{-i\omega t}\,dt\right|$$

$$\le \int_{-\infty}^{\infty}\left|f^{(r)}(t)\right|\,dt < \infty\,,$$

da f stark abfallend ist. Für

$$\omega^r\widehat{f^{(k)}}(\omega) = \omega^r(-i)^k\widehat{t^k f}(\omega)$$

können wir genauso argumentieren, und wegen des Zusammenhangs zwischen FOURIER-Transformation und inverser FOURIER-Transformation folgt daraus auch das Ergebnis für $\check{g}$.

Als nächstes wollen wir uns überlegen, daß für $f \in \mathcal{S}(\mathbb{R})$

$$\check{\widehat{f}}(t) = f(t)$$

ist. Dazu benutzen wir zwei zunächst beliebige weitere Funktionen $g, h \in \mathcal{S}(\mathbb{R})$, die wir im Laufe der Rechnung nach Bedarf genauer festlegen werden.

Nach Definition ist

$$\check{\widehat{f}}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\widehat{f}(\omega)e^{i\omega t}\,d\omega\,;$$

wir betrachten das etwas allgemeinere Integral

$$\int_{-\infty}^{\infty} \widehat{f}(\omega) e^{i\omega t} g(\omega)\, d\omega = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(s) e^{-i\omega s} ds \right) e^{i\omega t} g(\omega)\, d\omega \,,$$

das wir nach dem Satz von FUBINI weiter ausrechnen können als

$$\int_{-\infty}^{\infty} f(s) \cdot \left(\int_{-\infty}^{\infty} g(\omega) e^{-i\omega(s-t)}\, d\omega \right) ds = \int_{-\infty}^{\infty} f(s) \widehat{g}(s-t)\, ds \,.$$

Nun sei a eine positive reelle Konstante und $g(\omega) = h(a\omega)$, wobei die Funktion $h \in \mathcal{S}(\mathbb{R})$ im Augenblick noch beliebig ist. Dann führt die Substitution $\nu = a\omega$ auf

$$\widehat{g}(s) = \int_{-\infty}^{\infty} h(a\omega) e^{-i\omega s}\, d\omega = \int_{-\infty}^{\infty} h(\nu) e^{-i\nu \frac{s}{a}} \frac{d\nu}{a}$$

$$= \frac{1}{a} \int_{-\infty}^{\infty} h(\nu) e^{-i\nu \frac{s}{a}}\, d\nu = \frac{1}{a} \widehat{h}\left(\frac{s}{a}\right) \,.$$

Nach obiger Rechnung ist daher

$$\int_{-\infty}^{\infty} \widehat{f}(\omega) e^{i\omega t} h(a\omega)\, d\omega = \int_{-\infty}^{\infty} f(s) \cdot \frac{1}{a} \cdot \widehat{h}\left(\frac{s-t}{a}\right) ds \,.$$

Mit der neuen Variablen

$$u \underset{\text{def}}{=} \frac{s-t}{a}$$

ist $s = t + au$, und wir können diese Formel auch kürzer schreiben als

$$\int_{-\infty}^{\infty} \widehat{f}(\omega) e^{i\omega t} h(a\omega)\, d\omega = \int_{-\infty}^{\infty} f(t+au) \cdot \widehat{h}(u)\, du \,.$$

Beide Seiten sind stetig in a; für $a \to 0$ erhalten wir auf der linken Seite

$$\int_{-\infty}^{\infty} \widehat{f}(\omega) e^{i\omega t} h(0)\, d\omega = h(0) \int_{-\infty}^{\infty} \widehat{f}(\omega) e^{i\omega t}\, d\omega = 2\pi \cdot h(0) \cdot \check{\widehat{f}}(t)$$

und rechts

$$\int_{-\infty}^{\infty} f(t) \cdot \widehat{h}(u)\, du = f(t) \int_{-\infty}^{\infty} \widehat{h}(u)\, du = 2\pi \cdot f(t) \cdot \check{\widehat{h}}(0) \,.$$

Also ist für zwei beliebige Funktionen $f, h \in \mathcal{S}(\mathbb{R})$ stets

$$h(0) \cdot \check{\widehat{f}}(t) = f(t) \cdot \check{\widehat{h}}(0) \,.$$

Setzen wir nun für h speziell eine GAUSS-Funktion ein, etwa

$$h(\omega) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\omega^2}{2}} \,,$$

so wissen wir bereits aus dem obigem Beispiel, daß $\check{\widehat{h}}$ und h übereinstimmen; insbesondere haben beide an der Stelle $\omega = 0$ den von null verschiedenen Wert $\frac{1}{\sqrt{2\pi}}$, so daß wir durch diesen Wert dividieren können und die gewünschte Formel

$$\check{\widehat{f}}(t) = f(t)$$

erhalten. Wegen der Beziehungen

$$\widehat{\check{f}}(\omega) = 2\pi \check{\widehat{f}}(-\omega) \quad \text{und} \quad \check{f}(\omega) = \frac{1}{2\pi} \widehat{f}(-\omega)$$

ist dann auch

$$\widehat{\check{f}}(\omega) = \frac{1}{2\pi} \cdot 2\pi f\big(-(-\omega)\big) = f(\omega) \,.$$

Zu beweisen bleibt noch, daß die beiden Transformationen auch das HERMITEsche Skalarprodukt auf $\mathcal{S}(\mathbb{R})$ respektieren. Dazu wiederholen wir einfach die Rechnung zu Beginn des Beweises ohne den Faktor $e^{i\omega t}$: Für eine beliebige Funktion $g(\omega)$ aus $\mathcal{S}(\mathbb{R})$ ist

$$\int_{-\infty}^{\infty} \widehat{f}(\omega) g(\omega)\, d\omega = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt \right) g(\omega)\, d\omega$$

nach dem Satz von FUBINI gleich

$$\int_{-\infty}^{\infty} f(t) \cdot \left(\int_{-\infty}^{\infty} g(\omega) e^{-i\omega t}\, d\omega \right) dt = \int_{-\infty}^{\infty} f(t) \widehat{g}(t)\, dt \,,$$

wir haben also die Beziehung

$$\int_{-\infty}^{\infty} \widehat{f}(\omega) g(\omega)\, d\omega = \int_{-\infty}^{\infty} f(t) \widehat{g}(t)\, dt\,. \qquad (*)$$

Um daraus Aussagen über das HERMITEsche Skalarprodukt herzuleiten, benutzen wir die Beziehungen

$$\widehat{\widehat{f}}(t) = 2\pi \check{f}(-t) = 2\pi f(-t) \quad \text{oder} \quad f(t) = \frac{1}{2\pi} \widehat{\widehat{f}}(-t) \qquad (**)$$

und

$$\begin{aligned} \overline{\widehat{g}}(\omega) &= \int_{-\infty}^{\infty} \overline{g}(t)\, \overline{e^{-i\omega t}}\, dt = \int_{-\infty}^{\infty} \overline{g}(t)\, e^{i\omega t}\, dt \\ &= -\int_{\infty}^{-\infty} \overline{g}(-t) e^{-i\omega t}\, dt = \int_{-\infty}^{\infty} \overline{g}(-t) e^{-i\omega t}\, dt \qquad (***) \\ &= \widehat{\overline{g}}(-\omega)\,, \end{aligned}$$

wobei der Übersichtlichkeit halber $\overline{g}$ für diejenige Funktion steht, die jedem Wert t den Funktionswert $\overline{g}(t) = \overline{g(t)}$ zuordnet; entsprechend ist $\widehat{\overline{g}}(t) = \widehat{\overline{g}(t)}$.

Damit läßt sich das HERMITEsche Skalarprodukt folgendermaßen umformen:

$$\begin{aligned} (f, g) &= \int_{-\infty}^{\infty} f(t)\overline{g(t)}\, dt \underset{(**)}{=} \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \widehat{\widehat{f}}(-t)\widehat{\overline{g}}(-t)\, dt \\ &\underset{(*)}{=} \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \widehat{\widehat{\widehat{f}}}(-t)\widehat{\overline{g}}(-t)\, dt \underset{(**)}{=} \frac{1}{4\pi^2} \cdot 2\pi \int_{-\infty}^{\infty} \widehat{f}(t)\widehat{\overline{g}}(-t)\, dt \\ &\underset{(***)}{=} \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(t)\overline{\widehat{g}}(t)\, dt = \frac{1}{2\pi}(\widehat{f}, \widehat{g})\,. \end{aligned}$$

Die Aussage über das Produkt der inversen FOURIER-Transformierten folgt nun einfach daraus, daß die beiden Transformationen zueinander invers sind:

$$(\check{f}, \check{g}) = \frac{1}{2\pi}(\widehat{\widehat{f}}, \widehat{\widehat{g}}) = \frac{1}{2\pi}(f, g)\,.$$

■

Bemerkung: Falls wir für beide Transformationen den Vorfaktor $1/\sqrt{2\pi}$ gewählt hätten, würden beide das HERMITEsche Skalarprodukt respektieren, allerdings müßten wir uns dann ständig mit dieser Wurzel vor den Integralen herumschlagen. So hat jede Normierung ihre Vor- und Nachteile.

§8: Die Fourier-Transformation auf $L^2(\mathbb{R}, \mathbb{C})$

Wie wir im vorigen Paragraphen gesehen haben, verhält sich die FOURIER-Transformation auf dem SCHWARTZ-Raum genau so, wie wir es erwarten. Leider sind aber die Funktionen aus dem SCHWARTZ-Raum für die meisten Anwendungen zu schön, um nützlich zu sein. Wir brauchen daher einen größeren Funktionenraum, auf dem wir die FOURIER-Transformation immer noch gut verstehen können. Darum geht es in diesem Paragraphen.

a) Quadratintegrierbare Funktionen

Definition: Eine integrierbare Funktion $f: \mathbb{R} \to \mathbb{C}$ heißt *quadratintegrierbar,* wenn

$$\int_{-\infty}^{\infty} |f(t)|^2\, dt$$

existiert und konvergiert. Der Vektorraum aller quadratintegrierbarer Funktionen wird mit $L^2(\mathbb{R}, \mathbb{C})$ bezeichnet.

Nach Aussage *c)* des ersten Lemmas aus §7*a)* ist jede stark abfallende Funktion quadratintegrierbar, d.h. der SCHWARTZ-Raum $\mathcal{S}(\mathbb{R})$ ist ein Untervektorraum von $L^2(\mathbb{R}, \mathbb{C})$. Er ist allerdings deutlich kleiner als

$L^2(\mathbb{R}, \mathbb{C})$, denn beispielsweise ist auch jeder Rechteckimpuls quadratintegrierbar und allgemeiner jede stückweise stetige Funktion, die außerhalb eines endlichen Intervalls $[a, b]$ identisch verschwindet. Auch Funktionen wie $e^{-|t|}$ liegen in $L^2(\mathbb{R}, \mathbb{C})$, denn

$$\int_{-\infty}^{\infty} \left|e^{-|t|}\right|^2 dt = 2\int_0^{\infty} e^{-2t}\, dt = 1\,.$$

Funktionen wie $\sin \omega t$ sind natürlich nicht quadratintegrierbar; aber bei periodischen Funktionen betrachtet man ohnehin sinnvollerweise nur Integrale über eine Periode, nicht solche über die gesamte reelle Achse. (Das ist der aus der Elektrotechnik bekannte Unterschied zwischen Energie- und Leistungssignalen; die Energiesignale sind genau die quadratintegrierbaren.)

Auf dem SCHWARTZ-Raum haben wir ein HERMITEsches Produkt, bezüglich dessen wir das Integral über $|f|^2$ kurz als $\sqrt{(f, f)}$ schreiben können; wir wollen uns als nächstes überlegen, daß zumindest die Definition

$$(f, g) = \int_{-\infty}^{\infty} f(t)\overline{g(t)}\, dt$$

auch für $f, g \in L^2(\mathbb{R}, \mathbb{C})$ sinnvoll ist:

Da $\big(|f(t)| - |g(t)|\big)^2 \geq 0$ für alle $t \in \mathbb{R}$, ist nach der binomischen Formel auch

$$|f(t)| \cdot |g(t)| \leq \frac{1}{2}\Big(|f(t)|^2 + |g(t)|^2\Big)\,,$$

also

$$\int_N^M |f(t)g(t)|\, dt \leq \frac{1}{2}\int_N^M f(t)\overline{g(t)}\, dt + \frac{1}{2}\int_N^M g(t)\overline{g(t)}\, dt$$

für alle $N \leq M \in \mathbb{R}$. Rechts konvergieren beide Integrale für $N \to -\infty$ und $M \to \infty$, also auch links, und damit konvergiert das Integral zu (f, g) sogar absolut.

Es hat alle Eigenschaften eines HERMITEschen Produkts mit Ausnahme der positiven Definitheit – genau wie wir es vom periodischen Fall her gewohnt sind. Wie dort bezeichnen wir

$$\|f\|_2 \underset{\text{def}}{=} \sqrt{(f, f)}$$

kurz, wenn auch schlampig als L^2-Norm von f, denn – wie schon bei den periodischen Funktionen – können die Funktionen $f \not\equiv 0$ mit $\|f\|_2 = 0$ für die meisten Anwendungen *praktisch* vernachlässigt werden.

Definition: $f \in L^2(\mathbb{R}, \mathbb{C})$ heißt *Nullfunktion*, wenn $\|f\|_2 = 0$ ist.

Nach der CAUCHY-SCHWARZschen Ungleichung, die wir in **[HM1]**, *Kap. 1*, §6*c)* aus gutem Grund auch für Produkte bewiesen haben, die nur bis auf die positive Definitheit HERMITEsch sind, ist dann für eine beliebige Funktion $g \in L^2(\mathbb{R}, \mathbb{C})$

$$|(f, g)| \leq \|f\|_2 \cdot \|g\|_2 = 0\,,$$

für eine Nullfunktion f verschwindet also jedes Produkt (f, g), und umgekehrt ist auch jede Funktion mit dieser Eigenschaft eine Nullfunktion, denn $\|f\|_2$ ist ja die Wurzel aus (f, f).

b) Distributionen auf dem Schwartz-Raum

Jede quadratintegrierbare Funktion $f \in L^2(\mathbb{R}, \mathbb{C})$ definiert eine lineare Abbildung

$$\widetilde{T}_f\colon \begin{cases} L^2(\mathbb{R}, \mathbb{C}) \to \mathbb{C} \\ \qquad g \mapsto \displaystyle\int_{-\infty}^{\infty} f(t)g(t)\, dt \end{cases}$$

Man beachte, daß hier trotz der komplexwertigen Funktionen keine komplexe Konjugation steht! Vergleich mit dem Produkt

$$(f, g) = \int_{-\infty}^{\infty} f(t)\,\overline{g(t)}\, dt$$

zeigt, daß

$$\widetilde{T}_f(g) = (f, \overline{g}) = (g, \overline{f})$$

ist. Insbesondere ist $\widetilde{T}_f$ genau dann gleich der Nullabbildung, wenn f eine Nullfunktion ist.

Da der SCHWARTZ-Raum ein Untervektorraum von $L^2(\mathbb{R}, \mathbb{C})$ ist, können wir $\widetilde{T}_f$ auf $\mathcal{S}(\mathbb{R})$ einschränken und die lineare Abbildung

$$T_f: \begin{cases} \mathcal{S}(\mathbb{R}) \to \mathbb{C} \\ \varphi \mapsto \displaystyle\int_{-\infty}^{\infty} f(t)\varphi(t)\,dt \end{cases}$$

betrachten. Mit Hilfe dieser Abbildung wollen wir im folgenden Eigenschaften von f und seiner FOURIER-Transformierten (über deren Existenz wir noch nichts wissen) auf Eigenschaften stark abfallender Funktionen zurückführen.

Die Abbildung T_f existiert nicht nur für quadratintegrierbare Funktionen f, sondern allgemeiner für *jede* Funktion, deren Betrag höchstens polynomial ansteigt:

Lemma: Falls es zu einer stückweise stetigen Funktion $f: \mathbb{R} \to \mathbb{C}$ Konstanten $k \in \mathbb{N}_0$ und $c \in \mathbb{R}_{>0}$ gibt, so daß

$$|f(t)| \le c \cdot |t|^k$$

ist, existiert $T_f(\varphi)$ für alle $\varphi \in \mathcal{S}(\mathbb{R})$.

Beweis: Für $\varphi \in \mathcal{S}(\mathbb{R})$ ist $t^\ell \varphi(t)$ beschränkt für alle $\ell \in \mathbb{N}_0$, insbesondere also für $\ell = k$ und $\ell = k+2$. Damit ist auch deren Summe beschränkt, es gibt also eine Konstante $M > 0$, für die

$$\left|t^k(1+t^2)\varphi(t)\right| = \left|t^k\varphi(t) + t^{k+2}\varphi(t)\right| \le M$$

ist. Damit folgt

$$\left|(1+t^2)f(t)\varphi(t)\right| \le \left|(1+t^2)ct^k\varphi(t)\right| \le cM$$

und

$$|f(t)\varphi(t)| \le \frac{cM}{1+t^2}\,.$$

Da

$$\int_{-\infty}^{\infty} \frac{cM}{1+t^2}\,dt = cM \arctan t \Big|_{-\infty}^{\infty} = cM\pi$$

konvergiert, ist auch das Integral $T_f(\varphi)$ über die linke Seite der Gleichung absolut konvergent. ∎

Außerdem hat T_f eine Stetigkeitseigenschaft, die wir im Hinblick auf spätere Anwendungen gleich etwas allgemeiner formulieren wollen:

Lemma: $\varphi_n: \mathbb{R} \to \mathbb{C}$ und $f, \varphi: \mathbb{R} \to \mathbb{C}$ seien Funktionen derart, daß die Integrale

$$\int_{-\infty}^{\infty} f(t)\,\varphi_n(t)\,dt \quad \text{und} \quad \int_{-\infty}^{\infty} f(t)\,\varphi(t)\,dt$$

existieren. Außerdem sei f beschränkt und

$$\int_{-\infty}^{\infty} |\varphi(t) - \varphi_n(t)|\,dt \to 0 \quad \text{für} \quad n \to \infty$$

Dann ist

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} f(t)\,\varphi_n(t)\,dt = \int_{-\infty}^{\infty} f(t)\,\varphi(t)\,dt\,.$$

Beweis: Ist $|f(t)| \le M$ für alle $t \in \mathbb{R}$, so ist

$$\left|\int_{-\infty}^{\infty} f(t)\varphi(t)\,dt - \int_{-\infty}^{\infty} f(t)\varphi_n(t)\,dt\right|$$

$$\le \int_{-\infty}^{\infty} |f(t)| \cdot |\varphi(t) - \varphi_n(t)|\,dt$$

$$\leq M \int_{-\infty}^{\infty} |\varphi(t) - \varphi_n(t)|\, dt\,,$$

und letztere ist das M-fache einer Nullfolge, also selbst Nullfolge. ∎

Wir wollen dies anwenden auf Funktionen φ aus dem SCHWARTZ-Raum und Funktionen f, die höchstens polynomial ansteigen, die aber nicht notwendigerweise beschränkt sind. Um das zu kompensieren, führen wir für Folgen aus dem SCHWARTZ-Raum einen stärkeren Konvergenzbegriff ein, wobei wir (wie schon bei der Definition einer stark abfallenden Funktion) gleich so viel wie nur irgendwie möglich fordern:

Definition: Eine Folge $(\varphi_n)_{n\in\mathbb{N}}$ von Funktionen $\varphi_n \in \mathcal{S}(\mathbb{R})$ konvergiert gegen $\varphi \in \mathcal{S}(\mathbb{R})$, wenn für alle $r, k \in \mathbb{N}_0$ gilt:

$$\sup_{t\in\mathbb{R}} \left| t^k \left(\varphi^{(r)}(t) - \varphi_n^{(r)}(t)\right)\right| \to 0 \quad \text{für} \quad n \to \infty\,.$$

Wir fordern also, daß *alle* Produkte von t-Potenzen und Ableitungen von φ_n gegen die entsprechende Konstruktion für φ konvergieren. Unter dieser extrem starken Voraussetzung verwundert nicht

Lemma: f sei eine stückweise stetige Funktion, die einer Abschätzung der Form $|f(t)| \leq ct^k$ genüge. Dann ist für jede gegen $\varphi \in \mathcal{S}(\mathbb{R})$ konvergente Folge von Funktionen $\varphi_n \in \mathcal{S}(\mathbb{R})$

$$\lim_{n\to\infty} T_f(\varphi_n) = T_f(\varphi)\,.$$

Beweis: Wir gehen ähnlich vor wie beim Beweis der Existenz von $T_f(\varphi)$: Für jedes $\ell \in \mathbb{N}_0$ ist

$$\sup_{t\in\mathbb{R}} \left| t^\ell \left(\varphi(t) - \varphi_n(t)\right)\right|$$

eine Nullfolge, es gibt also zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$, so daß

$$\sup_{t\in\mathbb{R}} \left| t^\ell \left(\varphi(t) - \varphi_n(t)\right)\right| < \varepsilon \quad \text{für alle } n > N\,.$$

Insbesondere gibt es solche Werte für $\ell = k$ und für $\ell = k+2$, und damit gibt es auch zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$, so daß

$$\sup_{t\in\mathbb{R}} \left| t^k(1+t^2)\left(\varphi(t) - \varphi_n(t)\right)\right| < \varepsilon \quad \text{für alle } n > N\,,$$

d.h.

$$\left| t^k \left(\varphi(t) - \varphi_n(t)\right)\right| < \frac{\varepsilon}{1+t^2} \quad \text{für alle } n > N \text{ und } t \in \mathbb{R}\,.$$

Damit ist

$$\left|T_f(\varphi) - T_f(\varphi_n)\right| \leq \int_{-\infty}^{\infty} \left| f(t)\left(\varphi(t) - \varphi_n(t)\right)\right| dt$$

$$\leq \int_{-\infty}^{\infty} \left| ct^k \left(\varphi(t) - \varphi_n(t)\right)\right| dt \leq \int_{-\infty}^{\infty} \frac{c\varepsilon}{1+t^2}\, dt = c\pi \cdot \varepsilon$$

für alle $n > N_0$. Da c und π konstant sind und wir ε beliebig klein machen können, folgt die Behauptung. ∎

Definition: Eine *Distribution* auf dem SCHWARTZ-Raum ist eine lineare Abbildung $T\colon \mathcal{S}(\mathbb{R}) \to \mathbb{C}$, so daß für jede gegen ein $\varphi \in \mathcal{S}(\mathbb{R})$ konvergente Folge stark abfallender Funktionen φ_n gilt:

$$\lim_{n\to\infty} T_f(\varphi_n) = T_f(\varphi)\,.$$

Nach dem gerade bewiesenen Lemma ist also T_f für jede stückweise stetige Funktion f, die nicht stärker als ein Polynom wächst, eine Distribution auf dem SCHWARTZ-Raum.

Das sind allerdings bei weitem noch nicht alle Distributionen auf dem SCHWARTZ-Raum: Beispielsweise ist für jedes $a \in \mathbb{R}$ auch

$$\Delta_a\colon \begin{cases} \mathcal{S}(\mathbb{R}) \to \mathbb{C} \\ \quad \varphi \mapsto \varphi(a) \end{cases}$$

eine Distribution: Die Linearität ist klar, und für eine gegen $\varphi \in \mathcal{S}(\mathbb{R})$ konvergente Folge stark abfallender Funktionen φ_n ist insbesondere

$$\sup_{t\in\mathbb{R}} |\varphi(t) - \varphi_n(t)|$$

eine Nullfolge, erst recht also $|\varphi(a) - \varphi_n(a)|$, so daß es auch mit der Stetigkeit keine Probleme gibt.

Diese Distribution bezeichnet man als DIRACsche Delta-Distribution. Nicht ganz korrekt spricht man auch von einer DIRACschen Delta-Funktion und schreibt, gerade so als sei Δ_a von der Form T_δ,

$$\Delta_0(f) = \int_{-\infty}^{\infty} \delta(t) f(t)\,dt \quad \text{und} \quad \Delta_a(f) = \int_{-\infty}^{\infty} \delta(t-a) f(t)\,dt\,.$$

Die Schreibweise $\int_{-\infty}^{\infty} \delta(t-a) f(t) = f(a)$ findet man nicht nur für Funktionen f aus dem SCHWARTZ-Raum, sondern oft auch für beliebige stetige Funktionen f.

PAUL ADRIEN MAURICE DIRAC (1902–1984) wuchs auf in England als Sohn eines Schweizer Vaters und einer englischen Mutter. Trotz großem Interesse an der Mathematik studierte er von 1918–1921 Elektrotechnik an der Universität Bristol, da er auf keinen Fall Lehrer werden wollte. 1921 erhielt er ein Stipendium der Universität Cambridge; da dieses aber nicht zum Leben gereicht hätte, blieb er in Bristol, wo ihn die Universität von Studiengebühren befreite und seinen Wechsel in die Mathematik erlaubte. Ab 1923 arbeitete er in Cambridge an seiner Dissertation über Quantenmechanik, die er 1926 abschloß. 1930 folgte ein Buch über Quantenmechanik, für das er 1933 mit dem Nobelpreis für Physik ausgezeichnet wurde. 1932 bekam er einen Lehrstuhl für Mathematik an der Universität Cambridge. Nach seiner Emeritierung lebte er in Florida, wo er 1971 Physikprofessor an der Florida State University wurde. Zentrales Thema seiner Arbeiten war die Anwendung mathematischer Methoden auf die Quantenmechanik und die Relativitätstheorie sowie auch Ansätze zur (bis heute nicht befriedigend gelösten) Vereinheitlichung dieser beiden Theorien.

Wenn es wirklich eine *Funktion* $\delta\colon \mathbb{R} \to \mathbb{C}$ gäbe, für die

$$\int_{-\infty}^{\infty} \delta(t) f(t)\,dt = f(0)$$

wäre für jede (stark abfallende oder auch einfach stetige) Funktion f, so müßte $\delta(t)$ für $t \neq 0$ verschwinden – abgesehen eventuell von einigen

isolierten Punkten, die für die Integration bedeutungslos sind. Damit müßte aber unabhängig vom Funktionswert $\delta(0)$ und unabhängig von der Funktion f das Integral verschwinden.

Die „Lösung“, $\delta(0) = \infty$ zu setzen, führt nicht zu einer sinnvollen Interpretation des linksstehenden Integrals, denn wenn man einen Ausdruck wie $2\cdot\infty$ überhaupt sinnvoll interpretieren kann, dann wohl nur im Sinne von $2 \cdot \infty = \infty$, und damit wäre $2\delta(t) = \delta(t)$, obwohl die Distributionen

$$\begin{cases} \mathcal{S}(\mathbb{R}) \to \mathbb{C} \\ \varphi \mapsto \varphi(0) \end{cases} \quad \text{und} \quad \begin{cases} \mathcal{S}(\mathbb{R}) \to \mathbb{C} \\ \varphi \mapsto 2\varphi(0) \end{cases}$$

wohldefiniert und offensichtlich verschieden sind.

Die Schreibweise mit einer „Funktion“ δ ist also in mehrfacher Hinsicht problematisch, hat sich aber gerade in der technischen Literatur eingebürgert und soll daher auch hier verwendet werden. Man sollte sich aber klar machen, daß man nur Ausdrücke wie

$$\int_{-\infty}^{\infty} \delta(t-x) f(t)\,dt = f(x)$$

sinnvoll interpretieren kann, in anderen Zusammenhängen hat $\delta(t)$ keine vernünftige Bedeutung.

Problemlos unter einem Integralzeichen sind auch Linearkombinationen der Art

$$\sum_{k=1}^{n} a_k \delta(t - t_k)\,,$$

denn Linearkombinationen von Distributionen sind wieder Distributionen. Im vorliegenden Fall wäre dies die Distribution

$$\sum_{k=1}^{n} a_k \Delta_{t_k}\,,$$

für eine stark abfallende Funktion φ ist also

$$\int_{-\infty}^{\infty} \left(\sum_{k=1}^{n} a_k \delta(t - t_k) \right) \varphi(t)\,dt = \sum_{k=1}^{n} a_k \Delta_{t_k}(\varphi) = \sum_{k=1}^{n} a_k \varphi(t_k)\,,$$

und da man zumindest die DIRAC-Distribution auch einfach als lineare Abbildung auf $\mathcal{C}^0(\mathbb{R},\,\mathbb{C})$ betrachtet, kann man dies auch für eine beliebige stetige Funktion φ sinnvoll interpretieren. So ist beispielsweise

$$\int_{-\infty}^{\infty} \delta(t-1)e^{i\omega t}\,dt = e^{i\omega} \quad \text{und}$$

$$\int_{-\infty}^{\infty} \frac{1}{2}\big(\delta(\omega-1)+\delta(\omega+1)\big)e^{i\omega t}\,d\omega = \frac{e^{it}+e^{-it}}{2} = \cos t\,.$$

Da wir nur Distributionen auf dem SCHWARTZ-Raum betrachten, sind auch viele unendliche Linearkombinationen wie etwa

$$\sum_{k=1}^{\infty}\Delta_k \quad \text{oder} \quad \sum_{k=1}^{\infty} k\Delta_k$$

wohldefiniert, denn für eine stark abfallende Funktion φ konvergieren sowohl

$$\sum_{k=1}^{\infty}\varphi(k) \quad \text{als auch} \quad \sum_{k=1}^{\infty} k\varphi(k)\,.$$

Wir können eine Distribution T auf dem SCHWARTZ-Raum nicht nur mit Konstanten multiplizieren, sondern allgemeiner auch mit mit einer beliebig oft stetig differenzierbaren Funktion g, die höchstens polynomiales Wachstum hat: Für eine Distribution der Form T_f ist

$$T_{gf}(\varphi) = \int_{-\infty}^{\infty} g(t)f(t)\varphi(t)\,dt = \int_{-\infty}^{\infty} f(t)\big(g(t)\varphi(t)\big)\,dt = T_f(g\varphi)\,,$$

da auch $g\varphi$ eine stark abfallende Funktion ist. Somit können wir für eine beliebige Distribution T auf dem SCHWARTZ-Raum das Produkt

$$gT\colon \begin{cases} \mathcal{S}(\mathbb{R}) \to \mathbb{C} \\ \quad \varphi \mapsto T(g\varphi) \end{cases}$$

definieren. Beispielsweise gehört $t\delta(t)$ zur Distribution

$$t\Delta_0\colon \begin{cases} \mathcal{S}(\mathbb{R}) \to \mathbb{C} \\ \quad \varphi \mapsto \Delta_0(t\varphi) = (t\varphi)(0) = 0\cdot\varphi(0) = 0\,, \end{cases}$$

d.h. $t\delta = 0$. Man überlegt sich leicht, daß für jede Funktion g wie oben gilt $g\delta = g(0)\delta$.

Problematischer ist die Definition eines Produkts von Distributionen: Die obige Rechnung drückt T_{gf} aus durch T_f und g, nicht aber durch T_f und T_g, wie wir es bräuchten, um ein Produkt zweier Distributionen zu definieren. Auch sonstige Versuche, den Ausdruck $T_{gf}(\varphi)$ umzuformen, führen nicht zu brauchbareren Ergebnissen, und in der Tat kann man in der Theorie der Distributionen ein Produkt nur als Linearform auf dem SCHWARTZ-Raum der stark abfallenden Funktionen *zweier* Veränderlicher definieren. Dieser Raum wird weiter hinten zwar kurz erwähnt werden, es würde aber zu weit führen, ihn wirklich zu behandeln. Wir wollen daher nur festhalten, daß Produkte von δ-„Funktionen“ nicht sinnvoll als Distributionen auf $\mathcal{S}(\mathbb{R})$ definiert werden können und.

Ähnlich ist es mit Ausdrücken der Form $e^{\delta(t)}$ oder $\sin\delta(t)$: Da beispielsweise

$$\int_{-\infty}^{\infty} e^{f(t)}\varphi(t)\,dt \quad \text{und} \quad \int_{-\infty}^{\infty} f(t)e^{\varphi(t)}\,dt$$

nichts miteinander zu tun haben (und $e^{\varphi(t)}$ nicht einmal eine stark abfallende Funktion ist), können wir hier nicht einfach die Exponentialfunktion ins Argument von T_f schieben, und es ist gibt auch keine sonstige Art und Weise, Ausdrücken wie $e^{\delta(t)}$ oder $\sin\delta(t)$ einen Sinn zu geben. Bei der Funktionsschreibweise von Distributionen muß man sich also stets sorgfältig überlegen, ob ein gegebener Ausdruck wirklich sinnvoll interpretiert werden kann oder nicht.

c) Die Fourier-Transformierte einer Distribution

f sei eine absolut integrierbare Funktion, d.h. das Integral

$$\int_{-\infty}^{\infty} |f(t)|\,dt$$

konvergiere gegen einen endlichen Wert. Dann ist auch $f(t)e^{-i\omega t}$ absolut integrierbar, da diese Funktion den gleichen Betrag hat wie $f(t)$, und

damit konvergiert auch das FOURIER-Integral

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,dt$$

absolut. Da Multiplikation des Integranden mit einer stark abfallenden Funktion φ nichts an der absoluten Integrierbarkeit ändert, ist auch die lineare Abbildung

$$T_{\widehat{f}}\colon \begin{cases} \mathcal{S}(\mathbb{R}) \to \mathbb{C} \\ \quad \varphi \mapsto \displaystyle\int_{-\infty}^{\infty} \widehat{f}(t)\varphi(t)\,dt \end{cases}$$

wohldefiniert, und nach dem Satz von FUBINI gilt für alle stark abfallenden Funktionen $\varphi \in \mathcal{S}(\mathbb{R})$

$$\begin{aligned} T_{\widehat{f}}(\varphi) &= \int_{-\infty}^{\infty} \widehat{f}(\omega)\,\varphi(\omega)\,d\omega = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,dt \right) \varphi(\omega)\,d\omega \\ &= \int_{-\infty}^{\infty} f(t) \left(\int_{-\infty}^{\infty} \varphi(\omega)e^{-i\omega t}\,d\omega \right) dt = \int_{-\infty}^{\infty} f(t)\widehat{\varphi}(t)\,dt = T_f(\widehat{\varphi})\,. \end{aligned}$$

Dies legt folgende Definition nahe:

Definition: Die FOURIER-Transformierte der Distribution $T\colon \mathcal{S}(\mathbb{R}) \to \mathbb{C}$ ist die Distribution

$$\widehat{T}\colon \begin{cases} \mathcal{S}(\mathbb{R}) \to \mathbb{C} \\ \quad \varphi \mapsto T(\widehat{\varphi}) \end{cases};$$

die inverse FOURIER-Transformierte von T ist

$$\check{T}\colon \begin{cases} \mathcal{S}(\mathbb{R}) \to \mathbb{C} \\ \quad \varphi \mapsto T(\check{\varphi}) \end{cases}.$$

Zunächst müssen wir uns überlegen, ob das überhaupt sinnvoll ist:

Lemma: Für eine Distribution $T\colon \mathcal{S}(\mathbb{R}) \to \mathbb{C}$ sind auch $\widehat{T}$ und $\check{T}$ wieder Distributionen und $\check{\widehat{T}} = \widehat{\check{T}} = T$.

Beweis: Die letzte Aussage folgt sofort aus den Definitionen sowie der entsprechenden Aussage für starkabfallende Funktionen in §7*c)*. Auch die Linearität von $\widehat{T}$ ist klar, da die FOURIER-Transformation auf dem SCHWARTZ-Raum eine lineare Operation ist, d.h. die FOURIER-Transformierte von $\lambda\varphi + \mu\psi$ für $\lambda, \mu \in \mathbb{C}$ ist $\lambda\widehat{\varphi} + \mu\widehat{\psi}$.

Für die Stetigkeit von $\widehat{T}$ genügt es wegen der Stetigkeit von T, wenn wir zeigen, daß für eine konvergente Folge von Funktionen $\varphi_n(t) \in \mathcal{S}(\mathbb{R})$ mit Grenzwert $\varphi \in \mathcal{S}(\mathbb{R})$ auch die Folge der FOURIER-Transformierten $\widehat{\varphi_n}$ gegen $\widehat{\varphi}$ konvergiert. Wir müssen also zeigen, daß für je zwei Zahlen $k, r \in \mathbb{N}_0$ gilt

$$\lim_{n\to\infty} \sup_{\omega\in\mathbb{R}} \left| \omega^k \widehat{\varphi_n}^{(r)}(\omega) - \omega^k \widehat{\varphi}^{(r)}(\omega) \right| = 0\,.$$

Nach den Formeln aus §6*b)* ist

$$\begin{aligned} \omega^k \widehat{\varphi}^{(r)}(\omega) &= \omega^k \cdot (-i)^r \widehat{t^r\varphi}(\omega) = (-i)^r \cdot \omega^k \widehat{t^r\varphi}(\omega) \\ &= (-i)^r \cdot (-i)^k \widehat{\psi}(\omega) = (-i)^{r+k} \widehat{\psi}(\omega) \quad \text{mit} \quad \psi = \frac{d^k\left(t^r\varphi(t)\right)}{dt^k}\,. \end{aligned}$$

Durch k-fache Anwendung der Produktregel folgt, daß ψ Linearkombination von Termen der Form $t^\ell\varphi^{(s)}$ ist. Wegen der Dreiecksungleichung reicht es also, zu zeigen, daß für alle $\ell, s \in \mathbb{N}_0$ gilt

$$\lim_{n\to\infty} \sup_{\omega\in\mathbb{R}} \left| \widehat{t^\ell\varphi_n^{(s)}}(\omega) - \widehat{t^\ell\varphi^{(s)}}(\omega) \right| = 0\,.$$

Nach Definition der Konvergenz in $\mathcal{S}(\mathbb{R})$ gibt es zu ℓ, s und jedem $\varepsilon > 0$ ein $N_1 \in \mathbb{N}$, so daß für $n \geq N_1$ gilt

$$\sup_{t\in\mathbb{R}} \left| t^\ell\varphi_n^{(s)}(t) - t^\ell\varphi^{(s)}(t) \right| < \frac{\varepsilon}{2}\,.$$

Genauso gibt es auch ein $N_2 \in \mathbb{N}$, so daß für $n \geq N_2$

$$\sup_{t\in\mathbb{R}} \left| t^{\ell+2}\varphi_n^{(s)}(t) - t^{\ell+2}\varphi^{(s)}(t) \right| < \frac{\varepsilon}{2}$$

ist; für n größer oder gleich dem Maximum N_0 von N_1 und N_2 gilt also

$$\left|t^\ell \varphi_n^{(s)}(t) - t^\ell \varphi^{(s)}(t)\right| < \frac{\varepsilon}{1+t^2} \qquad \text{und}$$

$$\left|\widehat{t^\ell \varphi_n^{(s)}}(\omega) - \widehat{t^\ell \varphi^{(s)}}(\omega)\right| = \left|\int_{-\infty}^{\infty} \left(t^\ell \varphi_n^{(s)}(t) - t^\ell \varphi^{(s)}(t)\right) e^{-i\omega t}\, dt\right|$$

$$\leq \int_{-\infty}^{\infty} \left|t^\ell \varphi_n^{(s)}(t) - t^\ell \varphi^{(s)}(t)\right| dt \leq \varepsilon \int_{-\infty}^{\infty} \frac{dt}{1+t^2} = \varepsilon\pi\,,$$

der Limes für $n \to \infty$ ist also gleich null, wie behauptet. ∎

Um zu sehen, was die neue Definition bringt, wollen wir die FOURIER-Transformierte des Sinus berechnen: Im klassischen Sinne als

$$\widehat{\sin}\,\omega = \int_{-\infty}^{\infty} \sin t \cdot e^{-i\omega t}\, dt$$

existiert sie bekanntlich nicht. Im Distributionensinne ist

$$\widehat{T}_{\sin}(\varphi) = T_{\sin}(\widehat{\varphi}) = \int_{-\infty}^{\infty} \sin\omega\, \widehat{\varphi}(\omega)\, d\omega = \frac{1}{2i} \int_{-\infty}^{\infty} (e^{i\omega} - e^{-i\omega})\, \widehat{\varphi}(\omega)\, dt$$

$$= \frac{1}{2i} \int_{-\infty}^{\infty} \widehat{\varphi}(\omega) e^{i\omega}\, d\omega - \frac{1}{2i} \int_{-\infty}^{\infty} \widehat{\varphi}(\omega) e^{i\omega}\, d\omega$$

$$= \frac{2\pi}{2i} \left(\check{\widehat{\varphi}}(1) - \check{\widehat{\varphi}}(-1)\right) = -\pi i \big(\varphi(1) - \varphi(-1)\big)\,,$$

denn für jede Funktion g ist

$$\check{g}(1) = \frac{1}{2\pi} \int_{-\infty}^{\infty} g(\omega) e^{i\omega}\, d\omega \quad \text{und} \quad \check{g}(-1) = \frac{1}{2\pi} \int_{-\infty}^{\infty} g(\omega) e^{-i\omega}\, d\omega\,.$$

Für die oben eingeführte DIRAC-Distribution gilt

$$\Delta_a(\varphi) = \int_{-\infty}^{\infty} \delta(t-a)\varphi(t)\, dt = \varphi(a)\,,$$

und damit ist

$$\widehat{T}_{\sin} = -\pi i \big(\Delta_1 - \Delta_{-1}\big) = \pi i \big(\Delta_{-1} - \Delta_1\big)\,.$$

Kurz, wenn auch etwas kriminell, können wir dies als

$$\widehat{\sin}(\omega) = \pi i \big(\delta(\omega+1) - \delta(\omega-1)\big)$$

schreiben.

Falls diese Rechnung auf ein sinnvolles Ergebnis führte, sollte

$$\sin t = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{\sin}\,\omega \cdot e^{i\omega t}\, d\omega$$

sein, und in der Tat ist

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} (-\pi i)\big(\delta(\omega-1) - \delta(\omega+1)\big) e^{i\omega t}\, d\omega$$

$$= -\frac{i}{2} \int_{-\infty}^{\infty} \delta(\omega-1) e^{i\omega t}\, d\omega + \frac{i}{2} \int_{-\infty}^{\infty} \delta(\omega+1) e^{i\omega t}\, d\omega$$

$$= -\frac{i}{2}\big(e^{it} - e^{-it}\big) = \frac{e^{it} - e^{-it}}{2i} = \sin t\,.$$

d) Der Satz von Riesz

Der letzte Abschnitt hat gezeigt, daß die FOURIER-Transformation auf dem Niveau der Distributionen weitgehend unproblematisch ist. Was uns wirklich interessiert, sind aber Aussagen über die FOURIER-Transformation auf dem Niveau der *Funktionen;* wir müssen also wissen, wie wir von Distributionen wieder zurückkommen zu Funktionen. Wie das Beispiel der DIRAC-Distribution zeigt, ist das nicht immer möglich; wir müssen uns daher als erstes überlegen, was Distributionen der Form T_f mit $f \in \mathrm{L}^2(\mathbb{R}, \mathbb{C})$ auszeichnet.

Betrachten wir dazu für $f \in L^2(\mathbb{R}, \mathbb{C})$ zunächst die lineare Abbildung

$$\widetilde{T}_f\colon \begin{cases} L^2(\mathbb{R}, \mathbb{C}) \to \mathbb{C} \\ \qquad g \mapsto \displaystyle\int_{-\infty}^{\infty} f(t)g(t)\,dt \end{cases} .$$

Wie im Fall von T_f rechnet man auch hier schnell nach, daß $\widetilde{T}_f$ der Stetigkeitsbedingung

$$\lim_{n\to\infty} \widetilde{T}_f(g_n) = \widetilde{T}_f\big(\lim_{n\to\infty} g_n\big)$$

genügt, hier allerdings für bezüglich der L^2-Norm konvergente Folgen (g_n).

Außerdem ist nach der CAUCHY-SCHWARZschen Ungleichung

$$\big|T_f(g)\big| = |(f, \overline{g})| \le \|f\|_2 \cdot \|\overline{g}\|_2 = \|f\|_2 \cdot \|g\|_2 ,$$

für jede Funktion $g \in L^2(\mathbb{R}, \mathbb{C})$ läßt sich $\big|T_f(g)\big|$ also abschätzen durch ein von g unabhängiges Vielfaches der L^2-Norm von g.

Diese Eigenschaft hat nicht jede stetige lineare Abbildung von $L^2(\mathbb{R}, \mathbb{C})$ nach $\mathbb{C}$: Beispielsweise ist für die Fortsetzung

$$\widetilde{\Delta}_0\colon \begin{cases} L^2(\mathbb{R}, \mathbb{C}) \to \mathbb{C} \\ \qquad g \mapsto g(0) \end{cases}$$

der DIRAC-Distribution Δ_0 auf $L^2(\mathbb{R}, \mathbb{C})$ und den Rechteckimpuls

$$g_a\colon \begin{cases} \mathbb{R} \to \mathbb{C} \\ \quad t \mapsto \begin{cases} a & \text{falls } |t| \le \frac{1}{a^2} \\ 0 & \text{sonst} \end{cases} \end{cases}$$

die L^2-Norm unabhängig von a gleich

$$\|g_a\|_2 = \sqrt{\int_{-\infty}^{\infty} |g_a|^2\,(t)\,dt} = \sqrt{\int_{-1/a^2}^{1/a^2} a^2\,dt} = \sqrt{2} ,$$

aber $\Delta_0(g_a) = g_a(0) = a$ kann beliebig große Werte annehmen. Hier kann $|\Delta_0(g)|$ also nicht durch ein konstantes Vielfaches von $\|g\|_2$ abgeschätzt werden.

Definition: Eine lineare Abbildung $T\colon L^2(\mathbb{R}, \mathbb{C}) \to \mathbb{C}$ heißt *beschränkt*, wenn es eine Konstante $c \in \mathbb{R}$ gibt, so daß

$$|T(g)| \le c\,\|g\|_2 \quad \text{für alle } g \in L^2(\mathbb{R}, \mathbb{C}) .$$

Das Infimum aller Zahlen c, die diese Eigenschaft haben, bezeichnen wir als die *Norm* $\|T\|$ von T.

Für $f \in L^2(\mathbb{R}, \mathbb{C})$ ist $\widetilde{T}_f$ also beschränkt und hat die Norm $\|f\|_2$, denn wie wir gerade gesehen haben, ist $\big|\widetilde{T}_f(g)\big| \le \|f\|_2 \cdot \|g\|_2$ für alle g, und speziell für $g = \overline{f}$ ist $\big|\widetilde{T}_f(g)\big| = (f, \overline{\overline{f}}) = (f, f) = \|f\|_2^2$.

Das Schöne an quadratintegrierbaren Funktionen ist, daß sich diese Aussage auch umkehren läßt: Zu jeder beschränkten Distribution T gibt es eine Funktion $f \in L^2(\mathbb{R}, \mathbb{C})$, so daß $T = \widetilde{T}_f$ ist.

Zum Beweis brauchen wir unter anderem, daß $L^2(\mathbb{R}, \mathbb{C})$ bis auf das Problem mit den Nullfunktionen ein HILBERT-Raum ist, d.h.

Lemma: In $L^2(\mathbb{R}, \mathbb{C})$ hat jede CAUCHY-Folge einen Grenzwert.

Dieses Lemma ist, so wie wir $L^2(\mathbb{R}, \mathbb{C})$ definiert haben, leider falsch; es gilt nur, wenn wir $L^2(\mathbb{R}, \mathbb{C})$ ersetzen durch den etwas größeren Raum aller LEBESGUE-integrierbarer Funktionen, für die das Integral über das Betragsquadrat endlich bleibt. Da LEBESGUE-Integrale in dieser Vorlesung nicht definiert wurden, muß also hier eine Lücke bleiben; wo es Probleme gibt, zeigt der

„Beweis“: g_n sei eine CAUCHY-Folge von Funktionen aus $L^2(\mathbb{R}, \mathbb{C})$, d.h. zu jedem $\varepsilon > 0$ gibt es ein $N > 0$, so daß für $n, m \ge N$ gilt

$$\|g_n - g_m\|_2 < \varepsilon .$$

Offensichtlicher Kandidat für eine Grenzfunktion ist jene Funktion g, die jedem Wert t den Limes der $g_n(t)$ zuordnet; leider gibt es aber

zunächst keinen Grund, warum diese Folge von Funktionswerten für jedes t konvergieren sollte. Wir müssen daher etwas härter arbeiten.

Wir verschaffen uns zunächst eine Folge von Werten $\varepsilon_\nu > 0$, für die

$$\sum_{\nu=1}^{\infty} \varepsilon_\nu < \infty$$

konvergiert – beispielsweise können wir $\varepsilon_\nu = \frac{1}{\nu^2}$ setzen. Da (g_n) eine CAUCHY-Folge ist, gibt es zu jedem dieser ε_ν ein n_ν, so daß

$$\|g_n - g_m\|_2 \le \varepsilon_\nu \quad \text{für alle } n, m \ge n_\nu \,.$$

Insbesondere ist also

$$\left\|g_{n_{\nu+1}} - g_{n_\nu}\right\|_2 \le \varepsilon_\nu \,.$$

Damit ist für jede natürliche Zahl k

$$\begin{aligned}\sum_{\nu=k}^{\infty} \varepsilon_\nu &\ge \lim_{\ell\to\infty} \sum_{\nu=k}^{\ell} \left\|g_{n_{\nu+1}} - g_{n_\nu}\right\|_2 \ge \lim_{\ell\to\infty} \left\|\sum_{\nu=k}^{\ell} (g_{n_{\nu+1}} - g_{n_\nu})\right\|_2 \\ &= \lim_{\ell\to\infty} \left\|g_{n_{\ell+1}} - g_{n_k}\right\|_2 \,.\end{aligned}$$

Da die linke Seite für $k \to \infty$ wegen der Konvergenz der Summe der ε_ν gegen null geht, gilt dies auch für die rechte. Daher muß es eine Funktion g geben, die fast überall mit

$$t \mapsto \lim_{\nu\to\infty} g_{n_\nu}(t)$$

übereinstimmt und für die

$$\lim_{\nu\to\infty} \left\|g - g_{n_\nu}\right\|_2 = 0$$

ist. Da alle g_{n_ν} in $\mathrm{L}^2(\mathbb{R}, \mathbb{C})$ liegen, zeigt die Dreiecksungleichung, daß auch g dort liegen muß – falls g integrierbar ist. Man kann zeigen, daß g in jedem Fall LEBESGUE-integrierbar ist, auch wenn die g_n ebenfalls nur LEBESGUE-integrierbar sind; g muß aber nicht RIEMANN-integrierbar sein. Eine letzte Anwendung der Dreiecksungleichung zeigt noch, daß nicht nur die Teilfolge der g_{n_k}, sondern die Folge aller g_n in der L^2-Norm gegen g konvergiert.

■

Damit kommen wir zum eigentlich interessanten

Satz von Riesz: Zu jeder beschränkten und stetigen linearen Abbildung $T\colon \mathrm{L}^2(\mathbb{R}, \mathbb{C}) \to \mathbb{C}$ gibt es eine Funktion $f \in \mathrm{L}^2(\mathbb{R}, \mathbb{C})$, so daß $T = \widetilde{T}_f$ ist und $\|f\|_2 = \|T\|$. Die Funktion f ist bis auf Nullfunktionen eindeutig bestimmt.

Der *Beweis* ist etwas langwierig, aber seine Grundidee ist einfach:

Angenommen, wir betrachteten anstelle von $\mathrm{L}^2(\mathbb{R}, \mathbb{C})$ den endlichdimensionalen Vektorraum $\mathbb{R}^3$ und eine lineare Abbildung $T\colon \mathbb{R}^3 \to \mathbb{R}$. Dann wissen wir natürlich, daß sich $T(\vec{x})$ schreiben läßt als

$$T(\vec{x}) = a_1 x_1 + a_2 x_2 + a_3 x_3$$

mit geeigneten reellen Zahlen a_1, a_2 und a_3. Diese können wir zusammenfassen zu einen Vektor $\vec{a} \in \mathbb{R}^3$, für den

$$T(\vec{x}) = \vec{a} \cdot \vec{x}$$

ist. Dieser Vektor $\vec{a}$ entspricht der gesuchten Funktion f; er steht offensichtlich senkrecht auf dem Untervektorraum

$$E = \left\{\vec{x} \in \mathbb{R}^3 \;\middle|\; T(\vec{x}) = 0\right\},$$

der außer für $\vec{a} = \vec{0}$ eine Ebene beschreibt, und er ist durch E bis auf eine Proportionalitätskonstante eindeutig bestimmt.

In Analogie dazu betrachten wir auch für den Satz von RIESZ den Kern

$$N \underset{\text{def}}{=} \left\{g \in \mathrm{L}^2(\mathbb{R}, \mathbb{C}) \;\middle|\; T(g) = 0\right\}$$

von T. Falls $N = \mathrm{L}^2(\mathbb{R}, \mathbb{C})$ ist, sind wir fertig: Dann verschwindet $T(g)$ überall, und $f \equiv 0$ erfüllt alle Behauptungen.

Andernfalls gibt es eine Funktion $h \in \mathrm{L}^2(\mathbb{R}, \mathbb{C})$, die nicht in N liegt.

Der erste und umständlichste Beweisschritt besteht darin, daß wir uns überlegen, daß es in $\mathrm{L}^2(\mathbb{R}, \mathbb{C}) \setminus N$ auch eine Funktion $\widetilde{f}$ gibt, die auf N senkrecht steht, für die also $(\widetilde{f}, g) = 0$ ist für alle $g \in N$.

Dazu betrachten wir den Abstand $d \underset{\text{def}}{=} \inf_{g\in N} \|h - g\|_2$ von g und h.

Zur Erinnerung: In der Schule definiert man den Abstand eines Punkts von einer Ebenen als den Abstand zum nächstgelegenen Punkt der Ebenen. Dieser Punkt ist der Fußpunkt des Lots vom gegebenen Punkt auf die Ebene; der Verbindungsvektor steht also senkrecht auf der Ebenen. Bei einem unendlichdimensionalen Raum wie N können wir nicht sicher sein, daß es so etwas wie einen „Lotfußpunkt“ gibt – in der Tat besteht die Hauptarbeit des ersten Beweisschritts genau darin, dies zu zeigen. Deshalb können wir (noch) nicht von einem minimalen Abstand reden, sondern müssen uns zunächst mit einem Infimum begnügen. Wir hoffen aber (zu recht, wie sich bald zeigen wird), daß der „Lotfußpunkt“ auch in unserem Fall existiert und daß der „Lotvektor“ senkrecht auf N steht.

Obwohl h nicht in N liegt, können wir zumindest *a priori* nicht sicher sein, daß obiges Infimum positiv ist – wenn wir anstelle einer *beschränkten* stetigen linearen Abbildung T beispielsweise die stetige lineare Abbildung $\widetilde{\Delta}_0$ betrachten würden, wäre $d = 0$.

Da unser T aber beschränkt ist, haben wir eine Konstante $c > 0$, so daß

$$|T(g)| \le c\,||g||_2 \quad \text{für alle } g \in \mathrm{L}^2(\mathbb{R},\,\mathbb{C})\,.$$

Insbesondere ist für jedes $g \in N$

$$|T(h)| = |T(h) - T(g)| = |T(h-g)| \le c\,||h-g||_2\ .$$

$T(h)$ verschwindet nicht, da h nicht in N liegt; folglich ist

$$||h-g||_2 \ge \left|\frac{T(h)}{c}\right| \quad \text{für alle } g \in N\,.$$

Damit ist auch das Infimum d aller dieser Werte größer oder gleich $|T(h)|\,/c$, also positiv.

Ein Infimum muß nicht angenommen werden, man kann ihm aber beliebig nahekommen. Somit gibt es eine Folge (g_n) von Funktionen aus N, so daß $\lim_{n\to\infty} ||h-g_n||_2 = d$ ist.

Eine (ziemlich langweilige) Abschätzung zeigt, daß diese Folge eine CAUCHY-Folge ist, d.h. für jedes $\varepsilon > 0$ gibt es eine natürliche Zahl n_0, so daß

$$||g_m - g_n||_2 \le \varepsilon \quad \text{für } n, m > n_0\,.$$

Die Einzelheiten seien zum leichteren Überlesen im Kleindruck angegeben:

Zunächst ist für beliebige Funktionen p und q

$$||p+q||_2^2 = (p+q,\,p+q) = (p,\,p) + (p,\,q) + (q,\,p) + (q,\,q)$$

und

$$||p-q||_2^2 = (p-q,\,p-q) = (p,\,p) - (p,\,q) - (q,\,p) + (q,\,q)\,,$$

also

$$||p+q||_2^2 + ||p-q||_2^2 = 2\left(||p||_2^2 + ||q||_2^2\right)\ .$$

Damit erhalten wir

$$\begin{aligned}
||g_m - g_n||_2^2 &= ||(h-g_m) - (h-g_n)||_2^2 \\
&= 2\left(||h-g_m||_2 + ||h-g_n||_2^2\right) - ||2h - g_m - g_n||_2^2 \\
&= 2\left(||h-g_m||_2 + ||h-g_n||_2^2\right) - 4\left\|h - \frac{g_m+g_n}{2}\right\|_2^2 \\
&\le 2\left(||h-g_m||_2 + ||h-g_n||_2^2\right) - 4d^2\,,
\end{aligned}$$

denn mit g_m und g_n liegt auch $(g_m + g_n)/2$ in N, hat also mindestens Abstand d von h.

Da für die Folge der g_n die Abstände $||h-g_n||_2$ gegen d konvergiert, konvergiert auch die Folge der Abstandsquadrate gegen d^2, und es gibt zu jedem $\varepsilon > 0$ ein n_0, so daß

$$||h-g_n||_2^2 \le d^2 + \frac{\varepsilon}{4} \quad \text{für } n > n_0\,.$$

Für $n, m > n_0$ ist daher

$$||g_m - g_n||_2^2 \le 2\left(d^2 + \frac{\varepsilon}{4} + d^2 + \frac{\varepsilon}{4}\right) - 4d^2 = \varepsilon\,,$$

wie behauptet.

Da in $\mathrm{L}^2(\mathbb{R},\,\mathbb{C})$ nach dem vorigem Lemma jede CAUCHY-Folge konvergiert, folgt daß der Grenzwert

$$\widetilde{g} = \lim_{n\to\infty} g_n$$

in $\mathrm{L}^2(\mathbb{R},\,\mathbb{C})$ existiert. Da

$$T\big(\widetilde{g}\big) = T\Big(\lim_{n\to\infty} g_n\Big) = \lim_{n\to\infty} T(g_n) = \lim_{n\to\infty} 0 = 0$$

ist, liegt $\widetilde{g}$ in N.

Die Funktion $\widetilde{g}$ entspricht dem „Lotfußpunkt“; der „Lotvektor“

$$\widetilde{f} = h - \widetilde{g}\,,$$

von dem wir bislang nur wissen, daß $\|\widetilde{f}\|_2 = d$ ist, sollte also orthogonal zu N sein.

Für eine beliebige Funktion $g \in N$ und eine reelle Zahl $\lambda \neq 0$ betrachten wir den Abstand

$$\|h - (\widetilde{g} + \lambda g)\|_2 .$$

Da $\widetilde{g} + \lambda g$ in N liegt, ist dieser Abstand mindestens gleich d, d.h.

$$\begin{aligned} d^2 &\leq \|h - (\widetilde{g} + \lambda g)\|_2^2 = \|(h - \widetilde{g}) - \lambda g\|_2^2 = \left\|\widetilde{f} - \lambda g\right\|_2^2 \\ &= (\widetilde{f} - \lambda g,\ \widetilde{f} - \lambda g) = \|\widetilde{f}\|_2^2 + \lambda^2 \|g\|_2^2 - \lambda(g,\ \widetilde{f}) - \overline{\lambda}(\widetilde{f},\ g) . \end{aligned}$$

Da $\|\widetilde{f}\|_2^2 = d^2$ und $\overline{\lambda} = \lambda$ ist, folgt nach Division durch λ, daß

$$\begin{aligned} 0 &\leq \lambda \|g\|_2^2 - \Big((g,\ \widetilde{f}) + (\widetilde{f},\ g)\Big) = \lambda \|g\|_2^2 - \Big((g,\ \widetilde{f}) + \overline{(g,\ \widetilde{f})}\Big) \\ &= \lambda \|g\|_2^2 - 2\,\mathfrak{Re}(g,\ \widetilde{f}) \end{aligned}$$

für alle reellen $\lambda \neq 0$. Lassen wir λ, sowohl von links, als auch von rechts, gegen null gehen, folgt also

$$\mathfrak{Re}(g,\ \widetilde{f}) = 0 .$$

Die Funktion $g \in N$ war beliebig; da mit g auch ig in N liegt, ist daher auch

$$\mathfrak{Re}(ig,\ \widetilde{f}) = \mathfrak{Re}\Big(i \cdot (g,\ \widetilde{f})\Big) = -\,\mathfrak{Im}(g,\ \widetilde{f}) = 0 ,$$

also verschwindet auch der Imaginärteil von $(g,\ \widetilde{f})$ und damit $(g,\ \widetilde{f})$ selbst. $\widetilde{f}$ steht daher in der Tat senkrecht auf allen $g \in N$.

$\widetilde{f}$ ist nur bis auf eine Konstante bestimmt; wir wollen uns überlegen, daß

$$f \underset{\text{def}}{=} \frac{T(\widetilde{f})}{\|\widetilde{f}\|_2^2} \cdot \overline{\overline{\widetilde{f}}}$$

dasjenige Vielfache von $\widetilde{f}$ mit $\widetilde{T}_f = T$ ist:

Für $g \in L^2(\mathbb{R},\ \mathbb{C})$ ist

$$\begin{aligned} \widetilde{T}_f(g) = (g,\ \overline{f}) &= \left(g,\ \frac{\overline{T(\widetilde{f})}}{\|\widetilde{f}\|^2} \cdot \overline{\overline{\widetilde{f}}}\right) = \left(g,\ \frac{\overline{T(\widetilde{f})}}{\|\widetilde{f}\|^2} \cdot \widetilde{f}\right) \\ &= \frac{T(\widetilde{f})}{\|\widetilde{f}\|^2}(g,\ \widetilde{f}) . \end{aligned}$$

Insbesondere ist also $\widetilde{T}_f(g) = 0$ für alle $g \in N$ nach Konstruktion von $\widetilde{f}$.

Für ein Vielfaches $\lambda\widetilde{f}$ von $\widetilde{f}$ ist

$$\widetilde{T}_f(\lambda\widetilde{f}) = \frac{T(\widetilde{f})}{\|\widetilde{f}\|^2}\left(\lambda\widetilde{f},\ \overline{\overline{\widetilde{f}}}\right) = \frac{T(\widetilde{f})}{\|\widetilde{f}\|^2}\lambda\|\overline{f}\|_2^2 = \lambda T(\widetilde{f}) = T(\lambda\widetilde{f}) ,$$

auch in diesem Fall stimmen $\widetilde{T}_f$ und T also überein. Wegen der Linearität von T und von $\widetilde{T}_f$ ist daher

$$\widetilde{T}_f(g + \lambda\widetilde{f}) = T(g + \lambda\widetilde{f}) \quad \text{für alle } g \in \mathbb{N},\ \lambda \in \mathbb{C} .$$

Eine beliebige Funktion $h \in L^2(\mathbb{R},\ \mathbb{C})$ können wir in der Form

$$h = \left(h - \frac{T(h)}{T(\widetilde{f})}\widetilde{f}\right) + \frac{T(h)}{T(\widetilde{f})}\widetilde{f}$$

darstellen. Da

$$T\left(h - \frac{T(h)}{T(\widetilde{f})}\widetilde{f}\right) = T(h) - \frac{T(h)}{T(\widetilde{f})}T(\widetilde{f}) = T(h) - T(h) = 0$$

verschwindet, liegt der erste Summand in N, und der zweite ist natürlich ein Vielfaches von $\widetilde{f}$. Also läßt sich jede quadratintegrierbare Funktion darstellen als Summe einer Funktion aus N und einem Vielfachen von $\widetilde{f}$, die linearen Abbildungen T und $\widetilde{T}_f$ stimmen also überein.

Damit sind wir fast fertig: Wenn $T = \widetilde{T}_f$ ist, haben beide Abbildungen natürlich auch dieselbe Norm, und wir wissen bereits, daß $\widetilde{T}_f$ dieselbe Norm hat wie f, d.h.

$$\|T\| = \|\widetilde{T}_f\| = \|f\|_2 .$$

■

FRIGYES RIESZ (1880–1956) studierte Mathematik in Budapest, Göttingen und Zürich. 1902 promovierte er in Budapest mit einer Arbeit über Geometries, 1911 wurde er Professor an der damals ungarischen Universität Kolozsvár. Nachdem Kolozsvár 1920 rumänisch wurde, zog er mit der Universität um nach Szeged. 1945 bekam er einen Lehrstuhl an der Universität Budapest.

RIESZ ist einer der Väter der *Funktionalanalysis*, jener mathematischen Disziplin also, die Funktionenräume mit analytischen Methoden untersucht und insbesondere auch fundamental für die FOURIER-Analysis ist. Den obigen Satz bewies er 1907.

e) Die Plancherel-Formel

Der Satz von RIESZ sagt uns, wann lineare Funktionen auf $L^2(\mathbb{R}, C)$ in der Form $\widetilde{T}_f$ geschrieben werden können mit einer Funktion f aus $L^2(\mathbb{R}, C)$. Da wir die FOURIER-Theorie auf $L^2(\mathbb{R}, C)$ zurückführen wollen auf die für Distributionen auf dem SCHWARTZ-Raum, müssen wir wir daher versuchen, solche Distributionen auf $L^2(\mathbb{R}, C)$ fortzusetzen. Als erstes wollen wir uns dazu überlegen, daß wir jede Funktion aus $L^2(\mathbb{R}, C)$ als Grenzwert einer Folge von Funktionen aus dem SCHWARTZ-Raum $\mathcal{S}(\mathbb{R})$ schreiben können.

Wir beginnen mit dem Beispiel des Rechteckimpulses

$$f(t) = \begin{cases} 1 & \text{für } a \le t \le b \\ 0 & \text{sonst} \end{cases},$$

der offensichtlich in $L^2(\mathbb{R}, C)$ liegt, wegen der beiden Unstetigkeitsstellen aber natürlich nicht in $\mathcal{S}(\mathbb{R})$.

Wir kennen bereits eine Funktion in $\mathcal{S}(\mathbb{R})$, die auch außerhalb des Intervalls $[a, b]$ verschwindet und in dessen Innern positiv ist, nämlich die Funktion

$$g\colon \begin{cases} \mathbb{R} \to \mathbb{R} \\ t \mapsto \begin{cases} e^{\dfrac{-1}{(t-a)(b-t)}} & \text{falls } a < t < b \\ 0 & \text{sonst} \end{cases} \end{cases}.$$

Allgemeiner hat für jede reelle Zahl $r > 0$ auch

$$g_r\colon \begin{cases} \mathbb{R} \to \mathbb{R} \\ t \mapsto \begin{cases} e^{\dfrac{-r}{(t-a)(b-t)}} & \text{falls } a < t < b \\ 0 & \text{sonst} \end{cases} \end{cases}$$

dieselbe Eigenschaft. Da $(t-a)(b-t)$ bei $t = (a+b)/2$ maximal wird, hat g_r dort sein einziges Maximum und

$$g_r\left(\frac{a+b}{2}\right) = e^{\dfrac{-4r}{(b-a)^2}}.$$

Unser Rechteckimpuls hat eins als Maximalwert; deshalb betrachten wir besser anstelle der g_r die mit dem Kehrwert des Maximums multiplizierten Funktionen

$$f_r\colon \begin{cases} \mathbb{R} \to \mathbb{R} \\ t \mapsto \begin{cases} e^{\dfrac{4r}{(b-a)^2}} e^{\dfrac{-r}{(t-a)(b-t)}} & \text{falls } a < t < b \\ 0 & \text{sonst} \end{cases} \end{cases},$$

die alle bei $(a+b)/2$ ihren Maximalwert eins annehmen.

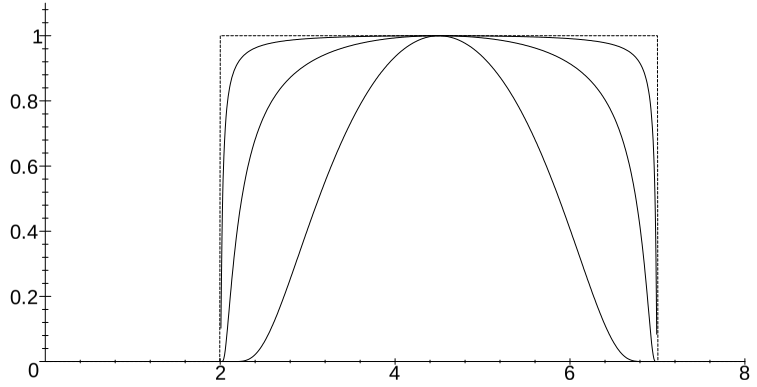


Abb. 24: Approximation des Rechteckimpulses durch stark abfallende Funktionen

Abbildung 24 zeigt für $a = 2$ und $b = 7$ die entsprechenden Funktionen mit $r = 10$, $r = 1$ und für $r = 0{,}1$. Die innerste Kurve für $r = 10$ zeigt noch ein klar ausgeprägtes Maximum, die Kurve für $r = 1$ ist schon deutlich flacher im mittleren Teil, und die für $r = 0{,}1$ schließlich erinnert schon recht gut an den Rechteckimpuls f.

Diese Abbildung legt die Vermutung nahe, daß die f_r für $r \to 0$ in der L^2-Norm gegen f konvergieren. Leider können wir aber $\|f - f_r\|_2$ nicht ausrechnen, da wir keine Stammfunktion von f_r kennen. (Schon e^{-t^2} ist schließlich nicht elementar integrierbar.) Deshalb müssen wir uns mit Abschätzungen begnügen. Wir erwarten, daß f_r im mittleren Bereich immer besser mit der Geraden auf Höhe eins übereinstimmt, während es am Rand des Intervalls immer steiler gegen null geht. Daher wählen wir ein $\delta > 0$ und betrachten getrennt denn mittleren Teil $[a + \delta, b - \delta]$ des Intervalls und die beiden Randintervalle $[a, a + \delta]$ und $[b - \delta, b]$.

Über das Verhalten von f_r in den Randintervallen können wir so gut wie nichts sagen; wir wissen nur, daß auf jeden Fall $0 \le f_r(t) \le 1$ ist und schätzen die Differenz zwischen $f(t) = 1$ und $f_r(t)$ daher ab durch eins.

Im mittleren Intervall ist die Differenz zwischen $f(t)$ und $f_r(t)$ im Intervallmittelpunkt $(a + b)/2$ gleich null und wächst dann zu den Intervallenden hin monoton, da f_r selbst dort monoton fällt. Tatsächlich sieht man leicht, daß f_r monoton wachsend sowohl in $t - a$ als auch in $b - t$ ist; da beide Ausdrücke im Intervall $[a + \delta, b - \delta]$ durch δ nach unten beschränkt sind, ist $f_r(t)$ in diesem Intervall daher überall mindestens gleich

$$e^{\frac{4r}{(b-a)^2}} e^{\frac{-r}{\delta(1-\delta)}} = e^{-r\left(\frac{1}{\delta(1-\delta)} - \frac{4}{(b-a)^2}\right)}.$$

Für alle $t \in [a + \delta, b - \delta]$ ist daher

$$f(t) - f_r(t) \le 1 - e^{-r\left(\frac{1}{\delta(1-\delta)} - \frac{4}{(b-a)^2}\right)}.$$

Dieser Ausdruck ist noch nicht sehr angenehm; wir wollen ihn weiter abschätzen. Nach Konstruktion von f_r ist der Exponent negativ, und für alle $x \ge 0$ ist $1 - e^{-x} \le x$, denn dies gilt für $x = 0$, und die Ableitung

e^{-x} von $1 - e^{-x}$ ist für jedes positive x kleiner als die Ableitung eins von x. Daher ist für $t \in [a + \delta, b - \delta]$

$$\begin{aligned} f(t) - f_r(t) &\le 1 - e^{-r\left(\frac{1}{\delta(1-\delta)} - \frac{4}{(b-a)^2}\right)} \\ &\le r\left(\frac{1}{\delta(1-\delta)} - \frac{4}{(b-a)^2}\right). \end{aligned}$$

Wir interessieren uns vor allem für kleine Werte von δ; deshalb betrachten wir im folgenden nur noch Werte $\delta \le \frac{1}{2}$. Dann ist $1 - \delta \ge \frac{1}{2}$ und

$$f(t) - f_r(t) \le r\left(\frac{2}{\delta} - \frac{4}{(b-a)^2}\right).$$

Damit können wir die L^2-Norm der Differenz abschätzen:

$$\begin{aligned} \|f - f_r\|_2^2 &= \int_{-\infty}^{\infty} |f(t) - f_r(t)|^2 \, dt \\ &= \int_a^{a+\delta} |f(t) - f_r(t)|^2 \, dt + \int_{a+\delta}^{b-\delta} |f(t) - f_r(t)|^2 \, dt \\ &\quad + \int_{b-\delta}^{b} |f(t) - f_r(t)|^2 \, dt \\ &\le \delta + r^2(b-a)\left(\frac{2}{\delta} - \frac{4}{(b-a)^2}\right)^2 + \delta. \end{aligned}$$

Setzen wir hier speziell $\delta = \sqrt{r}$, was wir für hinreichend kleine r dürfen, so wird dies zu

$$\begin{aligned} &\sqrt{r} + r^2(b-a)\left(\frac{2}{\sqrt{r}} - \frac{4}{(b-a)^2}\right)^2 + \sqrt{r} \\ &= 2\sqrt{r} + r(b-a)\left(2 - \frac{4\sqrt{r}}{(b-a)^2}\right)^2, \end{aligned}$$

und dieser Ausdruck geht gegen null für $r \to 0$. Also konvergieren die f_r für $r \to \infty$ in der L^2-Norm gegen f.

Der Vollständigkeit halber wollen wir uns noch überlegen, daß auch die Fläche zwischen den Graphen von f_r und von f für $r \to \infty$ gegen null geht: Wenn wir wie eben vorgehen, erhalten wir die Ungleichung

$$\int_{-\infty}^{\infty} |f(t) - f_r(t)|\, dt \leq \delta + r(b-a)\left(\frac{2}{\delta} - \frac{4}{(b-a)^2}\right) + \delta\,,$$

und wenn wir hier wieder spezialisieren auf $\delta = \sqrt{r}$ wird dies zu

$$\sqrt{r}\left(2 + (b-a)\left(2 - \frac{4\sqrt{r}}{(b-a)^2}\right)\right)\,,$$

was für $r \to \infty$ gegen null geht.

Da f sowie sämtliche f_r außerhalb des Intervalls $[a, b]$ verschwinden, geht damit auch

$$\int_{-\infty}^{\infty} |f(t) - f_r(t)|\, dt$$

für $r \to \infty$ gegen null. Dieses Integral bezeichnet man als die L^1-Norm von $f - f_r$; die Folge der Funktionen f_r konvergiert also auch in der L^1-Norm gegen f.

Diese Annäherung des Rechteckimpulses durch stark abfallende Funktionen wollen wir im nächsten Lemma auf beliebige quadratintegrierbare Funktionen ausdehnen:

Lemma: Zu jeder Funktion $f \in L^2(\mathbb{R}, \mathbb{C})$ gibt es eine Folge von Funktionen $\varphi_n \in \mathcal{S}(\mathbb{R})$, so daß

$$\lim_{n\to\infty} \|f - \varphi_n\|_2 = 0$$

ist; f läßt sich also bezüglich der L^2-Norm beliebig gut durch stark abfallende Funktionen annähern.

Beweis: In einem ersten Schritt sollten wir uns überlegen, daß f bezüglich der L^2-Norm als Grenzwert einer Folge von Treppenfunktionen τ_n mit jeweils nur endlich vielen Sprungstellen dargestellt werden kann.

Da f nach Voraussetzung integrierbar ist, können wir die Funktion zumindest auf jedem endlichen Intervall durch solche Treppenfunktionen annähern, und indem wir die Intervallgrenzen gegen unendlich gehen lassen, gilt dasselbe für ganz f. Der Beweis, daß wir so eine Folge von Treppenfunktionen bekommen, die *bezüglich der* L^2*-Norm* gegen f konvergiert ist ziemlich technisch und muß die ganze Konstruktion des RIEMANN-Integrals nachvollziehen; wir wollen daher auf die Einzelheiten verzichten und obige Aussage einfach glauben.

Jede der Treppenfunktionen τ_n ist eine Summe von endlich vielen Rechteckimpulsen R_{ni}, von denen wiederum jeder als Grenzwert einer Folge $(\psi_{nij})_{j\in\mathbb{N}}$ stark abfallender Funktionen geschrieben werden kann. Mit

$$\varphi_{nj} = \sum_i \psi_{nij}$$

ist dann auch

$$\lim_{j\to\infty} \varphi_{nj} = \lim_{j\to\infty} \sum_i \psi_{nij} = \sum_i R_{ni} = \tau_n\,,$$

denn die Summen über i sind endlich. Genau deshalb liegen die Funktionen φ_{ni} auch in $\mathcal{S}(\mathbb{R})$, und damit ist

$$\lim_{n\to\infty} \varphi_{nn} = \lim_{n\to\infty} \tau_n = f\,,$$

wie behauptet. ∎

Korollar: Zu jeder Distribution $T\colon \mathcal{S}(\mathbb{R}) \to \mathbb{C}$ gibt es eine stetige lineare Abbildung $\widetilde{T}\colon L^2(\mathbb{R}, \mathbb{C}) \to \mathbb{C}$, die auf $\mathcal{S}(\mathbb{R})$ mit T übereinstimmt.

Beweis: Jede Funktion $f \in L^2(\mathbb{R}, \mathbb{C})$ läßt sich als Limes einer Folge φ_n stark abfallender Funktionen schreiben; wir setzen einfach

$$\widetilde{T}(f) = \lim_{n\to\infty} T(\varphi_n)\,.$$

∎

Damit haben wir alle Vorbereitungen zusammen und können endlich beweisen, worauf es wirklich ankommt:

Satz von Plancherel: *a)* Zu jeder Funktion $f \in L^2(\mathbb{R}, \mathbb{C})$ gibt es Funktionen $\widehat{f}$ und $\check{f}$ in $L^2(\mathbb{R}, \mathbb{C})$, so daß

$$\widehat{T}_f = T_{\widehat{f}} \quad \text{und} \quad \check{T}_f = T_{\check{f}}$$

ist; FOURIER-Transformierte und inverse FOURIER-Transformierte von f existieren also als Funktionen.
b) Falls die rechten Seiten existieren, kann man

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t}\, dt \quad \text{und} \quad \check{f}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\omega) e^{i\omega t}\, d\omega$$

setzen.
c) $\check{\widehat{f}}$ und f unterscheiden sich höchstens durch eine Nullfunktion.
d) $\left\| \widehat{f} \right\|_2 = \sqrt{2\pi}\, \|f\|_2$ und $\left\| \check{f} \right\|_2 = \frac{1}{\sqrt{2\pi}}\, \|f\|_2$.

Beweis: Nach der CAUCHY-SCHWARZschen Ungleichung ist für jede stark abfallende Funktion $\varphi \in \mathcal{S}(\mathbb{R})$

$$\left| \widehat{T}_f(\varphi) \right| = \left| T_f\left(\widehat{\varphi}\right) \right| = \left| \left(f, \overline{\widehat{\varphi}}\right) \right| \le \|f\|_2\, \|\widehat{\varphi}\|_2 \ .$$

Wie wir aus §7*c)* wissen, ist $\|\widehat{\varphi}\|_2 = \sqrt{2\pi}\, \|\varphi\|_2$, also

$$\left| \widehat{T}_f(\varphi) \right| \le \sqrt{2\pi}\, \|f\|_2\, \|\varphi\|_2 \ .$$

Nach dem Satz von RIESZ gibt es daher eine Funktion $\widehat{f} \in L^2(\mathbb{R}, \mathbb{C})$, für die

$$\widehat{T}_f = T_{\widehat{f}}$$

ist. Falls das FOURIER-Integral existiert, wissen wir aus der Diskussion zu Beginn von §8*c)*, daß wir für $\widehat{f}$ die so definierte Funktion nehmen können.

In jedem Fall ist die Norm von $\widehat{f}$ gleich der von $\widehat{T}_f$, also ist nach obiger Abschätzung

$$\left\| \widehat{f} \right\|_2 \le \sqrt{2\pi}\, \|f\|_2 \ .$$

Genauso zeigt man auch die Existenz von $\check{f}$ und daß gilt

$$\left\| \check{f} \right\|_2 \le \frac{1}{\sqrt{2\pi}}\, \|f\|_2 \ .$$

Da die Distributionen $\check{\widehat{T}}$ und T übereinstimmen, unterscheiden sich $\check{\widehat{f}}$ und f höchstens durch eine Nullfunktion, haben also insbesondere dieselbe Norm. Nach den bislang bewiesenen Ungleichungen ist

$$\left\| \check{\widehat{f}} \right\|_2 \le \frac{1}{\sqrt{2\pi}} \left\| \widehat{f} \right\|_2 \le \|f\|_2 \ ;$$

da links und rechts dieselbe Zahl steht, muß in beiden Ungleichungen das Gleichheitszeichen gelten, und der Satz ist bewiesen. ∎

MICHEL PLANCHEREL (1885–1967) war Professor für höhere Mathematik an der Eidgenössischen Technischen Hochschule Zürich, publizierte seine Arbeiten aber in französischer Sprache. Diese befassen sich nicht nur mit der der FOURIER-Theorie einer und mehrerer Verränderlicher, sondern enthalten beispielsweise auch wichtige Sätze aus der sogenannten Ergodentheorie, der allgemeinen Theorie dynamischer Systeme. Seine letzte, 1962 erschienene Arbeit, befaßt sich mit dem Einfluß der Steuergesetze auf die Stabilität einer Volkswirtschaft. Den obigen Satz bewies er 1910; oft wird auch nur dessen letzte Aussage als PLANCHEREL-Formel bezeichnet.

Der gerade bewiesene Satz sagt uns also, daß die FOURIER-Transformation auch auf $L^2(\mathbb{R}, \mathbb{C})$ zumindest bis auf Nullfunktionen wohldefiniert ist, was für die meisten Zwecke genügt. Außerdem gibt er uns eine Aussage über die Normen, die dem Satz von PARSEVAL aus der Theorie der FOURIER-Reihen periodischer Funktionen entspricht, und die Aussage, daß FOURIER-Transformation und inverse FOURIER-Transformation zumindest bis auf Nullfunktionen tatsächlich invers zueinander sind.

Gelegentlich wollen wir aber die FOURIER-Transformation an einer bestimmten Stelle wirklich kennen, und dazu ist der obige Satz zu schwach: Da die Distribution T_f die Funktion F nur bis auf Nullfunktionen eindeutig bestimmt, legt T_f für kein einziges Argument t den Wert $f(t)$ wirklich fest.

Im Rest dieses Abschnitts wollen wir uns überlegen, daß auch der Funktionswert von f an allen Stetigkeitsstellen von f durch T_f eindeutig bestimmt ist.

Wir gehen dazu aus von zwei stückweise stetige Funktionen f und g mit $T_f = T_g$ ist. Für jede stark abfallende Funktion φ ist dann

$$T_f(\varphi) = T_g(\varphi) \quad \text{oder} \quad \int_{-\infty}^{\infty} f(t)\varphi(t)\,dt = \int_{-\infty}^{\infty} g(t)\varphi(t)\,dt\,.$$

Dies wollen wir anwenden auf die zu Beginn dieses Abschnitt betrachteten Funktionen und dort einfach mit f_r bezeichneten Funktionen

$$\varphi_{a,b,r}\colon \begin{cases} \mathbb{R} \to \mathbb{R} \\ t \mapsto \begin{cases} e^{\frac{4r}{(b-a)^2}} e^{\frac{-r}{(t-a)(b-t)}} & \text{falls } a < t < b \\ 0 & \text{sonst} \end{cases} \end{cases},$$

von denen wir dort gezeigt hatten, daß sie für feste Werte von a, b und ein variables $r > 0$ für $r \to 0$ gegen den Rechteckimpuls

$$R_{a,b}\colon \begin{cases} \mathbb{R} \to \mathbb{C} \\ t \mapsto \begin{cases} 1 & \text{falls } a \le t \le b \\ 0 & \text{sonst} \end{cases} \end{cases}$$

konvergieren. Diese Konvergenz haben wir sowohl bezüglich der L^2-Norm als auch bezüglich der L^1-Norm nachgerechnet. Wegen letzterer können wir aus den Gleichungen

$$T_f(\varphi_{a,b;r}) = T_g(\varphi_{a,b;r})$$

oder

$$\int_{-\infty}^{\infty} f(t)R_{a,b}(t)\,dt = \int_{-\infty}^{\infty} g(t)R_{a,b}(t)\,dt$$

schließen, daß auch

$$\int_a^b f(t)\,dt = \int_a^b g(t)\,dt \quad \text{für alle } a, b\,.$$

Als integrierbare Funktionen haben f und g Stammfunktionen F und G; damit ausgedrückt ist

$$F(b) - F(a) = G(b) - G(a) \quad \text{für alle } a, b \in \mathbb{R}\,.$$

(Strenggenommen haben wir das nur gezeigt für $b > a$, aber im Falle $b < a$ können wir einfach obige Überlegung für das Intervall $[b, a]$ wiederholen.) Setzen wir $b = a + h$, so gilt daher auch

$$F(a + h) - F(a) = G(a + h) - G(a)$$

und

$$\frac{F(a + h) - F(a)}{h} = \frac{G(a + h) - G(a)}{h}$$

für alle $a, h \in \mathbb{R}$ mit $h \neq 0$.

Lassen wir in dieser Gleichung h gegen null gehen, erhalten wir, sofern F *bzw.* G im Punkt a differenzierbar ist, den Wert der jeweiligen Ableitung im Punkt a.

Falls die Funktionen f und g in der Umgebung eines Punkts stetig sind, habe sie dort differenzierbare Stammfunktionen und sind gleich deren Ableitungen; damit ist

$$f(t) = g(t) \quad \text{falls } f \text{ und } g \text{ im Punkt } t \text{ stetig sind.}$$

Die Funktionen f und g können sich also höchstens an ihren Unstetigkeitsstellen unterscheiden.

Sind f und g sogar stetig, ist also $f = g$, und das gilt auch, wenn sowohl f als auch g nur stückweise stetig sind und zusätzlich die in §4*e)* betrachtete Mittelwerteigenschaft

$$f(t) = \frac{f(t^+) + f(t^-)}{2} \quad \text{und} \quad g(t) = \frac{g(t^+) + g(t^-)}{2}$$

erfüllen, denn die links- und rechtsseitigen Grenzwerte hängen nur von den Werten ab, die f und g Funktionen an den Stellen annehmen, an denen sie stetig sind.

Das wird uns in den meisten Fällen reichen, insbesondere wenn wir uns auf absolut integrierbare Funktionen beschränken:

Lemma: Ist die Funktion $f \in L^2(\mathbb{R}, \mathbb{C})$ absolut integrierbar, so existiert die FOURIER-Transformierte von f als Funktion; diese Funktion ist stetig und beschränkt.

Beweis: Nach Definition ist

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\,dt\,.$$

Der Betrag des Integranden ist $|f(t)|$; da f absolut integrierbar ist, konvergiert das Integral absolut und ist damit insbesondere konvergent.

Der Integrand $f(t)e^{-i\omega t}$ ist als Funktion von ω für jeden Wert von t stetig und als Funktion von t immerhin noch stückweise stetig. Daher zeigt das Lemma aus §6*a)* zunächst, daß für Intervalle $[a, b]$, in denen f stetig ist, auch

$$\int_a^b f(t)e^{-i\omega t}\,dt$$

eine stetige Funktion von ω ist. Damit gilt dasselbe für jedes endliche Intervall, denn endliche Summen stetiger Funktionen sind wieder stetig. Für $a \to -\infty$ und $b \to \infty$ schließlich konvergiert das Integral nach Voraussetzung absolut, also auch gleichmäßig, und damit ist auch die Grenzfunktion $\widehat{f}(\omega)$ stetig und beschränkt. ∎

Damit folgt insbesondere der z.B. für die Identifikation von Lösungen von Differentialgleichungen wichtige

Satz: $f, g \in \mathrm{L}^2(\mathbb{R}, \mathbb{C})$ seien stetige Funktionen.
a) Falls die FOURIER-Transformierten $\widehat{f}$ und $\widehat{g}$ übereinstimmen, ist $f = g$.
b) Falls es ein $r > 0$ gibt, so daß $\mathcal{L}\{f(t)\}(s) = \mathcal{L}\{g(t)\}(s)$ für alle $s \in \mathbb{C}$ mit $\mathfrak{Re}\, s = r$, ist $f(t) = g(t)$ für alle $t > 0$.

Beweis: a) folgt unmittelbar aus dem gerade bewiesenen Lemma, und *b)* folgt daraus, daß man dieses Lemma auf die Funktionen

$$f_r\colon \begin{cases} \mathbb{R} \to \mathbb{C} \\ t \mapsto \begin{cases} 0 & \text{falls } t < 0 \\ f(t)e^{-rt} & \text{sonst} \end{cases} \end{cases}$$

und

$$g_r\colon \begin{cases} \mathbb{R} \to \mathbb{C} \\ t \mapsto \begin{cases} 0 & \text{falls } t < 0 \\ g(t)e^{-rt} & \text{sonst} \end{cases} \end{cases}$$

anwendet, die zumindest für $t > 0$ stetig sind, und deren FOURIER-Transformationen gerade die LAPLACE-Transformationen von f und g für $\mathfrak{Re}\, s = r$ sind. ∎

f) Ableitungen von Distributionen

Wie wir in §6*b)* gesehen haben, ist für alle mindestens r-fach stetig differenzierbare Funktionen f, sofern alle vorkommenden FOURIER-Transformierten existieren,

$$\frac{d^r}{d\omega^r}\widehat{f}(\omega) = (-i)^r\widehat{t^r f}(\omega)$$

und

$$\omega^r\,\widehat{f}(\omega) = (-i)^r\widehat{f^{(r)}}(\omega)\,;$$

eine ähnliche, leicht komplexere Formel gilt auch für die LAPLACE-Transformation. Dies hatten wir im weiteren Verlauf von §6 zur Lösung erster Differentialgleichungen verwendet.

Inzwischen können wir die Voraussetzungen etwas präziser formulieren; insbesondere ist klar, daß diese Formeln für alle stark abfallenden Funktionen und alle $r \in \mathbb{N}$ gelten. Auch wissen wir, daß sie für quadratintegrierbare Funktionen gelten, *falls* auch alle Ableitungen bis zur jeweils betrachteten quadratintegrierbar sind.

In diesem Paragraphen wollen wir uns überlegen, wie man diesen Formeln auch für *beliebige* quadratintegrierbare Funktionen mit Hilfe von Distributionen zumindest bis auf Nullfunktionen einen Sinn geben kann.

Dazu überlegen wir uns zunächst, was Ableitungen auf dem Niveau der Distributionen bedeuten, wie man also beispielsweise eine Ableitung der DIRACschen δ-Distribution definieren kann. Es ist klar, daß ein Ansatz wie

$$\dot{\delta}(t) = \lim_{h\to 0}\frac{\delta(t+h) - \delta(t)}{h}$$

zu keinem vernünftigen Ergebnis führen kann; wir müssen unserer alten Strategie folgen und für eine differenzierbare Funktion f die Distribution $T_{\dot f}$ ausrechnen in der Hoffnung, daß dies zu einer Formel führt, die sich auf beliebige Distributionen verallgemeinern läßt.

Für eine differenzierbare Funktion f mit der Eigenschaft, daß sowohl f als auch die Ableitung $\dot f$ höchstens polynomiales Wachstum haben, ist

$$T_{\dot f}(\varphi) = \int_{-\infty}^{\infty} \dot f(t)\varphi(t)\,dt\,.$$

definiert und nach der Regel für partielle Integration ist

$$T_{\dot f}(\varphi) = f(t)\varphi(t)\,\Big|_{-\infty}^{\infty} - \int_{\infty}^{\infty} f(t)\dot\varphi(t)\,dt\,.$$

Da f höchstens polynomiales Wachstum hat, ist $|f(t)|$ kleiner oder gleich einem Ausdruck der Form $c\,|t|^k$ für eine reelle Zahl $c > 0$ und eine natürliche Zahl k. Da außerdem φ eine stark abfallende Funktion ist, bleibt $\left|t^{k+1}\varphi(t)\right|$ beschränkt für alle t, d.h.

$$|\varphi(t)| \le \frac{M}{|t|^{k+1}}$$

für eine reelle Zahl $M > 0$. Damit ist

$$|f(t)\varphi(t)| \le \left|ct^k \cdot \frac{M}{t^{k+1}}\right| \le \frac{cM}{|t|}\,.$$

Somit geht das Produkt $\varphi(t)f(t)$ gegen Null für $t \to \pm\infty$ und

$$T_{\dot f}(\varphi) = -\int_{\infty}^{\infty} f(t)\dot\varphi(t)\,dt = -T_f(\dot\varphi)\,.$$

Damit ist klar, wie wir die Ableitung einer Distribution definieren:

Definition: Die Ableitung einer Distribution $T\colon \mathcal{S}(\mathbb{R}) \to \mathbb{C}$ ist die Distribution

$$\dot T\colon \begin{cases} \mathcal{S}(\mathbb{R}) \to \mathbb{C} \\ \quad \varphi \mapsto -T_f(\dot\varphi) \end{cases},$$

die n-te Ableitung entsprechend

$$T^{(n)}\colon \begin{cases} \mathcal{S}(\mathbb{R}) \to \mathbb{C} \\ \quad \varphi \mapsto (-1)^n T_f\left(\varphi^{(n)}\right) \end{cases}.$$

Zum Nachweis, daß $\dot T$ und allgemeiner auch $T^{(n)}$ Distributionen sind, müssen wir zeigen, daß dies lineare Abbildungen sind – angesichts der Linearität der Differentiation ist das klar. Zum Nachweis der Stetigkeit aber müssen wir wissen, daß für eine konvergente Folge von Funktionen aus dem SCHWARTZ-Raum auch die Folge der abgeleiteten Funktionen konvergiert; dies gilt nur deshalb, weil wir die Konvergenz im SCHWARTZ-Raum so definiert haben, daß auch alle Ableitungen und deren Produkte mit t-Potenzen konvergieren müssen.

Beispielsweise ist also für die DIRAC-Distribution

$$\Delta_a^{(n)}(\varphi) = (-1)^n\varphi^{(n)}(a)$$

oder, mit der δ-„Funktion“ ausgedrückt

$$\int_{-\infty}^{\infty} \delta^{(n)}(t-a)\,\varphi(t)\,dt = (-1)^n\varphi^{(n)}(a)\,.$$

Auch Sprungfunktionen wie

$$\vartheta\colon \begin{cases} \mathbb{R} \to \mathbb{C} \\ \quad t \mapsto \begin{cases} 0 & \text{für } t < 0 \\ 1 & \text{für } t \ge 0 \end{cases} \end{cases}$$

lassen sich in der schönen neuen Welt der Distributionen problemlos differenzieren:

$$T_\vartheta(\varphi) = \int_{-\infty}^{\infty} \vartheta(t)\varphi(t)\,dt = \int_0^{\infty} \varphi(t)\,dt$$

hat als Ableitung die Distribution $\dot T_\vartheta$ mit

$$\dot T_\vartheta(\varphi) = -\int_{-\infty}^{\infty} \vartheta(t)\dot\varphi(t)\,dt = -\int_0^{\infty} \dot\varphi(t)\,dt = -\varphi(t)\,\Big|_0^{\infty} = \varphi(0)\,,$$

da $\varphi(t)$ bei einer stark abfallenden Funktion für $t \to \infty$ gegen null geht. Diese Distribution kennen wir aber: Es ist gerade die DIRAC-Distribution Δ_0. Also ist

$$\dot{T}_\vartheta = \Delta_0 \,,$$

was sich in Funktionen ausgedrückt (mit aller gebotenen Vorsicht) auch als

$$\dot{\vartheta}(t) = \delta(t)$$

schreiben läßt. Entsprechend lassen sich im Distributionensinne auch andere Sprungfunktionen differenzieren; die Ableitung an einer Sprungstelle $t = t_0$ ist jeweils Sprunghöhe mal $\delta(t - t_0)$.

Auch mit der Ableitung der Betragsfunktion haben wir auf Distributionenniveau keine Probleme: Für $f(t) = |t|$ zeigt partielle Integration, daß

$$\begin{aligned}
\dot{T}_f(\varphi) &= -\int_{-\infty}^{\infty} |t| \cdot \varphi(t)\,dt = \int_{-\infty}^{0} t \cdot \varphi(t)\,dt - \int_{0}^{\infty} t \cdot \varphi(t)\,dt \\
&= t\varphi(t)\,\Big|_{-\infty}^{0} - \int_{-\infty}^{0} \varphi(t)\,dt - t\varphi(t)\,\Big|_{0}^{\infty} + \int_{0}^{\infty} \varphi(t)\,dt \\
&= -\int_{-\infty}^{0} \varphi(t)\,dt + \int_{0}^{\infty} \varphi(t)\,dt = T_g(\varphi)
\end{aligned}$$

mit

$$g(t) = \begin{cases} -1 & \text{für } t < 0 \\ 1 & \text{für } t > 0 \end{cases} .$$

An der Stelle $t = 0$ können wir einen beliebigen Funktionswert wählen, denn T_g hängt nicht von diesem Wert ab. Wir bekommen also für $t \neq 0$, wo $f(t) = |t|$ differenzierbar ist, die erwarteten Ergebnisse, und für $t = 0$ keine Aussage. Nichtsdestoweniger ist die *Distribution* $\dot{T}_f$ wohldefiniert. Auf dem Niveau der Distributionen sind Ableitungen also auch für nur stückweise differenzierbare Funktionen problemlos.

Das Produkt einer Distribution mit einer beliebig oft stetig differenzierbaren Funktion mit höchstens polynomialem Wachstum haben wir

bereits definiert; das können wir insbesondere anwenden auf die Funktion

$$\Pi_r \colon \mathbb{R} \to \mathbb{C}; \quad t \mapsto t^r \,.$$

Wir erwarten

Lemma: Für jede Distribution $T \colon \mathcal{S}(\mathbb{R}) \to \mathbb{C}$ und jede natürliche Zahl r ist

$$\widehat{T}^{(r)} = (-i)^r \widehat{\Pi_r T} \quad \text{und} \quad \Pi_r \widehat{T} = (-i)^r \widehat{T^{(r)}} \,.$$

Beweis: Für $\varphi \in \mathcal{S}(\mathbb{R})$ gilt nach den entsprechenden Formeln für stark abfallende Funktionen aus §6*b)*

$$\begin{aligned}
\widehat{T}^{(r)}(\varphi) &= (-1)^r \widehat{T}(\varphi^{(r)}) = (-1)^r T\big(\widehat{\varphi^{(r)}}\big) = (-1)^r T\big(i^r \Pi_r \widehat{\varphi}\big) \\
&= (-i)^r T(\Pi_r \widehat{\varphi}) = (-i)^r \Pi_r T(\widehat{\varphi}) = (-1)^r \widehat{\Pi_r T}(\varphi) \,,
\end{aligned}$$

denn aus der Formel $\omega^r \widehat{\varphi}(\omega) = (-i)^r \widehat{\varphi^{(r)}}(\omega)$ folgt

$$\widehat{\varphi^{(r)}}(\omega) = i^r \omega^r \widehat{\varphi}(\omega) \,.$$

Entsprechend zeigt man auch die zweite Formel

$$\begin{aligned}
\Pi_r \widehat{T}(\varphi) &= \widehat{T}(\Pi_r \varphi) = T\big(\widehat{\Pi_r \varphi}\big) = T\big(i^r \widehat{\varphi}^{(r)}\big) \\
&= i^r T\big(\widehat{\varphi}^{(r)}\big) = (-i)^r T^{(r)}(\widehat{\varphi}) = (-i)^r \widehat{T^{(r)}} \,.
\end{aligned}$$

■

Rückübersetzt für Funktionen heißt das, daß die Formeln

$$\frac{d^r}{d\omega^r} \widehat{f}(\omega) = (-i)^r \widehat{t^r f}(\omega)$$

und

$$\omega^r \widehat{f}(\omega) = (-i)^r \widehat{f^{(r)}}(\omega) \,;$$

zumindest bis auf Nullfunktionen auch dann für quadratintegrierbare Funktionen gelten, wenn diese nur im Distributionensinn differenzierbar sind. Die entsprechende Formel für die LAPLACE-Transformation, die zusätzlich die Funktions- und Ableitungswerte an der Stelle Null enthält, ist natürlich (auch modulo Nullfunktionen) nur dann sinnvoll, wenn diese Werte wohldefiniert sind.

g) Faltungen

Bei der Untersuchung von FOURIER-Reihen in §4*a)* erwies sich die (periodische) Faltung zweier Funktionen als wichtiges Instrument zum Nachweis der Konvergenz; außerdem war sie oft nützlich, um ohne großen Aufwand neue FOURIER-Reihen aus bekannten herzuleiten.

Hier im nichtperiodischen Fall ist sie einfacher und anschaulicher zu verstehen als im periodischen Fall: $f(\star g)(t)$ ist einfach das gewichtete Mittel der Funktionswerte von f in der Umgebung von t, wobei g die Gewichtsfunktionen ist. Am einfachsten ist es, wenn man sich g als eine Funktion vorstellt, die im Punkt Null ein Maximum hat und dann nach beiden Seiten monoton abfällt; dann kann man sich $f \star g$ als eine „verschmierte“ (oder auch geglättete) Version von f vorstellen. Indem man für $g(t)$ GAUSSsche Glockenkurven nimmt, kann man beispielsweise unscharfe (oder weichgezeichnete) Photographien simulieren – je größer der Parameter σ, desto unschärfer ist das Resultat.

Für die formale Definition lassen wir allerdings beliebige Funktionen f und g zu; später werden wir sogar Faltungen von Funktionen mit Distributionen betrachten.

Definition: Für zwei Funktionen $f, g\colon \mathbb{R} \to \mathbb{C}$ heißt

$$f \star g\colon \begin{cases} \mathbb{R} \to \mathbb{C} \\ t \mapsto \displaystyle\int_{-\infty}^{\infty} f(t-s)g(s)\,ds \end{cases},$$

falls dieses Integral existiert, *Faltung* von f mit g.

Lemma: Für $f, g \in L^2(\mathbb{R}, \mathbb{C})$ existiert die Faltung $f \star g$.

Beweis: Mit f liegt für jedes $t \in \mathbb{R}$ auch die Funktion $s \mapsto f(t-s)$ in $L^2(\mathbb{R}, \mathbb{C})$; die Abschätzungen aus §8*a)* zeigen daher die Existenz des Integrals $f \star g(t)$. ∎

Ebenfalls in völliger Analogie zum periodischen Fall gilt

Lemma: Falls die FOURIER-Transformationen von f, g und von $h(t) = (f \star g)(t)$ als Funktionen existieren, ist $\widehat{h}(\omega) = \widehat{f}(\omega) \cdot \widehat{g}(\omega)$.

Beweis: Nach dem Satz von FUBINI ist

$$\begin{aligned}
\widehat{h}(\omega) &= \int_{-\infty}^{\infty} h(t)e^{-i\omega t}\,dt = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(t-s)\,g(s)\,ds \right) e^{-i\omega t}\,dt \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(t-s)\,g(s)e^{-i\omega t}\,dt \right) ds \\
&\underset{u=t-s}{=} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(u)\,g(s)e^{-i\omega(u+s)}\,du \right) ds \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(u)e^{-i\omega u}\,du \right) g(s)e^{-i\omega s}\,ds \\
&= \left(\int_{-\infty}^{\infty} f(u)e^{-i\omega u}\,du \right) \cdot \left(\int_{-\infty}^{\infty} g(s)e^{-i\omega s}\,ds \right) = \widehat{f}(\omega) \cdot \widehat{g}(\omega)\,,
\end{aligned}$$

wie behauptet. ∎

Als erste Anwendung hiervon können wir die FOURIER-Transformierte eines Produkts durch die FOURIER-Transformierten der Faktoren ausdrücken:

Korollar: $\widehat{fg}(\omega) = \dfrac{1}{2\pi}(\widehat{f} \star \widehat{g})(\omega)\,.$

Beweis: Wir wenden das gerade bewiesenen Lemma an auf die FOURIER-Transformierten von f und g; dann ist

$$\widehat{\widehat{f} \star \widehat{g}}(t) = \widehat{\widehat{f}}(t) \cdot \widehat{\widehat{g}}(t)\,.$$

Wie wir wissen, unterscheiden sich FOURIER-Transformation und inverse FOURIER-Transformation durch den Faktor $1/2\pi$ vor der inversen

Transformation und das Vorzeichen des Argument, d.h.

$$\widehat{\widehat{f}}(t) = 2\pi \cdot f(-t), \quad \widehat{\widehat{g}}(t) = 2\pi \cdot g(-t) \quad \text{und} \quad \widehat{\widehat{f \star g}}(t) = 4\pi^2 \cdot f(-t)g(-t)\,.$$

Aus dem gleichen Grund ist $\widehat{\widehat{f \star g}}(t) = 2\pi \cdot (\widehat{f} \star \widehat{g})(-t)$, also

$$2\pi \cdot (\widehat{f} \star \widehat{g})(-t) = 4\pi^2 \cdot \widehat{fg}(-t) \quad \text{oder} \quad \widehat{fg}(-t) = \frac{1}{2\pi}(\widehat{f} \star \widehat{g})(-t)\,.$$

Dies gilt für alle reellen Zahlen t, deshalb können wir das Minuszeichen links und rechts auch weglassen und haben dann die Behauptung des Korollars. ∎

Wie im periodischen Fall folgt auch, daß die Faltung (abgesehen von eventuell vorhandenen Unstetigkeitsstellen) kommutativ und assoziativ ist:

$$f \star g = g \star f \quad \text{und} \quad f \star (g \star h) = (f \star g) \star h\,,$$

denn für die FOURIER-Transformationen der beiden Seiten sind jeweils gleich nach dem Kommutativitätsgesetz und Assoziativitätsgesetz für die Multiplikation komplexer Zahlen.

Eine weitere interessante Konsequenz dieses Lemmas ist, daß sich Faltungen gelegentlich rückgängig machen lassen: $f \star g$ ist durch seine FOURIER-Transformation $\widehat{f} \cdot \widehat{g}$ (fast überall) bestimmt; falls $g(\omega)$ keine Nullstellen hat, kann man die Multiplikation mit $g(\omega)$ durch eine Division rückgängig machen. Eine Grundidee zum Rückgängigmachen der Faltung wäre also die folgende: Ist $h(t)$ die inverse FOURIER-Transformation von $1/\widehat{g}(\omega)$, so hat $(f \star g) \star h$ FOURIER-Transformierte

$$\widehat{f}(\omega) \cdot \widehat{g}(\omega) \cdot \widehat{h}(\omega) = \widehat{f}(\omega) \cdot \widehat{g}(\omega) \cdot \frac{1}{\widehat{g}(\omega)} = \widehat{f}(\omega)\,,$$

$(f \star g) \star h$ stimmt also fast überall mit f überein.

Leider ist die Sache aber doch nicht ganz so einfach, denn die Existenz von h ist alles andere als klar: Für eine stark abfallende Funktion $g(\omega)$ ist $1/g(\omega)$ „stark ansteigend“, und natürlich gibt es auch Probleme mit den Nullstellen von g. Die Mathematik kennt jedoch eine ganze Reihe von Regularisierungstechniken, mit denen man solche Probleme umgehen kann. Insbesondere kann man für praktische Zwecke sowohl

den Frequenzbereich, über den integriert wird, als auch den Zeit- oder Ortsbereich oft abschneiden, so daß nur ein Integral über ein endliches Intervall betrachtet werden muß.

Die Formel, die wir gerade benutzt haben, gelten, wenn man solche Techniken benutzt, natürlich nicht mehr exakt, aber doch oft mit einer Genauigkeit, die für praktische Zwecke völlig ausreicht. So konnte beispielsweise die NASA die Bilder des falsch fokusierten HUBBLE-Teleskops durch digitale Nachbehandlung so deutlich verbessern, daß die Bildqualität auch vor der Reperatur nicht viel schlechter war als bei einem korrekt fokusierten Teleskop.

Eine neuere Anwendung ist die sogenannte *brennpunktfreie Optik*, die von CMD Optics in Boulder, Colorado entwickelt wurde. Dort benutzt man eine (von Zeiss speziell zu diesem Zweck konstruierte) Linse ohne Brennpunkt; parallele einfallende Strahlen gehen also *nie* durch denselben Punkt der Bildebene, so daß grundsätzlich jedes Bild unscharf ist. Diese Unschärfe wird durch digitale Nachbearbeitung in der oben skizzierten Weise so gut es geht kompensiert.

Zweck dieser auf den ersten Blick unsinnigen Vorgehensweise ist die Erhöhung der Tiefenschärfe: Ein klassisches optisches System bildet, insbesondere wenn es mit wenig Licht auskommen muß und daher eine große Blende braucht, nur in einem sehr kleinen Entfernungsbereich scharf ab. Die brennpunktfreie Linse bildet natürlich überhaupt nirgends scharf ab, aber das Gesamtsystem aus Linse und digitaler Nachbearbeitung liefert scharfe Bilder aus einem deutlich größeren Entfernungsbereich als dies mit konventioneller Optik möglich ist.

Besonders einfach sind Faltungen mit δ-Funktionen zu berechnen: Für $\eta(t) = \delta(t - t_0)$ zeigt die Substitutionsregel mit $u = t - t_0 - s$, daß

$$\eta \star f = \int_{-\infty}^{\infty} \delta(t - t_0 - s)g(s)\,ds = \int_{-\infty}^{\infty} \delta(u)g(t - t_0 - u)\,du = g(t - t_0)$$

ist, Faltung mit $\delta(t - t_0)$ verschiebt also einfach das Argument um t_0. Insbesondere ist $\delta \star f = f$.

Im Falle einer Funktion, die außerhalb eines gewissen Intervalls null (oder praktisch null) ist, läßt sich durch Faltung mit einer Summe von

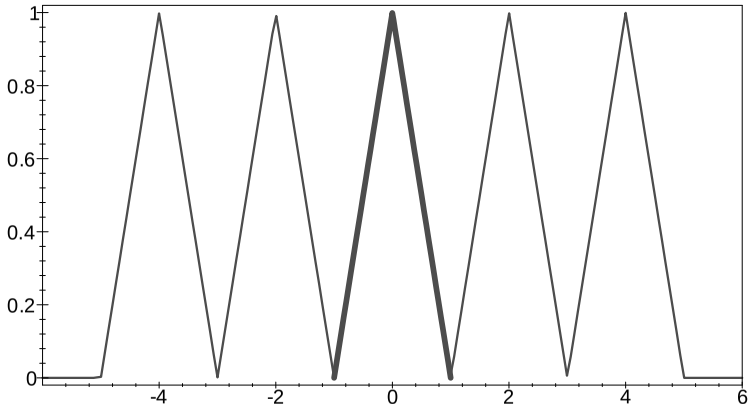


Abb. 25: Faltung eines Dreiecksimpuls mit einer Summe von δ-Distributionen

δ-Funktionen der Graph an verschiedene Stellen verschieben; Abbildung 25 zeigt dies für die Faltung eines (fett eingezeichneten) Dreiecksimpulses auf $[-1,\ 1]$ und die Distribution

$$\delta(t-4)+\delta(t-2)+\delta(t)+\delta(t+2)+\delta(t+4)\,.$$

h) Der Abtastsatz von Nyquist

Egal ob es um die automatische Erfassung von Meßwerten geht oder um die Aufzeichnung von Musik: Die digitale Darstellung analoger Daten ist wesentlicher Bestandteil der Informationsverarbeitung. Nun ist aber eine beliebige Funktion $f\colon \mathbb{R} \to \mathbb{R}$ sicherlich nicht durch ihre Funktionswerte an endlich vielen Stellen oder auch an ein einer diskreten Menge von Stellen bestimmt: Auch wenn wir wissen, daß $f(t) = 0$ ist für jedes ganzzahlige Vielfache von 0,001, wissen wir noch nicht, daß f die Nullfunktion ist: Auch die Funktionen $f(t) = \sin(1000\pi t)$ und $f(t) = -3\sin(5000\pi t)$ haben diese Eigenschaft. Auch bei von null verschiedenen Abtastwerten tritt dieses Problem auf: Beispielsweise stimmen auch die Funktionen $f(t) = \cos(500\pi t)$ und $g(t) = \cos(1500\pi t)$ für alle ganzzahligen Vielfachen von 0,001 überein, aber sie nehmen hier abwechselnd die Werte $1, 0, -1, 0$ an; siehe Abbildung 26. Da der Frequenzunterschied zwischen den beiden Schwingungen fast dem zwischen Baß und Sopran entspricht, ist klar, daß man die beiden Schwingungen zumindest auf einer Musik-CD nicht miteinander verwechseln darf.

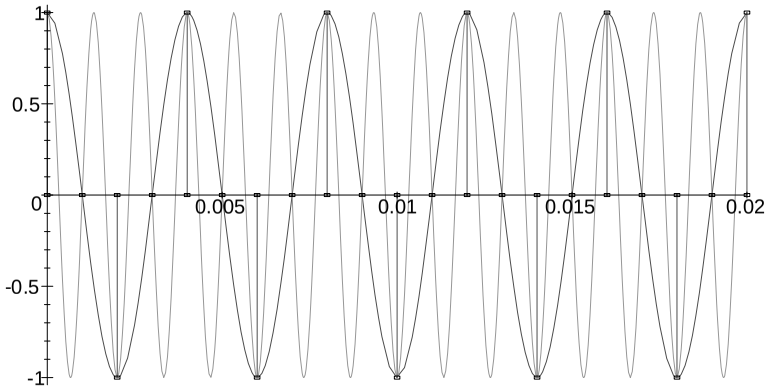


Abb. 26: Abtastung zweier Schwingungen

Die Probleme bei den obigen Beispielen beruhen offensichtlich darauf, daß es zu jedem gegebenen Signal auch höherfrequente Signale gibt, die an vorgegebenen Abtastpunkten mit ihm übereinstimmen; eine eindeutige Rekonstruktion ist höchstens dann möglich, wenn man eine Grenze festlegt, oberhalb derer Frequenzen nicht mehr berücksichtigt werden sollen. Der Abtastsatz von NYQUIST sagt, daß dann in der Tat eine Rekonstruktion möglich ist, und er sagt auch, wo die Grenze liegen soll, oberhalb derer man die Frequenzen abschneiden muß: Die Abtastfrequenz muß mehr als doppelt so hoch sein als die höchste im Signal vorkommende Frequenz.

Die genaue Formulierung des Satzes ist etwas technischer; insbesondere müssen wir berücksichtigen, daß die Kreisfrequenz ω, mit der wir immer arbeiten, etwas anderes ist, als die Frequenz: Eine reine Schwingung mit einer Frequenz von 1000 Hz ist nicht gegeben durch eine Funktion wie $\sin 1000t$, sondern – bei in Sekunden gemessener Zeit – durch $\sin 2000\pi t$. Entsprechend kommt auch jetzt bei der Formulierung des Abtastsatzes von NYQUIST ein Faktor 2π ins Spiel:

Satz: $f \in L^2(\mathbb{R}, \mathbb{C})$ habe die Eigenschaft, daß $\widehat{f}(\omega)$ außerhalb eines Intervalls der Länge Ω verschwinde. Dann ist f eindeutig bestimmt durch die Werte $f(2k\pi/\Omega)$ mit $k \in \mathbb{Z}$.

Beweis: (ω_1, ω_2) sei ein Intervall der Länge Ω derart, daß $\widehat{f}(\omega)$ außerhalb dieses Intervalls verschwindet. Dann ist bis auf eine Nullfunktion

$$f(t) = \check{\widehat{f}}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\omega) e^{i\omega t}\, d\omega = \frac{1}{2\pi} \int_{\omega_1}^{\omega_2} \widehat{f}(\omega) e^{i\omega t}\, d\omega \,.$$

Indem wir f durch die rechte Seite ersetzen (was nichts wesentliches ändert) können wir annehmen, daß diese Gleichung wirklich gilt. Also ist insbesondere

$$f\left(\frac{2k\pi}{\Omega}\right) = \frac{1}{2\pi} \int_{\omega_1}^{\omega_2} \widehat{f}(\omega) e^{2k\pi i\omega/\Omega}\, d\omega \,. \qquad (*)$$

Nun betrachten wir jene Funktion $g(\omega)$, die auf dem Intervall $[\omega_1, \omega_2)$ mit $\widehat{f}(\omega)$ übereinstimmt und die periodisch mit Periode Ω in ω auf $\mathbb{R}$ fortgesetzt ist. Für diese Funktion ist natürlich auch

$$f\left(\frac{2k\pi}{\Omega}\right) = \frac{1}{2\pi} \int_{\omega_1}^{\omega_2} g(\omega) e^{2k\pi i\omega/\Omega}\, d\omega \,,$$

denn im Integrationsintervall stimmen $\widehat{f}$ und g überein.

g als periodische Funktion in ω mit Periode Ω hat eine Darstellung als FOURIER-*Reihe*

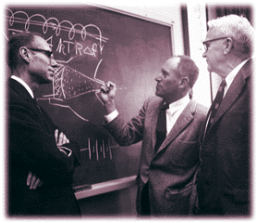
$$\sum_{k=-\infty}^{\infty} c_k e^{ik\lambda\omega} \quad \text{mit} \quad \lambda = \frac{2\pi}{\Omega} \,;$$

der k-te FOURIER-Koeffizient ist

$$c_k = \frac{1}{\Omega} \int_{\omega_1}^{\omega_2} g(\omega) e^{-ik\lambda\omega}\, d\omega = \frac{1}{\Omega} \int_{\omega_1}^{\omega_2} g(\omega) e^{-2\pi i\omega k/\Omega}\, d\omega$$
$$= \frac{2\pi}{\Omega} f\left(\frac{-2k\pi}{\Omega}\right) \,,$$

wobei das letzte Gleichheitszeichen wegen $(*)$ gilt.

Durch die Werte $f\left(\frac{2k\pi}{\Omega}\right)$ sind also alle FOURIER-Koeffizienten von g bestimmt, damit auch (fast überall) die Funktion $g(\omega)$, und damit auch die Funktion $\widehat{f}(\omega)$, die im Intervall $[\omega_1, \omega_2)$ mit $g(\omega)$ übereinstimmt und außerhalb (außer eventuell im Punkte ω_2) verschwindet. Damit ist auch $f(t) = \check{\widehat{f}}(t)$ fast überall durch diese Werte bestimmt. ∎



HARRY NYQUIST (1889–1976) wurde in Schweden geboren, arbeitete aber ab Anfang der zwanziger Jahre bei den Bell Labaratories; das Bild zeigt ihn um 1960 mit seinen dortigen Kollegen JOHN PIERCE *(links)* und RUDOLF KOMPFNER *(Mitte)*. Seine Arbeit von 1924 über die Übertragungsgeschwindigkeit von Telegraphen gilt als eine der Begründungen der Informationstheorie. Den Abtastsatz, den CAUCHY bereits 1841 postuliert hatte, bewies er 1928. Weitere wichtige Arbeiten befassen sich mit der quantitativen Erforschung des thermischen Rauschens und der Stabilität von Verstärkern.

Bei praktischen Anwendungen dieses Satzes wird $f(t)$ im allgemeinen eine reelle Funktion sein; dann verschwindet

$$\widehat{f}(-\omega) = \int_{-\infty}^{\infty} f(t) e^{-i(-\omega)t}\, dt = \int_{-\infty}^{\infty} f(t) e^{i\omega t}\, dt = \int_{-\infty}^{\infty} \overline{f(t) e^{-i\omega t}}\, dt$$
$$= \overline{\int_{-\infty}^{\infty} f(t) e^{-i\omega t}\, dt} = \overline{\widehat{f}(\omega)}$$

genau dann, wenn auch $\widehat{f}(\omega)$ verschwindet. Daher wird in diesem Fall alles einfacher, wenn man das Intervall, außerhalb dessen $\widehat{f}(\omega)$ verschwindet, symmetrisch zum Nullpunkt wählen kann, also von der Form $(-\omega_0, \omega_0)$. Die zur Kreisfrequenz $\omega_0 = 2\pi\nu_0$ gehörende Frequenz ν_0 wird in diesem Zusammenhang oft als *Bandbreite* bezeichnet. Hier ist $\Omega = 2\omega_0$, zur Rekonstruktion der Funktion f brauchen wir also die Funktionswerte

$$f\left(\frac{2k\pi}{\Omega}\right) = f\left(\frac{k\pi}{\omega_0}\right) \quad \text{mit} \quad k \in \mathbb{Z} \,.$$

Ein Signal der Bandbreite ν_0 muß also mit einer Frequenz von mindestens $2\nu_0$ abgetastet werden, damit man es eindeutig rekonstruieren kann.

Bekanntestes Beispiel hierfür sind Musik-CDs: Praktisch niemand kann Töne mit Frequenzen von mehr als 20kHz hören; für Aufnahmen auf CD wird 44 100 Mal pro Sekunde der Schalldruck gemessen und gespeichert, für Signale die nicht allzuweit oberhalb von 20kHz abgeschnitten werden, ist also eine perfekte Rekonstruktion möglich.

Auch in der Computergraphik spielt der Satz von NYQUIST eine wichtige Rolle, denn Pixelgraphik ist schließlich nichts anderes als die (zweidimensionale) diskrete Abtastung eines kontinuierlichen Bilds. Falls das Bild zu hochfrequente Anteile enthält, entstehen sogenannte *alias-Effekte*, da das Auge diese Anteile anhand des Pixelbilds als niedrigerfrequente Strukturen mit gleichen Abtastwerten interpretiert. Vor der Abtastung muß das Bild daher tiefpaßgefiltert werden; da die Funktion $\frac{\sin ax}{ax}$, die FOURIER-Transformierte des Rechteckimpulses, einigermaßen schnell abfällt, wendet man dazu meist das Lemma aus dem letzten Abschnitt an und faltet mit einer geeigneten solchen Funktion. Falls das Ursprungsbild auch schon als (höher aufgelöste) Pixelgraphik gegeben war, wird die Faltung hier einfach zu einer Summation über nicht garzu viele Nachbarpixel, was sehr effizient durchgeführt werden kann.

§9: Ausblick: Mehrdimensionale Fourier-Theorie

a) Faltungen und Fourier-Integrale

In völliger Analogie zum eindimensionalen Faltungsintegral läßt sich auch ein n-dimensionales definieren: Für zwei Funktionen $f, g\colon \mathbb{R}^n \to \mathbb{C}$ definieren wie die Faltung als

$$f \star g = \int_{\mathbb{R}^n} \cdots \int f(x_1 - y_1, \ldots, x_n - y_n) g(y_1, \ldots, y_n)\, dy_1 \ldots dy_n$$

– sofern dieses Integral existiert.

Auch die anschauliche Interpretation ist dieselbe wie im eindimensionalen Fall: Wenn wir f als eine Gewichtsfunktion auffassen, ist $f \star g$ ein gewichtetes Mittel über Werte von g; für

$$f(x_1, \ldots, x_n) = \frac{1}{\pi^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^n x_i^2}$$

etwa, die n-dimensionale GAUSS-Funktion, entspricht das im Fall $n = 2$ einem je nach Größe von σ mehr oder weniger defokusierten Bild.

Durch mehrdimensionale Faltungen mit δ-Distributionen lassen sich Verschiebungungen realisieren: Beispielsweise wäre, wenn der Satz von FUBINI in einer solchen Situation anwendbar wäre,

$$\iint_{\mathbb{R}^2} \delta(x-a)\delta(y-b) f(x,y)\, dx\, dy = f(x-a, y-b)\,,$$

und genau so *definieren* wir die Interpretation der *a priori* sinnlosen linken Seite.

(Man beachte, daß Ausdrücke wir $\delta(x-a)\delta(x-b)$ oder $\delta(x)^2$ weiterhin sinnlos bleiben, egal ob sie unter einem oder mehreren Integralzeichen stehen.)

Ist also $f\colon \mathbb{R}^2 \to \mathbb{R}$ eine Funktion, die (z.B. durch Grauwerte) ein Bild definiert und die außerhalb des Bereichs $0 \le x, y \le 1$ verschwindet, so ist mit der Distribution

$$\eta(x,y) = \sum_{k=1}^{N}\sum_{\ell=1}^{M} \delta(x-k)\delta(y-\ell)$$

die Faltung $\eta \star f$ ein Bilderbogen aus NM Exemplaren dieses Bildes. Abbildung 27 zeigt dies für den Graph einer zweidimensionalen Normalverteilung.

Auch die FOURIER-Transformation läßt sich in völliger Analogie zum eindimensionale Fall auf beliebige Dimensionen verallgemeinern: Für $f\colon \mathbb{R}^n \to \mathbb{C}$ definieren wir

$$\widehat{f}\colon \begin{cases} \mathbb{R}^n \to \mathbb{C} \\ (\omega_1, \ldots, \omega_n) \mapsto \displaystyle\int_{\mathbb{R}^n} \cdots \int f(x_1, \ldots, x_n) e^{-i\sum_{k=1}^n \omega_i x_i}\, dx_1 \cdots dx_n \end{cases}.$$

Da es nur eine Zeit gibt, läßt sich dies nicht als Zerlegung eines zeitlichen Signals in seine Frequenzen interpretieren; die x_i sollte man sich hier

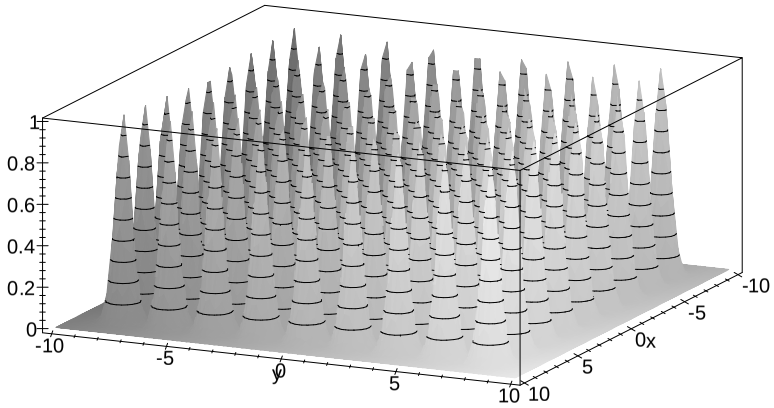


Abb. 27: Eine zweidimensionale Faltung

als *räumliche* Koordinaten vorstellen. Beispiele dazu folgen im nächsten Abschnitt, wo wir eine Anwendung solcher räumlicher FOURIER-Transformationen betrachten.

Wenigsten kurz sei noch angedeutet, wie man auch die mehrdimensionale FOURIER-Theorie über stark abfallende Funktionen mehrerer Veränderlicher exakt begründen kann:

Eine Funktion $\varphi\colon \mathbb{R}^n \to \mathbb{C}$ heißt *stark abfallend*, wenn *alle* Ausdrücke der Form

$$x_1^{e_1} \cdots x_n^{e_n} \frac{\partial^{r_1+\cdots+r_n}}{\partial x_1^{r_1} \cdots \partial x_n^{r_n}} \varphi(x_1, \ldots, x_n)$$

auf ganz $\mathbb{R}^n$ beschränkt sind. Der Vektorraum aller dieser Funktionen ist der SCHWARTZ-Raum $\mathcal{S}(\mathbb{R}^n)$.

Für Funktionen aus diesem Raum ist wieder alles relativ problemlos; zur Verallgemeinerungen auf interessantere Funktionen führt auch hier der Umweg über Distributionen $T\colon \mathcal{S}(\mathbb{R}^n) \to \mathbb{C}$, die in der naheliegenden Weise als Verallgemeinerungen eindimensionaler Distribibutionen definiert werden. Beispielsweise kann man dem gerade *ad hoc* betrachteten

Produkt $\delta(x-a)\delta(y-b)$ über die Distribution

$$\Delta_{a,b}\colon \begin{cases} \mathbb{R}^2 \to \mathbb{C} \\ \varphi \mapsto \varphi(a,b) \end{cases}$$

einen präzisen Sinn geben – solange es in einem sinnvollen Kontext unter zwei Integralzeichen steht.

b) Fraunhofer-Beugung

Wenn Licht auf Strukturen trifft, in Vergleich zu deren Größe seine Wellenlänge nicht mehr vernachlässigbar klein ist, lassen sich die Gesetze der geometrischen Optik bekanntlich nicht mehr anwenden; man beobachtet dann Beugungsphänomene.

Beugung ist ein sehr komplexes Gebiet; für ein Beispiel im Rahmen einer Vorlesung über Höhere Mathematik müssen wir uns auf den allereinfachsten Fall beschränken. Wir gehen daher aus von einem Lichtstrahl, der aus sehr großer Entfernung kommt oder der zumindest (z.B. dank einer Linse, aus deren Brennpunkt er kommt) so aussieht, und beobachten auch die Beugungsfigur in großer Entfernung. Diese Situation bezeichnet man als FRAUNHOFER-Beugung.

JOSEPH VON FRAUNHOFER (1787–1826) wurde in Straubing als elftes und letztes Kind eines Glasermeisters geboren; er machte auch selbst eine Lehre als Glasschleifer und Spiegelmacher. Daneben besuchte er die Feierabendschule, wo er zumindest primitive Grundkenntnisse im Rechnen erwarb. 1806 kam er an das optische Institut von UTZSCHNEIDER, der ihm Bücher über Optik und Mathematik besorgte. FRAUNHOFER entwickelte Präzisionsmaschinen zur Fertigung optischer Instrumente von bis dahin nicht gekannter Qualität und erfand auch das optische Gitter. Durch seine Versuche zur Lichtbeugung bewies er die Wellennatur des Lichts. 1824 wurde er zum Professor ernannt; er berichtete unter anderem in öffentlichen Sonntagsvorlesungen über seine Arbeit. Im gleichen Jahr wurde er vom bayrischen König LUDWIG I. in den Adelsstand erhoben. Zwei Jahre später starb er an Tuberkulose.

Zur mathematischen Behandlung der optischen Beugung brauchen wir zunächst ein physikalisches Modell für Lichtwellen. Für eine physikalisch korrekte Beschreibung müssen wir Licht als zeitlich veränderliches

elektromagnetisches Feld betrachten, d.h. wir brauchen zwei räumlich und zeitlich variable Vektorfelder $\vec{E}(x, y, z; t)$ und $\vec{B}(x, y, z; t)$, die den MAXWELLschen Gleichungen genügen. Glücklicherweise muß man in der Optik aber nur selten so weit gehen: Zwar hängt die Beugung an einem Spalt theoretisch durchaus von der Leitfähigkeit des verwendeten Materials ab, aber diese Abhängigkeit ist so gering, daß man sie für alle praktischen Zwecke vernachlässigen kann.

Man arbeitet daher in der Wellenoptik gerne mit einer sogenannten *skalaren Welle*, über deren physikalische Bedeutung man sich keine sonderlichen Gedanken macht. Aus rechnerischen Gründen betrachtet man sie als komplexwertige Funktion; falls man sich unbedingt etwas darunter vorstellen will, kann man beispielsweise den Realteil dieser Funktion als die x-Komponente des elektrischen Felds interpretieren, muß dann aber beachten, daß eine skalare Welle im Gegensatz zu einem elektrischen Feld *keine* Wechselwirkung mit Materie egal welcher Leitfähigkeit zeigt – das ist eine der Idealisierungen hinter dem Konzept der skalaren Welle. Wichtig für uns ist nur, daß die Intensität der Welle (also z.B. die Intensität der Beugungslinien, die wir auf einem Schirm beobachten) gleich dem Betragsquadrat der Wellenfunktion sein soll.

Eine Welle hat eine räumliche wie auch zeitliche Periodizität. Zeitlich periodische Vorgänge kennen wir bereits: Das sind Schwingungen, die mathematisch durch Funktionen der Art

$$f(t) = A_0 e^{i\omega t} \quad \text{oder etwas allgemeiner} \quad f(t) = A_0 e^{i(\omega t+\varphi)}$$

beschrieben werden, wobei die Phasenverschiebung φ dafür sorgt, daß wir auch Schwingungen behanden können, die ihre maximale Auslenkung nicht zur Zeit $t = 0$ erreichen. Wir wollen dies jedoch im folgenden ignorieren und mit der einfacheren ersten Funktion arbeiten.

Für räumlich periodische Vorgänge haben wir entsprechend zur Periode T einer Schwingung eine Wellenlänge λ; das Analogon zur Kreisfrequenz $\omega = 2\pi/T$ bezeichnen wir als

$$\textit{Wellenzahl} \quad k = \frac{2\pi}{\lambda} .$$

Ein eindimensionaler periodischer Vorgang kann somit beschrieben werden durch eine Funktion $g(x) = A_0 e^{ikx}$.

Im Mehrdimensionalen müssen wir die Wellenzahl k ersetzen durch einen Vektor $\vec{k}$ der Länge k, den *Wellenzahlvektor*, und betrachten die Funktion

$$g(\boldsymbol{x}) = A_0 e^{i\vec{k}\cdot\vec{x}} .$$

(Die Wellenlänge betrachten wir weiterhin nur als Skalar.)

Eine Welle soll zeitlich *und* räumlich periodisch sein; dies leistet die Funktion

$$\psi(\boldsymbol{x}, t) = A_0 e^{i(\omega t-\vec{k}\cdot\vec{x})}$$

oder natürlich auch die entsprechende Funktion mit einem Pluszeichen im Exponenten. Der Grund, warum wir das Minuszeichen bevorzugen, ist folgender:

Im eindimensionalen Fall ist

$$\psi(x, t)A_0 e^{i(\omega t-\vec{k}\cdot\vec{x})} = A_0 e^{ik\left(\frac{\omega}{k}t-x\right)} ,$$

$\psi(x, t)$ hängt also nur ab von $x - \frac{\omega}{k}t$. Dies können wir auch so interpretieren, daß

$$v = \frac{\omega}{k} = \frac{\lambda}{T} = \frac{\lambda\omega}{2\pi}$$

die Ausbreitungsgeschwindigkeit der Welle ist; denn eine Änderung der Zeit um Δt hat denselben Effekt wie eine Änderung des Orts um $v \cdot \Delta t$.

Im Falle mehrerer räumlicher Dimensionen ist alles grundsätzlich genauso, nur die Schreibweise ist etwas komplizierter: Ist $\vec{k}_0$ ein Einheitsvektor in Richtung von $\vec{k}$, d.h. $\vec{k} = k \cdot \vec{k}_0$, so ist

$$\psi(\boldsymbol{x}, t) = A_0 e^{ik\left(\frac{\omega}{k}t-\vec{k}_0\cdot\vec{x}\right)} ;$$

dabei ist $\vec{k}_0 \cdot \vec{x}$ die $\vec{x}$-Komponente in Richtung von $\vec{k}$. Damit ist $\vec{k}_0$ die *Richtung* des Geschwindigkeitsvektors; der Wellenzahlvektor zeigt also in Richtung der Ausbreitungsgeschwindigkeit der Welle, und der Betrag v des Geschwindigkeitsvektors ist durch obige Formel gegeben.

Die Annahme einer konstanten Amplitude A_0 in obigen Formeln ist nur in seltenen Fällen realistisch: Licht kommt meist aus einer (zumindest in erster Näherung) punktförmigen Lichtquelle, und seine Intensität nimmt

mit dem Quadrat der Entfernung ab. Da die Intensität das Betragsquadrat der Wellenfunktion sein soll, müssen wir eine solche Kugelwelle also in der Form

$$\psi(\boldsymbol{x}, t) = \frac{A_0}{|\vec{x}|} e^{i(\omega t - \vec{k}\cdot\vec{x})}$$

ansetzen, sofern die Lichtquelle im Nullpunkt des Koordinatensystems sitzt.

Im Falle einer weit entfernten Lichtquelle, wie wir sie bei der FRAUNHOFER-Beugung annehmen und auch von der Sonne her kennen, spielt allerdings die Ortsabhängigkeit der Amplitude praktisch keine Rolle, so daß wir keinen nennenswerten Fehler machen, wenn wir sie als konstant annehmen. In diesem Fall sprechen wir von einer *ebenen* Welle.

Ausgangspunkt für die Berechnung von Beugungsbilder ist das HUYGENSsche Prinzip: Jeder Punkt des Hindernisses ist Quelle einer Kugelwelle, deren Amplitude gleich der Amplitude der einfallenden Welle mal der Durchlässigkeitsfunktion α des Hindernisses im betrachteten Punkt ist. Letztere gibt an, welcher Teil des Lichts durchgelassen wird; sie ist also eins an den Stellen, an denen alles Licht durchkommt, und null dort, wo nichts durchkommt. An Stellen, an denen ein Teil des Lichts durchgelassen wird, kann sie auch Zwischenwerte annehmen.

CHRISTIAAN HUYGENS (1629–1695) kam aus einer niederländischen Diplomatenfamilie. Dadurch und später auch durch seine Arbeit hatte er Kontakte zu führenden europäischen Wissenschaftlern wie DESCARTES und PASCAL. Nach seinem Studium der Mathematik und Jura arbeitete er teilweise auch selbst als Diplomat, interessierte sich aber bald vor allem für Astronomie und den Bau der dazu notwendigen Instrumente. Er entwickelte eine neue Methode zum Schleifen von Linsen und erhielt ein Patent für die erste Pendeluhr. Trotz des französisch-niederländischen Kriegs arbeitete er einen großen Teil seines Lebens an der *Académie Royale des Sciences* in Paris, wo beispielsweise LEIBNIZ viel Mathematik bei ihm lernte. HUYGENS war ein scharfer Kritiker sowohl von NEWTONS Theorie des Lichts als auch seiner Gravitationstheorie, die er für absurd und nutzlos hielt. Gegen Ende seines Lebens beschäftigte er sich mit der Möglichkeit außerirdischen Lebens.

Bei der FRAUNHOFER-Beugung betrachten wir auch die gebeugten Wellen nur aus sehr großer Entfernung und können daher statt von Kugelwellen von ebenen Wellen ausgehen. Außerdem können wir die Zeitabhängigkeit der Welle ignorieren, denn die Frequenzen, mit denen das sichtbare Licht schwingt, liegen um Größenordnungen jenseits sowohl unserer Reaktionszeit als auch der unserer Meßinstrumente, so daß wir nur die Amplituden messen können. Schreiben wir die einfallende Welle als

$$\psi(\boldsymbol{x}, t) = A_0 e^{-i\vec{k}\vec{x}} \cdot e^{i\omega t},$$

ist also der zweite Faktor eine zeitabhängige Phasenvariation, die wir bei der Berechnung des räumlichen Intensitätsverteilung des Beugungsbilds ignorieren können; es reicht also, den Faktor $A_0 e^{-i\vec{k}\vec{x}}$ zu betrachten.

Eine weitere Konsequenz des (auch im Vergleich zur Größe des Hindernisses) weit entfernten Betrachtungspunkts ist, daß wir das Hindernis vom Schirm aus praktisch nur als Punkt sehen; was an einer gegebenen Stelle des Schirms ankommt, hängt also im wesentlichen nur ab vom Winkel θ oder (im Zweidimensionalen) den Winkeln θ und φ, unter dem (oder denen) die Strahlen von diesem „Punkt“ ausgehen.

Um die Intensität des Beugungsbilds in einem gegebenen Punkt zu berechnen, müssen wir also alle vom Hindernis in einem festen Winkel ausgehenden Strahlen aufsummieren, und *hierbei* müssen wir auch die Phasen berücksichtigen, da diese Strahlen miteinander interferieren. Abbildung 28 zeigt, wie sich die Laufwege zweier benachbarter Strahlen unterscheiden, und diese Differenzen können wir nicht vernachlässigen, da sie in der Größenordnung des Hindernisses und damit auch der Wellenlänge des Lichts liegen.

Betrachten wir zunächst den (in Abbildung 28 dargestellten) eindimensionalen Fall. Verglichen mit dem Strahl, der von einem (irgendwie gewählten) Nullpunkt des Hindernisses ausgeht, hat der Strahl mit Ausgangspunkt in Entfernung x einen Laufwegunterschied von $x \sin\theta$; dies entspricht einem Phasenfaktor von $e^{-ikx\sin\theta}$. Wählen wir also die Phase im Nullpunkt als Referenz (die wir in den zu ignorierenden Phasenfaktor der einfallenden Welle hineinziehen können), ist die Summe aller unter

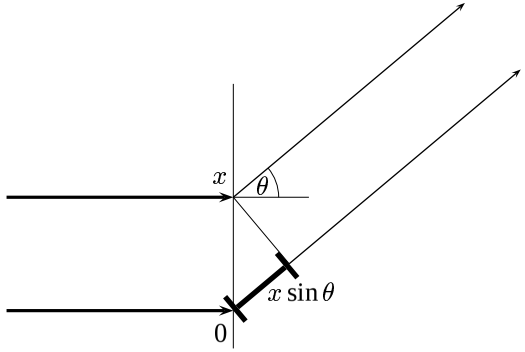


Abb. 28: Laufwegunterschied zweier paralleler Strahlen

dem Winkel θ abgehenden Strahlen gleich

$$\int_{-\infty}^{\infty} \alpha(x) e^{-ikx\sin\theta}\, dx\,;$$

das ist gleich der FOURIER-Transformierten von $\alpha(x)$, ausgewertet im Punkt $u = k\sin\theta$.

Bei einem zweidimensionalen Hindernis müssen entsprechend zwei Winkelvariablen θ und ϕ berücksichtigt werden, und auch die Durchlässigkeitsfunktion α hängt von zwei Variablen x, y ab; außerdem müssen wir nun vom Wellenzahlvektor sowohl die x- als auch die y-Komponente berücksichtigen. Wir erhalten daher als Summe aller Strahlen unter den beiden gegeben Winkeln das Integral

$$\iint_{\mathbb{R}^2} \alpha(x,y) e^{-i(k_1 x\sin\theta + k_2 y\sin\phi)}\, dx\, dy\,,$$

d.h. die zweidimensionale FOURIER-Transformierte von α, ausgewertet im Punkt $(u,\ v) = (k_1\sin\theta,\ k_2\sin\varphi)$.

Zur Vereinfachung der Schreibweise drückt man das Beugungsbild meist einfach in der Variablen u *bzw.* den Variablen u und v aus statt in den Winkelvariablen; dann ist das Beugungsbild eines Hindernisses mit Durchlässigkeitsfunktion α einfach die FOURIER-Transformierte von α.

Die Größen u und v lassen sich zwar als Strecken interpretieren, sind aber *nicht* proportional zu den Strecken, die man auf einem ebenen Schirm messen kann: Deren Längen sind proportional zu $\tan\theta$ und $\tan\varphi$. Für kleine Winkel, auf die man sich bei der FRAUNHOFER-Beugung wegen des großen Abstands zum Schirm notwendigerweise beschränken muß, unterscheiden sich allerdings Sinus, Tangens und Bogenmaß nur wenig, so daß man auch ohne Umrechnung ein gutes Bild des Beugungsmusters erhält.

Als erstes Beispiel wollen wir das Beugungsbild eines eindimensionalen Spalts berechnen. Dieser habe die Breite a; seine Durchlässigkeitsfunktion kann also beispielsweise geschrieben werden als

$$\alpha\colon \begin{cases} \mathbb{R} \to \mathbb{R} \\ x \mapsto \begin{cases} 1 & \text{für } -\frac{a}{2} \le x \le \frac{a}{2} \\ 0 & \text{sonst} \end{cases} \end{cases}$$

Damit ist

$$\widehat{\alpha}(u) = \int_{-\infty}^{\infty} \alpha(x) e^{-uix}\, dx = \int_{-\frac{a}{2}}^{\frac{a}{2}} e^{-iux} dx$$

$$= \frac{e^{-\frac{iau}{2}} - e^{\frac{iau}{2}}}{-iu} = \frac{2\sin\frac{au}{2}}{u} = a\frac{\sin\frac{au}{2}}{\frac{au}{2}} = a\,\mathrm{sinc}\,\frac{au}{2}\,.$$

Dies erklärt, warum die Funktion $\mathrm{sinc}\, x$ auch als *Spaltfunktion* bezeichnet wird.

Die Lichtintensitäten, die man im Beugungsbild beobachtet, sind allerdings *nicht* durch diese Funktion gegeben: $\widehat{\alpha}(u)$ ist die Amplitude einer skalaren Welle; die Intensität ist gleich dem Betragsquadrat davon, bei einer reellen Funktion wie hier also einfach das Quadrat

$$\widehat{\alpha}(u)^2 = 4\frac{\sin^2\frac{au}{2}}{u^2} = a^2\,\mathrm{sinc}^2\,\frac{au}{2}\,.$$

Als nächstes Beispiel betrachten wir Beugung an einem regelmäßigen Strichgitter. Der Abstand zweier Striche sei d und es gebe insgesamt $2N+1$ Striche. Wenn wir in erster Näherung die Breite der Striche

vernachlässigen, können wir die Durchlässigkeitsfunktion α als Summe von δ-Distributionen schreiben:

$$\alpha(t) = \sum_{k=-N}^{N} \delta(x - kd)\,.$$

Das Beugungsbild ist somit gegeben durch

$$\begin{aligned}
\widehat{\alpha}(u) &= \int_{-\infty}^{\infty} \alpha(x) e^{-iux}\,dx = \sum_{k=-N}^{N} \int_{-\infty}^{\infty} \delta(x - kd) e^{-iux}\,dx \\
&= \sum_{k=-N}^{N} e^{-iudk} = \sum_{k=-N}^{N} e^{iukd} = e^{-iuNd} \sum_{k=0}^{2N} e^{iukd} \\
&= e^{-iuNd} \frac{1 - e^{iu(2N+1)d}}{1 - e^{iud}} = \frac{e^{-iuNd} - e^{iu(N+1)d}}{1 - e^{iud}} \\
&= \frac{e^{iu(N+\frac{1}{2})d} - e^{-iu(N+\frac{1}{2})d}}{e^{iu\frac{d}{2}} - e^{-iu\frac{d}{2}}} = \frac{\sin u\big(N + \frac{1}{2}\big)d}{\sin \frac{ud}{2}}\,.
\end{aligned}$$

Abbildung 29 zeigt diese Funktion; man sieht ihr das charakteristische Linienmuster an, das man bei der Beugung am Gitter beobachtet.

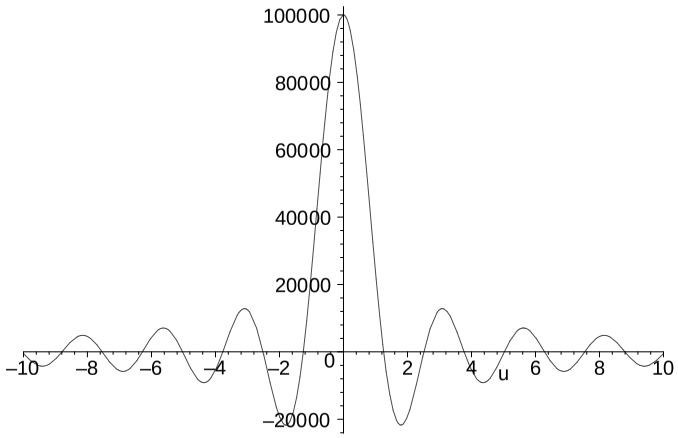


Abb. 29: Beugung am Gitter

Trotzdem wird vielleicht einigen Lesern unwohl sein beim Gedanken an eine Summe von δ-Distributionen als Durchlässigkeitsfunktion. Deshalb wollen wir zur Sicherheit nachrechnen, was sich ändert, wenn wir stattdessen die Striche als Spalte der Breite a annehmen.

Für einen einzelnen solchen Spalt haben wir dann die oben betrachtete Durchlässigkeitsfunktion

$$\alpha_a\colon \begin{cases} \mathbb{R} \to \mathbb{R} \\ x \mapsto \begin{cases} 1 & \text{für } -\frac{a}{2} \le x \le \frac{a}{2} \\ 0 & \text{sonst} \end{cases} \end{cases}$$

eines Spalts der Breite a, und die Durchlässigkeitsfunktion des gesamten Strichgitters ist die Faltung $\alpha \star \alpha_a$ dieser Funktion mit der oben betrachteten Funktion α. Das Beugungsbild ist somit gegeben durch das Produkt des gerade berechneten Beugungsbilds mit dem Beugungsbild eines Spalts, also durch

$$\frac{\sin u\big(N + \frac{1}{2}\big)d}{\sin \frac{ud}{2}} \cdot \frac{\sin \frac{au}{2}}{\frac{u}{2}}\,.$$

Da der Abstand zwischen zwei Spaltmitten gleich d ist, muß die Spaltbreite a echt kleiner als d sein, und die Anzahl N der Striche im Gitter liegt typischerweise bei mindestens einigen Zehntausend. Somit hat der Sinus im zweiten Term eine Kreisfrequenz, die um einen mindestens fünfstelligen Faktor gröser ist als die im ersten; der zweite Term zeigt also erst dann eine nennenswerte Variation, wenn wir im ersten Faktor mehrere Tausend Linien betrachten. In dem Bereich, den wir realistischerweise beobachten können, ist der zweite Term daher für alle praktischen Zwecke konstant. Der Betrag dieser Konstanten ist irrelevant, denn da wir bei der FRAUNHOFER-Beugung das Beugungsbild in „sehr großer“ Entfernung vom Gitter betrachten, können wir sinnvollerweise ohnehin nur von relativen, nicht aber von absoluten Helligkeiten reden.

Als letztes Beispiel zur eindimensionalen Beugung möchte ich eines betrachten, bei dem das Licht nicht als konstante Wellenfront einfällt: Je nach Wahl der Randbedingungen im optischen Resonator entsteht nicht immer ein Strahl, der näherungsweise als ebene Welle betrachtet werden kann (die sogenannte TEM_{00}-Mode); bringt man Hindernisse in

den Strahlengang, können auch höhere TEM-Moden angeregt werden (TEM = *transversal elektromagnetisch*). Bei einem dünnen Hindernis wie etwa einem Haar genau in der Mitte des Strahls beispielsweise entsteht die TEM_{01}-Mode, die aus einem linken und einem rechten Halbstrahl besteht, deren Phasen sich um 180° unterscheiden, und die man ansonsten wieder näherungsweise als ebene Wellen betrachten kann. Trifft ein solcher Strahl auf einen Spalt, dessen Mitte mit der Grenze zwischen den beiden Halbstrahlen zusammenfällt, ist also in der linken Hälfte des Spalts die Phase um 180° gegenüber der rechten verschoben; dies können wir formal dadurch beschreiben, daß wir die Durchlässigkeitsfunktion des Spalts multiplizieren mit einer Funktion, die in der linken Hälfte $+1$ und in der rechten gleich -1 ist. Für einen Spalt der Breite a erhalten wir als Beugungsbild

$$\int_{-\frac{a}{2}}^{0} e^{-iux}\,dx + \int_{0}^{\frac{a}{2}} -e^{-iux}\,dx = \frac{1 - e^{iua/2} - e^{-iua/2} + 1}{-iu}$$

$$= \frac{2i}{u}\left(1 - \cos\frac{ua}{2}\right) .$$

Alternativ läßt sich dies auch über die Beziehung

$$e^{iua/2} + e^{-iua/2} - 2 = (e^{iua/4} - e^{-iua/4})^2 = -4\sin^2\frac{ua}{4}$$

als

$$\frac{4i}{u}\sin^2\frac{ua}{4}$$

schreiben. Daß hier imaginäre Größen auftreten, braucht uns natürlich nicht zu stören: Die beobachteten Intensitäten sind bekanntlich die Betragsquadrate der hier berechneten Funktionen, also reell und positiv.

Zum Abschluß möchte ich noch zumindest ein Beispiel eines zweidimensionalen Beugungsbilds betrachten. Leider sind die zugehörigen FOURIER-Integrale schon in so einfachen Fällen wie dem einer scheibenförmigen Blende nicht mehr elementar auswertbar; wir beschränken uns daher auf den extrem einfachen Fall einer rechteckigen Blende. Deren

Durchlässigkeitsfunktion ist

$$\alpha\colon \begin{cases} \mathbb{R}^2 \to \mathbb{R} \\ (x,y) \mapsto \begin{cases} 1 & \text{falls } -\frac{a}{2} \le x \le \frac{a}{2} \text{ und } -\frac{b}{2} \le y \le \frac{b}{2} \\ 0 & \text{sonst} \end{cases} \end{cases} ,$$

die Beugungsfigur ist also gegeben durch

$$\widehat{\alpha}(u,v) = \iint_{\mathbb{R}^2} \alpha(x,y) e^{-i(ux+vy)}\,dx\,dy = \iint\limits_{\substack{-\frac{a}{2}\le x\le\frac{a}{2} \\ -\frac{b}{2}\le y\le\frac{b}{2}}} e^{-i(ux+vy)}\,dx\,dy$$

$$= \int_{-\frac{a}{2}}^{\frac{a}{2}} \int_{-\frac{b}{2}}^{\frac{b}{2}} e^{-iux} e^{-ivy}\,dx\,dy = \int_{-\frac{a}{2}}^{\frac{a}{2}} e^{-iux} \left(\int_{-\frac{b}{2}}^{\frac{b}{2}} e^{-ivy}\,dy \right) dx$$

$$= \int_{-\frac{a}{2}}^{\frac{a}{2}} e^{-iux} \cdot \frac{\sin\frac{bv}{2}}{v}\,dx = 4 \cdot \frac{\sin\frac{au}{2}}{u} \cdot \frac{\sin\frac{bv}{2}}{v} ,$$

da wir das Rechteck als Normalbereich betrachten können und somit das zweidimensionale Integral über zwei eindimensionale Integrationen berechnen können.

Als Beugungsfigur erhalten wir, nicht gerade überraschenderweise, das Produkt einer vertikalen und einer horizontalen Beugungsfigur eines eindimensionalen Spalts.

Kapitel 4
Differentialgleichungen

Differentialgleichungssysteme sind so ziemlich *das* wichtigste mathematischen Hilfsmittel der Naturwissenschaften und der Technik. Die dahinterstehende Grundidee ist einfach: Man kann zwar nur selten *a priori* sagen, wie sich ein System über einen längeren Zeitraum hinweg entwickeln wird, aber man hat oft aufgrund von Naturgesetzen eine klare Vorstellung über die Zustandsänderung *im nächsten Augenblick*, d.h. also über den Wert der zeitlichen Ableitung der Zustandsgrößen in Abhängigkeit vom gegenwärtigen Zustand des Systems.

§ 1: Definitionen und erste Beispiele

a) Wurfparabel

Ein einfaches Beispiel hierfür liefert das zweite NEWTONsche Gesetz, wonach die zeitliche Ableitung des Impuls eines Teilchens gleich der auf das Teilchen wirkenden Kraft ist.

Ein in die Luft geworfener Gegenstand bewegt sich unter gewissen Bedingungen näherungsweise auf einer parabelförmigen Bahn. Wir wollen diese etwas vage Aussage präzisieren und mathematisch herleiten.

Es gibt viele Wurftechniken, und nur wenige davon können auf einfache Weise durch ein mathematisches Modell beschrieben werden; wir ignorieren daher den genauen Vorgang des Abwurfs und gehen davon aus, daß der Gegenstand *irgendwie* eine Anfangsgeschwindigkeit $\vec{v}$ erreicht hat im Abwurfpunkt mit Koordinaten (x_0, y_0, z_0); den Zeitpunkt des Abwurfs bezeichnen wir mit t_0.

Als nächstes nehmen wir an, daß wir den Luftwiderstand vernachlässigen können, eine Annahme, die beim Kugelstoßen kaum zu Fehlern führt, die aber beispielsweise für einen Fallschirmspringer (auch mit geschlossenem Fallschirm) oder einen Papierflieger völlig unrealistisch ist. Als nächstes wollen wir auch noch annehmen, daß wir nur relativ geringe Wurfhöhen erreichen, so daß die Erdanziehung als konstant angenommen werden kann.

Die Bewegung des Gegenstandes wird dann durch zwei Naturgesetze bestimmt: Das Gravitationsgesetz beschreibt den Effekt der Erdanziehung, und das zweite NEWTONsche Gesetz sagt uns, wie sich diese Kraft auf die Bewegung des Gegenstands auswirkt. Die Gravitation können wir aufgrund der gemachten Annahmen als konstant annehmen, d.h. auf einen Körper der Masse m wirkt die Kraft gm, wobei $g \approx 9{,}8\,m/s^2$ die Gravitationsbeschleunigung an der Erdoberfläche ist; bei „üblicher" Ausrichtung des Koordinatensystems wirkt sie in Richtung der negativen z-Achse. Diese Gravitationskraft ist nach dem zweiten NEWTONschen Gesetz gleich der Ableitung des Impulses nach der Zeit; wenn wir die Masse m als konstant voraussetzen, ist das also gleich m mal der Ableitung der Geschwindigkeit oder m mal der zweiten Ableitung des Orts. Wir haben somit das Differentialgleichungssystem

$$\ddot{x}(t) = 0, \quad \ddot{y}(t) = 0 \quad \text{und} \quad \ddot{z}(t) = -g\,.$$

Diese Gleichungen sind erfüllt, wann immer $x(t)$ und $y(t)$ lineare Funktionen von t sind und $z(t)$ eine quadratische Funktion mit führendem Koeffizienten $-g$. Die sechs noch fehlenden Koeffizienten dieser drei Polynomfunktionen geben uns die Anfangsbedingungen: Zum Zeitpunkt $t = t_0$ ist

$$x(t_0) = x_0, \quad y(t_0) = y_0 \quad \text{und} \quad z(t_0) = z_0\,,$$

und die Geschwindigkeit ist $\vec{v}$, d.h.

$$\dot{x}(t_0) = v_1, \quad \dot{y}(t_0) = v_2 \quad \text{und} \quad \dot{z}(t_0) = v_3\,.$$

Also ist

$$x(t) = v_1(t - t_0) + x_0, \quad y(t) = v_2(t - t_0) + y_0$$

und

$$z(t) = -g(t - t_0)^2 + v_3(t - t_0) + z_0\,.$$

Diese Gleichungen beschreiben in der Tat fast immer eine Parabel: Falls wir die x-Achse des Koordinatensystems so wählen, daß die Anfangsgeschwindigkeit $\vec{v}$ in der (x,z)-Ebene liegt, ist $v_2 = 0$. Falls auch v_1 verschwindet, falls wir den Gegenstand also senkrecht nach oben (oder gar unten) werfen, sind $x(t) = x_0$ und $y(t) = y_0$ konstant und nur

$$z(t) = -g(t-t_0)^2 + v_3(t-t_0) + z_0$$

hängt von der Zeit ab. Andernfalls können wir durch v_1 dividieren; wir erhalten

$$t - t_0 = \frac{x(t)-x_0}{v_1} \quad \text{und} \quad z(t) = \frac{-g}{v_1}\big(x(t)-x_0\big)^2 + \frac{v_3}{v_1}\big(x(t)-x_0\big) + z_0\,,$$

die Punkte $\big(x(t), z(t)\big)$ liegen also in der Tat auf einer Parabel.

b) Radioaktiver Zerfall

Das gerade durchgerechnete Beispiel war insofern untypisch für Differentialgleichungen, als auf den rechten Seite der Gleichungen nur Konstanten standen; üblicherweise wird man dort Funktionen erwarten, die nicht nur von t abhängen (so daß man sie einfach integrieren kann), sondern auch noch von den gesuchten Funktionen. Beim radioaktiven Zerfall etwa ist die pro (kleiner) Zeiteinheit zerfallende Masse proportional zur noch vorhandenen Masse, es gibt also eine Konstante $\lambda > 0$, die sogenannte Zerfallskonstante, so daß die zum Zeitpunkt t vorhandene Masse $m(t)$ der Gleichung

$$\dot{m}(t) = -\lambda m(t)$$

genügt – zumindest, wenn diese Masse hinreichend groß ist. (Im atomaren Bereich muß man auch statistische Effekte berücksichtigen, aber ab etwa 10^{10} Atomen können die für alle praktischen Fälle vernachlässigt werden.)

Wir kennen bereits eine Funktion, die sich so verhält, wie es die obige Differentialgleichung angibt, nämlich die Exponentialfunktion $e^{-\lambda t}$, und natürlich entspricht auch für jedes konstante Vielfache dieser Funktion die Differentiation einfach der Multiplikation mit $-\lambda$. Das sind dann aber bereits alle Funktionen mit dieser Eigenschaft, denn der Quotient

$$q(t) = \frac{m(t)}{e^{-\lambda t}} = m(t)\cdot e^{\lambda t}$$

einer Lösungsfunktion und der Funktion $e^{-\lambda t}$ hat die Ableitung

$$\dot{q}(t) = \dot{m}(t)\cdot e^{\lambda t} + m(t)\cdot \lambda e^{\lambda t} = -\lambda m(t)\cdot e^{\lambda t} + \lambda m(t)\cdot e^{\lambda t} = 0\,,$$

ist also gleich einer Konstanten c, so daß

$$m(t) = c\cdot e^{-\lambda t}$$

ist. Indem wir $t = 0$ setzen, sehen wir, daß die Konstante $c = m(0)$ gleich der zum Zeitpunkt 0 vorhandenen Masse ist; falls wir stattdessen die Masse $m_0 = m(t_0)$ zu einem anderen Zeitpunkt t_0 kennen, können wir analog zum obigen Beispiel auch schreiben

$$m(t) = m_0 e^{-\lambda(t-t_0)} = (m_0 e^{\lambda t_0})\cdot e^{-\lambda t}\,.$$

c) Differentialgleichungen und Differentialgleichungssysteme

Wir betrachten ein System, das durch n zeitlich veränderliche Größen $y_1(t), \ldots, y_n(t)$ beschrieben wird; unter einem System von Differentialgleichungen oder kurz einer Differentialgleichung verstehen wir eine Vorschrift, die die zeitlichen Ableitungen $\dot{y}_1(t), \ldots, \dot{y}_n(t)$ aus den Funktionswerten berechnet:

$$\begin{aligned}
\dot{y}_1(t) &= f_1\big(t, y_1(t), \ldots, y_n(t)\big)\\
\dot{y}_2(t) &= f_2\big(t, y_1(t), \ldots, y_n(t)\big)\\
&\vdots\ \ \vdots\\
\dot{y}_n(t) &= f_n\big(t, y_1(t), \ldots, y_n(t)\big)\,.
\end{aligned}$$

Falls die Funktionen f_i nur von $y_1(t), \ldots, y_n(t)$ abhängen und nicht auch noch direkt von der Zeit t spricht man von einem *autonomen* System. Da Naturgesetze nicht von der Zeit abhängen, hat man es in naturwissenschaftlich-technischen Anwendungen meist mit autonomen Systemen zu tun; man kann allerdings auch den Einfluß von Umgebungsgrößen in einem zeitabhängigen Term zusammenfassen und so ein nichtautonomes System erhalten.

Falls wir, wie in den Beispiel aus den vorangegangenen Abschnitten, die Werte der beteiligten Funktionen zu einem festen Zeitpunkt $t = t_0$ kennen, reden wir von einem *Anfangswertproblem.* Solche Probleme treten

typischerweise dann auf, wenn das weitere Verhalten *eines* konkreten Systems vorhergesagt werden soll.

Auch Differentialgleichungen, in denen wie im Beispiel der Wurfparabel höhere Ableitungen vorkommen, lassen sich so interpretieren: Wenn wir dort die drei Komponenten des Geschwindigkeitsvektors als neue Funktionen

$$u(t) = \dot{x}(t), \quad v(t) = \dot{y}(t) \quad \text{und} \quad w(t) = \dot{z}(t)$$

einführen, können wir das System schreiben als

$$\begin{aligned} \dot{x}(t) &= u(t), & \dot{y}(t) &= v(t), & \dot{z}(t) &= w(t), \\ \dot{u}(t) &= 0, & \dot{v}(t) &= 0, & \dot{w}(t) &= -g, \end{aligned}$$

und wir kennen für jede der sechs beteiligten Funktionen ihren Wert an der Stelle $t = t_0$.

Ein für die Informationstechnik wichtiger Spezialfall sind Gleichungen der Form

$$y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \cdots + a_1\dot{y}(t) + a_0y(t) = b(t)\,,$$

die sogenannten linearen Differentialgleichungen n-ter Ordnung mit konstanten Koeffizienten. Auch diese Gleichungen lassen sich leicht auf die obige Form bringen: Wir betrachten n neue Funktionen

$$y_0(t), y_1(t), \ldots, y_{n-1}(t)$$

mit der Idee, daß sich $y_i(t)$ so verhalten soll wie die i-te Ableitung von $y(t)$. Dazu bilden wir das Differentialgleichungssystem

$$\begin{aligned} \dot{y}_0(t) &= y_1(t) \\ \dot{y}_1(t) &= y_2(t) \\ &\vdots \quad \vdots \\ \dot{y}_{n-2}(t) &= y_{n-1}(t) \\ \dot{y}_{n-1}(t) &= b(t) - a_{n-1}y^{(n-1)}(t) - \cdots - a_1\dot{y}(t) - a_0y(t)\,. \end{aligned}$$

Für jede Lösung $y(t)$ der obigen Gleichung ist dann das n-tupel

$$\big(y(t), \dot{y}(t), \ddot{y}(t), \ldots, y^{(n-1)}(t)\big)$$

eine Lösung des Systems, und für jede Lösung

$$\big(y_0(t), y_1(t), y_2(t), \ldots, y_{n-1}(t)\big)$$

des Differentialgleichungssystems ist $y_0(t)$ eine Lösung der obigen Gleichung.

Auch wenn das Differentialgleichungssystem als Anfangswertproblem gegeben ist, läßt sich das leicht in Anfangswerte für die Gleichung höherer Ordung umschreiben: Hier werden die Werte $y(t_0), \dot{y}(t_0)$ *usw.* bis $y^{(n-1)}(t_0)$ vorgegeben.

Somit beschreiben das System von Differentialgleichungen erster Ordnung und die eine Differentialgleichung höherer Ordnung genau dasselbe Phänomen. Wie wir im vorigen Kapitel gesehen haben, läßt sich die Differentialgleichung höherer Ordnung recht gut mit Hilfe von LAPLACE-Transformationen lösen; in diesem Kapitel werden wir sehen, daß der im letzten Semester entwickelte (und im folgenden noch auszubauende) Apparat der linearen Algebra eine strukturelle Übersicht über die Lösungsmenge des Systems. Erst die Kombination beider Ansätze liefert ein vollständiges Bild.

d) Systeme linearer Differentialgleichungen

Wir betrachten in diesem Abschnitt Systeme von Differentialgleichungen, wie sie zu Beginn dieses Paragraphen definiert wurden, unter der (sehr) einschränkenden Voraussetzungen, daß die rechten Seiten linear in den gesuchten Funktionen $y_1(t), \ldots, y_n(t)$ sind; wir betrachten also ein System

$$\begin{aligned} \dot{y}_1(t) &= a_{11}(t)y_1(t) + a_{12}(t)y_2(t) + \cdots + a_{1n}(t)y_n(t) + b_1(t) \\ \dot{y}_2(t) &= a_{21}(t)y_1(t) + a_{22}(t)y_2(t) + \cdots + a_{2n}(t)y_n(t) + b_2(t) \\ &\vdots \quad\quad \vdots \\ \dot{y}_n(t) &= a_{n1}(t)y_1(t) + a_{n2}(t)y_2(t) + \cdots + a_{nn}(t)y_n(t) + b_n(t) \end{aligned} \qquad (*)$$

Eventuell haben wir noch Anfangsbedingungen der Form

$$y_1(t_0) = c_0, \quad y_2(t_0) = c_2, \quad \ldots, \quad y_n(t_0) = c_n \qquad (**)$$

für ein festes $t_0 \in \mathbb{R}$.

Für das im letzten Kapitel betrachtete Beispiel des elektrischen Schwingkreises mit angelegter Wechselspannung etwa haben wir bei dieser Sicht der Dinge die beiden Funktionen $Q(t)$, die Ladung des Kondensators zum Zeitpunkt t, und $I(t) = \dot{Q}(t)$, die resultierende Stromstärke; das Differentialgleichungssystem $(*)$ ist hier also

$$\begin{aligned} \dot{Q}(t) &= I(t) \\ \dot{I}(t) &= -\frac{R}{L} I(t) - \frac{Q(t)}{LC} + A_0 \cos \omega_0 t \end{aligned} .$$

Wir können die Funktionen $y_i(t)$, $b_i(t)$ und die Anfangswerte c_i jeweils zu Vektoren zusammenfassen und die Koeffizientenfunktionen $a_{ij}(t)$ zu einer Matrix: Mit

$$\vec{y}(t) = \begin{pmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{pmatrix}, \qquad \vec{b}(t) = \begin{pmatrix} b_1(t) \\ b_2(t) \\ \vdots \\ b_n(t) \end{pmatrix}, \qquad \vec{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$$

und

$$A(t) = \begin{pmatrix} a_{11}(t) & a_{12}(t) & \dots & a_{1n}(t) \\ a_{21}(t) & a_{22}(t) & \dots & a_{2n}(t) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(t) & a_{n2}(t) & \dots & a_{nn}(t) \end{pmatrix},$$

erhalten wir die übersichtlichere Form

$$\dot{\vec{y}}(t) = A(t) \cdot \vec{y}(t) + \vec{b}(t) ,$$

wobei die Ableitung eines Vektors von Funktionen natürlich der Vektor der abgeleiteten Funktionen sein soll. Falls es Anfangsbedingungen gibt, können sie nun in der kompakte Form $\vec{y}(t_0) = \vec{c}$ geschrieben werden.

In Analogie zu den linearen Gleichungssystemen bezeichnen wir das System $(*)$ als *homogen*, wenn $\vec{b}(t)$ der Nullvektor ist, wenn also alle Funktionen $b_i(t)$ verschwinden; andernfalls bezeichnen wir es als *inhomogen*. Das homogene System zu einem gegebenen inhomogenen System soll einfach dasjenige System sein, in dem alle $b_i(t)$ durch null ersetzt wurden.

Analog zum Fall linearer Gleichungssystemen gilt auch hier

Lemma: *a)* Die Menge aller Lösungen eines *homogenen* Differentialgleichungssystems der Form $(*)$ ist ein $\mathbb{R}$-Vektorraum.
b) Ist das System nicht homogen und ist $\vec{y}(t)$ eine feste Lösung, so läßt sich jede andere Lösung $\vec{z}(t)$ schreiben als $\vec{z}(t) = \vec{y}(t) + \vec{x}(t)$ mit einer Lösung $\vec{x}(t)$ des zugehörigen homogenen Systems; die Lösungsmenge ist also ein affiner Raum.

Beweis: a) Wir müssen zeigen, daß für zwei Lösungen $\dot{\vec{x}}(t)$ und $\dot{\vec{y}}(t)$ eines homogenen Systems auch jede Linearkombination $\lambda\dot{\vec{x}}(t) + \mu\dot{\vec{y}}(t)$ mit $\lambda, \mu \in \mathbb{R}$ wieder eine Lösung ist. Das ist aber klar, denn wenn

$$\dot{\vec{x}}(t) = A(t) \cdot \vec{x}(t) \quad \text{und} \quad \dot{\vec{y}}(t) = A(t) \cdot \vec{y}(t)$$

ist, gilt auch

$$\begin{aligned} \frac{d}{dt}\left(\lambda\vec{x}(t) + \mu\vec{y}(t)\right) &= \lambda\dot{\vec{x}}(t) + \mu\dot{\vec{y}}(t) = \lambda A(t) \cdot \vec{x}(t) + \mu A(t) \cdot \vec{y}(t) \\ &= A(t) \cdot \left(\lambda\vec{x}(t) + \mu\vec{y}(t)\right) . \end{aligned}$$

b) Sind $\vec{y}(t)$ und $\vec{z}(t)$ zwei Lösungen von $(*)$, so ist

$$\dot{\vec{y}}(t) = A(t) \cdot \vec{y}(t) + \vec{b}(t) \quad \text{und} \quad \dot{\vec{z}}(t) = A(t) \cdot \vec{z}(t) + \vec{b}(t) ;$$

die Differenz $x(t) = z(t) - y(t)$ hat somit die Ableitung

$$\begin{aligned} \dot{\vec{x}}(t) = \dot{\vec{z}}(t) - \dot{\vec{y}}(t) &= \left(A(t) \cdot \vec{z}(t) + \vec{b}(t)\right) - \left(A(t) \cdot \vec{z}(t) + \vec{b}(t)\right) \\ &= A(t) \cdot \vec{z}(t) - A(t) \cdot y(t) = A(t) \cdot \left(\vec{z}(t) - \vec{y}(t)\right) = A(t) \cdot \vec{x}(t) \end{aligned}$$

und $x(t)$ löst also in der Tat das zugehörige homogene System. ∎

Um die Lösungsmenge des Differentialgleichungssystems $(*)$ zu verstehen, müssen wir nach diesem Lemma zwei Teilaufgaben lösen:

1.) Wir müssen den Vektorraum der Lösungen des homogenen Systems bestimmen.
2.) Wir müssen uns wenigstens eine Lösung des inhomogenen Systems verschaffen – oder zumindest wissen, daß eine existiert.

Um im einfachsten Fall zu sehen, wie so etwas funktionieren könnte, betrachten wir ein „System“ aus genau einer Gleichung

$$\dot{y}(t) = a(t) \cdot y(t) + b(t) ;$$

dabei nehmen wir an, daß y eine differenzierbare Funktion von $\mathbb{R}$ nach $\mathbb{R}$ sei und $a, b\colon \mathbb{R} \to \mathbb{R}$ stetige Funktionen.

Wir beginnen mit der Lösung des homogenen Systems

$$\dot{y}(t) = a(t) \cdot y(t)\,.$$

Unter der Annahme, daß wir das dürfen, dividieren wir durch $y(t)$ und erhalten

$$\frac{\dot{y}(t)}{y(t)} = a(t)\,.$$

Der Quotient links ist bekanntlich die logarithmische Ableitung von $y(t)$; falls dies nicht mehr bekannt sein sollte, zeigt eine einfache Anwendung der Kettenregel, daß in einem Intervall, in dem $y(t)$ positiv ist,

$$\frac{d}{dt} \ln y(t) = \frac{\dot{y}(t)}{y(t)} = a(t)$$

ist. In einem Intervall, in dem $y(t)$ negativ ist, gilt entsprechend

$$\frac{d}{dt} \ln\bigl(-y(t)\bigr) = \frac{-\dot{y}(t)}{-y(t)} = \frac{\dot{y}(t)}{y(t)} = a(t)\,,$$

und allgemein haben wir somit

$$\frac{d}{dt} \ln |y(t)| = \frac{\dot{y}(t)}{y(t)} = a(t)$$

in jedem Intervall, in dem $y(t)$ nirgends verschwindet.

Integration beider Seiten führt auf

$$\ln |y(t)| = \int a(t)\,dt + C \quad \text{oder} \quad y(t) = e^{\int a(t)\,dt + C} = e^C \cdot e^{\int a(t)\,dt}$$

oder

$$y(t) = \pm e^C \cdot e^{\int a(t)\,dt}\,,$$

wobei das Vorzeichen wegen der Stetigkeit von y im gesamten Intervall konstant ist, da die Exponentialfunktion nie null wird.

Damit ist in diesem Fall das erste Problem auf eine einfache Integration zurückgeführt.

Bleibt noch die Frage, was passiert, wenn $y(t)$ an irgendeinem Punkt t_0 eine Nullstelle haben sollte. Wir wollen uns überlegen, daß $y(t)$ dann auch für jedes $t > t_0$ verschwinden muß.

Falls nicht, gibt es einen Punkt $t_1 > t_0$, so daß $y(t_1) \neq 0$ ist. Wegen der Stetigkeit von $y(t)$ ist die Funktion dann auch in einer Umgebung von t_1 von Null verschieden, d.h. dort können wir die obigen Argumente anwenden und sehen, daß $y(t)$ dort die Form $e^{h(t)}$ hat mit irgendeiner Funktion h. Da y als differenzierbare Funktion insbesondere überall stetig sein muß und $e^{h(t)}$ nirgends verschwindet, ist das nicht möglich. Genauso überlegt man sich, daß $y(t)$ für jedes $t < t_0$ verschwinden muß, $y(t)$ ist also gleich der Nullfunktion. Diese ist somit die einzige Lösung, die noch zusätzlich betrachtet werden muß. Insbesondere folgt daraus auch, daß eine Lösungsfunktion, die in irgendeinem Punkt positiv *bzw.* negativ ist, überall positiv *bzw.* negativ sein muß, denn eine stetige Funktion kann ihr Vorzeichen nur wechseln, wenn sie in irgendeinem Punkt null wird; wie wir gerade gesehen haben, ist das genau dann der Fall, wenn sie überall verschwindet.

Somit hat jede Lösung die Form

$$y(t) = a e^{\int a(t)\,dt} \quad \text{mit einem } a \in \mathbb{R}$$

und umgekehrt ist auch jede dieser Funktionen eine Lösung. Insbesondere ist die Lösungsmenge ein eindimensionalen Vektorraum.

Es wäre schön, wenn wir im mehrdimensionalen Fall genauso vorgehen könnten: In Analogie zu

$$\dot{y}(t) = a(t) \cdot y(t) \Longrightarrow y(t) = c e^{\int a(t)\,dt} \quad \text{mit} \quad c \in \mathbb{R}$$

könnte vielleicht gelten

$$\dot{\vec{y}}(t) = A(t) \cdot \vec{y}(t) \Longrightarrow y(t) = e^{\int A(t)\,dt} \cdot \vec{c} \quad \text{mit} \quad \vec{c} \in \mathbb{R}^n \;\; ?$$

Das Problem dabei ist nur, daß wir hier nicht die geringste Ahnung haben, was die rechte Seite bedeuten soll; unser nächstes Ziel wird sein, ihr eine Bedeutung zu geben und uns dann zu überlegen, ob *bzw.* unter welchen Bedingungen die obige Formel korrekt ist.

e) Die Matrixexponentialfunktion

Wir orientieren uns wieder am Eindimensionalen: Für eine reelle Zahl x ist

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!},$$

also setzen wir analog für eine $n \times n$-Matrix X

$$e^X \underset{\text{def}}{=} \sum_{i=0}^{\infty} \frac{1}{i!} \cdot X^i .$$

Damit ist klar, daß e^X eine $n \times n$-Matrix sein soll, und das erklärt auch, warum oben der Konstantenvektor $\vec{y}_0$ auf der rechten Seite steht. Was wir uns noch überlegen müssen, ist die Konvergenz der Reihe.

Dazu müssen wir die Größe der Einträge in den Matrizen X^i abschätzen: Sind allgemein A, B zwei $n \times n$-Matrizen und sind die Beträge aller Einträge von A kleiner oder gleich a und die von B kleiner oder gleich b, so kann es in AB offensichtlich keinen Eintrag geben, dessen Betrag größer ist als nab: Schließlich ist jeder Eintrag in der Produktmatrix eine Summe von n Summanden, deren jeder Produkt je eines Eintrags von A und von B ist.

Um diese Formel leichter anwenden zu können, machen wir sie mutwillig schlechter und begnügen uns damit, daß jeder Eintrag von AB höchstens den Betrag $(na) \cdot (nb)$ hat.

Ist nun x der Betrag des größten Eintrags in der Matrix X, so folgt induktiv sofort, daß in X^i höchstens Zahlen bis zum Betrag $(nx)^i$ stehen können; in der endlichen Teilsumme

$$\sum_{i=0}^{M} \frac{1}{i!} X^i$$

hat daher jeder Eintrag einen Betrag kleiner

$$\sum_{i=0}^{M} \frac{(nx)^i}{i!} .$$

Letztere Summe konvergiert für $M \to \infty$ gegen e^{nx}, und damit muß auch die Matrixsumme absolut konvergieren, denn die Reihe für e^{nx} ist konvergente Majorante des Betrags eines jeden Eintrags. Insbesondere hat jeder Eintrag von e^X höchstens den Betrag e^{nx}.

Damit wissen wir also, daß die Matrix e^X für jede $n \times n$-Matrix X existiert; somit ist die Funktion $t \mapsto e^{At}$ wohldefiniert.

f) Eigenschaften der Matrixexponentialfunktion

Wir können natürlich nicht erwarten, daß die Matrixexponentialfunktion alle schönen Eigenschaften der gewöhnlichen Exponentialfunktion erbt. Beispielsweise ist nur schwer vorstellbar, daß für beliebige Matrizen A und B gelten sollte $e^{A+B} = e^A \cdot e^B$: Da für zwei Matrizen A und B stets $A + B = B + A$ ist, müßte dann auch $e^A \cdot e^B = e^B \cdot e^A$ sein, was zumindest unwahrscheinlich aussieht. In der Tat ist etwa für

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{und} \quad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

sowohl A^2 als auch B^2 gleich der Nullmatrix, d.h.

$$e^A = E + A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad e^B = E + B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

und

$$e^A \cdot e^B = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} .$$

Das Quadrat von $C = A + B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ ist aber gleich der Einheitsmatrix und daher ist

$$\begin{aligned} e^{A+B} &= E + C + \frac{1}{2!}E + \frac{1}{3!}C + \frac{1}{4!}E + \frac{1}{5!}C + \cdots \\ &= \left(\sum_{i=0}^{\infty} \frac{1}{(2i)!} \right) \cdot E + \left(\sum_{i=0}^{\infty} \frac{1}{(2i+1)!} \right) \cdot C \\ &= \cosh 1 \cdot E + \sinh 1 \cdot C = \begin{pmatrix} \cosh 1 & \sinh 1 \\ \sinh 1 & \cosh 1 \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} e + e^{-1} & e - e^{-1} \\ e - e^{-1} & e + e^{-1} \end{pmatrix} . \end{aligned}$$

Allgemeiner ist

$$e^{At} = E + tA = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \quad \text{und} \quad e^{Bt} = E + tB = \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}$$

und

$$e^{Ct} = \sum_{i=0}^{\infty} \frac{1}{i!} C^i = \left(\sum_{i=0}^{\infty} \frac{1}{(2i)!} \right) \cdot E + \left(\sum_{i=0}^{\infty} \frac{1}{(2i+1)!} \right) \cdot C = \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix} .$$

Zum Glück gilt aber wenigstens

Lemma: Für zwei Matrizen $A, B \in \mathbb{R}^{n \times n}$ mit $AB = BA$ ist

$$e^{A+B} = e^A \cdot e^B = e^B \cdot e^A .$$

Insbesondere ist für $s, t \in \mathbb{R}$

$$e^{A(s+t)} = e^{As} \cdot e^{At} .$$

Beweis: Für zwei reelle Zahlen x, y ist $e^{x+y} = e^x e^y = e^y e^x$; damit gilt dieselbe Formel auch für zwei reellwertige Variablen x und y. Wenn wir in allen Potenzreihen alle x- und y-Potenzen oberhalb der N-ten ignorieren, sagt die Gleichung aus, daß drei Polynome in x und y *als Polynome* identisch sind.

Beim Rechnen mit Polynomen in x und y verwendet man keine speziellen Eigenschaften dieser Variablen *außer, daß sie kommutieren*. Damit kann man in so eine Polynomidentität auch kommutierende Matrizen A und B einsetzen: Beispielsweise führt die Polynomidentität

$$(x + y)^2 = x^2 + 2xy + y^2$$

zur Identität

$$(A + B)^2 = A^2 + 2AB + B^2 ,$$

die wegen der für beliebige Matrizen gültigen Gleichung

$$(A + B)^2 = A^2 + AB + BA + B^2$$

für kommutierende Matrizen in der Tat erfüllt ist.

Damit ist für kommutierende Matrizen A, B speziell stets

$$e^{A+B} = e^A \cdot e^B = e^B \cdot e^A .$$

Da zwei skalare Vielfache derselben Matrix stets miteinander kommutieren, folgt damit auch die letzte Aussage des Lemmas. ∎

Die für uns wichtigste Anwendung hiervon ist

Satz: Für jede $n \times n$-Matrix A ist die Funktion

$$\begin{cases} \mathbb{R} \to \mathbb{R}^{n \times n} \\ t \mapsto e^{At} \end{cases}$$

stetig differenzierbar mit Ableitung $t \mapsto A \cdot e^{At} = e^{At} \cdot A$.

Beweis: Die Ableitung ist definiert als

$$\lim_{h \to 0} \frac{e^{A(t+h)} - e^{At}}{h} .$$

Da die Matrizen At und Ah miteinander vertauschbar sind, ist nach dem gerade bewiesenen Lemma $e^{A(t+h)} = e^{At} \cdot e^{Ah}$, also

$$\frac{e^{A(t+h)} - e^{At}}{h} = e^{At} \cdot \frac{e^{Ah} - E}{h} .$$

Dabei ist

$$\frac{e^{Ah} - E}{h} = \frac{1}{h} \cdot \sum_{i=1}^{\infty} \frac{(Ah)^i}{i!} = A + A^2 h \sum_{i=2}^{\infty} \frac{(Ah)^{i-2}}{i!} = A + A^2 h \sum_{i=0}^{\infty} \frac{(Ah)^i}{(i+2)!} .$$

Nach der obigen Diskussion ist jeder Eintrag der Matrix in der rechtsstehenden Summenmatrix höchstens gleich

$$\sum_{i=0}^{\infty} \frac{(ah)^i}{(i+2)!} = \frac{e^{ah} - (1 + ah)}{a^2 h^2} ,$$

bleibt also insbesondere beschränkt. Dies gilt auch für $h \to 0$, denn wie zweimalige Anwendung der DE L'HOSPITALschen Regel oder TAYLOR-Entwicklung zeigen, ist der Grenzwert dann $\frac{1}{2}$. Damit existiert

$$\lim_{h \to 0} \sum_{i=1}^{\infty} \frac{(Ah)^i}{(i+2)!} ,$$

und somit ist $\frac{d}{dt}e^{At} = \lim_{h\to 0}\frac{e^{A(t+h)} - e^{At}}{h} = e^{At}\cdot A$, da der Vorfaktor A^2h gegen Null geht. Dies ist auch gleich $A\cdot e^{At}$, denn da A mit jeder seiner Potenzen vertauschbar ist, ist es auch mit jeder endlichen Teilsumme der Reihe von e^{At} vertauschbar, also auch mit e^{At} selbst. ∎

Am Ende des vorigen Abschnitts hatten wir gehofft, daß vielleicht auch für jede matrixwertige Funktion $A(t)$ gelten könnte, daß

$$\frac{d}{dt}e^{A(t)} = \dot{A}(t)e^{A(t)}$$

ist; dies war offensichtlich zu optimistisch: Da

$$e^{A(t+h)} = e^{A(t)+h\dot{A}(t)+o(h)}$$

ist, bräuchten wir für einen Beweis nach obigem Vorbild, daß $A(t)$ und $\dot{A}(t)$ miteinander kommutieren; dies ist aber im allgemeinen nicht der Fall. Für

$$A(t) = \begin{pmatrix} 1 & t \\ 1 & 1 \end{pmatrix}$$

beispielsweise ist

$$\dot{A}(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

also

$$A(t)\cdot\dot{A}(t) = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \text{aber} \quad \dot{A}(t)\cdot A(t) = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}.$$

Dies ist natürlich kein Beweis dafür, daß die Ableitung von $e^{A(t)}$ ungleich $\dot{A}(t)\cdot e^{A(t)}$ ist, aber im vorliegenden Fall ist die Ableitung in der Tat verschieden sowohl von $\dot{A}(t)e^{A(t)}$ als auch von $e^{A(t)}\cdot A(t)$: Mit den Methoden, die wir im nächsten Abschnitt kennenlernen werden, können wir durch eine (alles andere als angenehme) Rechnung zeigen, daß

$$e^{A(t)} = \begin{pmatrix} \frac{e^{1+\sqrt{t}} + e^{1-\sqrt{t}}}{2} & \frac{\sqrt{t}\left(e^{1+\sqrt{t}} - e^{1-\sqrt{t}}\right)}{2} \\ \frac{e^{1+\sqrt{t}} - e^{1-\sqrt{t}}}{2\sqrt{t}} & \frac{e^{1+\sqrt{t}} + e^{1-\sqrt{t}}}{2} \end{pmatrix},$$

$$\frac{d}{dt}e^{A(t)} = \begin{pmatrix} \frac{\left(e^{2\sqrt{t}}-1\right)e^{1-\sqrt{t}}}{4\sqrt{t}} & \frac{\left(-1+\sqrt{t}+\sqrt{t}e^{2\sqrt{t}}+e^{2\sqrt{t}}\right)e^{1-\sqrt{t}}}{4\sqrt{t}} \\ \frac{\left(1-e^{2\sqrt{t}}+\sqrt{t}+\sqrt{t}e^{2\sqrt{t}}\right)e^{1-\sqrt{t}}}{4t\sqrt{t}} & \frac{\left(e^{2\sqrt{t}}-1\right)e^{1-\sqrt{t}}}{4\sqrt{t}} \end{pmatrix},$$

aber

$$\dot{A}(t)\cdot e^{A(t)} = \begin{pmatrix} \frac{e^{1+\sqrt{t}} - e^{1-\sqrt{t}}}{2\sqrt{t}} & \frac{e^{1+\sqrt{t}} + e^{1-\sqrt{t}}}{2} \\ 0 & 0 \end{pmatrix}$$

und

$$e^{A(t)}\cdot\dot{A}(t) = \begin{pmatrix} 0 & \frac{e^{1+\sqrt{t}} + e^{1-\sqrt{t}}}{2} \\ 0 & \frac{e^{1+\sqrt{t}} - e^{1-\sqrt{t}}}{2\sqrt{t}} \end{pmatrix}$$

ist. Wir müssen uns bei diesem Ansatz also begnügen mit linearen homogenen Differentialgleichungen mit *konstanten* Koeffizienten.

Alles was uns zu deren theoretischer Lösung jetzt noch fehlt sind Rechenregeln für den Umgang mit Ableitungen von Matrixfunktionen; für die praktische Lösung fehlen natürlich auch noch Verfahren zur effizienten Berechnung der Matrixexponentialfunktion.

Zumindest für Summen und Produkte gelten, wenn man von der Nichtkommutativität der Multiplikation absieht, für matrixwertige Funktionen die üblichen Regeln:

Lemma: *a)* $F, G\colon (a,b) \to \mathbb{R}^{n\times n}$ seien zwei matrixwertige Funktionen auf dem offenen Intervall (a,b). Dann ist

$$\frac{d}{dt}\big(F(t) + G(t)\big) = \dot{F}(t) + \dot{G}(t).$$

b) Für $F\colon (a,b) \to \mathbb{R}^{n\times m}$ und $G\colon (a,b) \to \mathbb{R}^{m\times p}$ ist

$$\frac{d}{dt}\big(F(t)G(t)\big) = \dot{F}(t)\cdot G(t) + F(t)\cdot\dot{G}(t).$$

c) Für einen konstanten Vektor $\vec{v} \in \mathbb{R}^n$ ist

$$\frac{d}{dt}\big(F(t)\cdot\vec{v}\big) = \dot{F}(t)\cdot\vec{v}.$$

Beweis: a) Sind $f_{ij}(t)$ und $g_{ij}(t)$ die Komponenten der Matrizen F und G, so sind die Summen $f_{ij}(t) + g_{ij}(t)$ die Komponenten von $F+G$, und deren Ableitung ist die Summe der Ableitungen.

b) Die (i,j)-Komponente von FG ist die Funktion $\sum_{\nu=1}^{m} f_{i\nu}(t) \cdot g_{\nu j}(t)$, und deren Ableitung ist

$$\sum_{\nu=1}^{m} \left(\dot{f}_{i\nu}(t) \cdot g_{\nu j}(t) + f_{i\nu}(t) \cdot \dot{g}_{\nu j}(t) \right) ,$$

die (i,j)-Komponente von $\dot{F}(t) \cdot G(t) + F(t) \cdot \dot{G}(t)$.

c) Ist der Spezialfall $n+m$ und $p=1$ von *b)*, wobei zusätzlich noch G eine konstante Funktion ist, so daß alle Terme mit $\dot{G}(t)$ verschwinden. ∎

Damit haben wir alles zusammen und können zeigen

Satz: Die sämtlichen Lösungen des homogenen Differentialgleichungssystems $\dot{y} = Ay$ sind genau die Funktionen $t \mapsto e^{At}\vec{y}_0$ mit $\vec{y}_0 \in \mathbb{R}^n$.

Beweis: Da e^{At} die Ableitung Ae^{At} hat, ist die Ableitung der vektorwertigen Funktion $\vec{f}(t) = e^{At} \cdot \vec{y}_0$ nach der zuletzt bewiesenen Formel gleich $A \cdot e^{At} \cdot \vec{y}_0$, also in der Tat gleich $A \cdot \vec{f}(t)$.

Nun sei $\vec{y}\colon (a,b) \to \mathbb{R}^n$ irgendeine differenzierbare vektorwertige Funktion mit der Eigenschaft, daß $\dot{\vec{y}}(t) = A \cdot \vec{y}(t)$ ist. Wir betrachten die Funktion $\vec{g}(t) = e^{-At} \cdot \vec{y}(t)$. Deren Ableitung ist

$$\dot{\vec{g}}(t) = -A \cdot e^{-At} \cdot \vec{y}(t) + e^{-At} \cdot \dot{\vec{y}}(t) = -A \cdot e^{-At} \cdot \vec{y}(t) + e^{-At} \cdot A\vec{y}(t) = \vec{0} ,$$

denn die Matrix A ist mit e^{-At} vertauschbar. Damit haben alle Komponenten von $\vec{g}$ die Ableitung Null, sind also konstant, und somit ist $\vec{g}(t) \underset{\text{def}}{=} \vec{y}_0$ ein konstanter Vektor mit der Eigenschaft, daß

$$\vec{y}_0 = e^{-At} \cdot \vec{y}(t), \quad d.h. \quad \vec{y}(t) = e^{At} \cdot \vec{y}_0 ,$$

wie behauptet. ∎

Korollar: Für jeden Vektor $\vec{y}_0 \in \mathbb{R}^n$ und jede reelle Zahl $t_0 \in \mathbb{R}$ gibt es genau eine differenzierbare Funktion $\vec{y}(t)$ mit den Eigenschaften, daß $\dot{\vec{y}}(t) = A\vec{y}(t)$ und $\vec{y}(t_0) = \vec{y}_0$ ist; dies ist $\vec{y}(t) = e^{A(t-t_0)} \cdot \vec{y}_0$. ∎

Wir wußten bereits, daß die Lösungen einen Vektorraum bilden; obiges Korollar sagt uns, daß dieser Vektorraum die Dimension n hat und daß für jede reelle Zahl t_0 die Abbildung $\vec{y} \mapsto \vec{y}(t_0)$ ein Isomorphismus auf den $\mathbb{R}^n$ ist.

Als erstes Beispiel betrachten wir das Anfangswertproblem

$$\dot{x}(t) = y(t) \quad \text{und} \quad \dot{y}(t) = x(t) \quad \text{mit} \quad x(0) = a \quad \text{und} \quad y(0) = b$$

oder

$$\begin{pmatrix} \dot{x}(t) \\ \dot{y}(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} \quad \text{mit} \quad \begin{pmatrix} x(0) \\ y(0) \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix} .$$

Die Koeffizientenmatrix ist gleich der zu Beginn des Abschnitts betrachteten Beispielmatrix C, deren Exponentialfunktion wird dort berechnet haben; die Lösung des Anfangswertproblems ist also

$$e^{Ct} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a\cosh t + b\sinh t \\ a\sinh t + b\cosh t \end{pmatrix} .$$

§ 2: Eigenwerte, Eigenvektoren und Hauptvektoren

Die Matrixexponentialfunktion ist zwar wohldefiniert, aber eine matrixwertige Potenzreihe ist für allgemeine Matrizen A nicht gerade einfach zu berechnen. Wir brauchen daher alternative Rechenverfahren.

Zumindest ein Fall ist problemlos: Ist nämlich

$$D = \begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_n \end{pmatrix}$$

eine Diagonalmatrix, ist offensichtlich

$$e^{Dt} = \begin{pmatrix} e^{d_1 t} & 0 & \dots & 0 \\ 0 & e^{d_2 t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & e^{d_n t} \end{pmatrix}$$

wieder eine Diagonalmatrix, wobei die Exponentialfunktion einfach komponentenweise auf die Diagonaleinträge ihres Arguments angewandt wird.

Noch ein weiterer Fall ist relativ unproblematisch: für eine obere (oder untere) Dreiecksmatrix N mit Nullen in der Hauptdiagonalen. Eine solche Matrix definiert eine lineare Abbildung $\mathbb{C}^n \to \mathbb{C}^n$, die den k-ten Einheitsvektor $\vec{e}_k$ in den von $\vec{e}_{k+1}$ bis $\vec{e}_n$ erzeugten Untervektorraum abbildet. Das Quadrat von N bildet ihn entsprechend in den von $\vec{e}_{k+2}$ bis $\vec{e}_n$ erzeugten Untervektorraum ab und so weiter, spätestens N^n ist also die Nullmatrix. Damit wird die Potenzreihe der Exponentialfunktion zu einer endlichen Summe, die, wir wir bereits in zwei Beispielen gesehen haben, zumindest für kleine n leicht berechnet werden kann.

Wir wissen bereits aus dem letzten Semester (*Kap. I, §3j1)*), welche Bedingung eine Basis von $\mathbb{R}^n$ *bzw.* $\mathbb{C}^n$ erfüllen muß, damit eine Matrix A bezüglich dieser Basis Diagonalgestalt hat: Die Basisvektoren müssen allesamt Eigenvektoren von A sein. Aus *Kap. I, §6i)* wissen wir auch, wie man Eigenwerte und ausgehend davon Eigenvektoren mit Hilfe von Determinanten bestimmen kann. In diesem Paragraphen wollen wir wir die entsprechende Theorie noch etwas weiterentwickeln und sehen, daß sich die Berechnung einer beliebigen Matrixexponentialfunktion auf die beiden gerade diskutierten Spezialfälle zurückführen läßt.

Wir arbeiten dabei wieder, wie im ersten Kapitel, über einem beliebigen Körper k, denn auch wenn uns im Augenblick zur Anwendung auf Differentialgleichungen nur die Fälle $k = \mathbb{R}$ und $k = \mathbb{C}$ interessieren, hat die hier entwickelte Theorie doch auch interessante Anwendungen über anderen Körpern: Eigenvektoren über endlichen Körpern spielen beispielsweise bei einigen Problem der Signalverarbeitung eine Rolle.

a) Mehr über Eigenwerte und Eigenvektoren

Zur Bequemlichkeit der Leser sei die Definition noch einmal wiederholt:

Definition: *a)* V sei ein k-Vektorraum. Ein Vektor $\vec{v} \in V \setminus \{\vec{0}\}$ heißt *Eigenvektor* der linearen Abbildung $\varphi: V \to V$ zum *Eigenwert* $\lambda \in k$, wenn $\varphi(\vec{v}) = \lambda\vec{v}$ ist.

b) $\lambda \in k$ heißt *Eigenwert* von φ, falls φ einen Eigenvektor zum Eigenwert λ hat.

c) Eigenwerte und Eigenvektoren einer Matrix $A \in k^{n\times n}$ sind die Eigenwerte und Eigenvektoren der linearen Abbildung

$$\varphi: \begin{cases} k^n \to k^n \\ \vec{v} \mapsto A\vec{v} \end{cases} .$$

Offensichtlich ist mit einem Vektor $\vec{v}$ auch jedes Vielfache (außer dem nach Definition ausgeschlossenen Nullvektor) ein Eigenvektor zum selben Eigenwert; allgemeiner ist sogar jede Linearkombination (außer $\vec{0}$) von Eigenvektoren zum Eigenwert λ wieder ein Eigenvektor zum Eigenwert λ, d.h. die Eigenvektoren zu einem festen Eigenwert λ bilden zusammen mit dem Nullvektor einen Untervektorraum von V, den sogenannten *Eigenraum* von λ.

Definition: Die Dimension des Eigenraums von λ heißt *geometrische Vielfachheit* des Eigenwerts λ.

Lemma: Sind $\vec{v}_1, \ldots, \vec{v}_r \in V$ Eigenvektoren der linearen Abbildung $\varphi: V \to V$ zu verschiedenen Eigenwerten $\lambda_1, \ldots, \lambda_r$, so sind diese Vektoren linear unabhängig.

Beweis: Angenommen, $\vec{v}_1, \ldots, \vec{v}_r$ seien linear abhängig. Dann können wir eine Zahl $2 \le s \le r$ finden, so daß zwar $\vec{v}_1, \ldots, \vec{v}_s$ linear abhängig sind, nicht aber $\vec{v}_1, \ldots, \vec{v}_{s-1}$. Es gibt daher Skalare $\alpha_i \in k$, so daß

$$\alpha_1\vec{v}_1 + \cdots + \alpha_s\vec{v}_s = \vec{0}$$

ist. Wenden wir auf beide Seiten dieser Gleichung die Abbildung φ an und beachten, daß $\varphi(\vec{v}_i) = \lambda_i\vec{v}_i$ ist, folgt, daß auch

$$\alpha_1\lambda_1\vec{v}_1 + \cdots + \alpha_s\lambda_s\vec{v}_s = \vec{0}$$

ist. Andererseits können wir obige Gleichung auch einfach mit λ_s multiplizieren mit dem Ergebnis, daß

$$\lambda_s\alpha_1\vec{v}_1 + \cdots + \lambda_s\alpha_s\vec{v}_s = \vec{0} .$$

Durch Subtraktion der letzten beiden Gleichungen voneinander erhalten wir eine lineare Abhängigkeit

$$\alpha_1(\lambda_s - \lambda_1)\vec{v}_1 + \cdots + \alpha_{s-1}(\lambda_s - \lambda_{s-1})\vec{v}_{s-1} = \vec{0}$$

zwischen $\vec{v}_1, \ldots, \vec{v}_{s-1}$. Da diese Vektoren linear unabhängig sind, müssen alle Koeffizienten verschwinden. Da die Eigenwerte $\lambda_1, \cdots, \lambda_s$ aber allesamt verschieden sind, ist dies nur möglich, wenn α_1 bis α_{s-1} verschwinden. Wegen $\vec{v}_s \neq \vec{0}$ muß dann aber auch α_s verschwinden, im Widerspruch zur angenommenen linearen Unabhängigkeit von $\vec{v}_1, \ldots, \vec{v}_s$.

Also sind die Vektoren $\vec{v}_1, \ldots, \vec{v}_r$ linear unabhängig. ∎

Eigenwerte und Eigenvektoren sind auch interessant für Selbstabbildungen eines unendlichdimensionalen Vektorraums: Ist $V = \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ beispielsweise der Vektorraum aller beliebig oft stetig differenzierbarer reeller Funktionen, so sind $\sin \omega t$ und $\cos \omega t$ Eigenvektoren der linearen Abbildung

$$\varphi: \begin{cases} \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}) \to \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}) \\ f \mapsto \ddot{f} \end{cases}$$

zum Eigenwert $-\omega^2$; genauso sind $\sinh \omega t$ und $\cosh \omega t$ Eigenvektoren zum Eigenwert ω^2. Für $\omega = 0$ degenerieren diese beiden Eigenvektoren jeweils zu null und eins, wodurch ein Eigenwert verschwindet; dafür kommt die Identität als neuer Eigenvektor hinzu. Damit ist also jede reelle Zahl Eigenwert von φ mit einer geometrischen Vielfachheit von mindestens zwei. (Wir werden im nächsten Paragraphen sehen, daß die Vielfachheit immer gleich zwei ist.)

Eigenwertprobleme für lineare Abbildungen, die durch Differentialoperatoren gegeben sind, spielen in vielen Anwendungen eine wichtige Rolle; im Hinblick auf solche Anwendungen bezeichnet man die Menge aller Eigenwerte einer linearen Abbildung oder Matrix auch als deren *Spektrum.* Dieses Wort kommt daher, daß z.B. beim (mehrdimensionalen) Differentialoperator, der die Schwingungen des Fells einer Trommel beschreibt, die Eigenwerte gerade die Frequenzen sind, die die Trommel produzieren kann.

Uns interessieren hauptsächlich Eigenwerte und Eigenvektoren in endlichdimensionalen Vektorräumen. Dort können wir konkret mit Matrizen rechnen; ist A die Abbildungsmatrix zu $\varphi: V \to V$, so ist $\varphi(\vec{v}) = \lambda\vec{v}$ äquivalent dazu, daß $A\vec{v} = \lambda\vec{v}$ oder $(A - \lambda E)\vec{v} = \vec{0}$ ist, wobei E wie üblich die Einheitsmatrix bezeichnet.

In letzterer Form ist dies jenes homogene lineare Gleichungssystem für die Komponenten von $\vec{v}$, das wir bereits in *Kap. I, §6i1)* betrachtet haben. Wie jedes homogene lineare Gleichungssystem hat es den Nullvektor als Lösung, der allerdings nach Definition genau aus diesem Grund *nicht* als Eigenvektor betrachtet wird. Weitere Lösungen gibt es genau dann, wenn die Matrix $A - \lambda E$ des Gleichungssystems singulär ist, wenn also $\det(A - \lambda E)$ verschwindet. Somit ist $\lambda \in k$ genau dann ein Eigenwert, wenn $\det(A - \lambda E) = 0$ ist; die zugehörigen Eigenvektoren sind die nichttrivialen Lösungen des homogenen linearen Gleichungssystems $(A - \lambda E)\vec{v} = \vec{0}$.

Damit ist klar, wie man Eigenwerte und Eigenvektoren berechnen kann: Man löse die Gleichung $\det(A - \lambda E) = 0$ und dann für jede Nullstelle λ_i dieser Gleichung das lineare Gleichungssystem $(A - \lambda E)\vec{v} = \vec{0}$. Dieses homogene lineare Gleichungssystem hat *nie* maximalen Rang, da es nach Definition eines Eigenwerts nichttriviale Lösungen geben muß; kommt man also auf ein eindeutig lösbares Gleichungssystem (und damit auf den Nullvektor als einzige Lösung), ist das immer ein Zeichen für einen Rechenfehler.

b) Ein erstes Beispiel

Wir wollen e^A *bzw.* e^{At} berechnen für die bereits in *Kap. I, §6i)* betrachtete Matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 8 & 7 & 6 & 5 \\ 4 & 3 & 2 & 1 \end{pmatrix}.$$

Wie wir dort nachgerechnet haben, ist hier

$$\det(A - \lambda E) = \lambda^2(\lambda^2 - 14)(\lambda - 72)$$

mit Nullstellen $\lambda_1 = \lambda_2 = 0, \lambda_3 = -4$ und $\lambda_4 = 18$. Als Eigenvektoren dazu hatten wir die Vektoren

$$\vec{b}_1 = \lambda \begin{pmatrix} 1 \\ -2 \\ 1 \\ 0 \end{pmatrix}, \quad \vec{b}_2 = \begin{pmatrix} 2 \\ -3 \\ 0 \\ 1 \end{pmatrix}, \quad \vec{b}_3 = \begin{pmatrix} -1 \\ -1 \\ 1 \\ 1 \end{pmatrix} \quad \text{und} \quad \vec{b}_4 = \begin{pmatrix} 5 \\ 13 \\ 13 \\ 5 \end{pmatrix}$$

gefunden, wobei $\vec{b}_i$ Eigenvektor zum Eigenwert λ_i ist, d.h.

$$\varphi(\vec{b}_1) = 0\,\vec{b}_1, \quad \varphi(\vec{b}_2) = 0\,\vec{b}_2, \quad \varphi(\vec{b}_3) = -4\,\vec{b}_3 \quad \text{und} \quad \varphi(\vec{b}_4) = 18\,\vec{b}_4\,.$$

Bezüglich der neuen Basis $(\vec{b}_1, \vec{b}_2, \vec{b}_3, \vec{b}_4)$ hat φ daher die Abbildungsmatrix

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 18 \end{pmatrix},$$

bei der die Eigenwerte von A in der Hauptdiagonalen stehen und alle sonstigen Einträge verschwinden. Bei dieser Matrix haben wir keinerlei Probleme mit der Berechnung der Exponentialfunktion: Offensichtlich ist

$$e^M = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & e^{-4} & 0 \\ 0 & 0 & 0 & e^{18} \end{pmatrix} \quad \text{und} \quad e^{Mt} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & e^{-4t} & 0 \\ 0 & 0 & 0 & e^{18t} \end{pmatrix}.$$

Damit läßt sich auch e^A berechnen: Sind nämlich

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad \vec{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \vec{e}_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

die vier Einheitsvektoren des $\mathbb{R}^4$, so ist

$$\vec{b}_i = B\vec{e}_i \quad \text{mit} \quad B = \begin{pmatrix} 1 & 2 & -1 & 5 \\ -2 & -3 & -1 & 13 \\ 1 & 0 & 1 & 13 \\ 0 & 1 & 1 & 5 \end{pmatrix},$$

und die Gleichung $A\vec{b}_i = \lambda_i \vec{b}_i$ mit $\lambda_{1/2} = 0, \lambda_3 = -4$ und $\lambda_4 = 18$ wird zu $AB\vec{e}_i = \lambda_i B\vec{b}_i$ oder $B^{-1}AB\,\vec{e}_i = \lambda_i \vec{e}_i$.

Also ist $B^{-1}AB = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & 18 \end{pmatrix} = M$ und $A = BMB^{-1}$.

Wir wollen uns überlegen, daß sich eine solche Relation auch in Potenzen sowie in die Exponentialfunktion hineinziehen läßt:

Lemma: Ist $B \in k^{n\times n}$ eine invertierbare Matrix, $M \in k^{n\times n}$ irgendeine Matrix und m eine ganze Zahl, so ist

$$\left(BMB^{-1}\right)^m = BM^mB \quad \text{und} \quad e^{BMB^{-1}} = Be^MB^{-1}.$$

Zum *Beweis* betrachten wir zunächst eine natürliche Zahl $m \in \mathbb{N}$; für diese ist

$$\left(BMB^{-1}\right)^m = \underbrace{BMB^{-1}BMB^{-1}\cdots BMB^{-1}}_{m \text{ mal}}$$
$$= B \cdot M \cdot E \cdot M \cdots M \cdot E \cdot MB^{-1} = BM^mB^{-1},$$

da BB^{-1} die Einheitsmatrix ist. Für $m = 0$ gibt es ebenfalls keine Probleme, da die nullte Potenz *jeder* $n \times n$-Matrix die Einheitsmatrix ist, und für negative m schließlich ist BM^mB^{-1} invers zu $BM^{-m}B^{-1}$, denn

$$BM^mB^{-1} \cdot BM^{-m}B^{-1} = B \cdot M^m \cdot M^{-m}B^{-1} = B \cdot B^{-1} = E.$$

Also ist

$$BM^mB^{-1} = \left(BM^{-m}B^{-1}\right)^{-1} = \left(\left(BMB^{-1}\right)^{-m}\right)^{-1} = \left(BMB^{-1}\right)^m,$$

da $-m$ eine natürliche Zahl ist, für die wir die Formel bereits bewiesen haben. Schließlich ist auch

$$e^{BMB^{-1}} = \sum_{m=0}^{\infty} \frac{1}{m!}\left(BMB^{-1}\right)^m = \sum_{m=0}^{\infty} \frac{1}{m!} BM^mB^{-1} = Be^MB^{-1},$$

wie behauptet. ∎

In unserem Fall erhalten wir

$$e^A = B \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & e^{-4} & 0 \\ 0 & 0 & 0 & e^{18} \end{pmatrix} B^{-1},$$

was sich zumindest im Prinzip ausrechnen läßt – auch wenn das Ergebnis

$$\frac{1}{72} \begin{pmatrix} 27e^{-4}+35+10e^{18} & 10e^{18}-19+9e^{-4} & 10e^{18}-1-9e^{-4} & 10e^{18}+17-27e^{-4} \\ -53+27e^{-4}+26e^{18} & 9e^{-4}+26e^{18}+37 & 26e^{18}-17-9e^{-4} & 26e^{18}+1-27e^{-4} \\ 26e^{18}+1-27e^{-4} & 26e^{18}-17-9e^{-4} & 9e^{-4}+26e^{18}+37 & -53+27e^{-4}+26e^{18} \\ 10e^{18}+17-27e^{-4} & 10e^{18}-1-9e^{-4} & 10e^{18}-19+9e^{-4} & 27e^{-4}+35+10e^{18} \end{pmatrix}$$

alles andere als angenehm ist. Dies zeigt wieder einmal, wieviel man sich ersparen kann, wenn man *vor* Beginn einer Rechnung eine gute Basis *bzw.* ein gutes Koordinatensystem wählt.

c) Das charakteristische Polynom und seine Nullstellen

Es ist kein Zufall, daß im obigen Beispiel die Gleichung $\det(A-\lambda E) = 0$ auf ein Polynom vierten Grades führte: Ist $A = (a_{ij})$ eine $n \times n$-Matrix, so hat die Matrix $A - \lambda E$ in der Diagonalen die Einträge $a_{ii} - \lambda$, ansonsten stimmen alle Einträge mit denen von A überein. Berechnet man daher $\det(A - \lambda E)$ gemäß der definierenden Formel, so gibt es genau ein Produkt, in dem n mit λ behaftete Faktoren vorkommen, nämlich das Produkt

$$(a_{11} - \lambda) \cdots (a_{nn} - \lambda) = (-1)^n \lambda^n + \textit{Terme niedrigerer Ordnung}$$

der Diagonaleinträge. Die restlichen Produkte, die zur Determinante aufsummiert werden, enthalten zwischen null und $n-1$ mit λ behaftete Faktoren, die Summe ist also ein Polynom vom Grad n mit höchstem Term $(-1)^n \lambda^n$.

Definition: Das Polynom $\det(A-\lambda E)$ heißt *charakteristisches Polynom* der Matrix A.

Demgemäß sind also die Eigenwerte von A gleich den Nullstellen des charakteristischen Polynoms von A, und wir sollten uns wenigsten kurz überlegen, wie man die Nullstellen eines solchen Polynoms bestimmen kann.

Zur Bestimmung der Eigenwerte muß man somit die Nullstellen des charakteristischen Polynoms finden. Für ein Polynom vom Grad höchstens zwei (oder aber ein Polynom, das man als Produkt solcher Polynome schreiben kann) ist das nicht schwer: Nullstellen eines linearen Polynoms erhält man durch eine einfache Division, solche eines quadratischen durch quadratische Ergänzung: Da für $a \neq 0$

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 + c - \frac{b^2}{4a}$$

ist, hat das linksstehende Polynom die Nullstellen

$$x_{1/2} = -\frac{b}{2a} \pm \sqrt{\frac{b^2}{4a} - c} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Bei Polynomen höherer Grade, die man nicht auf einfache Weise über binomische Formeln oder ähnliches in kleinere Faktoren zerlegen kann, ist es oft einen Versuch wert, einige der Lösungen zu *erraten*, um so den Grad des Polynoms zu reduzieren.

Bei Polynomen mit ganzzahligen (und eventuell auch rationalen) Nullstellen, ist dazu der Wurzelsatz von VIÈTE ein vielversprechender Ansatzpunkt: Angenommen, das Polynom n-ten Grades

$$f(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \ldots + a_2x^2 + a_1x + a_0$$

mit höchstem Koeffizient eins habe die Nullstellen $z_1, \ldots, z_n$. Dann ist

$$f(x) = (x - z_1)(x - z_2) \cdot \cdots \cdot (x - z_n).$$

Dies läßt sich ausmultiplizieren und liefert dann einen Zusammenhang zwischen Nullstellen und Koeffizienten: Beispielsweise ist

$$a_0 = (-1)^n z_1 z_2 \cdots z_n \quad \text{und} \quad a_{n-1} = -(z_1 + z_2 + \ldots + z_n),$$

und genauso zeigt man auch daß der allgemeine Koeffizient a_i die Summe aller möglicher Produkte aus $n-i$ Nullstellen z_j ist, multipliziert mit $(-1)^{n-i}$. Diese Aussage bezeichnet man als den Wurzelsatz von VIÈTE.

FRANÇOIS VIÈTE (1540–1603) studierte Jura an der Universität Poitiers, danach arbeitete er als Hauslehrer. 1573, ein Jahr nach dem Massaker an den Hugenotten, berief ihn CHARLES IX (obwohl VIÈTE Hugenotte war) in die Regierung der Bretagne; unter HENRI III wurde er geheimer Staatsrat. 1584 wurde er auf Druck der katholischen Liga vom Hofe verbannt und beschäftigte sich fünf Jahre lang mit Mathematik. Unter HENRI IV arbeitete er wieder am Hof und knackte u.a. verschlüsselte Botschaften an den spanischen König PHILIP II. In seinem Buch *In artem analyticam isagoge* rechnete er als erster systematisch mit symbolischen Größen.

Für das Erraten von Nullstellen einfacher Polynome, bei denen man (aus inhaltlichen Gründen oder aber weil so etwas in Übungs- und Klausuraufgaben fast die Regel ist) ganzzahlige Lösungen erwartet, ist vor allem die erstgenannte Beziehung wichtig: In der Form

$$(-1)^n a_0 = z_1 z_2 \cdots z_n$$

gibt sie das Produkt aller Nullstellen. Falls die a_i alle ganzzahlig sind, lohnt es sich also, die Teiler von a_0 zu testen. Ist beispielsweise

$$f(x) = x^4 + 14x^3 - 52x^2 - 14x + 51\,,$$

so ist

$$a_0 = 51 = 3 \cdot 17\,.$$

Da das Produkt aller Nullstellen gleich diesem Wert sein muß, kommen – falls *alle* Nullstellen ganzzahlig sind – für diese nur die Werte $\pm 1, \pm 3$ und ± 17 in Frage. Da das Produkt aller vier Nullstellen gleich 51 ist, gibt es jeweils genau eine Nullstelle vom Betrag 3 *bzw.* 17, sowie zwei Nullstellen vom Betrag eins. Welche Vorzeichen wirklich auftreten, läßt sich durch Einsetzen festellen oder aber auch dadurch, daß nach VIÈTE die *Summe* aller Nullstellen gleich -14 sein muß. Das ist offenbar nur möglich, wenn sowohl +1 als auch -1 Nullstellen sind, sowie -17 und +3. In der Tat zeigt Einsetzen, daß dies auch tatsächlich Nullstellen sind. (Das Einsetzen ist notwendig, da wir nicht sicher sein können, daß wirklich alle Nullstellen ganzzahlig sind.)

In diesem extrem einfachen (und konstrierten) Fall führt also die Primfaktorzerlegung direkt zur Lösung; in komplizierteren Fällen, wenn a_0

mehr Primfaktoren hat, muß man zunächst alle Kombinationsmöglichkeiten, die zum Produkt a_0 führen können, in Betracht ziehen und davon dann durch Einsetzen potentieller Nullstellen alle bis auf die tatsächlichen Nullstellen eliminieren.

Beim Polynom

$$f(x) = x^6 + 27x^5 - 318x^4 - 5400x^3 - 10176x^2 + 27648x + 32768$$

etwa ist $a_0 = 32768 = 2^{15}$; hier wissen wir also nur, daß – sofern alle Nullstellen ganzzahlig sind – jede Nullstelle die Form $\pm 2^i$ haben muß, wobei die Summe aller Exponenten gleich 15 sein muß und die Anzahl der negativen Vorzeichen gerade. Einsetzen zeigt, daß

$$-1, \quad 2, \quad -4, \quad -8, \quad 16, \quad -32$$

die Nullstellen sind.

Man beachte, daß diese Vorgehensweise nur funktioniert, wenn das Polynom höchsten Koeffizienten eins hat; andernfalls ist das Produkt der Nullstellen gleich dem Quotienten aus konstantem Koeffizienten und führendem Koeffizienten mal $(-1)^{\text{Grad}}$.

Falls man nicht sicher sein kann, daß alle Nullstellen ganzzahlig sind, gibt es immer noch eine ganze Reihe von Methoden, um Nullstellen *exakt* zu berechnen: Beispielsweise kennt die Computeralgebra Algorithmen, um ein Polynom (soweit dies möglich ist) in ein Produkt von Polynomen kleineren Grades zu zerlegen mit Koeffizienten aus einem vorgegebenen Körper, der (in einem hier nicht präzisierten) Sinne nicht *zu weit* vom Körper der rationalen Zahlen *bzw.* einem endlichen Körper entfernt ist, und es gibt auch, seit der ersten Hälfte des sechzehnten Jahrhunderts, allgemeine Formeln zur Lösung von Gleichungen dritten und vierten Grades. Diese Formeln spielen wegen ihrer Komplexität und numerischen Instabilität in der Praxis keine sonderlich große Rolle und sollen daher hier nur im Kleindruck behandelt werden:

Für die kubische Gleichung

$$ax^3 + bx^2 + cx + d = 0$$

wenden wir zunächst einen ähnlichen Trick an wie die quadratische Ergänzung beim Fall der quadratischen Gleichungen: Durch die Substitution

$$z = x + \frac{b}{3a}$$

wird die Gleichung zu

$$az^3 + \left(c - \frac{b^2}{3a}\right) z + \frac{2b^3}{27a^2} + d\,,$$

was wir auch kurz als

$$z^3 + pz + q = 0$$

schreiben können. Zur Lösung dieser Gleichung ersetzen wir z durch die Summe

$$z = u + v$$

zweier Variablen und erhalten

$$u^3 + 3u^2v + 3uv^2 + v^3 = u^3 + v^3 + 3uv(u+v) + p(u+v) + q = 0\,.$$

Da die Zerlegung von z in eine Summe äußerst willkürlich ist, können wir hoffen, daß diese Gleichung für die beiden Variablen u und v auch Lösungen hat, wenn wir zusätzliche Bedingungen stellen: Die obige Gleichung für z wird beispielsweise sicherlich dann gelöst, wenn

$$u^3 + v^3 = -q \quad \text{und} \quad 3uv = -p$$

ist. Dann ist

$$u^3 + v^3 = -q \quad \text{und} \quad u^3 \cdot v^3 = -\frac{p^3}{3}\,,$$

wir kennen also Summe und Produkt von u^3 und v^3.

Sind aber Summe und Produkt zweier Zahlen r und s bekannt, so können wir leicht die Zahlen selbst bestimmen: Aus

$$r + s = c \quad \text{und} \quad rs = d$$

folgt, daß

$$r(c-r) = -r^2 + cr = d \quad \text{oder} \quad r^2 - cr + d = 0$$

ist; wir müssen also einfach eine quadratische Gleichung lösen und erhalten

$$r = \frac{c}{2} \pm \sqrt{\frac{c^2}{4} - d}\,.$$

Da die Summe dieser beiden Lösungen gleich c ist, muß also die eine gleich r und die andere gleich s sein.

Auf die kubische Gleichung angewandt heißt das, daß u^3 und v^3 die beiden Zahlen

$$-\frac{q}{2} \pm \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}$$

sind, also

$$u = \sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} \quad \text{und} \quad v = \sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}$$

oder umgekehrt.

Uns interessiert nur die Summe der beiden Zahlen, also

$$z = \sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}\,.$$

Damit sind wir fast fertig. Das verbleibende Problem ist, daß hier formal eine Lösung steht, wohingegen wir für eine kubische Gleichung *drei* Lösungen erwarten. Dieses Problem verkehrt sich sofort in sein Gegenteil, wenn wir beachten, daß genauso, wie die Quadratwurzel nur bis aufs Vorzeichen bestimmt ist, die Kubikwurzel nur bis dritte Einheitswurzeln bestimmt ist: Da die drei komplexen Zahlen

$$1, \quad \frac{-1 + i\sqrt{3}}{2} \quad \text{und} \quad \frac{-1 - i\sqrt{3}}{2}$$

alle dritte Potenz eins haben, ist mit jeder Kubikwurzel w einer Zahl y auch w mal einer dieser drei Zahlen Kubikwurzel; es gibt also (für $y \neq 0$) drei verschiedene Kubikwurzeln, und somit hat obige Formel für z gleich *neun* mögliche Interpretationen.

Daraus können wir die drei richtigen herausfiltern, wenn wir beachten, daß wir nicht nur das Produkt von u^3 und v^3 kennen, sondern auch das von u und v, nämlich $-p/3$. Damit ist der zweite Summand in der Formel für z eindeutig durch den ersten bestimmt, und es gibt nur die zu erwartenden drei Lösungen: Ist

$$u = \sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}$$

irgendeiner der drei möglichen Werte der Wurzel, so ist

$$z = u - \frac{p}{3u}$$

eine Lösung der Gleichung $z^3 + pz + q = 0$ und

$$x = z - \frac{b}{3a}$$

eine Lösung der ursprünglichen Gleichung $ax^3 + bx^2 + cx + d = 0$.

Auch biquadratische Gleichungen lassen sich auflösen: Hier eliminiert man den kubischen Term von

$$ax^4 + bx^3 + cx^2 + dx + e = 0$$

durch die Substitution

$$z = x + \frac{b}{4a}\,;$$

dies führt auf eine Gleichung der Form $z^4 + pz^2 + qz + r = 0$. Für eine beliebige Zahl y folgt daraus für jede Nullstelle z dieser Gleichung die Beziehung

$$(z^2 + y)^2 = z^4 + 2yz^2 + y^2 = (2y - p)z^2 - qz + y^2 - r\,.$$

Falls rechts das Quadrat eines linearen Polynoms $sz + t$ steht, ist

$$(z^2 + y)^2 = (sz + t)^2 \Longrightarrow z = \pm\sqrt{-y \pm (sz + t)}\,,$$

wir können die Gleichung also auflösen.

Nun wird die rechte Seite $(2y-p)z^2 - qz + y^2 - r$ im allgemeinen kein Quadrat eines linearen Polynoms in z sein, wir können aber hoffen, daß es zumindest für gewisse spezielle Werte der bislang noch willkürlichen Konstante y eines ist.

Ein quadratisches Polynom $\alpha z^2 + \beta z + \gamma$ ist genau dann Quadrat eines linearen, wenn die beiden Nullstellen der quadratischen Gleichung $\alpha z^2 + \beta z + \gamma = 0$ übereinstimmen. Nach der obigen Lösungsformel für quadratische Gleichungen ist dies genau dann der Fall, wenn dort der Ausdruck unter der Wurzel verschwindet, d.h. wenn $\beta^2 - 4\alpha\gamma = 0$ ist. In unserem Fall muß also

$$q^2 - 4(2y-p)(y^2-r) = -8y^3 + 4py^2 + 8ry + q^2 - 4pr$$

verschwinden. Dies ist eine kubische Gleichung für y; indem wir diese Gleichung lösen und eine der Lösungen für y einsetzen, erhalten wir die vier Lösungen der biquadratischen Gleichung.

Die erste Lösung einer kubischen Gleichung geht wohl aus SCIPIONE DEL FERRO (1465–1526) zurück, der von 1496 bis zu seinem Tod an der Universität Bologna lehrte. 1515 fand er eine Methode, um die Nullstellen von $x^3+px = q$ für *positive* Werte von p und q zu bestimmen (Negative Zahlen waren damals in Europa noch nicht im Gebrauch). Er veröffentlichte diese jedoch nie, so daß NICCOLO FONTANA (1499–1557, oberes Bild), genannt TARTAGLIA (der Stotterer), dieselbe Methode 1535 noch einmal entdeckte und gleichzeitig auch noch eine Modifikation, um einen leicht verschiedenen Typ kubischer Gleichungen zu lösen. TARTAGLIA war mathematischer Autodidakt, war aber schnell als Fachmann anerkannt und konnte seinen Lebensunterhalt als Mathematiklehrer in Verona und Venedig verdienen.

Die Lösung allgemeiner kubischer Gleichungen geht auf den Mathematiker, Arzt und Naturforscher GIROLAMO CARDANO (1501–1576, unteres Bild) zurück, dem TARTAGLIA nach langem Drängen und unter dem Siegel der Verschwiegenheit seine Methode mitgeteilt hatte. LODOVICO FERRARI (1522–1565) kam 14-jährig als Diener zu CARDANO; als dieser merkte, daß FERRARI schreiben konnte, machte er ihn zu seinem Sekretär. 1540 fand er die Lösungsmethode für biquadratische Gleichungen; 1545 veröffentlichte CARDANO in seinem Buch *Ars magna* die Lösungsmethodem für kubische und biquadratische Gleichungen.

Nach der erfolgreichen Auflösung der kubischen und biquadratischen Gleichungen in der ersten Hälfte des sechzehnten Jahrhunderts beschäftigten sich natürlich viele Mathematiker mit dem nächsten Fall, der Gleichung fünften Grades. Hier gab es jedoch über 250 Jahre lang keinerlei Fortschritt, bis zu Beginn des neunzehnten Jahrhunderts ABEL glaubte, eine Lösung gefunden zu haben. Er entdeckte dann aber recht schnell seinen Fehler und bewies stattdessen 1824, daß es *unmöglich* ist, die Lösungen einer allgemeinen Gleichung fünften (oder höheren) Grades durch Grundrechenarten und Wurzeln auszudrücken.

Die Grundidee seines Beweises liegt in der Betrachtung von Symmetrien innerhalb der Lösungsmenge, ähnlich wie wir in einem späteren Abschnitt einige Differentialgleichungen durch Symmetriebetrachtungen lösen werden. Unmöglichkeitsbeweise sind allerdings deutlich aufwendiger als Lösungsversuche mit Hilfe von Symmetriebetrachtungen; daher kann über Einzelheiten des ABELschen Beweises hier nichts weiter gesagt werden. Interessenten finden ihn in fast jedem Algebralehrbuch im Kapitel über GALOIS-Theorie.

Der norwegische Mathematiker NILS HENRIK ABEL (1802–1829) ist trotz seinen frühen Todes (an Tuberkulose) Initiator vieler Entwicklungen der Mathematik des neunzehnten Jahrhunderts; Begriffe wie abelsche Gruppen, abelsche Integrale, abelsche Funktionen, abelsche Varietäten, die auch in der heutigen Mathematik noch allgegenwärtig sind, verdeutlichen seinen Einfluß. Zu seinem 200. Geburtstag stiftete die norwegische Regierung einen ABEL-Preises für Mathematik mit gleicher Ausstattung und Vergabebedingungen wie die Nobelpreise; erster Preisträger war 2003 JEAN-PIERRE SERRE (*1926) vom Collège de France für seine Arbeiten über algebraische Geometrie, Topologie und Zahlentheorie.

Der ABELsche Satz besagt selbstverständlich nicht, daß Gleichungen höheren als vierten Grades *unlösbar* seien; er sagt nur, daß es *im allgemeinen* nicht möglich ist, die Lösungen durch Wurzelausdrücke in den Koeffizienten darzustellen: Für eine allgemeine Lösungsformel muß man also außer Wurzeln und Grundrechenarten noch weitere Funktionen zulassen. Beispielsweise fanden sowohl HERMITE als auch KRONECKER 1858 Lösungsformeln für Gleichungen fünften Grades mit sogenannten elliptischen Modulfunktionen; 1870 löste JORDAN damit Gleichungen beliebigen Grades.

Für die Berechnung von Eigenvektoren sind schon die Lösungen einer kubischen Gleichung nach CARDANOs Formel im allgemeinen zu kompliziert, als daß man ohne Computer damit rechnen könnte; dasselbe gilt erst recht für höhere Grade. Insbesondere sind die Formeln in vielen Fällen numerisch instabil, da annähernd gleich große Zahlen voneinander subtrahiert werden. Die Numerik geht daher aus gutem Grund anders vor, wenn sie Nullstellen von Polynomen berechnet.

d) Vielfachheiten von Eigenwerten

Ist x eine Nullstelle eines Polynoms $f(X)$, so kann $f(X)$ bekanntlich durch $(X-x)$ geteilt werden, und x heißt *r-fache Nullstelle* von $f(X)$, wenn $f(X)$ durch $(X-x)^r$ teilbar ist, nicht aber durch $(X-x)^{r+1}$.

Definition: Wir sagen, der Eigenwert λ von φ *bzw.* A habe die *algebraische Vielfachheit* r, wenn λ eine r-fache Nullstelle des charakteristischen Polynoms ist.

Im obigen Beispiel hatte also der Eigenwert Null die algebraische Vielfachheit zwei, die anderen beiden hatten algebraische Vielfachheit eins. Die Dimension des jeweiligen Eigenraums, die geometrische Vielfachheit also, war genauso groß, jedoch muß dies im allgemeinen nicht der Fall sein: Für die Matrix

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

etwa hat das charakteristische Polynom

$$\det(A - \lambda E) = \begin{vmatrix} 1-\lambda & 1 \\ 0 & 1-\lambda \end{vmatrix} = (1-\lambda)^2$$

die doppelte Nullstelle eins, $\lambda = 1$ ist also ein Eigenwert mit algebraischer Vielfachheit zwei. Der zugehörige Eigenraum ist die Lösungsmenge des linearen Gleichungssystems

$$0x_1 + 1x_2 = 0$$
$$0x_1 + 0x_2 = 0\,,$$

also gerade die Menge aller Vektoren der Form $\binom{x}{0}$ und somit eindimensional. Die geometrische Vielfachheit des Eigenwerts eins ist daher nur eins.

Das Beispiel der Abbildung

$$\varphi\colon \mathbb{R}^2 \to \mathbb{R}^2; \quad \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x\cos\vartheta - y\sin\vartheta \\ y\cos\vartheta + x\sin\vartheta \end{pmatrix}$$

mit Abbildungsmatrix

$$A = \begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

zeigt, daß es überhaupt keine Eigenwerte geben muß, denn hier ist das charakteristische Polynom gleich

$$\begin{vmatrix} \cos\vartheta - \lambda & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta - \lambda \end{vmatrix} = (\cos\vartheta - \lambda)^2 + \sin^2\vartheta\,.$$

Abgesehen vom Fall $\sin\vartheta = 0$, wenn A gleich der positiven oder negativen Einheitsmatrix ist, hat dieses Polynom keine reelle Nullstelle, da es nur positive Werte annimmt. Es hat aber natürlich die beiden komplexen Nullstellen

$$\lambda_{1/2} = \cos\vartheta \pm i\sin\vartheta = e^{\pm i\vartheta}\,;$$

fassen wir φ als Abbildung von $\mathbb{C}^2$ nach $\mathbb{C}^2$ auf, gibt es also zwei Eigenwerte. Beide haben die algebraische und geometrische Vielfachheit eins; zugehörige Eigenvektoren sind etwa $\binom{1}{i}$ und $\binom{1}{-i}$. Wählen wir diese beiden Vektoren als Basis, so wird die Abbildungsmatrix von φ bezüglich dieser neuen Basis zur Diagonalmatrix

$$\begin{pmatrix} e^{i\vartheta} & 0 \\ 0 & e^{-i\vartheta} \end{pmatrix}\,.$$

Allgemein gilt für die algebraischen und geometrischen Vielfachheiten von Eigenvektoren

Satz: *a)* Die geometrische Vielfachheit eines Eigenwerts ist stets kleiner oder gleich der algebraischen Vielfachheit.
b) Die Summe der algebraischen Vielfachheiten der verschiedenen Eigenwerte einer linearen Abbildung ist kleiner oder gleich der Dimension des Vektorraums.

Beweis: a) Der Eigenwert λ der $n \times n$-Matrix A habe die geometrische Vielfachheit r, d.h. der zugehörige Eigenraum habe die Dimension r. Wir wählen eine Basis $\vec{b}_1, \ldots, \vec{b}_r$ dieses Eigenraums und ergänzen sie zu einer Basis des gesamten Vektorraums; bezüglich dieser Basis sei C die Abbildungsmatrix der linearen Abbildung

$$\varphi\colon \begin{cases} k^n \to k^n \\ \vec{v} \mapsto A\vec{v} \end{cases}\,.$$

Da $\vec{b}_1, \ldots, \vec{b}_r$ Eigenvektoren zum Eigenwert λ sind, ist $\varphi(\vec{b}_i) = \lambda\vec{b}_i$. In den ersten r Spalten von C steht also jeweils in der Diagonalen das Element λ und ansonsten überall die Null. A hat somit die Form

$$A = \begin{pmatrix} \lambda & 0 & \ldots & 0 & * & \ldots & * \\ 0 & \lambda & \ldots & 0 & * & \ldots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda & * & \ldots & * \\ 0 & 0 & \ldots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & M & \\ 0 & 0 & \ldots & 0 & & & \end{pmatrix},$$

wobei uns weder die mit $*$ bezeichneten Körperelemente noch die $(n-r)\times(n-r)$-Matrix M weiter zu interessieren brauchen.

Für $C - xE$ gilt dasselbe, nur daß jetzt $\lambda - x$ in der Diagonalen steht, d.h. diese Matrix hat die Form

$$\begin{pmatrix} \lambda - x & 0 & \ldots & 0 & * & \ldots & * \\ 0 & \lambda - x & \ldots & 0 & * & \ldots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda - x & * & \ldots & * \\ 0 & 0 & \ldots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & M - xE_{n-r} & \\ 0 & 0 & \ldots & 0 & & & \end{pmatrix},$$

wobei E_{n-r} die $(n-r)\times(n-r)$-Einheitsmatrix bezeichnet.

Zur Berechnung ihrer Determinanten verwenden wir den LAPLACEschen Entwicklungssatz: Da in der ersten Zeile (oder Spalte) nur an der ersten Stelle ein von Null verschiedener Eintrag steht, ist diese Determinante gleich $(\lambda - x)$ mal der Determinante jener Matrix, die durch Streichen der ersten Zeile und Spalte entsteht. Falls $r > 1$ ist, hat diese neue Matrix dieselbe Form, wir können den LAPLACEschen Entwicklungssatz also noch einmal anwenden *usw.;* wir erhalten schließlich

$$\det(C - xE) = (\lambda - x)^r \det(M - xE_{n-r}).$$

Somit ist $\det(C - xE)$ durch $(x - \lambda)^r$ teilbar.

Was uns wirklich interessiert, ist aber nicht $\det(C - xE)$, sondern $\det(A - xE)$. Ist B die Matrix des Basiswechsels von der Standardbasis des k^n auf die Basis $\{\vec{b}_1, \ldots, \vec{b}_n\}$, jene Matrix also, deren Spaltenvektoren die $\vec{b}_i$ sind, so ist $C = B^{-1}AB$ und

$$\begin{aligned} \det(C - xE) &= \det(B^{-1}AB - xE) = \det(B^{-1}AB - xB^{-1}EB) \\ &= \det\bigl(B(A - xE)B^{-1}\bigr) = \det B \det(A - xE)(\det B)^{-1} \\ &= \det(A - xE). \end{aligned}$$

A und C haben also dasselbe charakteristische Polynom, und somit ist auch das charakteristische Polynom von A durch $(x - \lambda)^r$ teilbar. Die algebraische Vielfachheit von λ ist daher mindestens r.

Unabhängig von diesem Ergebnis wollen wir noch festhalten, daß nach der gerade durchgeführten Rechnung für eine beliebige Matrix A und eine invertierbare Matrix B die beiden Matrizen A und BAB^{-1} dasselbe charakteristische Polynom haben; insbesondere haben also die Abbildungsmatrizen einer linearen Abbildung zu verschiedenen Basen dasselbe charakteristische Polynom.

b) Sind $\lambda_1, \ldots, \lambda_\ell$ die verschiedenen Eigenwerte von φ und sind $r_1, \ldots, r_\ell$ ihre algebraischen Vielfachheiten, so ist das charakteristische Polynom $\det(A - xE)$ teilbar durch

$$(x - \lambda_1)^{r_1} \cdots (x - \lambda_\ell)^{r_\ell}.$$

Dies ist ein Polynom vom Grad $r_1 + \cdots + r_\ell$, wohingegen das charakteristische Polynom Grad n hat; daher ist

$$r_1 + \cdots + r_\ell \le n,$$

denn der Grad eines Teilers kann nicht größer sein als der des Polynoms selbst. ∎

Zum Abschluß dieses Abschnitts sei noch ein Kriterium angegeben, wann es für eine lineare Abbildung φ eine Basis aus Eigenwerten gibt, wann also die Abbildungsmatrix bezüglich einer geeigneten Basis Diagonalgestalt hat:

Satz: Zur linearen Abbildung $\varphi: V \to V$ eines n-dimensionalen Vektorraums gibt es genau dann eine Basis aus Eigenvektoren von φ, wenn

1.) das charakteristische Polynom von φ als Produkt von Linearfaktoren geschrieben werden kann

2.) die geometrische Vielfachheit eines jeden Eigenwerts gleich der algebraischen ist.

Beweis: Zunächst sei $\varphi: V \to V$ eine lineare Abbildung derart, daß V eine Basis $\{\vec{b}_1, \ldots, \vec{b}_n\}$ aus Eigenvektoren von φ habe. Wir müssen zeigen, daß *1.)* und *2.)* erfüllt sind.

Da die Basisvektoren $\vec{b}_i$ Eigenvektoren sind, gibt es zu jedem $\vec{b}_i$ ein Körperelement λ_i, so daß $\varphi(\vec{b}_i) = \lambda\vec{b}_i$ ist; bezüglich dieser Basis hat die Abbildungsmatrix A von φ daher Diagonalgestalt, und das charakteristische Polynom

$$\det(A - \lambda E) = \begin{vmatrix} \lambda_1 - \lambda & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n - \lambda \end{vmatrix} = \prod_{i=1}^{n} (\lambda_i - \lambda)$$

zerfällt in der Tat in Linearfaktoren. Die algebraische Vielfachheit des Eigenwerts λ_i ist gleich der Anzahl jeder Indizes $j \in \{1, \ldots, n\}$, für die $\lambda_j = \lambda_i$ ist; dies ist auch die geometrische Vielfachheit, denn der Eigenraum wird aufgespannt von den Vektoren $\vec{b}_j$ zu diesen j. Also sind *1)* und *2)* erfüllt.

Umgekehrt erfülle die Abbildung φ die Bedingungen *1)* und *2);* wir müssen zeigen, daß es eine Basis aus Eigenvektoren von φ gibt.

Wegen *1)* läßt sich das charakteristische Polynom in der Form

$$(\lambda_1 - \lambda)^{r_1} \cdot \cdots \cdot (\lambda_s - \lambda)^{r_s}$$

schreiben, wobei wir annehmen können, daß die λ_i paarweise verschieden sind. Dann ist r_i die algebraische Vielfachheit von λ_i. Da das charakteristische Polynom den Grad n hat, folgt, daß

$$r_1 + \cdots + r_s = n$$

ist. Außerdem gibt es wegen *2.)* zu jedem λ_i einen r_i-dimensionalen Eigenraum, also r_i linear unabhängige Eigenvektoren. Da Eigenvektoren

zu verschiedenen Eigenwerten nach dem Lemma vom Anfang dieses Abschnitts stets linear unabhängig sind, ist auch das System all dieser Eigenvektoren linear unabhängig und somit eine Basis, denn es besteht aus $n = \dim V$ Vektoren. Damit ist eine Basis aus Eigenvektoren von φ gefunden.

■

e) Eigenwerte symmetrischer und Hermitescher Matrizen

Wie wir im letzten Paragraphen gesehen haben, kann die geometrische Vielfachheit eines Eigenwerts kleiner sein als die algebraische, und im Falle einer reellen Matrix müssen nicht auch die Eigenwerte reell sein. In diesem Abschnitt wollen wir sehen, daß solche Dinge bei symmetrischen (und auch den noch zu definierenden HERMITEschen) Matrizen nicht möglich sind.

Symmetrische und HERMITEsche Matrizen hängen eng mit (HERMITEschen) Skalarprodukten zusammen: Für zwei Vektoren

$$\vec{v} = \sum_{i=1}^{n} v_i \vec{b}_i \quad \text{und} \quad \vec{w} = \sum_{i=1}^{n} w_i \vec{b}_i$$

aus einem endlichdimensionalen EUKLIDischen Vektorraum V mit Basis $\{\vec{b}_1, \ldots, \vec{b}_n\}$ ist wegen der Linearität des Skalarprodukts in beiden Argumenten

$$\vec{v} \cdot \vec{w} = \left(\sum_{i=1}^{n} v_i \vec{b}_i \right) \cdot \left(\sum_{j=1}^{n} w_j \vec{b}_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} v_i w_j \, \vec{b}_i \cdot \vec{b}_j \, .$$

Setzen wir

$$c_{ij} \underset{\text{def}}{=} \vec{b}_i \cdot \vec{b}_j \, ,$$

so ist wegen der Symmetrie des Skalarprodukts $c_{ij} = c_{ji}$, wir haben also eine symmetrische $n \times n$-Matrix C.

Die Matrix C legt das Skalarprodukt eindeutig fest, denn für zwei beliebige Vektoren $\vec{v}, \vec{w}$ wie oben ist

$$\vec{v} \cdot \vec{w} = \sum_{i=1}^{n} \sum_{j=1}^{n} v_i w_j \cdot c_{ij} \, .$$

Diese Formel definiert umgekehrt auch für jede symmetrische Matrix $C \in \mathbb{R}^{n\times n}$ eine bilineare Abbildung $V \times V \to \mathbb{R}$, allerdings muß diese nicht positiv definit und damit kein Skalarprodukt sein.

Ist V ein HERMITEscher Vektorraum, wieder mit Basis $\{\vec{b}_1, \ldots, \vec{b}_n\}$, so ist jetzt für zwei Vektoren

$$\vec{v} = \sum_{i=1}^{n} v_i \vec{b}_i \quad \text{und} \quad \vec{w} = \sum_{i=1}^{n} w_i \vec{b}_i \quad \text{mit} \quad v_i, w_i \in \mathbb{C}$$

$$\vec{v} \cdot \vec{w} = \left(\sum_{i=1}^{n} v_i \vec{b}_i\right) \cdot \left(\sum_{j=1}^{n} w_j \vec{b}_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} v_i \overline{w}_j \, \vec{b}_i \cdot \vec{b}_j \,.$$

Setzen wir auch hier wieder

$$c_{ij} \underset{\text{def}}{=} \vec{b}_i \cdot \vec{b}_j \,,$$

so ist nun $c_{ij} = \overline{c_{ji}}$. Matrizen mit dieser Eigenschaft wollen wir als HERMITEsch bezeichnen.

Um dies etwas kompakter ausdrücken zu können, definieren wir

Definition: *a)* Für eine Matrix $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ bezeichnen wir die Matrix $\overline{A} = \left(\overline{a_{ij}}\right)$ als die zu A *konjugiert komplexe Matrix.*
b) $A \in \mathbb{C}^{n\times n}$ heißt HERMITEsch, falls ${}^tA = \overline{A}$ ist.
c) Zu einem Vektor $\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ heißt $\overline{\vec{v}} = \begin{pmatrix} \overline{v_1} \\ \vdots \\ \overline{v_n} \end{pmatrix}$ der *konjugiert komplexe* Vektor.

(Letztere Schreibweise sieht zwar grausam aus, läßt sich aber nicht vermeiden, wenn man Vektoren mit Pfeilen kennzeichnet. Alternativen wie der Fettdruck von Vektoren funktionieren weder an der Tafel noch in einer Mitschrift, und für Frakturbuchstaben wie $\mathfrak{u}, \mathfrak{v}, \mathfrak{w}$ können sich leider nur wenige Studenten begeistern.)

Schließlich wollen wir Vektoren hier mit $1 \times n$-Matrizen identifizieren; insbesondere rechnen wir mit dem „transponierten Vektor“

$${}^t\vec{v} = (v_1, \ldots, v_n) \,.$$

Mit dieser Bezeichnung kann das Standardskalarprodukt zweier Vektoren $\vec{v}, \vec{w} \in \mathbb{R}^n$ als Matrixprodukt ${}^t\vec{v}\vec{w}$ geschrieben werden; das Standard-HERMITEsche Produkt in $\mathbb{C}^n$ ist entsprechend ${}^t\vec{v}\,\overline{\vec{w}}$.

Da die komplexe Konjugation auf $\mathbb{R}$ keine Wirkung hat, ist eine HERMITEsche Matrix mit reellen Einträgen einfach eine symmetrische Matrix; wir können uns im folgenden bei den Beweisen daher auf HERMITEsche Matrizen beschränken und erhalten trotzdem Ergebnisse, die auch für reelle symmetrische Matrizen gelten.

Das Hauptziel dieses Abschnitts ist

Satz: A sei eine symmetrische reelle oder HERMITEsche (komplexe) Matrix.
a) Dann sind alle Eigenwerte von A reell.
b) Eigenvektoren zu verschiedenen Eigenwerten sind orthogonal bezüglich des Standard- *bzw.* HERMITEschen Skalarprodukts.
c) Für jeden Eigenwert von A ist die geometrische Vielfachheit gleich der algebraischen Vielfachheit.
d) $\mathbb{R}^n$ *bzw.* $\mathbb{C}^n$ hat eine Orthonormalbasis aus Eigenvektoren von A.

Beweis: a) Ist $\lambda \in \mathbb{C}$ ein Eigenwert von A, so gibt es nach Definition einen Vektor $\vec{v} \neq \vec{0}$, so daß $A\vec{v} = \lambda\vec{v}$ ist. Da die komplexe Konjugation mit sämtlichen Grundrechenarten vertauschbar ist, folgt, daß

$$\overline{A}\,\overline{\vec{v}} = \overline{\lambda}\,\overline{\vec{v}} \,, \text{ d.h.} \quad {}^t\vec{v}\overline{A}\,\overline{\vec{v}} = {}^t\vec{v}\overline{\lambda}\,\overline{\vec{v}} = \overline{\lambda}\,{}^t\vec{v}\,\overline{\vec{v}} \,.$$

Bislang gilt alles noch für beliebige $n \times n$-Matrizen; um die Symmetrie *bzw.* HERMITE-Eigenschaft von A ins Spiel zu bringen, betrachten wir den Vektor ${}^t(A\vec{v}) = {}^t\vec{v}\,{}^tA$. Da nach Voraussetzung ${}^tA = \overline{A}$ ist, können wir die rechte Seite der Gleichung auch als ${}^t\vec{v}\,\overline{A}$ schreiben, und die linke Seite als ${}^t(\lambda v) = \lambda {}^t\vec{v}$, da $\vec{v}$ Eigenvektor von A ist. Somit können wir die Zahl ${}^t\vec{v}\overline{A}\,\overline{\vec{v}}$ auch schreiben als

$${}^t\vec{v}\overline{A}\,\overline{\vec{v}} = \left({}^t\vec{v}\,\overline{A}\right)\overline{\vec{v}} = \lambda\,{}^t\vec{v}\,\overline{\vec{v}} \,.$$

Somit haben wir die beiden Darstellungen

$${}^t\vec{v}\overline{A}\,\overline{\vec{v}} = \lambda\,{}^t\vec{v}a\,\overline{\vec{v}} \quad \text{und} \quad {}^t\vec{v}\overline{A}\,\overline{\vec{v}} = \overline{\lambda}\,{}^t\vec{v}a\,\overline{\vec{v}} \,,$$

die nur dann beide richtig sein können, wenn $\lambda = \overline{\lambda}$ und somit reell ist; denn ${}^t\vec{v}\,\overline{\vec{v}}$ kann wegen der Definitheit HERMITEscher Skalarprodukte für einen Vektor $\vec{v} \neq \vec{0}$ nicht verschwinden.

b) $\vec{v}$ sei Eigenvektor zum Eigenwert λ, und $\vec{w}$ sei Eigenvektor zum davon verschiedenen Eigenwert μ, d.h.

$$A\vec{v} = \lambda\vec{v} \quad \text{und} \quad A\vec{w} = \mu\vec{w} \quad \text{und} \quad \lambda \neq \mu\,.$$

Dann ist

$$\lambda\,{}^t\vec{v}\overline{\vec{w}} = {}^t(\lambda\vec{v})\overline{\vec{w}} = {}^t(A\vec{v})\overline{\vec{w}} = {}^t\vec{v}\,{}^tA\overline{\vec{w}} = {}^t\vec{v}\,\overline{A}\,\overline{\vec{w}} = {}^t\vec{v}\overline{A\vec{w}} = {}^t\vec{v}\overline{\mu\vec{w}} = \overline{\mu}\,{}^t\vec{v}\overline{\vec{w}}\,.$$

Wie wir schon wissen, sind alle Eigenwerte reell, d.h. $\overline{\mu} = \mu \neq \lambda$. Die obige Gleichungskette kann daher nur richtig sein, wenn ${}^t\vec{v}\,\overline{\vec{w}}$ verschwindet, d.h. wenn $\vec{v}$ und $\vec{w}$ orthogonal sind.

Beim Beweis von *c)* gehen wir im wesentlichen genauso vor wie im vorigen Abschnitt, als wir zeigten, daß die geometrische Vielfachheit eines Eigenwerts stets kleiner oder gleich der algebraischen ist; die zusätzliche Annahme über die Matrix A wird zeigen, daß hier die beiden Vielfachheiten sogar gleich sind.

λ sei also ein Eigenwert von A mit geometrischer Vielfachheit r, d.h. der zugehörige Eigenraum habe die Dimension r. Wir wählen eine Basis $\{\vec{b}_1, \ldots, \vec{b}_r\}$ davon und ergänzen sie zu einer Basis $\mathcal{B} = \{\vec{b}_1, \ldots, \vec{b}_n\}$ des gesamten Vektorraums $V = \mathbb{R}^n$ oder $\mathbb{C}^n$. Indem wir nötigenfalls das GRAM-SCHMIDTsche Orthogonalisierungsverfahren anwenden und anschließend die Längen aller Vektoren auf eins normieren, können wir annehmen, daß es sich dabei um eine Orthonormalbasis handelt.

Nun betrachten wir die lineare Abbildung

$$\varphi\colon V \to V; \quad \vec{v} \mapsto A\vec{v}\,.$$

Bezüglich der Standardbasis hat sie A als Abbildungsmatrix; für uns interessanter ist aber die Abbildungsmatrix C bezüglich der neuen Basis $\mathcal{B}$. Dazu sei B die Matrix mit Spaltenvektoren $\vec{b}_i$; da der Eintrag an der Stelle (i,j) eines Matrixprodukts das (Standard-)Skalarprodukt des i-ten Zeilenvektors des ersten Faktors mit dem j-ten Spaltenvektor des zweiten Faktors ist, steht an der Stelle (i,j) der Matrix ${}^tB\overline{B}$

das (Standard) HERMITEsche Produkt der Vektoren $\vec{b}_i$ und $\vec{b}_j$. Da $\mathcal{B}$ als Orthonormalbasis gewählt wurde, ist daher

$${}^tB\overline{B} = E \quad \text{und damit} \quad {}^tB = \overline{B}^{-1} = \overline{B^{-1}}\,.$$

Aus dieser Formel folgt, daß mit A auch C eine HERMITEsche Matrix ist, denn

$${}^tC = {}^t(B^{-1}AB) = {}^tB\,{}^tA\,{}^tB^{-1} = \overline{B^{-1}}\,\overline{A}\,\overline{B} = \overline{B^{-1}AB} = \overline{C}\,.$$

Die ersten r Basisvektoren $\vec{b}_i$ sind Eigenvektoren von A zum Eigenwert λ; für $i \leq r$ ist daher $\varphi(\vec{b}_i) = \lambda\vec{b}_i$, d.h. in der i-ten Spalte von C steht an der i-ten Stelle die reelle Zahl λ und ansonsten überall die Null, genau wie auch im vorigen Abschnitt. Im Gegensatz zu dort haben wir nun aber eine HERMITEsche Matrix; da in der i-ten Spalte abgesehen von λ auf der Hauptdiagonalen nur Nullen stehen, muß daher dasselbe auch für die i-te Zeile gelten; die Matrix C hat also die Form

$$C = \begin{pmatrix} \lambda & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \lambda & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & M & \\ 0 & 0 & \ldots & 0 & & & \end{pmatrix},$$

wobei M eine $(n-r) \times (n-r)$-Matrix ist, die uns nicht weiter zu interessieren braucht. Damit hat $C - xE$ die Form

$$\begin{pmatrix} \lambda - x & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \lambda - x & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda - x & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & M - xE_{n-r} & \\ 0 & 0 & \ldots & 0 & & & \end{pmatrix},$$

wobei E_{n-r} die $(n-r) \times (n-r)$-Einheitsmatrix bezeichnet.

Wie wir uns schon im vorigen Abschnitt überlegten beim Beweis, daß die geometrische Vielfachheit eines Eigenwerts immer kleiner oder gleich der algebraischen ist, haben A und C dasselbe charakterische Polynom; da wir die Matrix C besser kennen, rechnen wir mit ihr.

Wie in Abschnitt *d)* folgt auf Grund der obigen Form der Matrix $C - xE$ aus dem LAPLACEschen Entwicklungssatz, daß

$$\det(A - xE) = \det(C - xE) = (\lambda - x)^r \det(M - xE_{n-r})$$

ist, wobei E_{n-r} die $(n-r) \times (n-r)$-Einheitsmatrix bezeichnet. Wir müssen zeigen, daß die algebraische Vielfachheit von λ *genau* gleich r ist, daß also λ keine Nullstelle von $\det(M - xE_{n-r})$ sein kann.

Wäre λ Nullstelle von $\det(M - xE_{n-r})$, so hätte M den Eigenwert λ, es gäbe also einen $(n-r)$-dimensionalen Eigenvektor $\vec{w}$ von M. Wegen der speziellen Form der Matrix C ist für jeden Eigenvektor

$$\vec{w} = \begin{pmatrix} w_{r-1} \\ \vdots \\ w_n \end{pmatrix} \quad \text{von } M \text{ der Vektor} \quad \vec{v} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ w_{r-1} \\ \vdots \\ w_n \end{pmatrix}$$

ein Eigenvektor von C und damit von A – Eigenvektoren hängen schließlich nur von der linearen Abbildung ab, nicht von einer speziellen Abbildungsmatrix. Dies widerspricht aber der Voraussetzung, daß der Eigenraum zum Eigenwert λ von $\vec{b}_1, \ldots, \vec{b}_r$ erzeugt wird, denn $\vec{v}$ ist linear unabhängig von diesen $\vec{b}_i$.

Also hat λ die algebraische Vielfachheit r, und *c)* ist gezeigt.

d) ist nun eine einfache Folgerung aus den übrigen Aussagen sogenannten *Fundamentalsatz der Algebra*, wonach jedes reelle oder komplexe Polynom über den komplexen Zahlen in Linearfaktoren zerfällt:

Wir wissen dann, daß die Summe der algebraischen Vielfachheiten aller Eigenwerte gleich der Dimension n des Vektorraums ist und daß alle Eigenwerte reell sind; da die algebraischen gleich den geometrischen

Vielfachheiten sind, gibt es also n Eigenvektoren, die eine Basis von V bilden.

Für jeden einzelnen Eigenraum können wir die Eigenvektoren nach GRAM-SCHMIDT so wählen, daß sie eine Orthonormalbasis bilden; da Eigenvektoren zu verschiedenen Eigenwerten stets orthogonal sind, ist die Vereinigungsmenge dieser Basen Orthonormalbasis von V. ∎

f) Hauptvektoren und und die Jordan-Zerlegung

Falls die lineare Abbildung $\varphi: V \to V$ Eigenwerte hat, deren geometrische Vielfachheit kleiner als die algebraische ist, haben wir keine Chance auf eine Basis, bezüglich derer die Abbildungsmatrix von φ Diagonalgestalt hat: Die Elemente einer solchen Basis wären allesamt Eigenvektoren, und bei zu kleiner geometrischer Vielfachheit gibt es nicht genügend linear unabhängige Eigenvektoren. Außerdem gibt es offensichtlich keine Chance auf eine Diagonalgestalt, wenn das charakteristische Polynom von φ nicht in Linearfaktoren zerfällt, denn dann ist schon die Summe der *algebraischen* Vielfachheiten der Eigenwerte kleiner als die Dimension von V.

Das zweite dieser Probleme konnten wir zumindest beim Beispiel der Matrix

$$\begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix}$$

dadurch lösen, daß wir zu einem größeren Körper übergegangen sind, nämlich von den reellen zu den komplexen Zahlen.

Tatsächlich läßt es sich *immer* dadurch lösen, daß man zu einem größeren Körper übergeht: Nach dem *Fundamentalsatz der Algebra*, zerfällt jedes Polynom mit komplexen (also insbeondere auch mit reellen) Koeffizienten über den komplexen Zahlen in Linearfaktoren. Für andere Körper als die reellen oder komplexen Zahlen zeigt die Algebra, daß es zu jedem Polynom über einem Körper stets einen Erweiterungskörper gibt, der als Vektorraum über dem Ausgangskörper endliche Dimension hat, so daß das gegebene Polynom dort in Linearfaktoren zerfällt. Mit Methoden, die im allgemeinen nicht konstruktiv sind, folgt sogar,

daß es stets einen (im allgemeinen unendlichdimensionalen) Erweiterungskörper gibt, über dem *jedes* Polynom in Linearfaktoren zerfällt, den sogenannte *algebraischen Abschluß* des Ausgangskörpers. Einzelheiten findet man in jedem Lehrbuch der Algebra.

Somit können wir das Problem, daß das charakteristische Polynom eventuell nicht genügend viele Nullstellen hat, im wesentlichen ignorieren. Ernster ist das Problem mit Eigenwerten, deren geometrische Vielfachheit kleiner ist als die algebraische. Damit wollen wir uns in diesem Abschnitt beschäftigen.

Die Lösung wird darin bestehen, daß wir solchen Eigenwerten Räume zuordnen, die größer sind als die Eigenräume, aber immer noch eine gut an die Abbildung angepaßte Basis haben. Insbesondere sollen sie, genau wie die Eigenräume, *invariant* sein unter der betrachteten Abbildung:

Definition: $\varphi\colon V \to V$ sei eine lineare Abbildung. Ein Untervektorraum $U \leq V$ heißt invariant unter φ oder kurz φ-invariant, wenn $\varphi(U) \leq U$ ist.

Die φ-Invarianz der Eigenräume im Sinne dieser Definition ist klar, denn auf einem Eigenraum ist φ einfach die Multiplikation mit dem zugehörigen Eigenwert.

Für das folgende wollen wir der Einfachheit halber annehmen, daß V endliche Dimension habe. Dann ist erst recht jeder φ-invariante Unterraum U endlichdimensional, wir können also eine endliche Basis $\{\vec{b}_1, \ldots, \vec{b}_r\}$ von U finden und diese ergänzen zu einer Basis $\{\vec{b}_1, \ldots, \vec{b}_n\}$ von V. Da $\varphi(U) \leq U$ ist, liegen die Bilder der ersten r Basisvektoren wieder in U, d.h. die Abbildungsmatrix bezüglich dieser Basis hat die Form

$$\begin{pmatrix} \boxed{A} & \boxed{C} \\ \mathbf{0} & \boxed{B} \end{pmatrix}$$

mit einer $r \times r$-Matrix A, der Abbildungsmatrix von $\varphi_{|U}\colon U \to U$, einer $(n-r)\times(n-r)$-Matrix B und einer $(n-r)\times r$-Matrix C. Die fette Null soll hier, wie auch in den noch folgenden Matrizen, stets eine Nullmatrix der jeweils korrekten Größe bezeichnen.

Noch besser wird die Situation, wenn U ein φ-invariantes Komplement hat, wenn es also einen weiteren φ-invarianten Untervektorraum W gibt, so daß $V = U + W$ ist und $U \cap W = \{\vec{0}\}$. (Wir sagen dann, $V = U \oplus W$ sei die *direkte Summe* von U und W.) In diesem Fall können wir für $\vec{b}_{r+1}$ bis $\vec{b}_n$ die Vektoren einer Basis von W nehmen, und da nun auch W auf sich selbst abgebildet wird, haben wir eine Abbildungsmatrix der Form

$$\begin{pmatrix} \boxed{A} & \mathbf{0} \\ \mathbf{0} & \boxed{B} \end{pmatrix}.$$

Allgemein sagen wir für s Untervektorräume $U_1, \ldots, U_s$ von V, daß V die direkte Summe

$$V = U_1 \oplus \cdots \oplus U_s = \bigoplus_{i=1}^{s} U_i$$

sei, wenn

$$V = U_1 + \cdots + U_s = \sum_{i=1}^{s} U_i \quad \text{und} \quad U_i \cap \sum_{j \neq i} U_j = \{\vec{0}\}$$

ist. Falls hierbei die U_i allesamt φ-invariant sind, können wir ihre Basen aneinandersetzen und erhalten eine Basis, bezüglich derer die Abbildungsmatrix die Gestalt

$$\begin{pmatrix} \boxed{A_1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boxed{A_2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boxed{A_s} \end{pmatrix}$$

hat, wobei die A_i die Abbildungsmatrizen der Einschränkungen $\varphi_{|U_i}$ zu Abbildungen von U_i nach U_i sind.

Kandidaten für Untervektorräume U_i liefern die Haupträume:

Definition: *a)* Ein Vektor $\vec{v} \in V$ heißt *Hauptvektor* von φ zum Eigenwert λ, wenn es ein $\ell \in \mathbb{N}_0$ gibt, so daß $(\varphi - \lambda\,\mathrm{id})^{\ell}(\vec{v}) = \vec{0}$ ist. Falls $(\varphi - \lambda\,\mathrm{id})^{\ell-1} \neq \vec{0}$ ist, bezeichnen wir ℓ als die *Stufe* des Hauptvektors.

b) Die Menge aller Hauptvektoren von φ zum Eigenwert λ heißt *Hauptraum* zu λ und wird mit H_λ bezeichnet.

Insbesondere sind die Hauptvektoren der Stufe eins genau die Eigenvektoren zum Eigenwert λ: Der Nullvektor ist nämlich kein Hauptvektor erster Stufe, da er bereits von $(\varphi - \lambda\,\mathrm{id})^0 = \mathrm{id}$ auf $\vec{0}$ abgebildet wird.

Es ist klar, daß die Hauptvektoren einen Untervektorraum bilden, denn mit $(\varphi - \lambda\,\mathrm{id})$ sind auch dessen Schachtelungen

$$(\varphi - \lambda\,\mathrm{id})^\ell = (\varphi - \lambda\,\mathrm{id}) \circ \cdots \circ (\varphi - \lambda\,\mathrm{id})$$

lineare Abbildungen, und die Hauptvektoren der Stufe höchstens ℓ sind gerade die Elemente des Kerns dieser Abbildung. Da wir von einem endlichdimensionalen Vektorraum V ausgehen, kann die Folge dieser Kerne nicht unbeschränkt wachsen, es gibt also ein maximales ℓ, das als Stufe eines Hauptvektors auftreten kann. Mit diesem ℓ ist der Hauptraum H_λ gerade der Kern von $(\varphi - \lambda\,\mathrm{id})^\ell$.

Der Nutzen der Haupträume ergibt sich aus folgendem

Lemma: H_λ ist ein φ-invarianter Unterraum von V. Bezeichnet ℓ die größte Stufe eines Hauptvektors aus H_λ, so ist $\mathrm{Bild}(\varphi - \lambda\,\mathrm{id})^\ell$ ein φ-invariantes Komplement.

Beweis: Beginnen wir mit der Invarianz von H_λ unter φ.

Ist $\vec{v}$ ein Hauptvektor der Stufe j, so ist

$$\begin{aligned}(\varphi - \lambda\,\mathrm{id})^j(\vec{v}) &= (\varphi - \lambda\,\mathrm{id})^{j-1}\big((\varphi - \lambda\,\mathrm{id})(\vec{v})\big)\\ &= (\varphi - \lambda\,\mathrm{id})^{j-1}\big(\varphi(\vec{v}) - \lambda\vec{v}\big) = \vec{0}\,,\end{aligned}$$

$\varphi(\vec{v}) - \lambda\vec{v}$ ist also ein Hauptvektor der Stufe höchstens $j-1$ und somit insbesondere ein Element von H_λ. Da mit $\vec{v}$ auch $\lambda\vec{v}$ in H_λ liegt, ist damit auch $\varphi(\vec{v}) = \big(\varphi(\vec{v}) - \lambda\vec{v}\big) + \lambda\vec{v} \in H_\lambda$ ein Hauptvektor.

Die Invarianz von $\mathrm{Bild}(\varphi - \lambda\,\mathrm{id})^\ell$ folgt genauso: Für $\vec{v} = (\varphi - \lambda\,\mathrm{id})^\ell(\vec{w})$ ist

$$\varphi(\vec{v}) - \lambda\vec{v} = (\varphi - \lambda\,\mathrm{id})(\vec{v}) = (\varphi - \lambda\,\mathrm{id})^{\ell+1}(\vec{w}) = (\varphi - \lambda\,\mathrm{id})^\ell(\varphi(\vec{w}) - \lambda\vec{w})$$

wieder ein Element des Bilds und damit auch $\varphi(\vec{v})$ selbst.

Als nächstes müssen wir zeigen, daß der Durchschnitt der beiden Räume nur aus dem Nullvektor besteht. Dazu sei $\vec{v}$ ein Vektor aus diesem Durchschnitt. Dann liegt $\vec{v}$ sowohl im Kern als auch im Bild der linearen Abbildung $(\varphi - \lambda\,\mathrm{id})^\ell$, es gibt also einen Vektor $\vec{w} \in V$ derart, daß $\vec{v} = (\varphi - \lambda\,\mathrm{id})^\ell(\vec{w})$ ist, und $(\varphi - \lambda\,\mathrm{id})^\ell(\vec{v}) = (\varphi - \lambda\,\mathrm{id})^{2\ell}(\vec{w}) = \vec{0}$. Damit liegt $\vec{w}$ aber im Hauptraum zu λ, d.h. $\vec{v} = (\varphi - \lambda\,\mathrm{id})^\ell(\vec{w}) = \vec{0}$.

Nach der Dimensionsformel ist

$$\dim \mathrm{Bild}(\varphi - \lambda\,\mathrm{id})^\ell = \dim V - \dim \mathrm{Kern}(\varphi - \lambda\,\mathrm{id})^\ell\,,$$

also ist

$$\dim \mathrm{Kern}(\varphi - \lambda\,\mathrm{id})^\ell + \dim \mathrm{Bild}(\varphi - \lambda\,\mathrm{id})^\ell = \dim V\,,$$

die beiden Untervektorräume erzeugen somit ganz V. ∎

Die Zerlegung von V nach diesem Lemma heißt FITTING-Zerlegung.

Der deutsche Mathematiker HANS FITTING (1906–1938) beschäftigte sich vor allem mit der Untersuchung von Operatoren (und Operatorenringen). Trotz seines frühen Todes konnte er damit wesentliche Beiträge zur Algebra leisten, vor allem auch zur Erforschung der Struktur von Gruppen.

Das Schöne an der FITTING-Zerlegung ist, daß sie rekursiv fortgesetzt werden kann: Da $\mathrm{Bild}(\varphi - \lambda\,\mathrm{id})^\ell$ auch φ-invariant ist, können wir für die Einschränkung von φ auf diesen Unterraum einen Hauptraum zu einem anderen Eigenwert abspalten *usw.* Bevor wir uns das genauer überlegen, wollen wir uns aber zunächst eine gute Basis für den Hauptraum H_λ verschaffen.

Lemma: Der Hauptraum H_λ hat eine Basis, bezüglich derer die Abbildungsmatrix von φ eine obere Dreiecksmatrix ist. Alle Hauptdiagonaleinträge dieser Matrix sind gleich λ.

Beweis: Wir beginnen mit einer Basis $\{\vec{b}_1, \ldots, \vec{b}_{r_1}\}$ des Eigenraums zu λ und ergänzen diese zu einer Basis $\{\vec{b}_1, \ldots, \vec{b}_{r_2}\}$ des Raums aller Hauptvektoren der Stufe höchstens zwei und so weiter, bis eine Basis $\{\vec{b}_1, \ldots, \vec{b}_r\}$ des gesamten Hauptraums erreicht ist.

Der Vektor $\vec{b}_i$ sei Hauptvektor der Stufe ℓ; dann ist

$$\begin{aligned}(\varphi - \lambda\,\mathrm{id})^{\ell}(\vec{b}_i) &= (\varphi - \lambda\,\mathrm{id})^{\ell-1}\big((\varphi - \lambda\,\mathrm{id})(\vec{b}_i)\big)\\ &= (\varphi - \lambda\,\mathrm{id})^{\ell-1}\big(\varphi(\vec{b}_i) - \lambda\vec{b}_i\big) = \vec{0}\,,\end{aligned}$$

$\varphi(\vec{b}_i) - \lambda\vec{b}_i$ ist also ein Hauptvektor der Stufe höchstens $\ell - 1$. Nach Konstruktion der Basis ist $\varphi(\vec{b}_i) - \lambda\vec{b}_i$ daher eine Linearkombination von Basisvektoren $\vec{b}_j$ mit Indizes echt kleiner i, d.h.

$$\varphi(\vec{b}_i) = \lambda\vec{b}_i + \sum_{j=1}^{i-1} \alpha_{ij}\vec{b}_j\,.$$

Bezüglich der Basis $\{\vec{b}_1, \ldots, \vec{b}_r\}$ hat die Abbildungsmatrix daher in der Tat die gewünschte Form. ∎

Abspaltung immer weiterer Haupträume auch vom invarianten Komplement führt schließlich zum

Satz: Zu einer linearen Abbildung $\varphi\colon V \to V$ eines endlichdimensionalen k-Vektorraums V gibt es genau dann eine Basis von V, bezüglich derer die Abbildungsmatrix von φ eine Dreiecksmatrix ist, wenn das charakteristische Polynom von φ über k als Produkt von Linearfaktoren geschrieben werden kann. Alsdann kann die Basis so gewählt werden, daß die Abbildungsmatrix die Form

$$\begin{pmatrix} \boxed{A_1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boxed{A_2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boxed{A_s} \end{pmatrix}$$

hat mit Dreiecksmatrizen A_i, die auf der Hauptdiagonalen den i-ten Eigenwert λ_i stehen haben.

Beweis: Falls es zu φ eine Basis gibt, bezüglich derer die Abbildungsmatrix A von φ eine Dreiecksmatrix ist, ist bezüglich dieser Basis auch

$A - \lambda E$ eine Dreiecksmatrix. Da die Determinante einer Dreiecksmatrix gerade das Produkt der Diagonaleinträge ist, bekommen wir als charakteristisches Polynom $\det(A - \lambda E)$ ein Produkt von Linearfaktoren.

Falls umgekehrt das charakteristische Polynom in Linearfaktoren zerfällt, gibt es auf jeden Fall Eigenwerte; λ_1 sei einer davon, und H_1 sei der zugehörige Hauptraum. Dazu gibt es nach dem gerade bewiesenen Lemma eine Basis, bezüglich derer die Abbildungsmatrix von $\varphi_{|H_{\lambda_1}}$ eine obere Dreiecksmatrix ist, deren sämtliche Hauptdiagonaleinträge gleich λ_1 sind.

Nach dem Lemma von der FITTING-Zerlegung gibt es zu H_{λ_1} ein φ-invariantes Komplement V_1, so daß $V = H_{\lambda_1} \oplus V_1$ ist. Wir ergänzen die Basis von H_{λ_1} durch eine Basis von V_1 zu einer Basis von V; bezüglich dieser Basis hat φ dann eine Abbildungsmatrix der Form

$$A_1 = \begin{pmatrix} \boxed{D_1} & \mathbf{0} \\ \mathbf{0} & \boxed{B_1} \end{pmatrix}$$

mit einer oberen Dreiecksmatrix

$$D_1 = \begin{pmatrix} \lambda_1 & * & \ldots & * \\ 0 & \lambda_1 & \ldots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_1 \end{pmatrix}.$$

Entwicklung des charakteristischen Polynoms $\det(A - \lambda E)$ nach der ersten Spalte, gefolgt von der Entwicklung des Rests nach seiner ersten Spalte und so weiter, bis die ersten $r_1 = \dim H_{\lambda_1}$ Spalten aufgebraucht sind, zeigt, daß

$$\det(A_1 - \lambda E) = (\lambda_1 - \lambda)^{r_1} \cdot \det(B_1 - \lambda E)$$

ist, das charakteristische Polynom von $\varphi_{|V_1}$ ist also ein Teiler des charakteristischen Polynoms von φ und zerfällt somit auch in Linearfaktoren.

Insbesondere hat es (mindestens) eine Nullstelle λ_2; wir können deren Hauptraum H_{λ_2} in V_1 betrachten und damit V_1 genau wie oben weiter zerlegen in H_{λ_2} und dessen invariantes Komplement V_2. Nimmt man

nun als Basisvektoren von V zunächst die Basisvektoren von H_{λ_1} wie oben, dann entsprechende Basisvektoren für H_{λ_2} und schließlich noch solche für V_2, hat die Abbildungsmatrix A_2 bezüglich dieser neuen Basis die Form

$$A_2 = \begin{pmatrix} \boxed{D_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boxed{D_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boxed{B_2} \end{pmatrix}$$

mit einer neuen Dreiecksmatrix

$$D_2 = \begin{pmatrix} \lambda_2 & * & \dots & * \\ 0 & \lambda_2 & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_2 \end{pmatrix}.$$

Auf diese Weise lassen sich sukzessive immer weitere Haupträume abspalten, bis schließlich eine Basis erreicht ist, bezüglich derer die Abbildungsmatrix von φ die Form

$$A = \begin{pmatrix} \boxed{D_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boxed{D_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boxed{D_s} \end{pmatrix}$$

hat mit oberen Dreiecksmatrizen

$$D_i = \begin{pmatrix} \lambda_i & * & \dots & * \\ 0 & \lambda_i & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_i \end{pmatrix}$$

zu den Eigenwerten von φ. Ist D_i eine $r_i \times r_i$-Matrix, so ist das charakteristische Polynom von φ

$$\det(A - \lambda E) = (\lambda_1 - \lambda)^{r_1}(\lambda_2 - \lambda)^{r_2} \cdots (\lambda_1 - \lambda)^{r_s},$$

die r_i sind also gerade die algebraischen Vielfachheiten der λ_i. ∎

Für spätere Anwendungen wollen wir das gerade bewiesene Ergebnis noch etwas umformulieren:

Satz: Falls das charakteristische Polynom von $\varphi: V \to V$ in Linearfaktoren zerfällt, gibt es eine Basis von V, bezüglich derer die Abbildungsmatrix A von φ als $A = D + N$ geschrieben werden kann, wobei D eine Diagonalmatrix ist und N eine obere Dreiecksmatrix mit Nullen in der Hauptdiagonalen. Außerdem ist $DN = ND$.

Beweis: Wir nehmen natürlich die Basis aus dem gerade beendeten Beweis; die Diagonalmatrix D soll genau aus den Diagonalelementen der Abbildungsmatrix A bestehen, also die Eigenwerte entsprechend ihrer algebraischen Vielfachheiten als Diagonalelemente enthalten, und $N = A - D$. Für jede einzelne Dreiecksmatrix D_i aus dem obigen Beweis kommutiert der Diagonalanteil mit dem Rest, da der Diagonalanteil gerade das λ_i-fache der Einheitsmatrix ist. Damit ist auch $DN = ND$, denn bei beiden Multiplikationen treffen, abgesehen von den Nullen, immer nur Einträge aus einem D_i aufeinander. ∎

Diese Zerlegung aus diesem Satz bezeichnet man nach dem französischen Mathematiker CAMILLE JORDAN als JORDAN-Zerlegung.

MARIE ENNEMOND CAMILLE JORDAN (1838–1922) arbeitete bei der Herleitung dieser und weiterer Zerlegungen nicht mit komplexen Matrizen, sondern mit Matrizen über endlichen Körpern, motiviert durch Fragen aus der Gruppentheorie und Lösbarkeitsfragen für nichtlineare Gleichungen. Weitere Arbeiten beschäftigen sich mit der Anwendung gruppentheoretischer Methoden auf die Geometrie sowie mit der Topologie, wo er z.B. bewies, daß jede doppeltpunktfreie geschlossene Kurve die Ebene in zwei Gebiete zerlegt. Außerdem entwickelte er neue Methoden zum Nachweis der Konvergenz von FOURIER-Reihen.

Ziel unserer Betrachtungen in diesem Paragraphen war die Berechnung von Potenzen und Exponentialfunktionen einer Matrix. Mit der JORDAN-Zerlegung ist dies im wesentlichen erreicht: Da D und N miteinander

kommutieren, gilt für Potenzen der Summe $D + N$ der „übliche“ binomische Lehrsatz, d.h.

$$(D+N)^m = \sum_{\ell=0}^{m} \binom{m}{\ell} D^{m-\ell} N^\ell ,$$

und

$$e^{D+N} = e^D \cdot e^N .$$

Die Potenzen von D sind sehr einfach zu berechnen: D^j ist wieder eine Diagonalmatrix, ihre Diagonalelemente sind die j-ten Potenzen der Diagonalelemente von D; genauso ist e^D einfach die Diagonalmatrix mit den Exponentialfunktionen der Einträge von D als Einträgen.

N ist eine obere Dreiecksmatrix mit Nullen in der Hauptdiagonalen, wir wissen also bereits, daß es einen Exponenten gibt, ab dem alle Potenzen gleich der Nullmatrix sind, so daß die Exponentialreihe zu einer endlichen Summe wird und auch in der binomischen Formel selbst für große m nur relativ wenige Summanden auftreten.

Mit der JORDAN-Zerlegung können wir diese Aussage nun noch etwas präzisieren: Die lineare Abbildung ψ zu N bildet den i-ten Basisvektor $\vec{b}_i$ ab in den von Basisvektoren $\vec{b}_j$ mit $j \le i-1$ erzeugten Unterraum. Für diese Basisvektoren gilt eine analoge Aussage, $\psi^{(2)}(\vec{b}_i)$ liegt daher im Unterraum, den die $\vec{b}_j$ mit $j \le i-2$ aufspannen. Induktiv folgt, daß $\psi^{(\ell)}(\vec{b}_i)$ im von den $\vec{b}_j$ mit $j \le i-\ell$ aufgespannten Untervektorraum liegt; falls $i-\ell$ negativ wird, ist das natürlich der Nullraum.

Die Abbildungsmatrix N^ℓ von $\psi^{(\ell)}$ ist daher ebenfalls eine obere Dreiecksmatrix mit Nullen in der Hauptdiagonalen; zusätzlich stehen auch noch in den $\ell-1$ schrägen Reihen oberhalb und parallel zur Hauptdiagonale lauter Nullen, und spätestens wenn ℓ größer oder gleich der größten Stufe eines Hauptvektors wird, ist N^ℓ gleich der Nullmatrix.

Als Beispiel betrachten wir die Matrix

$$A = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} ,$$

die offensichtlich von der oben betrachteten Form ist; hier ist

$$D = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{und} \quad N = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 3 \\ 0 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} .$$

Eine kurze Rechnung zeigt, daß

$$N^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{und} \quad N^3 = 0$$

ist, also ist beispielsweise

$$A^{10} = D^{10} + 10\, D^9 N + 45\, D^8 N^2 = \begin{pmatrix} 1024 & 5120 & 0 & 0 & 0 \\ 0 & 1024 & 0 & 0 & 0 \\ 0 & 0 & 1 & 20 & 390 \\ 0 & 0 & 0 & 1 & 40 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} .$$

Mehr als Potenzen interessiert uns die Exponentialfunktion einer Matrix; auch diese läßt sich über die JORDAN-Zerlegung berechnen: Da D und N kommutieren, ist $e^A = e^{D+N} = e^D \cdot e^N$ mit

$$e^D = \begin{pmatrix} e^2 & 0 & 0 & 0 & 0 \\ 0 & e^2 & 0 & 0 & 0 \\ 0 & 0 & e & 0 & 0 \\ 0 & 0 & 0 & e & 0 \\ 0 & 0 & 0 & 0 & e \end{pmatrix}$$

und

$$e^N = E + N + \frac{1}{2} N^2 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 7 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} ,$$

also ist

$$e^A = e^D \cdot e^N = \begin{pmatrix} e^2 & e^2 & 0 & 0 & 0 \\ 0 & e^2 & 0 & 0 & 0 \\ 0 & 0 & e & 2e & 7e \\ 0 & 0 & 0 & e & 4e \\ 0 & 0 & 0 & 0 & e \end{pmatrix} .$$

Entsprechend läßt sich auch e^{At} berechnen:

$$e^{Nt} = E + Nt + \frac{1}{2}N^2t^2 = \begin{pmatrix} 1 & t & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2t & 3t + 4t^2 \\ 0 & 0 & 0 & 1 & 4t \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} ,$$

und da auch Dt und Nt kommutieren, ist

$$e^{At} = e^{Dt} \cdot e^{Nt} = \begin{pmatrix} e^{2t} & te^{2t} & 0 & 0 & 0 \\ 0 & e^{2t} & 0 & 0 & 0 \\ 0 & 0 & e^t & 2te^t & 3te^t + 4t^2e^t \\ 0 & 0 & 0 & e^t & 4te^t \\ 0 & 0 & 0 & 0 & e^t \end{pmatrix} .$$

Zur Vorsicht sei noch einmal ausdrücklich darauf hingewiesen, daß es für diese Rechnungen sehr wesentlich war, daß D und N miteinander kommutieren; es reicht nicht, wenn wir die Matrix A nur auf *irgendeine* Dreiecksgestalt bringen und dann als Summe einer Diagonalmatrix und einer oberen Dreiecksmatrix mit Nullen in der Hauptdiagonalen schreiben. Für

$$A = \begin{pmatrix} 1 & 3 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 0 & 3 \\ 0 & 0 \end{pmatrix} \underset{\text{def}}{=} D + N$$

beispielsweise ist

$$\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 0 & 3 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 3 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 6 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 3 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} ,$$

und in der Tat ist

$$D^2 + 2DN + N^2 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} + \begin{pmatrix} 0 & 6 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 6 \\ 0 & 4 \end{pmatrix}$$

verschieden von

$$A^2 = \begin{pmatrix} 1 & 9 \\ 0 & 4 \end{pmatrix} ,$$

und genauso ist

$$e^A = \begin{pmatrix} e & 3e^2 - 3e \\ 0 & e^2 \end{pmatrix} \neq e^D \cdot e^N = \begin{pmatrix} e & 3e \\ 0 & e^2 \end{pmatrix} .$$

Der Grund für die Verschiedenheit der Ergebnisse beim Quadrat liegt natürlich darin, daß wir im allgemeinen nur sagen können, daß

$$(D + N)^2 = D(D + N) + N(D + N) = D^2 + DN + ND + N^2$$

ist, aber wir können $DN + ND$ nicht zusammenfassen zu $2DN$. Mit wachsendem Exponenten verschlimmert sich die Situation drastisch; schon

$$(D + N)^3 = D^3 + D^2N + DND + ND^2 + DN^2 + NDN + N^2D + N^3$$

hat acht Summanden; die m-te Potenz hat 2^m, und von denen überleben viele auch dann, wenn N^r schon für relativ kleine r verschwindet. Für die Matrixexponentialfunktion, in die alle Potenzen eingehen, ist also ziemlich klar, daß es für nichtkommutierende Matrizen D und N keinen vernünftigen Zusammenhang zwischen e^{D+N} und $e^D \cdot e^N$ geben kann.

g) Ein Beispiel

Wir haben Eigenvektoren und Hauptvektoren in erster Linie eingeführt, um Differentialgleichungen zu lösen; daher soll das etwas ausführlichere Beispiel in diesem Abschnitt ebenfalls mit einer Differentialgleichung beginnen: Gesucht sind die Lösungen des Differentialgleichungssystems

$$\begin{aligned} \dot{x}(t) &= 2x(t) - \ y(t) - z(t) \\ \dot{y}(t) &= \ x(t) + 5y(t) + 2z(t) \\ \dot{z}(t) &= -x(t) - 2y(t) + z(t) . \end{aligned}$$

Hier ist

$$A = \begin{pmatrix} 2 & -1 & -1 \\ 1 & 5 & 2 \\ -1 & -2 & 1 \end{pmatrix} ,$$

und das charakteristische Polynom von A ist

$$\det(A-\lambda E) = -\lambda^3 + 8\lambda^2 - 21\lambda + 18 = -(\lambda-2)(\lambda-3)^2 .$$

Wir haben also den Eigenwert zwei mit algebraischer und somit auch geometrischer Vielfachheit eins und den Eigenwert drei mit algebraischer Vielfachheit zwei. In der Matrix

$$A - 2E = \begin{pmatrix} 0 & -1 & -1 \\ 1 & 3 & 2 \\ -1 & -2 & -1 \end{pmatrix}$$

ist die mittlere Spalte gleich der Summe der beiden äußeren,

$$\vec{v}_1 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

erzeugt also den Eigenraum. In

$$A - 3E = \begin{pmatrix} -1 & -1 & -1 \\ 1 & 2 & 2 \\ -1 & -2 & -2 \end{pmatrix}$$

stimmen die zweite und die dritte Spalte miteinander überein,

$$\vec{v}_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

ist also ein Eigenvektor und erzeugt auch den Eigenraum, denn da die erste Spalte kein Vielfaches der zweiten ist, hat die Matrix den Rang zwei. Die geometrische Vielfachheit des Eigenwerts drei ist also nur eins: Um zu einer Dreiecksmatrix zu kommen, müssen wir einen Hauptvektor zweiter Stufe berechnen. Aus

$$(A-3E)^2 = \begin{pmatrix} 1 & 1 & 1 \\ -1 & -1 & -1 \\ 1 & 1 & 1 \end{pmatrix}$$

sieht man, daß sich

$$\vec{v}_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$$

als von $\vec{v}_2$ linear unabhängiger Kandidat anbietet. Da

$$A\vec{v}_3 = \begin{pmatrix} 3 \\ -4 \\ 1 \end{pmatrix} = 3\vec{v}_3 - \vec{v}_2$$

ist, hat A bezüglich der Basis $\vec{v}_1, \vec{v}_2, \vec{v}_3$ die Form

$$M = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & -1 \\ 0 & 0 & 3 \end{pmatrix} ,$$

der erste „Kasten“ ist also einfach eine 1×1-Matrix und der zweite ist

$$\begin{pmatrix} 3 & -1 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} .$$

Da das Quadrat des zweiten Summanden verschwindet, ist

$$e^{\begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} t} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} t = \begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix}$$

und

$$e^{\begin{pmatrix} 3 & -1 \\ 0 & 3 \end{pmatrix} t} = e^{3t} \begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} e^{3t} & -te^{3t} \\ 0 & e^{3t} \end{pmatrix} .$$

Damit ist

$$e^{Mt} = \begin{pmatrix} e^{2t} & 0 & 0 \\ 0 & e^{3t} & -te^{3t} \\ 0 & 0 & e^{3t} \end{pmatrix} .$$

Um daraus e^{At} zu berechnen, müssen wir die Standardbasis des $\mathbb{R}^3$ durch die Hauptvektoren ausdrücken; man überzeugt sich leicht, daß

$$\vec{e}_1 = \vec{v}_1 + \vec{v}_2, \quad \vec{e}_2 = \vec{v}_1 + \vec{v}_2 - \vec{v}_3 \quad \text{und} \quad \vec{e}_3 = \vec{v}_1 - \vec{v}_3$$

ist. Bezüglich der Basis $\vec{v}_1, \vec{v}_2, \vec{v}_3$ ist also

$$e^{At}\vec{e}_1 = \begin{pmatrix} e^{2t} & 0 & 0 \\ 0 & e^{3t} & -te^{3t} \\ 0 & 0 & e^{3t} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} e^{2t} \\ e^{3t} \\ 0 \end{pmatrix} ,$$

was bezüglich der Standardbasis der Vektor

$$e^{2t}\vec{v}_1 + e^{3t}\vec{v}_2 = \begin{pmatrix} e^{2t} \\ -e^{2t} \\ e^{2t} \end{pmatrix} + \begin{pmatrix} 0 \\ e^{3t} \\ -e^{3t} \end{pmatrix} = \begin{pmatrix} e^{2t} \\ e^{3t} - e^{2t} \\ e^{2t} - e^{3t} \end{pmatrix} .$$

Also ist, bezüglich der Standardbasis ausgedrückt,

$$e^{At}\begin{pmatrix}1\\0\\0\end{pmatrix} = \begin{pmatrix}e^{2t}\\ e^{3t}-e^{2t}\\ e^{2t}-e^{3t}\end{pmatrix}.$$

Genauso überlegt man sich, daß

$$e^{At}\begin{pmatrix}0\\1\\0\end{pmatrix} = e^{2t}\vec{v}_1 + \big(e^{3t}+te^{3t}\big)\vec{v}_2 - e^{3t}\vec{v}_3 = \begin{pmatrix}e^{2t}-e^{3t}\\ 2e^{3t}-e^{2t}+te^{3t}\\ e^{2t}-e^{3t}-te^{3t}\end{pmatrix}$$

und

$$e^{At}\begin{pmatrix}0\\0\\1\end{pmatrix} = \begin{pmatrix}e^{2t}-e^{3t}\\ e^{3t}-e^{2t}+te^{3t}\\ e^{2t}-te^{3t}\end{pmatrix}$$

ist. Da in den Spalten einer Matrix die Bilder der Basisvektoren stehen, ist somit

$$e^{At} = \begin{pmatrix}e^{2t} & e^{2t}-e^{3t} & e^{2t}-e^{3t}\\ e^{3t}-e^{2t} & 2e^{3t}-e^{2t}+te^{3t} & e^{3t}-e^{2t}+te^{3t}\\ e^{2t}-e^{3t} & e^{2t}-e^{3t}-te^{3t} & e^{2t}-te^{3t}\end{pmatrix}.$$

Die Lösung, die den Anfangsbedingungen

$$x(0)=x_0,\quad y(0)=y_0\quad\text{und}\quad z(0)=z_0$$

genügt, ist also

$$\begin{pmatrix}x(t)\\y(t)\\z(t)\end{pmatrix} = \begin{pmatrix}e^{2t} & e^{2t}-e^{3t} & e^{2t}-e^{3t}\\ e^{3t}-e^{2t} & 2e^{3t}-e^{2t}+te^{3t} & e^{3t}-e^{2t}+te^{3t}\\ e^{2t}-e^{3t} & e^{2t}-e^{3t}-te^{3t} & e^{2t}-te^{3t}\end{pmatrix}\begin{pmatrix}x_0\\y_0\\z_0\end{pmatrix}$$
$$= \begin{pmatrix}(x_0+y_0+z_0)e^{2t}-(y_0+z_0)e^{3t}\\ \big(x_0+(t+2)y_0+(t+1)z_0\big)e^{3t}-(x_0+y_0+z_0)e^{2t}\\ (x_0+y_0+z_0)e^{2t}-\big(x_0+(t+1)y_0+tz_0\big)e^{3t}\end{pmatrix}.$$

h) Ergänzung: Die Jordan-Normalform

Die im vorigen Abschnitt konstruierte Normalform für Abbildungsmatrizen wird für alle Zwecke dieser Vorlesung ausreichen. Trotzdem ist sie nicht ganz befriedigend, da die Dreiecksmatrizen immer noch sehr willkürlich und damit komplizierter als notwendig sind. Für Interessenten sei in diesem Abschnitt gezeigt, wie sich die bislang erreichte

Dreiecksgestalt noch weiter vereinfachen läßt, indem man die bislang noch ziemlich willkürlichen Basen der Haupträume etwas geschickter wählt. Für das folgende werden wir die Ergebnisse dieses Abschnitts nicht benötigen; er kann also gefahrlos überlesen werden.

Die Potenzen einer oberen Dreiecksmatrix mit Nullen in der Hauptdiagonalen verschwinden, wie wir gesehen haben, ab einem meist überschaubar kleinen Exponenten, aber die Potenzen bis dahin muß man doch mühsam von Hand ausrechnen. Eine Ausnahme, bei der alles klar ist, bilden Matrizen der Form

$$N = \begin{pmatrix}0&1&0&\dots&0&0\\ 0&0&1&\dots&0&0\\ \vdots&\vdots&&\ddots&\vdots&\vdots\\ 0&0&0&\dots&0&1\\ 0&0&0&\dots&0&0\end{pmatrix},$$

bei denen direkt oberhalb der Hauptdiagonale lauter Einsen stehen, während alle anderen Einträge verschwinden, d.h.

$$N=(n_{ij})\quad\text{mit}\quad n_{ij}=\begin{cases}1 & \text{falls } j-i=1\\ 0 & \text{sonst}\end{cases}.$$

Bei der sukzessiven Potenzierung von N verschiebt sich einfach die Reihe von Einsen jeweils um eins weiter nach außen, d.h.

$$N^\ell=(n_{ij}^{(\ell)})\quad\text{mit}\quad n_{ij}^{(\ell)}=\begin{cases}1 & \text{falls } j-i=\ell\\ 0 & \text{sonst}\end{cases},$$

denn die zu N gehörige lineare Abbildung ψ bildet einfach den i-ten Basisvektor auf den $(i-1)$-ten ab oder auf den Nullvektor, falls es keinen $(i-1)$-ten Basisvektor mehr gibt, und entsprechend ist $\psi^{(\ell)}(\vec{b}_i)=\vec{b}_{i-\ell}$ beziehungsweise $\vec{0}$.

In diesem speziellen Fall sind die Potenzen von N also ohne jeden Aufwand zu berechnen, und tatsächlich genügen solche Matrizen N schon vollständig für eine Normalform der Abbildungsmatrix, die sogenannte JORDAN-Normalform:

Satz: Falls das charakteristische Polynom von $\varphi\colon V\to V$ als Produkt von Linearfaktoren geschrieben werden kann, gibt es eine Basis von V,

bezüglich derer die Abbildungsmatrix von φ die Form

$$A = \begin{pmatrix} \boxed{J_1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boxed{J_2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boxed{J_s} \end{pmatrix}$$

hat mit oberen Dreiecksmatrizen

$$J_i = \begin{pmatrix} \lambda_i & 1 & 0 & \dots & 0 \\ 0 & \lambda_i & 1 & \dots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & \lambda_i \end{pmatrix}$$

zu den Eigenwerten von φ. Die Anzahl der Kästchen J_i zu einem festen Eigenwert ist die geometrische Vielfachheit dieses Eigenwerts, die Summe ihrer Zeilenzahlen die algebraische.

Beweis: Wir gehen aus von der Zerlegung von V in die Haupträume zu den Eigenwerten von φ und betrachten einen festen Hauptraum H_λ. Die Einschränkung von φ auf diesen Untervektorraum läßt sich zerlegen in eine Summe

$$\varphi_{|H_\lambda} = \lambda\,\mathrm{id} + \psi\,;$$

dabei ist die Abbildungsmatrix von λ id bezüglich jeder beliebigen Basis gleich dem λ-fachen der Einheitsmatrix, und zumindest bezüglich der im vorigen Abschnitt konstruierten Basis $\{\vec{b}_1, \dots, \vec{b}_r\}$ ist die Abbildungsmatrix von ψ eine obere Dreiecksmatrix N mit Nullen in der Hauptdiagonalen.

ψ bildet den Basisvektor $\vec{b}_i$ daher ab in das Erzeugnis der Basisvektoren $\vec{b}_1$ bis $\vec{b}_{i-1}$; insbesondere geht $\vec{b}_1$ auf den Nullvektor. Wiederholte Anwendung von ψ zeigt, daß für jeden Basisvektor $\vec{b}_i$ gilt: $\psi^{(i-1)}(\vec{b}_i) = \vec{0}$, wobei der Exponent von ψ für die wiederholte Anwendung der Abbildung stehen soll. Insbesondere ist also $\psi^{(r)}(\vec{v}) = \vec{0}$ für alle $\vec{v} \in H_\lambda$.

Es könnte sein, daß es schon eine kleinere Zahl s gibt, so daß $\psi^{(s)}$ die Nullabbildung ist; die kleinste solche Zahl bezeichnen wir als den *Nilpotenzgrad* von ψ.

Hat der Nilpotenzgrad seinen größtmöglichen Wert r, so sind die Vektoren

$$\vec{b}_r,\ \psi(\vec{b}_r),\ \dots,\ \psi^{(s-1)}(\vec{b}_r)$$

allesamt ungleich dem Nullvektor. Sie sind auch linear unabhängig, denn ist

$$\alpha_0\vec{b}_r + \alpha_1\psi(\vec{b}_r) + \dots + \alpha_{r-1}\psi^{(r-1)}(\vec{b}_r) = \vec{0}\,,$$

so ist auch für jedes j

$$\begin{aligned}&\psi^{(j)}\big(\alpha_0\vec{b}_s + \alpha_1\psi(\vec{b}_s) + \dots + \alpha_{r-1}\psi^{(r-1)}(\vec{b}_r)\big)\\ &= \alpha_0\psi^{(j)}(\vec{b}_r) + \alpha_1\psi^{(j+1)}(\vec{b}_r) + \dots + \alpha_{r-1}\psi^{(j+r-1)}(\vec{b}_r)\big) = \vec{0}\,.\end{aligned}$$

Da $\psi^{(s)}$ für $s \geq r$ die Nullabbildung ist, treten hier nur die Summanden $\alpha_i\psi^{(i+j)}(\vec{b}_s)$ mit $i < r - j$ wirklich auf, für $j = r - 1$ also nur der Summand $\alpha_0\psi^{(r-1)}(\vec{b}_r)$. Da $\psi^{(r-1)}(\vec{b}_r)$ ungleich dem Nullvektor ist, muß also $\alpha_0 = 0$ sein. Anwendung von $\psi^{(r-2)}$ zeigt als nächstes, daß $\alpha_1 = 0$ ist, und genauso zeigt man sukzessive das Verschwinden aller α_i. Also können wir

$$\vec{c}_1 = \psi^{(r-1)}(\vec{b}_r), \quad \vec{c}_2 = \psi^{(r-2)}(\vec{b}_r), \quad \dots, \quad \vec{c}_r = \vec{b}_r$$

als Basisvektoren von H_λ wählen, und bezüglich dieser Basis ist

$$\psi(\vec{c}_i) = \begin{cases} c_{i-1} & \text{für } i > 1 \\ \vec{0} & \text{für } i = 1 \end{cases} \quad \text{und} \quad \varphi(\vec{c}_i) = \begin{cases} \lambda\vec{c}_i + \vec{c}_{i-1} & \text{für } i > 1 \\ \lambda\vec{c}_1 & \text{für } i = 1 \end{cases}.$$

Die Abbildungsmatrix von φ hat somit die einfache Gestalt

$$\begin{pmatrix} \lambda & 1 & 0 & \dots & 0 & 0 \\ 0 & \lambda & 1 & \dots & 0 & 0 \\ 0 & 0 & \lambda & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda & 1 \\ 0 & 0 & 0 & \dots & 0 & \lambda \end{pmatrix},$$

und das ist gerade eines der JORDAN-Kästchen aus der Formulierung des Satzes.

Falls der Nilpotenzgrad s von ψ kleiner als r ist, können wir nicht so argumentieren. Wir können aber immerhin einen Vektor $\vec{v} \in H_\lambda$ finden, so daß $\psi^{(s-1)}(\vec{v}) \neq \vec{0}$ ist, denn erst $\psi^{(s)}$ ist die Nullabbildung. Genau wie oben folgt, daß

$$\vec{c}_1 = \psi^{(s-1)}(\vec{v}), \quad \vec{c}_2 = \psi^{(s-2)}(\vec{v}), \quad \ldots, \quad \vec{c}_s = \vec{v}$$

linear unabhängig sind, allerdings spannen sie nur einen s-dimensionalen Teilraum U von H_λ auf. Dieser Teilraum ist ψ-invariant und damit auch φ-invariant, denn ψ bildet einfach die Basisvektoren aufeinander beziehungsweise auf den Nullvektor ab, und die Abbildungsmatrizen bezüglich dieser Basis sehen genauso aus wie oben; auch zu U gehört also ein JORDAN-Kästchen.

Um weitere Kästchen zu bekommen, brauchen wir ein invariantes Komplement von U in H_λ. Dazu wählen wir irgendeine lineare Abbildung $\omega\colon V \to k$, für die $\omega(\vec{v}) \neq 0$ ist und setzen

$$W = \left\{ \vec{w} \in H_\lambda \;\middle|\; \omega(\vec{w}) = \omega\big(\psi(\vec{w})\big) = \cdots = \omega\big(\psi^{(s-1)}(\vec{w})\big) = 0i \right\}.$$

Der Durchschnitt $U \cap W$ besteht nur aus dem Nullvektor, denn jeder Vektor aus U läßt sich als

$$\vec{w} = \alpha_1\vec{c}_1 + \ldots + \alpha_s\vec{c}_s$$

schreiben, und wenn $\vec{w}$ auch in W liegt, ist

$$\omega\big(\psi^{(j)}(\vec{w})\big) = \alpha_1\psi^{(j+s-1)}(\vec{v}) + \ldots + \alpha_{s-1}\psi^{j+1}(\vec{v}) + \alpha_s\psi^{(j)}(\vec{v}) = 0$$

für $j = 0, \ldots, s-1$. Da $\psi^{(\ell)}(\vec{v})$ für $\ell \geq s$ gleich dem Nullvektor ist, folgt für $j = s-1$, daß $\alpha_{s-1} = 0$ ist, und erniedrigt man j immer weiter, folgt nacheinander das Verschwinden aller Koeffizienten α_i. Somit ist $U \cap W$ in der Tat der Nullraum.

Die Dimension von W läßt sich zumindest nach unten leicht abschätzen: Bezüglich einer Basis von H_λ wird jede Gleichung $\omega\big(\psi^{(j)}(\vec{w})\big) = 0$ zu einer linearen Gleichung in den Koeffizienten von $\vec{w}$, der Untervektorraum W ist also die Lösungsmenge eines homogenen linearen Gleichungssystems aus s Gleichungen in $\dim H_\lambda$ Variablen. Daher ist

$$\dim W \geq \dim H_\lambda - s \quad \text{und} \quad \dim U \oplus W = \dim U + \dim W \geq \dim H_\lambda\,.$$

Da $U \oplus W$ Untervektorraum von H_λ ist, geht das nur, wenn das Gleichheitszeichen gilt, d.h. $H_\lambda = U \oplus W$.

Wir müssen uns noch überlegen, daß W unter ψ invariant ist. Dazu müssen wir zeigen, daß für alle $\vec{w} \in W$ gilt

$$\omega\big(\psi(\vec{w})\big) = \omega\Big(\psi\big(\big(\psi(\vec{w})\big)\big)\Big) = \cdots = \omega\Big(\psi^{(s-1)}\big((\vec{w})\big)\Big) = 0\,,$$

d.h.

$$\omega\big(\psi(\vec{w})\big) = \omega\big(\psi^{(2)}(\vec{w})\big) = \cdots = \omega\big(\psi^{(s)}(\vec{w})\big) = 0$$

falls

$$\omega(\vec{w}) = \omega\big(\psi(\vec{w})\big) = \cdots = \omega\big(\psi^{(s-1)}(\vec{w})\big) = 0$$

ist. Die einzige neue Bedingung ist $\omega\big(\psi^{(s)}(\vec{w})\big) = 0$, und die ist trivialerweise erfüllt, da $\psi^{(s)}$ die Nullabbildung ist. Also ist W invariant unter ψ und somit ein invariantes Komplement von U.

Auch $\psi_{|W}$ ist eine nilpotente Abbildung von einem Nilpotenzgrad $s' \leq s$; wenn wir also einen Vektor $\vec{w} \in W$ hernehmen, für den $\psi^{(s')}(\vec{w}) \neq \vec{0}$ ist, können wir die gleiche Kontruktion wie oben mit $\vec{v}$ noch einmal durchführen und erhalten einen neuen invarianten Unterraum $U' \leq W$ mit einer Basis, bezüglich derer φ ein JORDAN-Kästchen als Abbildungsmatrix hat.

Falls $U' = W$ ist, sind wir damit fertig; anderfalls können wir wieder wie oben ein invariantes Komplement W' von U' in W finden und einen weiteren Teilraum abspalten, *usw.* Jeder solche Teilraum führt auf ein JORDAN-Kästchen, und das Verfahren bricht schließlich ab, da wir in einem endlichdimensionalen Vektorraum arbeiten.

In jedem der konstruierten Teilräume liegt genau ein eindimensionaler Teilraum aus Eigenvektoren (und dem Nullvektor), nämlich der vom ersten Basisvektor aufgespannte. Der Eigenraum zu λ wird also von diesen ersten Basisvektoren aufgespannt und seine Dimension, die geometrische Vielfachheit von λ, ist damit gleich der Anzahl der JORDAN-Kästchen zu λ. Die algebraische Vielfachheit ist wegen der speziellen Gestalt der Abbildungsmatrix natürlich die Anzahl der λ

in der Hauptdiagonalen, d.h. gleich der Summe der Zeilenzahlen der JORDAN-Kästchen zu λ. ∎

Um wenigstens ein ganz einfaches Beispiel zu sehen, betrachten wir die Matrix

$$A = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

vom Ende des vorigen Abschnitts. Links oben steht schon ein JORDAN-Kästchen zum Eigenwert zwei, rechts unten müssen wir noch etwas arbeiten.

Die Basis des $\mathbb{R}^5$ sei $\{\vec{e}_1, \ldots, \vec{e}_5\}$; davon können wir $\vec{b}_1 = \vec{e}_1$ und $\vec{b}_2 = \vec{e}_2$ als Basis von H_2 gleich übernehmen. H_1 wird von $\vec{e}_3, \vec{e}_4$ und $\vec{e}_5$ aufgespannt; ein Vektor aus diesem dreidimensionalen Raum, der unter

$$N = \begin{pmatrix} 0 & 2 & 3 \\ 0 & 0 & 4 \\ 0 & 0 & 0 \end{pmatrix}$$

den maximalen Nilpotenzgrad hat, ist etwa $\vec{e}_5$, denn $\vec{e}_5$ wird abgebildet auf $3\vec{e}_3 + 4\vec{e}_4$; da e_3 auf den Nullvektor geht und $\vec{e}_4$ auf $2\vec{e}_3$, wird dieser Vektor weiter abgebildet auf $8\vec{e}_3$, was schließlich auf den Nullvektor abgebildet wird. Mit

$$\vec{b}_3 = 2\vec{e}_3, \quad \vec{b}_4 = 3\vec{e}_3 + 4\vec{e}_4 \quad \text{und} \quad \vec{b}_5 = \vec{e}_5$$

geht also $\vec{b}_5$ unter N auf $\vec{b}_4$ und weiter auf $\vec{b}_3$; daher ist

$$A\vec{b}_3 = \vec{b}_3, \quad A\vec{b}_4 = \vec{b}_4 + \vec{b}_3 \quad \text{und} \quad A\vec{b}_5 = \vec{b}_5 + \vec{b}_4\,,$$

und die Matrix A hat bezüglich der Basis $\{\vec{b}_1, \ldots, \vec{b}_5\}$ die Gestalt

$$A' = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

mit den beiden JORDAN-Kästchen

$$\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

§3: Lineare Differentialgleichungen und Differentialgleichungssysteme

Nach diesem längeren Einschub aus der linearen Algebra haben wir nun das Rüstzeug zusammen, um eine ganze Reihen von Differentialgleichungen und Differentialgleichungssystemen lösen zu können. Als erstes kehren wir zurück zu den Systemen homogener linearer Differentialgleichungen mit konstanten Koeffizienten, die ja der Ausgangspunkt für diesen Einschub waren.

In §2*g)* haben wir bereits ein Beispiel behandelt, und genauso können wir natürlich auch im allgemeinen Fall vorgehen:

a) Systeme homogener linearer Differentialgleichungen mit konstanten Koeffizienten

Wir betrachten ein Differentialgleichungssystem

$$\dot{\vec{y}}(t) = A\vec{y}(t)\,,$$

wobei A eine reelle oder komplexe $n \times n$-Matrix ist. Zumindest über den komplexen Zahlen zerfällt ihr charakteristisches Polynom in Linearfaktoren, es gibt also eine Basis $\mathcal{B}$ von $\mathbb{C}^n$, bezüglich derer A als obere Dreiecksmatrix geschrieben werden kann.

Ist B die komplexe $n \times n$-Matrix, deren Spalten die Vektoren aus $\mathcal{B}$ sind, ist dann also

$$A = BCB^{-1}$$

mit einer Dreiecksmatrix $C \in \mathbb{C}^{n\times n}$. Dann ist auch

$$e^{At} = Be^{Ct}B^{-1}\,,$$

und e^{Ct} kann berechnet werden, da $C = D + N$ Summe einer Diagonalmatrix und einer damit kommutierenden oberen Dreiecksmatrix mit Nullen in der Hauptdiagonalen ist. Insgesamt ist also

$$e^{At} = Be^{Dt}e^{Nt}B^{-1},$$

und jeder der vier Faktoren rechts ist in endlich vielen Schritten berechenbar.

Wie wir weiterhin wissen, hat jede Lösung der Differentialgleichung $\dot{\vec{y}}(t) = A\vec{y}(t)$ die Form

$$\vec{y}(t) = e^{At}\vec{y}_0 \quad \text{mit} \quad \boldsymbol{y}_0 = \vec{y}(0) \in \mathbb{C}^n ;$$

wir kennen also alle Lösungen des Systems. Falls Anfangsbedingungen vorgegeben sind, die sich auf einen beliebigen Zeitpunkt $t = t_0$ beziehen, können wir die Lösung entsprechend schreiben als

$$\vec{y}(t) = e^{A(t-t_0)}\vec{y}(t_0) .$$

Solange wir in der Lage sind, die Nullstellen des charakteristischen Polynoms von A zu berechnen, können wir also jedes Anfangswertproblem in Gestalt eines Systems linearer Differentialgleichungen mit konstanten Koeffizienten lösen.

b) Langzeitverhalten der Lösung

Nicht immer ist es notwendig oder auch nur möglich, ein Differentialgleichungssystem zur exakten numerischen Vorhersage der weiteren Entwicklung eines Systems zu verwenden; gelegentlich reicht auch ein qualitativer Überblick. Dabei geht es vor allem um das Langzeitverhalten des Systems: Nähert es sich einem Gleichgewicht, „explodiert“ es, oder wird es auf lange Sicht periodisch, wie wir es etwa vom Fall der erzwungenen Schwingung her kennen.

Für solche Aussagen reicht es im Falle eines linearen homogenen Differentialgleichungssystems $\dot{\vec{y}}(t) = A\vec{y}(t)$, die Eigenwerte der Matrix A zu kennen: Bezüglich einer Basis aus Hauptvektoren läßt sich A in der Form $D + N$ schreiben mit einer Diagonalmatrix D, für die e^{Dt} Diagonalmatrix ist mit den Funktionen $e^{\lambda t}$ als Einträgen, wobei λ die

Eigenwerte von A durchläuft. Die Matrix e^{Nt} hat Polynome in t als Einträge; das Produkt $e^{Dt}e^{Nt}$ hat also wegen der speziellen Formen von D und N Produkte von Polynomen in t mit $e^{\lambda t}$ als Einträge, wobei der Grad des Polynoms höchstens die um eins verminderte größte Stufe eines Hauptvektors zu λ ist. Bei der Rücktransformation auf die Ausgangsbasis entstehen Linearkombinationen solcher Funktionen, die bei der Multiplikation mit dem Vektor der Anfangswerte selbst wieder linear kombiniert werden. Insgesamt ist also jede Lösungsfunktion eine Linearkombination von Termen der Form $t^j e^{\lambda t}$.

Falls nur ein Eigenwert λ von A positiven Realteil hat, muß das System fast unweigerlich explodieren, da der Betrag von $e^{\lambda t}$ für hinreichend großes t jede vorgegebene Grenze überschreitet. Zwar sind eventuell Anfangsbedingungen möglich, bei denen $e^{\lambda t}$ nicht in der Lösung des Anfangswertproblems auftritt, aber da die Anfangswerte in der Praxis nie durch Naturgesetze gegeben sind, sondern durch fehlerbehaftete Meßwerte, kann schon eine kleine Störung den Term $e^{\lambda t}$ wieder ins Spiel bringen, und auch für sehr kleines c dominiert $ce^{\lambda t}$ langfristig jede beschränkte Funktion.

Haben dagegen *alle* Eigenwerte von A negativen Realteil, geht jeder Term $t^j e^{\lambda t}$ gegen Null für $t \to \infty$ und damit auch jede Lösungsfunktion; der Lösungsvektor nähert sich also immer mehr der Gleichgewichtslösung $\vec{y}(t) \equiv 0$.

Bleibt noch der Fall von Eigenwerten mit Realteil null. Hier werden die bei mehrfachen Eigenwerten möglichen Polynome wichtig, denn während der Betrag von $e^{\lambda t}$ konstant ist, geht der Betrag der mit einem nichtkonstanten Polynom multiplizierten Funktion gegen $\pm\infty$.

Schließlich sollten wir auch die Imaginärteile der Eigenwerte nicht ganz vergessen: Falls A, wie in den meisten Anwendungen, eine reelle Matrix ist, ist mit jedem nichtreellen Eigenwert λ auch die konjugiert komplexe Zahl $\overline{\lambda}$ ein Eigenwert derselben Vielfachheit wie λ. Mit $\lambda = a + ib$ ist

$$e^{\lambda t} = e^{at}(\cos bt + i \sin bt) \quad \text{und} \quad e^{\overline{\lambda} t} = e^{at}(\cos bt - i \sin bt) ,$$

jede Linearkombination von $e^{\lambda t}$ und $e^{\overline{\lambda} t}$ läßt sich also auch als Linear-

kombination der beiden reellen Funktionen

$$e^{at}\cos bt \quad \text{und} \quad e^{at}\sin bt$$

schreiben. Die Imaginärteile bringen also Schwingungsanteile in die Lösungsfunktionen.

Um einen ersten Eindruck vom Lösungsverhalten linearer homogener Differentialgleichungssysteme mit reellen Koeffizienten zu bekommen, wollen wir uns überlegen, was in niedrigen Dimensionen möglich ist.

Im Eindimensionalen besteht das „System“ einfach aus der Differentialgleichung $\dot{y}(t) = ay(t)$ mit Lösung $y(t) = y(0)e^{at}$; für $a > 0$ geht dies je nach Vorzeichen von $y(0)$ gegen plus oder minus unendlich für $t \to \infty$, für $a < 0$ gegen null, und für $a = 0$ ist $y(t) = y(0)$ konstant.

Etwas interessanter ist es im Zweidimensionalen: Hier hat die Matrix A einen oder zwei Eigenwerte.

Falls es nur einen gibt und dieser die geometrische Vielfachheit zwei hat, ändert sich nichts wesentliches gegenüber der eindimensionalen Situation: Die Lösungsfunktion ist $\binom{x(0)}{y(0)} \cdot e^{\lambda t}$.

Falls der Eigenwert nur die algebraische Vielfachheit eins hat, sind die Lösungsfunktionen Linearkombinationen von $e^{\lambda t}$ und $te^{\lambda t}$, also Produkte lineare Funktionen mit $e^{\lambda t}$. Jetzt können die Lösungen auch im Fall $\lambda = 0$ unbeschränkt werden.

Wenn es zwei verschiedene Eigenwerte gibt, können diese entweder beide reell oder aber konjugiert komplex sein.

Im reellen Fall können beide positiv sein; dann geht jede nichttriviale Lösung ins Unendliche, oder aber beide können negativ sein; dann nähert sich jede Lösung immer mehr dem Nullpunkt. Alternativ könnte einer positiv und einer negativ sein; in diesem Fall nähern sich alle Lösungskurven einer Geraden, gehen aber mit dieser ins Unendliche. Schließlich könnte noch ein Eigenwert null sein; dann liegen alle Lösungskurven auf Geraden und gehen dort je nach Vorzeichen des anderen Eigenwerts gegen einen festen Punkt oder aber ins Unendliche.

Bleibt noch der Fall zweier konjugiert komplexer Eigenwerte $a \pm ib$; in diesem Fall sind die Lösungsfunktionen Linearkombinationen von $e^{at}\cos t$ und $e^{at}\sin t$.

Für $a = 0$ haben wir einfach reine Schwingungen; falls die beiden Eigenvektoren senkrecht aufeinander stehen, bekommen wir als Lösungskurven Kreislinien, ansonsten affine Verzerrungen davon, also Ellipsen.

Für $a \neq 0$ ergeben sich entsprechend Spiralen, die je nach Vorzeichen von a entweder ins Unendliche gehen oder aber sich auf den Nullpunkt zusammenziehen.

Betrachten wir als Beispiel das System

$$\dot{x}(t) = -10{,}2x(t) - 25y(t) \quad \text{und} \quad \dot{y}(t) = 5x(t) + 9{,}8y(t)\,.$$

Die Matrix

$$A = \begin{pmatrix} -10\frac{1}{5} & -25 \\ 5 & 9\frac{4}{5} \end{pmatrix}$$

hat das charakteristische Polynom

$$\lambda^2 + \frac{2}{5}\lambda + 25\frac{1}{5}$$

mit Nullstellen $-\frac{1}{5} \pm 5i$, die Lösungskurven sind also sich nach innen zusammenziehende Spiralen. Mit den Anfangsbedingungen $x(0) = 0$ und $y(0) = 1$ erhalten wir die Lösung

$$x(t) = -5e^{-t/5}\sin 5t \quad \text{und} \quad y(t) = e^{-t/5}(\cos 5t + 2\sin 5t)\,,$$

die in Abbildung 30 zu sehen ist.

Im Dreidimensionalen gibt es entsprechend mehr Möglichkeiten; ich möchte auf die weniger interessanten Fälle mit lauter reellen Eigenwerten verzichten und nur den betrachten, daß zwei konjugiert komplexe auftreten, etwa $a \pm ib$. Der dritte Eigenwert λ muß dann natürlich reell sein.

Falls er null ist, sind wir im wesentlichen in der zweidimensionalen Situation: Da dann in Richtung des dritten Eigenvektors alles konstant ist, spielt sich alles in einer Ebenen ab, die parallel zu der von den ersten beiden Eigenvektoren aufgespannten liegt.

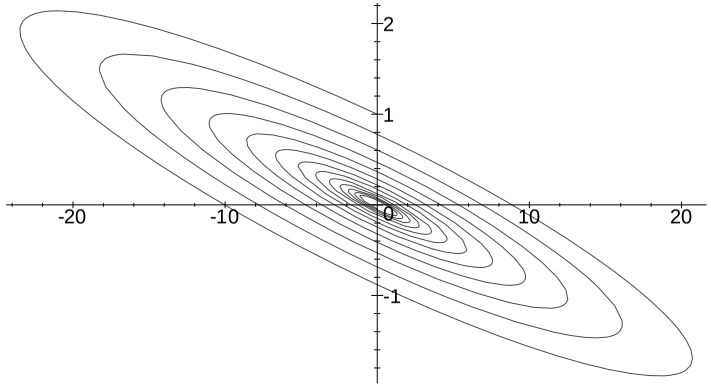


Abb. 30: Spiralförmige Annäherung an den Nullpunkt

Falls er negativ ist, nähert sich in Richtung des dritten Eigenvektors alles der Ebenen, in der dessen Koordinate null ist, die also von den beiden anderen Eigenvektoren aufgespannt wird, und dort ist die Dynamik je nach Vorzeichen von a spiralförmig nach innen oder außen oder einfach ellipsenförmig. Abbildung 31 zeigt den Fall $a = -1/10, b = 2$ und $\lambda = -1/6$; die Eigenvektoren zeigen in Richtung der Koordinatenachsen.

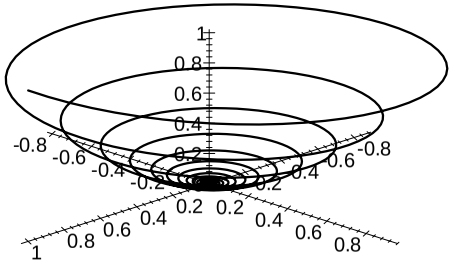


Abb. 31: Spiralförmige Annäherung an den Nullpunkt im Dreidimensionalen

In Abbildung 32 ist $a = +1/10$ und ansonsten alles unverändert; wie

man sieht, nähert sich nun die Lösungskurve zwar der (x, y)-Ebenen, geht in dieser aber spiralförmig ins Unendliche.

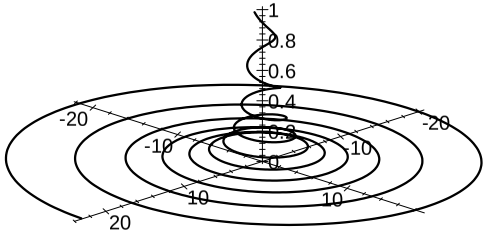


Abb. 32: Spirale, die sich der (x, y)-Ebenen nähert

Falls der dritte Eigenwert positiv ist, geht in Richtung seines Eigenvektors alles ins Unendliche; je nach Vorzeichen von a entfernen sich die Lösungskurven dabei spiralförmig von der durch diesen Eigenvektor aufgespannten Geraden oder aber gehen auf sie zu. Abbildung 33 zeigt den letzteren Fall mit $\lambda = +1/6$ und a, b wie in Abbildung 30.

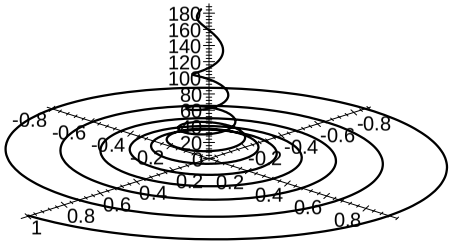


Abb. 33: Spirale, die sich der z-Achse nähert

Man beachte, daß die Abbildungen 32 und 33 zwar auf den ersten Blick sehr ähnlich aussehen, daß aber die Dynamik in Abbildung 32 in z-Richtung nach unten geht, in Abbildung 33 aber nach oben. Der Rotationssinn der Spiralen ist in beiden Fällen der Gegenuhrzeigersinn.

c) Lineare homogene Differentialgleichungen höherer Ordnung mit konstanten Koeffizienten

Eine der wichtigsten Anwendungen der Theorie aus den letzten Abschnitten in der Elektrotechnik sind lineare Differentialgleichungen höherer Ordnung mit konstanten Koeffizienten. Dieser Abschnitt soll zeigen, daß man in diesem Spezialfall weder Determinanten noch Eigen- und Hauptvektoren berechnen muß, um die Lösungsfunktionen zu finden.

Um zu sehen, was der allgemeine Ansatz in diesem Spezialfall ergibt, schreiben wir die Differentialgleichung

$$y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \cdots + a_1\dot{y}(t) + a_0y(t) = 0 \quad \text{mit} \quad a_i \in \mathbb{C}\,.$$

als ein System

$$\begin{aligned}
\dot{y}_0(t) &= y_1(t)\\
\dot{y}_1(t) &= y_2(t)\\
&\vdots \qquad \vdots\\
\dot{y}_{n-2}(t) &= y_{n-1}(t)\\
\dot{y}_{n-1}(t) &= -a_{n-1}y_{n-1}(t) - \cdots - a_1\dot{y}_1(t) - a_0y_0(t)
\end{aligned}$$

oder kurz $\dot{\vec{y}}(t) = A\vec{y}(t)$ mit

$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0\\ 0 & 0 & 1 & \dots & 0 & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \dots & 1 & 0\\ 0 & 0 & 0 & \dots & 0 & 1\\ -a_0 & -a_1 & -a_2 & \dots & -a_{n-2} & -a_{n-1} \end{pmatrix}$$

und

$$\vec{y}(t) = \begin{pmatrix} y_0(t)\\ \vdots\\ y_{n-1}(t) \end{pmatrix}.$$

Die Lösungen $\vec{y}(t)$ dieses Systems definieren Lösungen $y(t) = y_0(t)$ der obigen Gleichung, und umgekehrt ist auch für jede solche Lösung $y(t)$

der Vektor

$$\vec{y}(t) = \begin{pmatrix} y(t)\\ \dot{y}(t)\\ \ddot{y}(t)\\ \vdots\\ y^{(n-2)}(t)\\ y^{(n-1)}(t) \end{pmatrix}$$

eine Lösung des Systems $\dot{\vec{y}}(t) = A\vec{y}(t)$.

Aus dem vorigen Abschnitt wissen wir, daß die Lösungen dieses Systems einen n-dimensionalen Vektorraum V von vektorwertigen Funktionen $t \mapsto \vec{y}(t)$ bilden. Jede einzelne Komponente jeder Lösung aus diesem Vektorraum ist darstellbar als Linearkombination von Funktionen $t^j e^{\lambda_i t}$, wobei $\lambda_1, \dots, \lambda_r$ die verschiedenen Eigenwerte von A sind und die nichtnegative ganze Zahl j kleiner ist als die größte Stufe eines Hauptvektors zu λ_i, erst recht also kleiner als die algebraische Vielfachheit von λ_i. Insbesondere stehen daher bei einfachen Eigenwerten oder allgemeiner bei Eigenwerten, deren geometrische Vielfachheit gleich der algebraischen ist, überhaupt keine echten t-Potenzen vor den Exponentialfunktionen.

Außerdem wissen wir, daß jedes Anfangswertproblem eindeutig lösbar ist: Bei Anfangsbedingungen

$$y_0(t_0) = c_0, \quad \dots, \quad y_{n-1}(t_0) = c_{n-1}$$

ist $\vec{y}(t) = e^{A(t-t_0)}$ *die* Lösung. Vornehm ausgedrückt können wir auch sagen, daß für jedes $t_0 \in \mathbb{R}$ die Abbildung

$$V \to \mathbb{C}^n; \quad \vec{y}(t) \mapsto \vec{y}(t_0)$$

ein Isomorphismus ist.

Für jeden Lösungsvektor $\vec{y}(t)$ des Differentialgleichungssystems ist die nullte Komponenten $y_0(t)$ Lösung der Differentialgleichung

$$y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \cdots + a_1\dot{y}(t) + a_0y(t) = 0\,,$$

und diese Komponente bestimmt auch die anderen Komponenten $y_i(t)$, die ja gerade die Ableitungen von $y_0(t)$ sind. Daher bilden auch die

Lösungsfunktionen dieser Gleichung einen n-dimensionalen Vektorraum, d.h. auch die Abbildung $\vec{y}(t) \mapsto y_0(t)$ ist ein Isomorphismus, und wenn wir Anfangsbedingungen mit ins Spiel bringen folgt auch, daß es für jede Vorgabe von Werten $y(t_0), \dot{y}(t_0), \ddot{y}(t_0), \ldots, y^{(n-1)}(t_0)$ genau eine Lösung gibt.

Wie wir uns gerade überlegt haben, ist $y_0(t)$ Linearkombination von Funktionen der Art $t^j e^{\lambda_i t}$, wobei λ_i die Eigenwerte von A durchläuft und j kleiner als die algebraische Vielfachheit des Eigenwerts λ_i ist.

Die Summe der algebraischen Vielfachheiten der (komplexen) Eigenwerte einer komplexen $n \times n$-Matrix ist gleich dem Grad des charakteristischen Polynoms, also gleich n; damit gibt es genau n solche Funktionen. Andererseits wissen wir, daß der Lösungsraum die Dimension n hat; somit werden alle diese Funktionen wirklich gebraucht und sie bilden eine Basis des Lösungsraums.

Für eine wirklich befriedigende Kenntniss des Lösungsraums fehlt uns nun nur noch ein Verfahren, wie wir die Eigenwerte λ_i und deren algebraische Vielfachheiten α_i direkt aus den Koeffizienten der Gleichung berechnen können.

Zumindest die Eigenwerte λ_i lassen sich leicht aus der Gleichung ablesen: Für die Funktion $y(t) = e^{\lambda t}$ ist $y^{(i)}(t) = \lambda^i e^{\lambda t}$, sie ist also genau dann eine Lösung der Differentialgleichung, wenn

$$\begin{aligned} &\lambda^n e^{\lambda t} + a_{n-1}\lambda^{n-1} e^{\lambda t} + \ldots + a_1 \lambda e^{\lambda t} + a_0 e^{\lambda t} \\ &= (\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0) e^{\lambda t} \end{aligned}$$

verschwindet; da $e^{\lambda t}$ nirgends verschwindet, ist dies genau dann der Fall, wenn gilt

$$\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0 = 0\,. \qquad (*)$$

Definition: Die Gleichung $(*)$ heißt *charakteristische Gleichung* der Differentialgleichung

$$y^{(n)}(t) + a_{n-1} y^{(n-1)}(t) + \ldots + a_1 \dot{y}(t) + a_0 y(t) = 0\,.$$

Die Eigenwerte λ_i von A sind also genau die Nullstellen der charakteristischen Gleichung der Differentialgleichung. Da diese denselben Grad hat wie das charakteristische Polynom von A liegt die Vermutung nahe, daß sie (eventuell bis auf eine Konstante) damit übereinstimmt, und das ist auch tatsächlich der Fall:

Lemma: Die charakteristische Gleichung der Differentialgleichung ist das mit $(-1)^n$ multiplizierte charakteristische Polynom von A.

Beweis: Für $n = 1$ ist die Differentialgleichung $\dot{y}(t) = ay(t)$ identisch mit dem zugehörigen System; die charakteristische Gleichung ist $\lambda - a$, und die „Matrix“ $A = (a)$ hat das charakteristische Polynom $a - \lambda$. Für $n = 1$ stimmt die Behauptung also.

Für $n > 1$ entwickeln wir das charakteristische Polynom

$$\det(A - \lambda E) = \begin{vmatrix} -\lambda & 1 & 0 & \ldots & 0 & 0 \\ 0 & -\lambda & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & 0 & \ldots & -\lambda & 1 \\ -a_0 & -a_1 & -a_2 & \ldots & -a_{n-2} & -a_{n-1} - \lambda \end{vmatrix}$$

nach der ersten Spalte und erhalten

$$\begin{aligned} \det(A - \lambda E) = -\lambda &\begin{vmatrix} -\lambda & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & \ldots & -\lambda & 1 \\ -a_1 & -a_2 & \ldots & -a_{n-2} & -a_{n-1} - \lambda \end{vmatrix} \\ + (-1)^{n-1}(-a_0) &\begin{vmatrix} 1 & 0 & \ldots & 0 & 0 \\ -\lambda & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \\ 0 & 0 & \ldots & -\lambda & 1 \end{vmatrix}. \end{aligned}$$

Die erste Determinate auf der rechten Seite ist von derselben Form wie das betrachtete charakteristische Polynom, hat aber nur die Größe

$(n-1)\times(n-1)$. Wir können daher induktiv schließen, daß sie gleich

$$(-1)^{n-1}\left(a_n\lambda^{n-1}+a_{n-1}\lambda^{n-2}+\cdots+a_2\lambda+a_1\right)$$

ist. Die zweite Determinante läßt sich direkt ausrechnen: Wie man sich leicht überlegt (oder in **[HM1]**, Kapitel I, §4*f)* nachliest), ist die Determinante einer Dreiecksmatrix gleich dem Produkt ihrer Diagonalelemente; die zweite Determinante ist also gleich eins. Damit folgt

$$\begin{aligned}\det(A-\lambda E) &= (-\lambda)(-1)^{n-1}\left(a_n\lambda^{n-1}+a_{n-1}\lambda^{n-2}+\cdots+a_2\lambda+a_1\right)\\ &\quad+(-1)^{n-1}(-a_0)\\ &= (-1)^n\left(a_n\lambda^n+a_{n-1}\lambda^{n-1}+\cdots+a_2\lambda^2+a_1\lambda+a_0\right),\end{aligned}$$

wie behauptet. ∎

Insbesondere haben also die charakteristische Gleichung einer Differentialgleichung und das charakteristische Polynom der zugehörigen Matrix nicht nur dieselben Nullstellen, sondern auch die Vielfachheiten dieser Nullstellen sind gleich; die algebraischen Vielfachheiten der Eigenwerte lassen sich direkt aus der charakteristischen Gleichung ablesen.

Damit haben alle Bausteile zusammen und können das Ergebnis dieses Abschnitts im folgenden Satz zusammenfassen:

Satz: Die charakteristische Gleichung

$$\lambda^n+a_{n-1}\lambda^{n-1}+\cdots+a_1\lambda+a_0=0$$

der Differentialgleichung

$$y^{(n)}(t)+a_{n-1}y^{(n-1)}(t)+\cdots+a_1\dot y(t)+a_0y(t)=0$$

habe die Nullstellen $\lambda_1,\dots,\lambda_r$; die Vielfachheit der Nullstelle λ_i sei α_i. Dann bilden die Lösungen der Differentialgleichung einen n-dimensionalen Vektorraum V mit Basis

$$\left\{e^{\lambda_1 t},te^{\lambda_1 t},\dots,t^{\alpha_1-1}e^{\lambda_1 t},\quad\dots\quad e^{\lambda_r t},te^{\lambda_r t},\dots,t^{\alpha_r-1}e^{\lambda_r t}\right\}.$$

Für jedes $t_0\in\mathbb{R}$ ist die lineare Abbildung

$$V\to\mathbb{C}^n;\quad y\mapsto\left(y(t_0),\dot y(t_0),\ddot y(t_0),\dots,y^{(n-1)}(t_0)\right)$$

ein Isomorphismus; insbesondere ist jedes Anfangswertproblem

$$y(t_0)=c_0,\quad \dot y(t_0)=c_1,\quad\dots\quad y^{(n-1)}(t_0)=c_{n-1}$$

eindeutig lösbar. ∎

Sofern wir uns für komplexwertige Lösungen interessieren, liefert uns dieser Satz alles was wir brauchen. Bei vielen Anwendungen hat man es aber mit Gleichungen mit reellen Koeffizienten zu tun, und von der Natur des Problems her interessieren nur reelle Lösungen. Mit unserem bisherigen Ansatz kommen wir auch bei diesen Problemen nicht um komplexe Lösungen herum; beispielsweise hat die wohlbekannte Differentialgleichung

$$\ddot y(t)+y(t)=0$$

die charakteristische Gleichung

$$\lambda^2+1=0$$

mit den beiden rein imaginären Lösungen $\lambda=\pm i$, die oben angegebene Basis des Lösungsraums ist also

$$\left\{e^{it},\,e^{-it}\right\}.$$

Falls wir noch nie etwas vom obigen Satz gehört hätten, würden wir stattdessen sagen, daß die Lösungen dieser Differentialgleichung genau die Linearkombinationen von Sinus und Cosinus sind, Basis des Lösungsraums ist also

$$\left\{\sin t,\,\cos t\right\}.$$

Über $\mathbb{C}$ sind beide Aussagen äquivalent, denn auf Grund der EULERschen Formeln

$$\cos t=\frac{1}{2}\left(e^{it}+e^{-it}\right)\quad\text{und}\quad\sin t=\frac{1}{2i}\left(e^{it}-e^{-it}\right)$$

bzw.

$$e^{it}=\cos t+i\sin t\quad\text{und}\quad e^{-it}=\cos t-i\sin t$$

erzeugen beide Basen denselben $\mathbb{C}$-Vektorraum. Wenn wir an reellen Lösungen interessiert sind, ist aber die zweite Basis erheblich nützlicher, denn sie erzeugt auch den $\mathbb{R}$-Vektorraum der reellen Lösungen.

Entsprechend hätten wie auch für allgemeinere Differentialgleichungen mit reellen Koeffizienten gerne eine $\mathbb{C}$-Basis des (komplexen) Lösungsraums, die gleichzeitig $\mathbb{R}$-Basis des reellen Lösungsraums ist. Im Rest dieses Abschnitts wollen wir uns überlegen, wie wir diese Basis konstruieren können.

Wir gehen also aus von einer Differentialgleichung

$$y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \ldots + a_1\dot{y}(t) + a_0y(t) = 0 \quad \text{mit} \quad a_i \in \mathbb{R}$$

und betrachten zunächst wieder die charakteristische Gleichung

$$\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0 = 0\,.$$

Diese *kann* reelle Lösungen haben; diese bezeichnen wir mit λ_1 bis λ_p, und falls es keine gibt, setzen wir $p = 0$.

Falls p gleich der Gesamtzahl r der Nullstellen der charakteristischen Gleichung ist, sind wir fertig: Die Funktionen $t^j e^{\lambda_i t}$ sind allesamt reell, und die, bei denen j kleiner als die Vielfachheit der Nullstelle λ_i ist, spannen sowohl den komplexen als auch den reellen Lösungsraum auf.

Andernfalls gibt es auch noch nichtreelle Nullstellen. Für jede solche Nullstelle λ ist auch die konjugiert komplexe Zahl $\overline{\lambda}$ eine Nullstelle, denn da die Koeffizienten a_i reell sind, ist $\overline{a_i} = a_i$ und damit

$$\overline{\lambda}^n + a_{n-1}\overline{\lambda}^{n-1} + \ldots + a_1\overline{\lambda} + a_0 = \overline{\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0} = 0\,.$$

Nichtreelle Nullstellen treten also immer auf als Paare konjugiert komplexer Zahlen; außerdem haben λ und $\overline{\lambda}$ dieselbe Vielfachheit, denn eine Nullstelle ist genau dann eine r-fache Nullstelle eines Polynoms P, wenn sie Nullstelle von P und allen seinen Ableitungen bis zur $(r-1)$-ten ist. Da auch alle diese Ableitungen Polynome mit reellen Koeffizienten sind, zeigt die gleiche Rechnung wie oben, daß $P^{(j)}(\lambda)$ genau dann verschwindet, wenn auch $P^{(j)}(\overline{\lambda})$ verschwindet.

Wir fassen daher die nichtreellen Nullstellen zu Paaren $\left(\lambda_{p+j}, \overline{\lambda_{p+j}}\right)$ zusammen, wobei j von 1 bis zur Anzahl q dieser Paare läuft; die Gesamtzahl der Nullstellen ist also $r = p + 2q$.

Die Vielfachheit der Nullstelle λ_k sei weiterhin mit α_k bezeichnet; für $k > p$ ist das gleichzeitig auch die Vielfachheit der Nullstelle $\overline{\lambda}_k$.

Für $k > p$ wenden wir, genau wie im obigen Beispiel, die EULERschen Formeln an: Für

$$\lambda_k = \mu_k + i\nu_k \quad \text{mit} \quad \mu_k, \nu_k \in \mathbb{R}$$

ist

$$e^{\lambda_k t} = e^{\mu_k t}\left(\cos\nu_k t + i\sin\nu_k t\right) \quad \text{und} \quad e^{\overline{\lambda_k} t} = e^{\mu_k t}\left(\cos\nu_k t - i\sin\nu_k t\right),$$

genauso ist für jedes j

$$t^j e^{\lambda_k t} = t^j e^{\mu_k t}\left(\cos\nu_k t + i\sin\nu_k t\right)$$

und

$$t^j e^{\overline{\lambda_k} t} = t^j e^{\mu_k t}\left(\cos\nu_k t - i\sin\nu_k t\right).$$

Also spannt

$$\left\{t^j e^{\lambda_k t},\ t^j e^{\overline{\lambda_k} t}\right\}$$

denselben $\mathbb{C}$-Vektorraum auf wie

$$\left\{t^j e^{\mu_k t}\cos\nu_k t,\ t^j e^{\mu_k t}\sin\nu_k t\right\},$$

und letztere Basis spannt gleichzeitig den $\mathbb{R}$-Vektorraum aller reeller Funktionen aus diesem $\mathbb{C}$-Vektorraum auf.

Damit können wir auch eine Basis des reellen Lösungsraums angeben:

Satz: Die charakteristische Gleichung

$$\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0 = 0$$

der Differentialgleichung

$$y^{(n)}(t) + a_{n-1}y^{(n-1)}(t) + \ldots + a_1\dot{y}(t) + a_0y(t) = 0$$

habe die reellen Nullstellen $\lambda_1, \ldots, \lambda_p$ sowie die Paare konjugiert komplexer Nullstellen $\left(\lambda_{p+1}, \overline{\lambda_{p+1}}\right), \ldots, \left(\lambda_{p+q}, \overline{\lambda_{p+q}}\right)$ mit $\lambda_k = \mu_k + i\nu_k$; die Vielfachheit der Nullstelle λ_i sei α_i. Dann bilden die Lösungen der Differentialgleichung einen n-dimensionalen Vektorraum V, der aufgespannt wird von den Funktionen

$$t^j e^{\lambda_i t} \quad \text{für } i = 1, \ldots, p$$

und

$$t^j e^{\mu_i t} \cos \nu_i t, \quad t^j e^{\mu_i t} \sin \nu_i t \quad \text{für } i = p+1, \ldots, p+q\,,$$

wobei j jeweils von null bis $\alpha_i - 1$ läuft. Für jedes $t_0 \in \mathbb{R}$ ist die lineare Abbildung

$$V \to \mathbb{R}^n; \quad y \mapsto \big(y(t_0), \dot{y}(t_0), \ddot{y}(t_0), \ldots, y^{(n-1)}(t_0)\big)$$

ein Isomorphismus; insbesondere ist jedes Anfangswertproblem

$$y(t_0) = c_0, \quad \dot{y}(t_0) = c_1, \quad \ldots \quad y^{(n-1)}(t_0)$$

eindeutig lösbar.

■

d) Inhomogene Differentialgleichungen

Am Beispiel der Schwingungsdifferentialgleichungen haben wir gesehen, daß zur Beschreibung interessanter Phänomene homogene Differentialgleichungen oft nicht ausreichen; sobald etwa ein elektrischer Schwingkreis eine Stromquelle enthält, haben wir eine inhomogene Differentialgleichungen.

Wie schon im Falle der Schwingungsdifferentialgleichungen genügt es zur Lösung einer allgemeinen inhomogenen linearen Differentialgleichung, den Lösungsraum der zugehörigen homogenen Differentialgleichung zu kennen und dazu noch *eine* Lösung der inhomogenen, denn wegen der Linearität der linken Seite ist die Differenz zweier Lösungen der inhomogenen Differentialgleichung Lösung der homogenen.

Wir kennen bereits eine Methode, uns eine solche spezielle Lösung zu verschaffen: Sobald wir uns Anfangswerte vorgeben, können wir die entsprechende Lösung des inhomogenen Problems mittels LAPLACE-Transformation finden – sofern wir die LAPLACE-Transformierte des inhomogenen Anteils kennen und aus der LAPLACE-Transformierten der Lösungsfunktion diese Funktion selbst bestimmen können. In vielen praktisch relevanten Fällen wird dies mit Hilfe von Tabellen möglich sein, sofern man nur die grundlegenden Regeln für den Umgang mit LAPLACE-Transformationen kennt.

Eine zweite Methode ist das *Erraten* einer Lösung. In der Praxis betrachtet man Differentialgleichungen meist, um das künftige Verhalten eines

Systems vorauszusagen, und oft wird man (leider nicht immer richtige!) ungefähre Erwartungen haben, wie dieses Verhalten aussehen sollte. Falls man diese als Ansatz mit unbestimmten Parametern in die Differentialgleichung einsetzt und die Erwartungen richtig war, kann man die Parameter bestimmen und hat eine Lösung gefunden; andernfalls muß man sich etwas neues überlegen.

Als Beispiel betrachten wir nochmals die Differentialgleichung

$$\ddot{y}(t) + \rho \dot{y}(t) + \sigma y(t) = f(t)\,.$$

Wir kamen auf diese Differentialgleichung bei der Betrachtung eines elektrischen Schwingkreises, an den eine externe Spannung proportional $f(t)$ angelegt war, und für vernünftiges $f(t)$ sollte man erwarten, daß es Lösungen gibt, bei denen sich $y(t)$ völlig von $f(t)$ bestimmen läßt und daher von ähnlicher Gestalt ist.

Im Falle einer Gleichstromquelle $f(t) = c$ vermuten wir, daß es vielleicht eine Lösung gibt, bei der nur ein Gleichstrom fließt. Der Ansatz $y(t) = a$ führt auf

$$\sigma a = c \quad \text{oder} \quad a = \frac{c}{\sigma}\,,$$

falls $\sigma \neq 0$ ist.

Für $\sigma = 0$ haben wir *de facto* eine Differentialgleichung für $\dot{y}(t)$ statt für $y(t)$; man könnte es daher mit einem Ansatz der Form $\dot{y}(t) = b$ oder $y(t) = bt + a$ versuchen. Für $\rho \neq 0$ führt das zu

$$\rho b = c \quad \text{oder} \quad b = \frac{c}{\rho}\,;$$

a ist hier natürlich beliebig.

Entsprechend können wir auch bei einer angelegten Wechselspannung vorgehen: Der Schwingkreis wird sicherlich die Amplitude und die Phase verändern, aber die permanent angelegte Wechselspannung sollte doch dem Schwingkreis ihre Frequenz aufzwingen, so daß es zumindest eine Lösung geben sollte, die eine reine Schwingung mit dieser Frequenz ist. Für

$$\ddot{y}(t) + \rho \dot{y}(t) + \sigma y(t) = A \cos \omega_0 t$$

versuchen wir es also mit

$$y(t) = a\cos\omega_0 t + b\sin\omega_0 t$$

und berechnen die Ableitungen:

$$\dot{y}(t) = -a\omega_0\sin\omega_0 t + b\omega_0\cos\omega_0 t$$

und

$$\ddot{y}(t) = -a\omega_0^2\cos\omega_0 t - b\omega_0^2\sin\omega_0 t\,.$$

Einsetzen in die linke Seite der Differentialgleichung liefert

$$\begin{aligned}&\ddot{y}(t) + \rho\dot{y}(t) + \sigma y(t)\\ &= (-a\omega_0^2 + \rho b\omega_0 + \sigma a)\cos\omega_0 t + (-b\omega_0^2 - \rho a\omega_0 + \sigma b)\sin\omega_0 t\,,\end{aligned}$$

und das ist gleich $A\cos\omega_0 t$, falls

$$(\sigma - \omega_0^2)a + \rho\omega_0 b = A \quad \text{und} \quad (\sigma - \omega_0^2)b - \rho\omega_0 a = 0$$

ist.

Für $\sigma = \omega_0^2$ ist dies problemlos zu lösen: Dann ist einfach

$$a = 0 \quad \text{und} \quad b = \frac{A}{\rho\omega_0}\,.$$

Für $\sigma \neq \omega_0^2$ können wir die zweite Gleichung durch $(\sigma - \omega_0^2)$ dividieren und erhalten

$$b = \frac{\rho\omega_0}{\sigma - \omega_0^2}a\,.$$

Dies können wir in die erste Gleichung einsetzen:

$$(\sigma - \omega_0^2)a + \rho\omega_0\frac{\rho\omega_0}{\sigma - \omega_0^2}a = \frac{(\sigma - \omega_0^2)^2 + \rho^2\omega_0^2}{\sigma - \omega_0^2}a = A$$

oder

$$a = \frac{A(\sigma - \omega_0^2)}{(\sigma - \omega_0^2)^2 + \rho^2\omega_0^2} \quad \text{und} \quad b = \frac{A\rho\omega_0}{(\sigma - \omega_0^2)^2 + \rho^2\omega_0^2}\,.$$

In beiden Fällen gibt es also in der Tat eine Lösung der postulierten Form, die wir hiermit explizit bestimmt haben.

Im Prinzip haben wir diese Lösung bereits in §1*c)* berechnet, mit dem Ansatz hier ist die Situation nun aber sehr viel übersichtlicher: Im häufigsten Falle, in dem die homogene Gleichung eine gedämpfte Schwingung beschreibt etwa wissen wir nun, daß es eine feste reine Schwingung mit der erregenden Frequenz gibt, gegen die *alle* Lösungen der Differentialgleichung langfristig konvergieren.

Für allgemeinen Aussagen ist also dieser Ansatz besser geeignet; falls wir dagegen ein konkretes Anfangswertproblem lösen wollen, liefert die LAPLACE-Transformation direkt die Lösung, während wir beim Weg über die allgemeine Lösung noch ein lineares Gleichungssystem lösen müssen, um die freien Parameter der allgemeinen Lösung an vorgegebene Anfangswerte anzupassen. (Falls man freilich die inversen LAPLACE-Transformationen über Partialbruchzerlegung wirklich konkret berechnet, wird man oft auf genau dasselbe lineare Gleichungssystem stoßen, das man zur Anpassung der allgemeinen Lösung an konkrete Anfangsbedingungen lösen muß.)

Falls man einigermaßen konkrete Erwartungen über das Verhalten zumindest einer Lösungsfunktion hat, ist das Erraten einer speziellen Lösung also oft erfolgversprechend. Das in §1*c)* behandelte Beispiel der Resonanzkatastrophe zeigt aber, daß es auch Fälle gibt, wo man nur mit ziemlicher Erfahrung eine spezielle Lösung erraten kann: Die Gleichung

$$\ddot{y}(t) + \omega_0^2 y(t) = \cos\omega_0 t$$

hat schließlich *keine* Lösung, die eine reine Schwingung der Frequenz ω_0 ist.

Zum Glück gibt es außer LAPLACE-Transformation und Erraten noch ein drittes, rein formales Verfahren, das ebenfalls häufig zum Ziel führt, die bereits von den linearen Differentialgleichung erster Ordnung her bekannte *Variation der Konstanten.* Diese hatten wir dort als völlig unmotivierten Ansatz verwendet, der erstaunlicherweise zu einer Lösung führte. Im nächsten Abschnitt werden wir uns überlegen, daß auch diese Methode in ganz natürlicher Weise aus der genaueren Untersuchung von Differentialgleichungen auftritt und ihr Erfolg (oder Mißerfolg) in konkreten Beispielen damit verstanden werden kann.

e) Symmetriebetrachtungen

Viele Systeme haben eine natürliche oder konstruktionsbedingte Symmetrie; wenn man diese Symmetrie erkennt, hat man gleich zwei Werkzeuge in der Hand:

Einmal hat man eine Struktur des Lösungsraums erkannt und kann via Symmetrie eventuell aus einfachen, leicht erkennbaren Lösungen kompliziertere konstruieren. Dieser Aspekt spielt bei den linearen Differentialgleichungen, die wir hier betrachten, keine nennenswerte Rolle: Hier hat der Lösungsraum schließlich immer eine einfache Struktur als affiner Raum oder (im homogenen Fall) sogar Vektorraum – was andererseits natürlich gerade ein sehr einfacher Fall des gerade Gesagten ist.

Zum anderen kann man eventuell versuchen, die Symmetrie der Differentialgleichung durch Koordinatentransformationen auf eine *bekannte* Symmetrie zurückzuführen, die zu einem bekannten Typus von Differentialgleichungen gehört, von denen man weiß, wie sie gelöst werden können.

In diesem Abschnitt soll diese sehr vage Bemerkung anhand einiger konkreter Beispiele erläutert werden.

Die unproblematischste aller Differentialgleichungen ist

$$\dot{y}(t) = f(t)\,;$$

einfache Integration führt auf die Lösung

$$y(t) = \int f(t)\,dt + C\,. \qquad (*)$$

(Tatsächlich kann diese Integration alles andere als „einfach“ sein, aber wenn es um Lösungsformeln für Differentialgleichungen geht, wollen wir ein Problem auch dann als „gelöst“ betrachten, wenn die Formel noch Integrale enthält; das Auffinden von Stammfunktionen ist ein getrenntes Problem und hat seine eigenen Methoden.)

Wenn wir $(*)$ unter Symmetriegesichtspunkten betrachten, sehen wir, daß mit jeder Lösung $y(t)$ auch $y(t) + C$ eine Lösung ist.

Falls umgekehrt eine Differentialgleichung die Eigenschaft hat, daß mit $y(t)$ auch $y(t) + C$ für jede Konstante C eine Lösung ist, haben alle

Lösungen dieselbe Ableitung, die Differentialgleichung läßt sich also auf die Form

$$\dot{y}(t) = f(t)$$

bringen. Somit sind die Differentialgleichungen, die durch eine einfache Integration gelöst werden können, dadurch charakterisiert, daß mit jeder Lösungsfunktion $y(t)$ und jede Konstante C auch $y(t) + C$ eine Lösung ist.

Dies können wir dadurch ausnutzen, daß wir eine gegebene Differentialgleichung, in der wir eine Symmetrie erkennen können, so umformen, daß diese Symmetrie transformiert wird in die Addition einer Konstanten.

Ein Beispiel dafür kennen wir bereits, auch wenn wir damals anders vorgegangen sind: Die lineare homogene Differentialgleichung

$$\dot{y}(t) = f(t)y(t)$$

hat mit $y(t)$ auch jedes Vielfache $Cy(t)$ als Lösung. Die Modifikation, die aus Multiplikationen Additionen macht, ist der Logarithmus; wir müssen also schauen, welcher Differentialgleichung die Funktion

$$z(t) = \ln y(t)$$

genügt. Nach der Kettenregel ist

$$\dot{z}(t) = \frac{\dot{y}(t)}{y(t)}\,;$$

somit folgt aus der Ausgangsgleichung, daß

$$\dot{z}(t) = f(t)$$

ist, und das läßt sich in der Tat durch direkte Integration lösen:

$$z(t) = \int f(t)\,dt + C \quad \text{und} \quad y(t) = e^{\int f(t)\,dt}\,.$$

Auch die Lösung der inhomogenen linearen Differentialgleichung

$$\dot{y}(t) = f(t)y(t) + g(t)\,,$$

läßt sich leicht aus Symmetriebetrachtungen herleiten: Ist $y(t)$ eine Lösung dieser Differentialgleichung und $u(t)$ eine Lösung der zugehörigen homogenen Differentialgleichung

$$\dot{u}(t) = f(t)u(t)\,,$$

so ist für jede Konstante C auch

$$y(t) + Cu(t)$$

eine Lösung der gegebenen Differentialgleichung.

Auch hier gibt es eine offensichtliche Modifikation, die aus $y(t)$ eine Funktion macht, die bis auf eine additive Konstante bestimmt ist, nämlich

$$z(t) = \frac{y(t)}{u(t)}\,.$$

Für diese Funktion ist

$$\begin{aligned}\dot{z}(t) &= \frac{u(t)\dot{y}(t) - y(t)\dot{u}(t)}{u(t)^2} = \frac{\dot{y}(t)}{u(t)} - \frac{y(t)}{u(t)}\frac{\dot{u}(t)}{u(t)} \\ &= \frac{f(t)y(t) + g(t)}{u(t)} - z(t)f(t) = f(t)z(t) + \frac{g(t)}{u(t)} - z(t)f(t) \\ &= \frac{g(t)}{u(t)}\,.\end{aligned}$$

Die letztere Funktion kennen wir, sobald wir die homogene Differentialgleichung gelöst haben: Wegen

$$u(t) = e^{\int f(t)\,dt} \qquad \text{ist} \qquad \frac{g(t)}{u(t)} = e^{-\int f(t)\,dt}\,,$$

also

$$z(t) = \int ig(t)e^{-\int^t f(\tau)\,d\tau}\,dt$$

und

$$y(t) = u(t)z(t) = e^{\int f(t)\,dt}\int g(t)e^{-\int^t f(\tau)\,d\tau}\,dt\,.$$

Diese Methode der Variation der Konstanten kann gelegentlich auch dann mit Erfolg angewandt werden, wenn man die ihrer Ableitung zugrundeliegende Symmetrie nicht direkt sieht: Ohne eine solche Symmetrie verliert die Methode zwar ihre Erfolgsgarantie, aber als Ansatz der

vielleicht zum Erfolg führen kann, taugt sie allemal, und die Auflösung von Differentialgleichungen ist nunmal oft mehr Kunst als Wissenschaft.

Betrachten wir als Beispiel die Differentialgleichung

$$\ddot{y}(t) - y(t) = 6\sinh t\,.$$

Erraten einer speziellen Lösung nach dem Motto, daß diese so ähnlich aussehen sollte wie die rechte Seite, legt hier einen Ansatz der Form

$$y(t) = a\cosh t + b\sinh t\,,$$

nahe, aber der führt offensichtlich nicht zum Erfolg: Für jede solche Funktion die linke Seite identisch null ist.

Probieren wir es also mit Variation der Konstanten: Der obige Ansatz ist gleichzeitig die allgemeine Lösung der homogenen Gleichung (und konnte genau deshalb auch unmöglich zu einer Lösung der inhomogenen Gleichung führen); Variation der Konstanten macht daraus den Ansatz

$$y(t) = a(t)\cosh t + b(t)\sinh t\,.$$

Differenzieren führt auf

$$\dot{y}(t) = \dot{a}(t)\cosh t + a(t)\sinh t + \dot{b}(t)\sinh t + b(t)\cosh t$$

und

$$\begin{aligned}\ddot{y}(t) = {}& \ddot{a}(t)\cosh t + 2\dot{a}(t)\sinh t + a(t)\cosh t \\ &+ \ddot{b}(t)\sinh t + 2\dot{b}(t)\cosh t + b(t)\sinh t\,;\end{aligned}$$

Einsetzen in die Differentialgleichung ergibt

$$\big(\ddot{a}(t) + 2\dot{b}(t)\big)\cosh t + \big(\ddot{b}(t) + 2\dot{a}(t)\big)\sinh t = 6\sinh t\,.$$

Das sieht nicht gerade sehr vielversprechend aus: Diese Differentialgleichung ist eher komplizierter als die ursprüngliche.

Zum Glück brauchen wir aber nicht die allgemeine Lösung dieser Differentialgleichung, sondern nur *irgendeine* Lösung. Daher können wir versuchsweise umformen auf neue Gleichungen, die zwar nicht äquivalent sind zur obigen, deren Lösungen – so wir welche finden können – aber zu Lösungen dieser Differentialgleichung führen.

Ein offensichtlicher Kandidat für einen Ansatz ist das System

$$\ddot{a}(t) + 2\dot{b}(t) = 0 \quad \text{und} \quad \ddot{b}(t) + 2\dot{a}(t) = 6\,.$$

Dieses System können wir auffassen als lineares Differentialgleichungssystem für die Ableitungen $\dot{a}(t)$ und $\dot{b}(t)$ und somit nach den allgemeinen Methoden, die wir bereits entwickelt haben, lösen.

Da wir nur eine Lösung suchen, können wir uns diese Mühe allerdings sparen, indem wir heutistisch vorgehen und versuchen, eine spezielle einfache Lösung zu erraten.

Die einfachste Lösung der ersten Gleichung

$$\ddot{a}(t) + 2\dot{b}(t) = 0$$

ist natürlich $a(t) = b(t) = 0$, aber damit können wir nicht die zweite Gleichung lösen. Tatsächlich genügt es aber für eine Lösung der ersten Gleichung bereits, daß

$$\ddot{a}(t) = \dot{b}(t) = 0$$

ist, und wenn wir dies in die zweite Gleichung

$$\ddot{b}(t) + 2\dot{a}(t) = 6$$

einsetzen, folgt, daß $\dot{a}(t) = 3$ sein muß.

Eine Lösung davon ist $a(t) = 3t$ und $b(t) = 0$; damit erhalten wir die spezielle Lösung

$$y(t) = 3t\cosh t\,.$$

Die allgemeine Lösung der Ausgangsgleichung ist daher

$$y(t) = 3t\cosh t + a\cosh t + b\sinh t$$

mit zwei beliebigen Konstanten a und b.

Dieser Lösungsweg sieht sehr nach Trickserei und Glück aus; wenn wir die Gleichung aber umschreiben in ein nichthomogenes lineares Differentialgleichungssystem, können wir die Variation der Konstanten wieder durch Symmetriebetrachtungen rechtfertigen und erhalten ein Verfahren, das (wenn wir Stammfunktionen finden können) stets zu einer Lösung führt.

Wir betrachten dazu gleich das allgemeine System

$$\dot{\vec{y}}(t) = A\vec{y}(t) + \vec{b}(t)$$

mit einer Matrix $A \in \mathbb{C}^{n\times n}$ und einem Vektor $\vec{b}(t)$ von Funktionen $b_i(t)$.

Bei der Suche nach Symmetrien dieses Systems können wir genauso vorgehen wie im eindimensionalen Fall: Ist $\vec{u}(t)$ eine Lösung des homogenen Differentialgleichungssystems

$$\dot{\vec{u}}(t) = A\vec{u}(t)$$

und $\vec{y}(t)$ eine Lösung des inhomogenen Systems, so ist mit

$$\vec{y}(t) = \begin{pmatrix} y_1(t) \\ \vdots \\ y_n(t) \end{pmatrix} \quad \text{und} \quad \vec{u}(t) = \begin{pmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{pmatrix}$$

auch

$$\begin{pmatrix} y_1(t) + C_1u_1(t) \\ \vdots \\ y_n(t) + C_nu_n(t) \end{pmatrix}$$

für beliebige Konstanten $C_1, \ldots, C_n$ eine Lösung.

Beim System $\dot{\vec{y}}(t) = \vec{g}(t)$ ist mit $\vec{y}(t)$ auch $\vec{y}(t)+\vec{C}$ für jeden Vektor $\vec{C}$ von Konstanten eine Lösung und umgekehrt kann jedes System mit dieser Eigenschaft auf die Form $\dot{\vec{y}}(t) = \vec{g}(t)$ gebracht werden. Auch läßt sich, wie im Eindimensionalen, die Lösung auf Integrationen zurückführen: Die allgemeine Lösung ist

$$\vec{y}(t) = \int \vec{g}(t)\,dt + \vec{C}\,,$$

wobei das Integral über die vektorwertige Funktion $\vec{g}\colon \mathbb{R} \to \mathbb{R}^n$ einfach als Abkürzung dafür steht, daß wir jede der Komponenten g_i von $\vec{g}$ einzeln integrieren und die n Ergebnisse wieder zu einem Vektor zusammenfassen.

Für unser ursprüngliches Problem, die Lösung des inhomogenen Differentialgleichungssystems $\vec{y}(t) = A\vec{y}(t) + \vec{b}(t)$ bedeutet dies, daß wir den Vektor

$$\vec{z}(t) \quad \text{mit} \quad z_i(t) = \frac{y_i(t)}{u_i(t)}$$

betrachten sollten, denn der ist wohlbestimmt bis auf die Addition eines konstanten Vektors.

Die Lösungsmenge des homogenen Systems $\dot{\vec{u}}(t) = A\vec{u}(t)$ besteht aus den Funktionen $e^{At}\vec{C}$ mit einem beliebigen Vektor $\vec{C} \in \mathbb{C}^n$; der Lösungsvektor $\vec{y}(t)$ des inhomogenen Systems entsteht daraus durch komponentenweise Multiplikation mit dem noch zu bestimmenden Vektor $\vec{z}(t)$, von dem wir aber immerhin schon wissen, daß er direkt durch Integration einer vektorwertigen Funktion bestimmt werden kann.

Komponentenweise Multiplikation ist keine übliche Vektoroperation; wenn wir mit Vektoren und Matrizen arbeiten wollen, sollten wir sie also möglichst schnell eliminieren.

Für eine Matrix M und zwei Vektoren $\vec{v}$ und $\vec{w}$ zeigt die Multiplikationsformel sofort, daß $M\vec{v}$ komponentenweise multipliziert mit $\vec{w}$ das Gleiche ist wie M multipliziert mit dem komponentenweisen Produkt der Vektoren $\vec{v}$ und $\vec{w}$. In unserem Fall interessiert das komponentenweise Produkt $\vec{y}(t)$ von $\vec{u}(t) = e^{At}\vec{C}$ mit $\vec{z}(t)$, wobei die Wahl des Konstantenvektors $\vec{C}$ uns überlassen bleibt.

Speziell für den Vektor $\vec{C}$, dessen Komponenten allesamt Einsen sind, ist das komponentenweise Produkt von $\vec{C}$ mit einem beliebigen Vektor $\vec{v}$ gleich $\vec{v}$, also ist für diese Wahl von $\vec{C}$

$$\vec{y}(t) = e^{At}\vec{z}(t)\,.$$

Damit können wir auch hier die Gleichung durch Variation der Konstanten lösen: Wir ersetzen einfach den konstanten Vektor $\vec{C}$ durch eine vektorwertige Funktion $\vec{z}(t)$ und müssen nun diese durch Integration bestimmen.

Da die LEIBNIZsche Produktregel auch für matrixwertige Funktionen gilt, ist

$$\dot{\vec{y}}(t) = Ae^{At}\vec{z}(t) + e^{At}\dot{\vec{z}}(t) = A\vec{y}(t) + e^{At}\dot{\vec{z}}(t)\,;$$

$\vec{y}(t)$ erfüllt also genau dann die Differentialgleichung

$$\dot{\vec{y}}(t) = A\vec{y}(t) + \vec{b}(t)\,,$$

wenn

$$\vec{b}(t) = e^{At}\dot{\vec{z}}(t) \quad \text{oder} \quad \dot{\vec{z}}(t) = e^{-At}\vec{b}(t)$$

ist. Damit können wir die Komponenten von $\vec{z}(t)$ als Stammfunktionen der Komponenten des ganz rechts stehenden Vektors von Funktionen berechnen.

Zur Vereinfachung der Schreibweise vereinbaren wir, daß für eine vektorwertige Funktion $\vec{v}(t)$ gelten soll

$$\int \vec{v}(t)\,dt = \begin{pmatrix} \int v_1(t)\,dt \\ \vdots \\ \int v_n(t)\,dt \end{pmatrix} \quad \text{falls} \quad \vec{v}(t) = \begin{pmatrix} v_1(t) \\ \vdots \\ v_n(t) \end{pmatrix};$$

dann ist

$$\vec{z}(t) = \int e^{-At}\vec{b}(t)\,dt\,,$$

und wir haben eine spezielle Lösung gefunden. Die allgemeine Lösung des inhomogenen Systems ist daher

$$\vec{y}(t) = e^{At}\left(\int e^{-At}\vec{b}(t)\,dt\right) + e^{At}\vec{y}_0$$

mit einem beliebigen konstanten Vektor $\vec{y}_0$.

Zur Anwendung dieser Formel auf das obige Beispiel müssen wir die Differentialgleichung

$$\ddot{y}(t) - y(t) = 6\sinh t$$

umschreiben in ein System von Differentialgleichungen erster Ordnung, also in

$$\dot{y}(t) = z(t)$$
$$\dot{z}(t) = y(t) + 6\sinh t\,.$$

Hier ist

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{und} \quad \vec{b}(t) = \begin{pmatrix} 0 \\ 6\sinh t \end{pmatrix};$$

wir müssen zunächst die Matrix e^{At} berechnen.

Wie wir bereits zu Beginn von §1*e)* gesehen haben, ist A^2 die Einheitsmatrix, woraus folgt, daß

$$e^A = \begin{pmatrix} \cosh 1 & \sinh 1 \\ \sinh 1 & \cosh 1 \end{pmatrix} \quad \text{und} \quad e^{At} = \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix}$$

ist, also

$$e^{-At} = \begin{pmatrix} \cosh(-t) & \sinh(-t) \\ \sinh(-t) & \cosh(-t) \end{pmatrix} = \begin{pmatrix} \cosh t & -\sinh t \\ -\sinh t & \cosh t \end{pmatrix} .$$

Wir brauchen die Funktion

$$e^{-At}\vec{b}(t) = \begin{pmatrix} \cosh t & -\sinh t \\ -\sinh t & \cosh t \end{pmatrix} \begin{pmatrix} 0 \\ 6\sinh t \end{pmatrix} = \begin{pmatrix} -6\sinh^2 t \\ 6\sinh t \cosh t \end{pmatrix}$$

und müssen diese integrieren.

Beginnen wir mit dem ersten Eintrag:

$$\begin{aligned} \int -6\sinh^2 t\, dt &= -6\int \left(\frac{e^t - e^{-t}}{2}\right)^2 dt = -\frac{3}{2}\int (e^{2t} + e^{-2t} - 2)\, dt \\ &= -\frac{3}{2}\left(\frac{e^{2t}}{2} - \frac{e^{-2t}}{2} - 2t\right) + C \\ &= -\frac{3}{4}(e^t - e^{-t})(e^t + e^{-t}) + 3t + C' \\ &= -3\sinh t \cosh t + 3t + C' . \end{aligned}$$

Genauso finden wir auch eine Stammfunktion des zweiten Eintrags:

$$\begin{aligned} \int 6\sinh t \cosh t\, dt &= \frac{3}{2}\int (e^{2t} - e^{-2t})\, dt \\ &= \frac{3}{4}\left(e^{2t} + e^{-2t}\right) + C \\ &= \frac{3}{4}\left(e^{2t} + 2 + e^{-2t}\right) + C - \frac{3}{2} \\ &= 3\left(\frac{e^t + e^{-t}}{2}\right)^2 + C' = 3\cosh^2 t + C' . \end{aligned}$$

Da wir nur eine spezielle Lösung brauchen, können wir die beiden Integrationskonstanten unbesorgt auf null setzen; jede andere Wahl würde nur bedeuten, daß wir eine Lösung der homogenen Gleichung dazuaddieren. Also arbeiten wir mit

$$\vec{u}(t) = \begin{pmatrix} -3\sinh t \cosh t + 3t \\ 3\cosh^2 t \end{pmatrix}$$

und die gesuchte spezielle Lösung ist

$$\begin{aligned} e^{At}\vec{u}(t) &= \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix} \begin{pmatrix} -3\sinh t \cosh t + 3t \\ 3\cosh^2 t \end{pmatrix} \\ &= \begin{pmatrix} -3\sinh t \cosh^2 t + 3t\cosh t + 3\cosh^2 t \sinh t \\ -3\sinh^2 t \cosh t + 3t\cosh t + 3\cosh^3 t \end{pmatrix} . \end{aligned}$$

In der ersten Zeile dieses Ergebnisses heben sich der erste und der dritte Term gegenseitig weg; in der zweiten ist

$$3\cosh^3 t - 3\sinh^2 t \cosh t = 3\cosh t(\cosh^2 t - \sinh^2 t) = 3\cosh t .$$

Als Endergebnis erhalten wir somit die spezielle Lösung

$$\begin{pmatrix} y(t) \\ z(t) \end{pmatrix} = e^{At}\vec{u}(t) = \begin{pmatrix} 3t\cosh t \\ 3t\sinh t + 3\cosh t \end{pmatrix} .$$

Insbesondere ist

$$y(t) = 3t\cosh t$$

eine spezielle Lösung der Ausgangsgleichung

$$\ddot{y}(t) - y(t) = 6\sinh t ,$$

und die allgemeine Lösung dieser Differentialgleichung ist

$$y(t) = 3t\cosh t + a\cosh t + b\sinh t$$

mit beliebigen Konstanten a, b aus $\mathbb{R}$ oder $\mathbb{C}$ – je nachdem, über welchem der beiden Körper wir das Problem betrachten.

Die Beispiele aus diesem Abschnitt zeigen nur einen winzigen Ausschnitt der Möglichkeiten, wie Symmetriebetrachtungen zu Lösungen von Differentialgleichungen führen können; ihre volle Nützlichkeit entfalten sie erst bei nichtlinearen Differentialgleichungen und Differentialgleichungssystemen. Im Rahmen dieser Vorlesung bleibt keine Zeit,

näher darauf einzugehen; eine ausführliche Darstellung von Symmetriemethoden findet man etwa bei

G.W. BLUMAN, S. KUMEI: Symmetries and Differential Equations, *Springer,* 1989

Symmetriemethoden zur Lösung von Differentialgleichungen wurden ab etwa 1880 von dem norwegischen Mathematiker SOPHUS LIE (1842–1899) eingeführt. Insbesondere zeigte LIE auch, daß man nicht wirklich die (schwer bestimmbaren) Symmetrien eines Systems bestimmen muß, sondern daß bereits die mit linearer Algebra bestimmbaren sogenannten *infinitesimalen* Symmetrien ausreichen können, um Lösungen zu finden. Mit diesem Ansatz arbeiten auch die heutigen Computeralgebrasysteme; ein einfaches Beispiel dazu wird uns im nächsten Paragraphen bei der Suche nach integrierenden Faktoren begegnen.

Ausgebaut wurde die Methode von EMMY NOETHER (1882–1935), der Tochter des Mannheimer Mathematikers MAX NOETHER (1844-1921). Sie brachte Symmetrien mit den in den Naturwissenschaften allgegenwärtigen Erhaltungssätzen in Verbindung und bereitete damit auch EINSTEINs Relativitätstheorie vor. Bekannter ist sie allerdings als Mitbegründerin der modernen abstrakten Algebra. Nur dank der massiven Intervention HILBERTs durfte sie sich nach langem Kampf 1919 in Göttingen als erste Frau in Mathematik habilitieren. 1933 wurde sie als Jüdin von der Universität Göttingen entlassen und emigrierte nach USA, wo sie am Bryn Mawr College und dem Institute for Advanced Study in Princeton arbeitete.

f) Lineare homogene Differenzengleichungen

In einer analogen Schaltung ändern sich die relevanten physikalischen Größen kontinuierlich, so daß die Schaltung gut Differentialgleichungen beschrieben werden kann. In der Digitaltechnik dagegen hat man es mit getakteten Schaltungen zu tun, bei denen sich (idealerweise) während eines Takt gar nichts ändert, d.h. alle Ableitungen verschwinden. Somit sind Differentialgleichungen hier kein geeignetes Beschreibungsmittel.

Der Zustand der Schaltung während eines Takts sollte allerdings in deterministischer Weise von den Zuständen in den vorherigen Takten abhängen, und dies führt auf sogenannte *Differenzengleichungen*. Da alle Größen während eines Takts konstante Werte haben, können wir sie durch Funktionen auf $\mathbb{Z}$ oder $\mathbb{N}_0$ modellieren, wobei wir den Wert von y im Takt Nummer n als y_n bezeichnen.

Ein System von Differenzengleichungen erster Ordnung in den Variablen $y^{(1)}, \ldots, y^{(r)}$ ist ein System von Gleichungen

$$y_n^{(i)} = f_i(y_{n-1}^{(1)}, \ldots, y_{n-1}^{(r)}) ;$$

im linearen homogenen Fall gibt es also reelle Zahlen a_{ij}, so daß

$$y_n^{(i)} = a_{i1} y_{n-1}^{(1)} + \cdots + a_{ir} y_{n-1}^{(r)}$$

ist. Fassen wir die $y_n^{(i)}$ zusammen zu einem Vektor $\vec{y}_n$ und die a_{ij} zu einer Matrix A, wird dieses System zur Gleichung $\vec{y}_n = A\vec{y}_{n-1}$ mit der offensichtlichen Lösung $\vec{y}_n = A^n \vec{y}_{n-1}$.

Interessant ist auch hier vor allem der Fall von Gleichungen höherer Ordnung; im Falle von nur einer Variablen haben wir dann eine Gleichung der Form

$$y_n = a_1 y_{n-1} + a_2 y_{n-2} + \cdots + a_r y_{n-r} .$$

Wie im Falle der Differentialgleichungen können wir diese umschreiben in ein System, indem wir r Variablen $y^{(0)}, \ldots, y^{(r-1)}$ einführen mit $y^{(0)} = y$ und $y_n^{(i)} = y_{n-i}$ die um i Takte verschobene Variable y. Dann ist

$$\vec{y}_n = A\vec{y}_{n-1} \quad \text{mit} \quad A = \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 \\ a_r & a_{r-1} & a_{r-2} & a_{r-3} & \cdots & a_1 \end{pmatrix} .$$

Die Lösung y_n der Ausgangsgleichung ist dann die nullte Komponente des Vektors $A^n \vec{c}$, wobei $c_0, \ldots, c_{r-1}$ die Werte von y in den ersten r Takten sind.

Aus der entsprechenden Diskussion bei Differentialgleichungen wissen wir, daß A das charakteristische Polynom

$$(-1)^r(\lambda^r - a_1\lambda^{r-1} - a_2\lambda^{r-2} - \cdots - a_{r-2}\lambda - a_{r-1})$$

hat und daß es zu jeder s-fachen Nullstelle Hauptvektoren bis zur Stufe s gibt. Wählen wir eine Basis aus Hauptvektoren, bezüglich derer A obere Dreiecksgestalt $D + N$ hat, enthält der Block zum Eigenwert λ also Einträge in allen Potenzen N^j mit $j < s$. Somit sind die Einträge von $(D + N)^n$ im Block zum Eigenwert λ Linearkombinationen von $\lambda^n, \lambda^{n-1}, \ldots, \lambda^{n+1-s}$. Damit sind auch die Einträge von A^n sowie die von $A^n\vec{c}$ Linearkombinationen von Termen der Form λ^{n-j} mit Eigenwerten λ und j zwischen Null und eins weniger als die algebraische Vielfachheit von λ. Insbesondere ist auch y_n von dieser Form.

Falls alle Nullstellen einfach sind, ist y_n daher einfach eine Linearkombination der n-ten Potenzen der verschiedenen Nullstellen.

Als Beispiel können wir die FIBONACCI-Zahlen betrachten. Sie sind durch folgende Rekursionsformel definiert:

$$F_0 = 0, \qquad F_1 = 1 \qquad \text{und} \qquad F_i = F_{i-1} + F_{i-2} \quad \text{für } i \geq 2\,.$$

FIBONACCI führte sie ein, um die Vermehrung einer Karnickelpopulation durch ein einfaches Modell zu berechnen.. In seinem 1202 erschienenen Buch *Liber abaci* schreibt er:

Ein Mann bringt ein Paar Karnickel auf einen Platz, der von allen Seiten durch eine Mauer umgeben ist. Wie viele Paare können von diesem Paar innerhalb eines Jahres produziert werden, wenn man annimmt, daß jedes Paar jeden Monat ein neues Paar liefert, das vom zweiten Monat nach seiner Geburt an produktiv ist?

Wir haben somit die Differenzengleichung

$$F_n = F_{n-1} + F_{n-2} \quad \text{mit} \quad F_0 = 0 \quad \text{und} \quad F_1 = 1\,.$$

Diese führt auf das Polynom $\lambda^2 - \lambda - 1$ mit Nullstellen

$$\lambda_{1/2} = \tfrac{1}{2}(1 \pm \sqrt{5})\,,$$

es gibt also reelle Zahlen a, b, so daß $F_n = a\lambda_1^n + b\lambda_2^n$ ist. Aus $F_0 = 0$ folgt sofort, daß $b = -a$ sein muß, und die Bedigung $F_1 = 1$ zeigt dann, daß $a = 1/\sqrt{5}$ ist. Somit ist

$$F_n = \frac{\lambda_1^n - \lambda_2^n}{\sqrt{5}} = \frac{\lambda_1^n + |\lambda_2|^n}{\sqrt{5}} = \left[\frac{\lambda_1^n}{\sqrt{5}}\right]\,,$$

denn $|\lambda_2|^n/\sqrt{5}$ ist für alle $n \in \mathbb{N}_0$ kleiner als eins, und F_n ist natürlich eine ganze Zahl.

LEONARDO PISANO (1170–1250) ist heute vor allem unter seinem Spitznamen FIBONACCI bekannt; gelegentlich nannte er sich auch BIGOLLO, auf Deutsch *Tunichtgut* oder *Reisender.* Seine Bücher waren mit die ersten, die die indisch-arabischen Ziffern in Europa einführten. Er behandelt darin nicht nur Rechenaufgaben für Kaufleute, sondern auch zahlentheoretische Fragen, beispielsweise daß man die Quadratzahlen durch Aufaddieren der ungeraden Zahlen erhält. Auch betrachtet er Beispiele nichtlinearer Gleichungen, die er approximativ löst, und erinnert an viele in Vergessenheit geratene Ergebnisse der antiken Mathematik.

§4: Nichtlineare Differentialgleichungen

Lineare Differentialgleichungen spielen vor allem deshalb eine so große Rolle in den Anwendungen, weil es dafür (wie wir im letzten Paragraphen gesehen haben) eine gut ausgebaute Lösungstheorie gibt. Oft werden deshalb sogar für ihrem Wesen nach nichtlineare Probleme lineare Approximation betrachtet, um so wenigstens zu einem ungefähren Verständnis der Dynamik des Systems zu gelangen.

Nichtlineare Differentialgleichungen können nur selten explizit in geschlossener Form gelöst werden. Trotzdem lohnt sich die Suche nach einer Lösungsformel, denn eine solche Formel gestattet erstens bessere theoretische Aussagen über das Verhalten der Lösungen und zweitens kann eine einmal gefundene Lösungsformel immer wieder auf Systeme mit anderen Anfangswerten und/oder Parametern angewandt werden, wohingegen eine numerische Lösung für jede Konstellation von Anfangswerten neu berechnet werden muß. Auch ist die Auswertung einer

expliziten Lösungsformel zwar keinesfalls *immer* einfacher (oder numerisch stabiler) als eine direkte numerische Lösung, aber doch meistens.

Dieser Paragraph soll anhand einiger einfacher Beispiele und Sätze einen ersten Überblick über das sehr weite Gebiet der nichtlinearen Differentialgleichungen geben.

a) Eindeutigkeitsfragen

Wir hatten Differentialgleichungen eingeführt, um aus dem gegenwärtigen Zustand eines Systems Folgerungen über die künftige Entwicklung zu ziehen. Dies ist natürlich nur dann möglich, wenn das zugehörige Anfangswertproblem eindeutig lösbar ist. Bei linearen Differentialgleichungen mit konstanten Koeffizienten ist das, wie wir im vorigen Abschnitt gesehen haben, immer der Fall; in diesem Abschnitt wollen wir uns überlegen, welche Probleme im nichtlinearen Fall auftreten können.

Betrachten wir als erstes die Differentialgleichung

$$\dot{y}(t) = \frac{1}{2y(t)} ,$$

wobei wir uns hier wie bei allen Beispielen in diesem Abschnitt der Einfachheit halber auf *reelle* Lösungsfunktionen beschränken wollen.

Multiplikation beider Seiten mit $2y(t)$ führt auf

$$2\dot{y}(t)y(t) = 1 ,$$

die Ableitung von $y(t)^2$ ist also gleich 1. Damit ist

$$y(t)^2 = t + C \quad \text{oder} \quad y(t) = \pm\sqrt{t + C}$$

mit einer beliebigen Konstanten $C \in \mathbb{R}$. Fordern wir noch, daß $y(0) = 0$ sein soll, so hat das entsprechende Anfangswertproblem die beiden Lösungen

$$y(t) = \sqrt{t} \quad \text{und} \quad y(t) = -\sqrt{t} ,$$

von denen die eine für $t \to \infty$ gegen $+\infty$ geht, die andere gegen $-\infty$.

Als zweites Beispiel betrachten wir das Anfangswertproblem

$$\dot{y}(t) = 2\sqrt{|y(t)|} \quad \text{mit} \quad y(0) = 0 .$$

Es hat natürlich die Nullfunktion als Lösung, aber auch die stetig differenzierbare Funktion

$$y(t) = \begin{cases} t^2 & \text{für } t \geq 0 \\ -t^2 & \text{für } t \leq 0 \end{cases} .$$

($y(t) = t^2$ ist für $t < 0$ *keine* Lösung.) Da die Ableitung von $\pm t^2$ für $t = 0$ verschwindet, sind aber auch die beiden zusammengesetzten Funktionen

$$y(t) = \begin{cases} 0 & \text{für } t \geq 0 \\ -t^2 & \text{für } t \leq 0 \end{cases} \quad \text{und} \quad y(t) = \begin{cases} t^2 & \text{für } t \geq 0 \\ 0 & \text{für } t \leq 0 \end{cases}$$

stetig differenzierbar und Lösungen.

Es kommt noch schlimmer: Für jedes $a \in \mathbb{R}$ löst

$$y(t) = (t - a)^2$$

für $t \geq a$ (und nur dort) die Differentialgleichung, wenn auch nicht das Anfangswertproblem. Für $a \geq 0$ können wir dazu die zusammengesetzte Funktion

$$y(t) = \begin{cases} 0 & \text{für } t \leq a \\ (t - a)^2 & \text{für } t \geq a \end{cases}$$

betrachten; diese ist in jedem Punkt t stetig differenzierbar und erfüllt die Differentialgleichung samt Anfangsbedingung $y(0) = 0$.

Auf Grund der Differentialgleichung und der Anfangsbedingung wissen wir also nur, daß die Funktion entweder konstant gleich null ist oder aber nach einem Zeitraum $a \geq 0$ damit beginnt, quadratisch gegen unendlich zu gehen – nicht gerade viel Information.

Bei der Differentialgleichung

$$\dot{y}(t) + \frac{2y(t)}{t^3} = 0$$

schließlich überzeugt man sich leicht, daß für jedes $\lambda \in \mathbb{R}$ die Funktion $y(t) = \lambda e^{-1/t^2}$ eine Lösung ist. Diese Funktion ist zwar *a priori* für $t = 0$ nicht definiert, sie kann aber stetig ergänzt werden durch die Vorschrift, daß ihr Wert dort gleich null sein soll. Genauso verhält es sich mit ihrer Ableitung

$$\dot{y}(t) = -\frac{2y(t)}{t^3} = -\frac{2\lambda}{t^3} \cdot e^{-1/t^2} ,$$

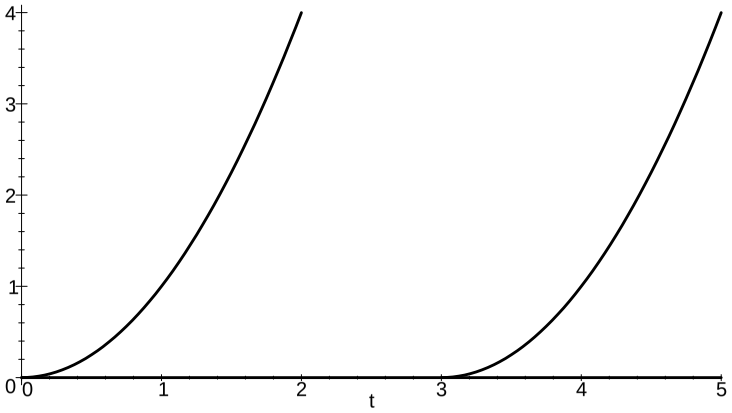


Abb. 34: Drei Lösungen von $\dot{y}(t) = \sqrt{|y(t)|}$ mit $y(0) = 0$

denn e^{-1/t^2} geht schneller gegen null als $\frac{2\lambda}{t^3}$ gegen unendlich. Also ist die Funktion stetig differenzierbar (sogar beliebig oft, wie man sich ohne große Schwierigkeiten überlegen kann). Ihr Wert an der Stelle $t = 0$ ist unabhängig von λ immer gleich null, d.h. das Anfangswertproblem

$$\dot{y}(t) + \frac{2y(t)}{t^3} = 0 \quad \text{und} \quad y(0) = 0$$

hat *jede* der Funktionen $y(t) = \lambda e^{-1/t^2}$ als Lösung, egal welchen Wert $\lambda \in \mathbb{R}$ man auch wählt. Über die Entwicklung eines Systems, das durch diese Gleichung beschrieben wird, läßt sich also *nichts* aussagen. Erst wenn $x(t_0)$ für einen Wert $t_0 \neq 0$ bekannt ist, lassen sich die Lösungen mit den verschiedenen Parameterwerten λ voneinander unterscheiden. Für $t \to \infty$ konvergiert die Lösungsfunktion gegen den Wert λ, der somit die entscheidende Größe für die Beschreibung des Langzeitverhaltens ist und trotzdem durch das Anfangswertproblem völlig unbestimmt gelassen wird.

Falls man Differentialgleichungen zur Beschreibung und vor allem zur Vorhersage realer Phänomene einsetzt, ist es also definitiv kein überflüssiger Luxus, wenn man sich zunächst darum kümmert, ob die Differentialgleichung zusammen mit ihrer Anfangsbedingung überhaupt

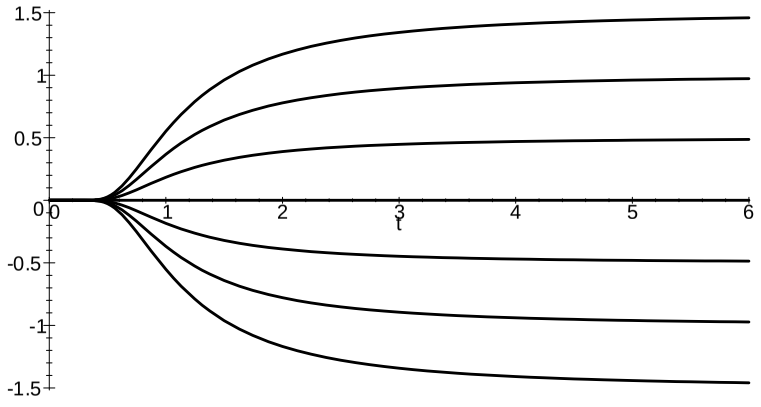


Abb. 35: Sieben Lösungen von $\dot{y}(t) + \frac{2y(t)}{t^3} = 0$ mit $y(0) = 0$

eine eindeutig bestimmte Lösungsfunktion festlegt. Ohne diese Vorüberlegung ist jede Lösungsfunktion wertlos.

b) Der Satz von Picard und Lindelöf

Wir betrachten in diesem Abschnitt das eindimensionale Anfangswertproblem

$$\dot{y}(t) = f\big(y(t), t\big) \quad \text{mit} \quad y(t_0) = c_0 \,.$$

Dabei sei

$$f\colon \begin{cases} \mathbb{R} \times [t_0,\, t_1] \to \mathbb{R} \\ \qquad (y,t) \mapsto f(y,t) \end{cases}$$

eine stetige Funktion, d.h. für jeden Punkt (y,t) aus dem Definitionsbereich und jedes $\varepsilon > 0$ gibt es ein $\delta > 0$, so daß

$$\big|f\left(\widetilde{y},\widetilde{t}\right) - f(y,t)\big| < \varepsilon \quad \text{falls} \quad |\widetilde{y} - y| < \delta \text{ und } \big|\widetilde{t} - t\big| < \delta \,.$$

Indem wir beide Seiten der Differentialgleichung ab t_0 integrieren, erhalten wir unter Berücksichtigung der Anfangsbedingung $y(t_0) = c_0$ die neue Gleichung

$$y(t) = c_0 + \int_{t_0}^{t} f\big(y(\tau), \tau\big)\, d\tau \,,$$

aus der durch Differenzieren sofort folgt, daß jede ihrer Lösungen das obige Anfangswertproblem löst.

Wir bezeichnen die Funktion auf der rechten Seite mit $T(y)$, d.h.

$$T(y)(t) \underset{\text{def}}{=} c_0 + \int_{t_0}^{t} f\big(y(\tau), \tau\big)\, d\tau \,.$$

Dies definiert einen Operator, der jeder stetigen Funktion auf $[t_0, t_1]$ wieder eine solche Funktion zuordnet, und wir können nun kurz sagen, daß wir eine Funktion suchen, die der Gleichung $y = T(y)$ genügt, oder, anders ausgedrückt, einen Fix„punkt“ des Operators T.

Um einen solchen Fixpunkt zu erhalten ‚starten wir mit einer beliebigen Funktion $y_0(t)$ auf dem Intervall $[t_0, t_1]$ und betrachten die Folge von Funktionen $y_n(t)$, die für $n \geq 1$ rekursiv durch $y_n = T(y_{n-1})$ definiert ist, d.h. $y_n = T^n(y_0)$. Falls diese Folge gegen eine Funktion y_∞ konvergiert, ist klar, daß diese Funktion Fixpunkt von T sein muß, denn natürlich ändert dann eine weitere Anwendung von T nichts mehr.

Als Beispiel betrachten wir das (wohlbekannte) Anfangswertproblem

$$\dot{y}(t) = y(t) \quad \text{mit} \quad y(0) = 1 \,.$$

Hier ist

$$T(y)(t) = 1 + \int_0^t y(\tau)\, d\tau \,;$$

falls wir also von der konstanten Funktion $y_0(t) = 1$ ausgehen, ist

$$y_1(t) = 1 + \int_0^t d\tau = 1 + t \,,$$

$$y_2(t) = 1 + \int_0^t (1 + \tau)\, d\tau = 1 + t + \frac{t^2}{2} \,,$$

$$y_3(t) = 1 + \int_0^t \left(1 + \tau + \frac{\tau^2}{2}\right) d\tau = 1 + t + \frac{t^2}{2} + \frac{t^3}{6} \,,$$

und so weiter. Ab hier wird man erraten, daß die Funktionen $y_n(t)$ gerade die TAYLOR-Polynome n-ten Grades der Exponentialfunktion sind, was sich dann leicht durch vollständige Induktion beweisen läßt. Somit ist

$$\lim_{n\to\infty} y_n(t) = e^t \,,$$

wie auch nicht anders zu erwarten war.

In diesem Beispiel geht also alles gut; in anderen Fällen kann es, wie die Beispiele aus dem vorigen Abschnitt zeigten, Probleme geben. Um zu positiven Ergebnisse zu kommen, wollen wir mehr von f verlangen als die bloße Stetigkeit und uns auch auf ein abgeschlossenes Intervall $[t_0, t_1]$ beschränken statt t auf der gesamten reellen Achse variieren zu lassen. Eine mögliche Verschärfung der Stetigkeitsforderung ist die

Lipschitz-Bedingung: Es gibt eine Konstante $L \in \mathbb{R}$, die sogenannt genannt LIPSCHITZ-Konstante, so daß für alle $t \in [t_0, t_1]$ und alle $y_1, y_2 \in \mathbb{R}$ gilt:

$$|f(y_2, t) - f(y_1, t)| \leq L\,|y_2 - y_1| \,.$$

RUDOLF OTTO SIGISMUND LIPSCHITZ (1832–1903) wurde in Königsberg geboren und starb in Bonn. Am besten bekannt ist er durch die gerade definierte LIPSCHITZ-Bedingung für die Existenz und Eindeutigkeit der Lösungen von Differentialgleichungen. Sein Hauptarbeitsgebiet waren die Differentialgleichungen der mathematischen Physik, jedoch beschäftigte er sich auch mit anderen Hilfsmitteln der mathematischen Physik wie etwa Matrizengruppen. Seine Arbeiten über dynamische Systeme haben wichtige Anwendungen in der Himmelsmechanik.

Die LIPSCHITZ-Bedingung muß glücklicherweise in vielen Fällen nicht explizit durch Abschätzungen nachgeprüft werden, denn es gilt

Lemma: Falls die partielle Ableitung von f nach y existiert und betragsmäßig beschränkt ist durch eine Konstante L, genügt f einer LIPSCHITZ-Bedingung mit L als LIPSCHITZ-Konstante.

Beweis: Nach dem Mittelwertsatz der Differentialrechnung gibt es für jedes feste t (das wir bezüglich der partiellen Ableitung wie eine Konstante betrachten können) und für je zwei Werte $y_1, y_2 \in \mathbb{R}$ einen Punkt $z \in [y_1,\ y_2]$ mit der Eigenschaft, daß

$$\frac{f(y_2,t) - f(y_1,t)}{y_2 - y_1} = \frac{\partial f}{\partial y}(z,t)$$

ist. Also ist

$$|f(y_2,t) - f(y_1,t)| = \left|\frac{\partial f}{\partial y}(z,t)\right| \ |y_2 - y_1| \ ,$$

und da der Betrag der partiellen Ableitung überall höchstens L ist, erfüllt f eine LIPSCHITZ-Bedingung mit Konstante L. ∎

Wir betrachten nun wieder den oben eingeführten Operator T, von dem uns im Augenblick nur interessieren soll, daß er stetige Funktionen auf dem Intervall $[t_0,\ t_1]$ in ebensolche Funktionen überführt. Wir suchen eine Funktion $y(t)$ mit $T(y) = y$, die wir wie oben als Limes einer Funktionenfolge konstruieren wollen.

Um von einem solchen Limes reden zu können, brauchen wir zunächst eine Norm; wir verwenden dazu die *Supremumsnorm*, die für auf $[t_0,\ t_1]$ stetige Funktionen h durch

$$||h|| \underset{\text{def}}{=} \sup_{\tau \in [t_0,\ t_1]} |h(\tau)|$$

definiert ist. Da eine stetige Funktion in einem abgeschlossenen Intervall ihr Maximum annimmt, existiert dieses Supremum und ist sogar ein Funktionswert von h.

Über dem Intervall $[0,\ 10]$ beispielsweise ist $||\sin|| = ||\cos|| = 1$, und für die Funktion $g(t) = t^2$ ist $||g|| = 10^2 = 100$.

Unser wichtigstes Hilfsmittel zur Untersuchung, wann eine Grenzfunktion existiert, ist der BANACHsche Fixpunktsatz, der in diesem Semester in der Numerik allgemein formuliert und bewiesen wird. Um keine neuen Begriffe einführen zu müssen, begnüge ich mich hier mit dem für unsere Zwecke notwendigen Spezialfall; der Beweis vereinfacht sich dadurch zwar nicht, aber die Formulierung des Satzes wird leichter verständlich.

Banachscher Fixpunktsatz: Für einen Operator T, der stetige Funktionen auf $[t_0,\ t_1]$ auf ebensolche Funktionen abbildet, gebe es eine Konstante $K < 1$ derart, daß für alle auf $[t_0,\ t_1]$ stetigen Funktionen $w(t), z(t)$ gilt

$$||T(z) - T(w)|| \le K\,||z - w|| \ .$$

Dann gibt es genau eine stetige Funktion $y\colon [t_0,\ t_1] \to \mathbb{R}$ mit der Eigenschaft $T(y) = y$.

Beweis: Zunächst ist klar, daß es *höchstens* eine solche Funktion gibt, denn sind $y(t)$ und $z(t)$ zwei Lösungen, so ist

$$||z - y|| = ||T(z) - T(y)|| \le K\,||z - y|| \ ,$$

was für $K < 1$ nur dann möglich ist, wenn $||z - y||$ verschwindet, wenn also überall $y(t) = z(t)$ ist.

Zum Nachweis der Existenz starten wir mit irgendeiner stetigen Funktion $y_0(t)$ und iterieren sie mit T, d.h. für $n > 0$ sei $y_n = T(y_{n-1})$.

Wir überlegen uns zunächst, daß für jedes feste $\tau \in [t_0,\ t_1]$ die Folge der reellen Zahlen $y_n(\tau)$ konvergiert. Nach dem CAUCHYschen Konvergenzkriterium müssen wir dazu zeigen, daß es zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$ gibt, so daß

$$|y_m(\tau) - y_n(\tau)| \le \varepsilon \quad \text{falls} \quad m > n > N$$

ist. Da wir mit der Supremumsnorm arbeiten, ist

$$|y_m(\tau) - y_n(\tau)| \le ||y_m - y_n|| \ ,$$

und nach Voraussetzung ist für jedes $r \in \mathbb{N}$

$$||y_{r+2} - y_{r+1}|| \le K\,||y_{r+1} - y_r|| \ .$$

Induktiv folgt daraus, daß

$$||y_{r+1} - y_r|| \le K^r\,||y_1 - y_0||$$

ist und damit nach der Dreiecksungleichung

$$\begin{aligned}\|y_m - y_n\| &\le \|y_m - y_{m-1}\| + \cdots + \|y_{n+1} - y_n\| \\ &\le K^m \|y_1 - y_0\| + \cdots + K^n \|y_1 - y_0\| \\ &= (K^m + \cdots + K^n)\, \|y_1 - y_0\| \\ &= K^n(1 + K + \cdots + K^{m-n})\, \|y_1 - y_0\| \\ &\le K^n \left(\sum_{i=0}^{\infty} K^i \right) \|y_1 - y_0\| = \frac{K^n}{1-K} \|y_1 - y_0\| \ .\end{aligned}$$

Wählen wir also N so, daß

$$\frac{K^N}{1-K} \|y_1 - y_0\| < \varepsilon$$

ist, gilt die Voraussetzung des CAUCHYschen Konvergenzkriteriums mit diesem N für alle $\tau \in [t_0,\, t_1]$. Somit konvergiert die Folge der Funktionen $y_n(t)$ *punktweise* gegen eine Funktion $y(t)$. Für jedes $\tau \in [t_0,\, t_1]$ ist dann

$$\begin{aligned}|y(\tau) - y_n(\tau)| &= \lim_{m\to\infty} |y_m(\tau) - y_n(\tau)| \le \lim_{m\to\infty} \|y_m - y_n\| \\ &\le \frac{K^n}{1-K} \|y_1 - y_0\| \ ,\end{aligned}$$

die Folge konvergiert also in $[t_0,\, t_1]$ *gleichmäßig* gegen $y(t)$. Damit ist $y(t)$ eine stetige Funktion, und der Satz ist bewiesen. ∎

STEFAN BANACH (1892–1945) wurde in Krakau geboren und ausgebildet, promovierte und arbeitete dann aber an der Universität von Lvov in der Ukraine, wo er unter schwierigen Bedingungen unter deutscher Besatzung den zweiten Weltkrieg verbrachte. Durch seine Arbeiten über lineare Operatoren und über Vektorräume von Funktionen wurde er zum Begründer der modernen Funktionalanalysis. Nach dem Krieg wollte er auf einen Lehrstuhl an der Universität Krakau wechseln, starb aber 1945 an Lungenkrebs. Das wichtigste mathematische Forschungsinstitut Polens, das Banach-Zentrum in Warschau, ist nach ihm benannt.

Als Anwendung erhalten wir den

Satz von Picard-Lindelöf: Die stetige Funktion

$$f\colon \begin{cases} \mathbb{R} \times [t_0,\, t_1] \to \mathbb{R} \\ \qquad (y,t) \mapsto f(y,t) \end{cases}$$

genüge einer LIPSCHITZ-Bedingung mit irgendeiner Konstanten $L \in \mathbb{R}$. Dann hat das Anfangswertproblem

$$\dot{y}(t) = f\big(y(t), t\big) \quad \text{mit} \quad y(t_0) = c_0$$

für jedes $c_0 \in \mathbb{R}$ eine in $[t_0,\, t_1]$ eindeutig bestimmte Lösung.

Beweis: Wir müssen zeigen, daß für den oben definierten Operator T mit

$$T(y)(t) \underset{\text{def}}{=} c_0 + \int_{t_0}^{t} f\big(y(\tau), \tau\big)\, d\tau$$

und zwei beliebige stetige Funktionen z, w auf $[t_0,\, t_1]$

$$\|T(z) - T(w)\| \le K \|z - w\|$$

ist mit einer Konstanten $K < 1$. Links steht die Norm jener Funktion, die an der Stelle t den Wert

$$T(z)(t) - T(w)(t) = \int_{t_0}^{t} \Big(f\big(z(\tau), \tau\big) - f\big(w(\tau), \tau\big) \Big)\, d\tau$$

annimmt, und auf Grund der LIPSCHITZ-Bedingung ist

$$\big| f\big(z(\tau), \tau\big) - f\big(w(\tau), \tau\big) \big| \le L\, |z(\tau) - w(\tau)| \le L\, \|z - w\| \ .$$

Also ist für $t \in [t_0,\, t_1]$

$$\begin{aligned}&\left| \int_{t_0}^{t} \Big(f\big(z(\tau), \tau\big) - f\big(w(\tau), \tau\big) \Big)\, d\tau \right| \le \int_{t_0}^{t} L\, \|z - w\|\, d\tau \\ &= L \cdot (t - t_0)\, \|z - w\| \le L \cdot (t_1 - t_0)\, \|z - w\| \ .\end{aligned}$$

Falls $L \cdot (t_1 - t_0)$ kleiner als eins ist, haben wir eine Konstante $K < 1$ gefunden und können den BANACHschen Fixpunktsatz anwenden; er

liefert uns eine stetige Funktion $y(t)$, die der Bedingung

$$y(t) = c_0 + \int_{t_0}^{t} f\big(y(\tau), \tau\big)\, d\tau$$

genügt. Da die rechte Seite wegen der Stetigkeit von f differenzierbar ist, haben wir sogar eine differenzierbare Funktion, und $\dot{y}(t) = f\big(y(t), t\big)$. Außerdem ist $y(t_0) = c_0$, da das Integral von t_0 bis t für $t = t_0$ verschwindet. Wir haben somit eine Lösung des Anfangswertproblems gefunden, und das ist die einzige Lösung, denn nach dem BANACHschen Fixpunktsatz hat T nur einen Fixpunkt.

Falls $L\cdot(t_1 - t_0)$ größer als eins ist, gibt es immerhin noch eine natürliche Zahl n, so daß $L \cdot (t_1 - t_0) < n$ ist. Wir unterteilen das Intervall $[t_0,\ t_1]$ in n gleich lange Teilintervalle $[\tau_i,\ \tau_{i+1}]$ mit Grenzen

$$t_0 = \tau_0 < \tau_1 < \cdots < \tau_{n-1} < \tau_n = t_1\,,$$

d.h. $\tau_i = t_0 + \frac{i}{n}(t_1 - t_0)$. Für jedes der Intervalle $[\tau_i,\ \tau_{i+1}]$ ist dann $L\cdot(\tau_{i+1} - \tau_i) < 1$; insbesondere ist also das Anfangswertproblem im Anfangsintervall $[\tau_0,\ \tau_1]$ eindeutig lösbar. Für seine Lösung sei $y(\tau_1) = c_1$. Dann betrachten wir im Intervall $[\tau_1,\ \tau_2]$ das entsprechende Anfangswertproblem mit $y(\tau_1) = c_1$ und erhalten eine in diesem Intervall eindeutige Lösung und so weiter, bis wir die Lösung auf ganz $[t_0, t_1]$ ausgedehnt haben.

■

ÉMILE PICARD (1856–1941) wurde in Paris geboren und hatte an der dortigen Universität auch seine erste Stelle, bis er 1878 eine Professur in Toulouse bekam. 1898 wurde er Professor an der Sorbonne und kehrte nach Paris zurück, wo er auch starb. Seine Arbeiten beschäftigen sich mit der Analysis (z.B. dem gerade bewiesenen Satz) und der Geometrie. Dort interessierte er sich für algebraische Flächen und die eng damit verbundenen algebraiscchen Funktionen zweier Veränderlichen sowie den zugehörigen Integralen. Weitere Arbeiten beschäftigen sich auch mit der Topologie von Flächen und wieder andere mit Anwendungen der Analysis in der Elektrodynamik, Wärmelehre und Elastizitätstheorie. Er war mit der Tochter von CHARLES HERMITE verheiratet.

ERNST LEONARD LINDELÖF (1870–1946) wurde in Helsinki geboren, als Finnland noch eine russische Provinz war. Sein Vater war Mathematikprofessor an der damals noch schwedischen (heute finnischen) Universität Helsingfors, an der später auch Lindelöf selbst sein Studium begann, unterbrochen durch Aufenthalte in Stockholm (1891), Paris (1893-1894) und Göttingen (1901). Der gerade bewiesene Satz stammt aus einer Arbeit von 1890; weitere Arbeiten beschäftigen sich mit analytischen Funktionen und Singularitäten. Nachdem er Professor in Helsingfors wurde, widmete er sich vor allem der Lehre und publizierte mehrere Bücher.

Die Anfangswertprobleme aus dem vorigen Abschnitt erfüllen offensichtlich allesamt *keine* LIPSCHITZ-Bedingung, denn sie sind ja nicht eindeutig lösbar. Im Falle von

$$\dot{y}(t) = \frac{1}{2y(t)} \quad \text{mit} \quad y(0) = 0$$

ist

$$f(y,t) = \frac{1}{2y}\,,$$

also

$$|f(y_2,t) - f(y_1(t)| = \left|\frac{1}{2y_2} - \frac{1}{2y_1}\right| = \frac{1}{|2y_1y_2|}\,|y_2 - y_1|\,,$$

und das kann man nur dann durch eine Schranke der Form $L\,|y_2 - y_1|$ abschätzen, wenn $1/\,|2y_1y_2|$ kleiner ist als L. In der Nähe des Nullpunkts kann $1/\,|2y_1y_2|$ aber beliebig groß werden, und somit kann hier keine LIPSCHITZ-Bedingung gelten.

Ähnlich verhält es sich beim Anfangswertproblem

$$\dot{y}(t) = 2\sqrt{|y(t)|} \quad \text{mit} \quad y(0) = 0\,.$$

Hier ist

$$f(y,t) = 2\sqrt{|y(t)|}$$

zwar überall stetig, aber die Ableitung

$$\frac{\partial f}{\partial y}(y,t) = 2\frac{d}{dy}\sqrt{|y|} = \pm\frac{1}{\sqrt{|y|}}$$

mit Pluszeichen für $y > 0$ und Minuszeichen für $y < 0$ wächst bei Annäherung an den Nullpunkt betragsmäßig unbeschränkt, die Kurve wird also immer steiler, und damit zeigt der Mittelwertsatz der Differentialrechnung, daß es keine LIPSCHITZ-Konstante geben kann.

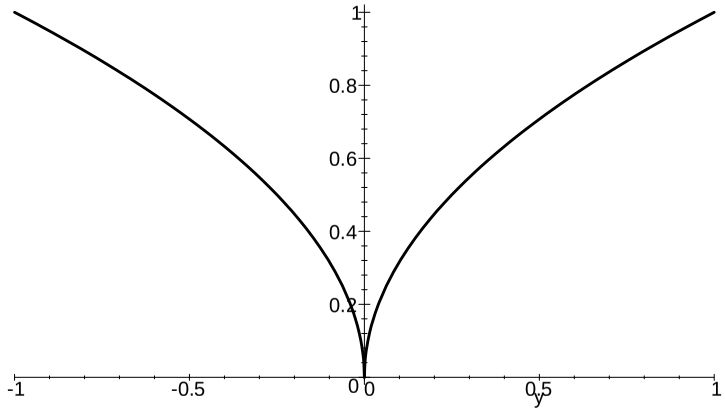


Abb. 36: Die Funktion $2\sqrt{|y|}$

Beim dritten Beispiel

$$\dot{y}(t) + \frac{y(t)}{2t^3} = 0 \quad \text{mit} \quad y(0) = 0$$

schließlich ist

$$f(y,t) = -\frac{y}{2t^3};$$

hier ist zwar die y-Abhängigkeit harmlos, aber

$$|f(y_2,t) - f(y_1,t)| = \frac{1}{|2t^3|}\,|y_2 - y_1|$$

läßt sich nicht durch durch eine Schranke der Form $L\,|y_2 - y_1|$ abschätzen, da $1/\,|2t^3|$ für t gegen null nicht beschränkt bleibt.

Obwohl in keinem der drei Fälle eine LIPSCHITZ-Bedingung erfüllt war, existierten doch immer Lösungen, und in der Tat reichen für die bloße

Existenz von Lösungen deutlich schwächere Voraussetzungen als für Existenz und Eindeutigkeit. Nach einem Satz von PEANO etwa hat jedes Anfangswertproblem

$$\dot{y}(t) = f\big(y(t),t\big) \quad \text{und} \quad y(t_0) = c_0$$

mit *stetigem* f mindestens eine Lösung. Auch PEANOs Beweis verwendet eine Iteration, nutzt dann allerdings die gleichgradige Stetigkeit einer Folge aus, um die Existenz und die Stetigkeit der Grenzfunktion zu zeigen. Da nichteindeutige Lösungen für Anwendungen meist nutzlos sind, wollen wir hier auf diesen Beweis verzichten.

GUISEPPE PEANO (1858–1932) war Sohn eines Landarbeiters und wuchs auf einem Bauernhof nahe Cuneo im Piemont auf. 1870 brachte ihn ein Bruder seiner Mutter nach Turin, wo er weiterführende Schulen und schließlich die Universität besuchte. Dort wurde er 1880 Assistent und 1890 Professor. Den gerade erwähnten Existenzsatz bewies er 1886, und 1890 zeigte er, daß das Anfangswertproblem $\dot{y}(t) = 3y^{2/3}$ mit $y(0) = 0$ mehrere Lösungen hat. Die berühmten PEANO-Axiome für die natürlichen Zahlen veröffentlichte er 1889, und zwar aus unerfindlichen Gründen in lateinischer Sprache. Später beschäftigte er sich vor allem mit Logik.

c) Eindeutigkeitsprobleme für Systeme

Natürlich können alle Probleme, die wir vom Eindimensionalen her kennen, auch im Mehrdimensionalen auftreten; auch bei Systemen von Differentialgleichungen muß man sich also um die Existenz und Eindeutigkeit von Lösungen Gedanken machen. Zum Glück geht alles fast genauso wie im eindimensionalen Fall.

Wir betrachten ein Anfangswertproblem

$$\dot{\vec{y}}(t) = f\big(\vec{y}(t),t\big) \quad \text{und} \quad \vec{y}(t_0) = \vec{y}_0$$

mit einer Funktion

$$f\colon \mathbb{R}^n \times [t_0,\, t_1] \to \mathbb{R}^n\,.$$

Ausgeschrieben ist also

$$\vec{y}(t) = \begin{pmatrix} y_1(t) \\ \vdots \\ y_n(t) \end{pmatrix} \quad \text{und} \quad f(\vec{y}, t) = \begin{pmatrix} f_1(y_1, \ldots, y_n, t) \\ \vdots \\ f_n(y_1, \ldots, y_n, t) \end{pmatrix}$$

mit Funktionen $f_i \colon \mathbb{R}^n \times [t_0,\, t_1] \to \mathbb{R}$.

Wir versehen den $\mathbb{R}^n$ mit der Maximumsnorm

$$\|\vec{v}\| = \max_{i=1}^{n} |v_i|$$

und sagen, die Funktion f oder auch das obige System erfülle eine LIPSCHITZ-Bedingung, wenn es eine Konstante $L \in \mathbb{R}$ gibt, so daß für alle $\vec{y}, \vec{z} \in \mathbb{R}^n$ und alle $t \in [t_0,\, t_1]$ gilt

$$\big\| f\big(\vec{z}, t\big) - f\big(\vec{y}, t\big) \big\| \le L \, \|\vec{z} - \vec{y}\| \ .$$

Dann gilt auch hier der

Satz von Picard und Lindelöf: Falls f stetig ist und eine LIPSCHITZ-Bedingung erfüllt, hat jedes Anfangswertproblem

$$\dot{\vec{y}}(t) = f\big(\vec{y}(t), t\big) \quad \text{mit} \quad \vec{y}(t_0) = \vec{y}_0$$

im Intervall $[t_0,\, t_1]$ eine eindeutig bestimmte Lösung.

Der *Beweis* geht fast wörtlich genauso wie im Eindimensionalen: Wir schreiben das System um als

$$\vec{y}(t) = \vec{y}_0 + \int_{t_0}^{t} f\big(\vec{y}(\tau), \tau\big) \, d\tau \ ,$$

wobei das Integral über einen Vektor von Funktionen einfach der Vektor der Integrale über die Komponenten sein soll:

$$\int_{t_0}^{t} f\big(\vec{y}(\tau), \tau\big) \, d\tau = \begin{pmatrix} \int_{t_0}^{t} f_1\big(\vec{y}(\tau), \tau\big) \, d\tau \\ \vdots \\ \int_{t_0}^{t} f_n\big(\vec{y}(\tau), \tau\big) \, d\tau \end{pmatrix} \ .$$

Sodann definieren wir eine Supremumsnorm durch

$$\|\vec{y}(t)\| \underset{\text{def}}{=} \max_{i=1}^{n} \sup_{t \in [t_0,\, t_1]} |y_i(t)| \ ,$$

beweisen auch hierfür einen BANACHschen Fixpunktsatz und folgern schließlich daraus den Satz von PICARD und LINDELÖF. ∎

d) Differentialgleichungen mit getrennten Veränderlichen

Es gibt eine ganze Reihe von Typen elementar integrierbarer nichtlinearer Differentialgleichungen; die meisten davon sind allerdings außerhalb von Lehrbüchern über Differentialgleichungen nur selten anzutreffen. Zwei Ausnahmen sind wichtig genug, um hier behandelt zu werden: Differentialgleichungen mit getrennten Veränderlichen und exakte Differentialgleichungen.

Wir betrachten zunächst Differentialgleichungen mit getrennten Veränderlichen.

Darunter versteht man Differentialgleichungen der Form

$$\dot{y}(t) = f\big(y(t), t\big) \ ,$$

bei denen sich die Funktion $f(y, t)$ als Produkt einer Funktion von y und einer Funktion von t schreiben läßt. Da die Funktion von y besser keine Nullstellen haben sollte, setzen wir ihren Kehrwert in den Nenner und schreiben die Differentialgleichung somit als

$$\dot{y}(t) = \frac{g(t)}{h\big(y(t)\big)} \ .$$

Multiplikation mit dem Nenner und Integration führen auf

$$h\big(y(t)\big)\dot{y}(t) = g(t) \quad \text{und} \quad \int h\big(y(t)\big)\dot{y}(t) \, dt = \int g(t) \, dt \ .$$

Das linke Integral ist nach der Substitutionsregel einfach $\int h(y) \, dy$, wir erhalten also mit

$$\int h(y) \, dy = \int g(t) \, dt + C$$

eine Beziehung zwischen y und t. Damit sind die Lösungsfunktionen zumindest implizit dargestellt.

Falls eine Anfangsbedingung der Form $y(t_0) = y_0$ gegeben ist, können wir die spezielle Lösung mit dieser Anfangsbedingung auch direkt darstellen durch die Gleichung

$$\int_{y_0}^{y} h(\eta) \, d\eta = \int_{t_0}^{t} g(\tau) \, d\tau \ ,$$

die für $t = t_0$ und $y = y_0$ offensichtlich erfüllt ist.

Ein erstes Beispiel kennen wir bereits: Eine lineare homogene Differentialgleichung erster Ordnung

$$\dot{y}(t) = a(t) \cdot y(t)$$

mit $y(t) \neq 0$ läßt sich auf diese Form bringen mit

$$g(t) = a(t) \quad \text{und} \quad h(y) = \frac{1}{y};$$

wir erhalten also die Beziehung

$$\int \frac{dy}{y} = \int a(t)\,dt \quad \text{oder} \quad \ln y = \int a(t)\,dt + C.$$

Diese Gleichung läßt sich durch Anwendung der Exponentialfunktion auflösen; wir erhalten das Ergebnis

$$y(t) = e^{\int a(t)\,dt + C}.$$

Als zweites Beispiel betrachten wir das Anfangswertproblem

$$\dot{y}(t) = -\frac{t}{y(t)} \quad \text{mit} \quad y(0) = 2.$$

Trennung der Veränderlichen und Integration führt auf

$$\int_2^y \eta\,d\eta = -\int_0^t d\tau \quad \text{oder} \quad \frac{y^2}{2} - \frac{2^2}{2} = -\frac{t^2}{2}.$$

Dies führt auf

$$y(t)^2 = 4 - t^2 \quad \text{oder} \quad y(t) = \pm\sqrt{4 - t^2},$$

aber tatsächlich ist die Anfangsbedingung $y(0) = 2$ nur dann erfüllt, wenn in Pluszeichen vor der Wurzel steht, d.h.

$$y(t) = \sqrt{4 - t^2}.$$

Beim Auflösen der Gleichung nach y muß also nochmals die Anfangsbedingung ins Spiel gebacht werden, was eigentlich niemanden verwundern sollte: Eine nichtlineare Gleichung hat nur selten eine eindeutig

bestimmte Lösung, wohingegen der Satz von PICARD-LINDELÖF zeigt, daß Anfangswertprobleme oft eindeutig lösbar sind. Von den mehreren Lösungen der impliziten Gleichung, die wir bei einer Differentialgleichung mit getrennten Veränderlichen bekommen, wird daher meist nur eine das Anfangswertproblem lösen.

Etwas komplizierter ist das Beispiel

$$\dot{y}(t) = \frac{1}{\cos^2 2t \cos^2 3y(t)};$$

Hier erhalten wir

$$\int \cos^2 3y\,dy = \int \frac{dt}{\cos^2 2t}.$$

Partielle Integration links und die Erinnerung an die Ableitung des Tangens für das rechte Integral führen auf die Beziehung

$$\frac{1}{6}\cos 3y \sin 3y + \frac{y}{2} = \frac{1}{2}\tan 2t + C,$$

die wir mittels der Beziehung $\sin x \cos x = \frac{1}{2}\sin 2x$ noch etwas vereinfachen können zu

$$y + \frac{1}{6}\sin 6y = \tan 2t + C'.$$

Trotzdem dürfte es ziemlich hoffnungslos sein, wenn wir versuchen, nach y aufzulösen. Die Auflösung nach t ist aber problemlos, so daß wir mit

$$t = \frac{1}{2}\arctan\left(\frac{\sin 6y}{6} + y - C'\right)$$

wenigstens die Umkehrfunktion der Lösungsfunktion explizit darstellen können.

Als letztes Beispiel schließlich wollen wir eine Differentialgleichung betrachten, die auch von unabhängigem Interesse ist, die *logistische Gleichung*

$$\dot{y}(t) = \lambda y(t)\big(K - y(t)\big) \quad \text{mit} \quad \lambda, K > 0.$$

Sie beschreibt das Wachstum einer Population $y(t)$, deren Wachstumsrate sich immer weiter verlangsamt je mehr sich $y(t)$ an die Grenze K annähert. Für $y(t) = K$ ist $\dot{y}(t) = 0$, die Bevölkerungszahl bleibt also stabil, und für $y(t) > K$ ist $\dot{y}(t)$ negativ, d.h. die Bevölkerungszahl sinkt.

Dieses Modell wurde 1846 von dem belgischen Mathematiker PIERRE FRANÇOIS VERHULST (1804–1849) vorgeschlagen und unter anderem auf die Bevölkerungsentwicklung in Belgien angewandt. Er kam damals auf einen Schätzwert $K \approx 9\,400\,000$, der sich nicht sehr von der derzeitigen Bevölkerungszahl 10 300 000 *(Stand 2004)* unterscheidet

Da auf der rechten Seite der Differentialgleichung nur eine Funktion von y steht, haben wir trivialerweise eine Differentialgleichung mit getrennten Veränderlichen, und im interessanten Bereich

$$0 < y(t) < K$$

gibt es auch keine Probleme mit Nullstellen des Nenners. Wir erhalten damit

$$\int \frac{dy}{\lambda y(t)\big(K - y(t)\big)} = \int dt = t - t_0 \,,$$

wobei der Grund für die Schreibweise der Integrationskonstante als $-t_0$ gleich klar werden wird.

Das Integral links kann durch Partialbruchzerlegung berechnet werden; da

$$\frac{1}{y} + \frac{1}{K-y} = \frac{K}{y(K-y)}$$

ist, folgt

$$\frac{1}{\lambda y(K-y)} = \frac{1}{\lambda K}\left(\frac{1}{y} + \frac{1}{K-y}\right) \,,$$

und da $1/y$ den Logarithmus als Stammfunktion hat und $1/(K-y)$ entsprechend $-\ln(K-y)$, erhalten wir

$$\int \frac{dy}{\lambda y(t)\big(K - y(t)\big)} = \frac{1}{\lambda K}\big(\ln y - \ln(K-y)\big) = \frac{1}{\lambda K} \ln \frac{y}{K-y} \,.$$

Die Umkehrfunktion der Lösungsfunktion ist somit

$$t = t_0 + \frac{1}{\lambda K} \ln \frac{y}{K-y} \,.$$

Wir können auch nach y auflösen:

$$\frac{y}{K-y} = e^{\lambda K(t-t_0)} \Longrightarrow \frac{K-y}{y} = e^{-\lambda K(t-t_0)} \Longrightarrow \frac{K}{y} = 1 + e^{-\lambda K(t-t_0)} \,,$$

und damit ist schließlich

$$y(t) = \frac{K}{1 + e^{-\lambda K(t-t_0)}} \,.$$

Speziell ist $y(t_0) = K/2$, die Integrationskonstante gibt also den Zeitpunkt an, zu dem die Population ihre halbe Maximalstärke erreicht hat.

Differenzieren der logistischen Differentialgleichung ergibt

$$\ddot{y}(t) = \lambda y(t)\big(-\dot{y}(t)\big) + \lambda \dot{y}(t)\big(K - y(t)\big) = \lambda \dot{y}(t)\big(K - 2y(t)\big) \,,$$

die Wachstumsrate $\dot{y}(t)$ steigt also, solange $y(t) < K/2$ oder $t < t_0$ ist, danach beginnt sie zu sinken. Dementsprechend ist der Graph der Lösungsfunktion $y(t)$ konkav für $t < t_0$ und konvex für $t > t_0$; bei $t = t_0$ sitzt der einzige Wendepunkt. Da diese Kurvenform an ein langestrecktes S erinnert, spricht man von einer *sigmoiden* Kurve; der Leser kann sich aussuchen, ob er sich eher an σ, ς oder Σ erinnert fühlt.

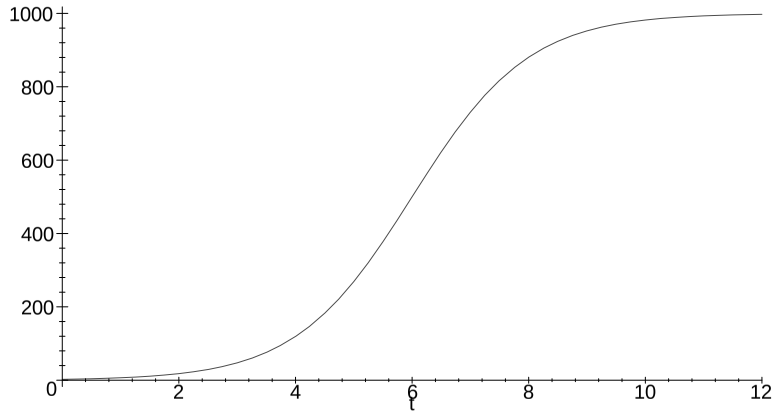


Abb. 37: Eine Lösungskurve der logistischen Differentialgleichung

Solche Kurven werden beispielsweise bei Erneuerungsprozessen oft beobachtet; ist etwa $y(t)$ = Prozentsatz aller in Frage kommender Personen oder Firmen, die zum Zeitpunkt t bereits eine technische oder

sonstige Neuerung eingeführt haben, so genügt auch $y(t)$ ungefähr einer logistischen Gleichung. Dies erscheint plausibel, denn ein potentieller Anwender erfährt von der Neuerung im Gespräch mit jemandem, der sie schon eingeführt hat; es gibt am Anfang also um so mehr Zuwachs bei $y(t)$ je größer $f(t)$ ist. Andererseits kann ein Prozentsatz nie größer als 100 werden, so daß die Steigerungsrate gegen Null gehen muß, wenn sich $y(t)$ der 100%-Grenze nähert. Vergleiche zwischen diesem Modell und vielen tatsächlichen Erneuerungsprozessen findet man etwa bei

C.T. FISCHER, R.H. PRY: A simple substitution model of technological change, *Technological Forecasting and Social Changes* **3** (1971), 75–88.

VERHULST führte die logistische Gleichung, wie bereits erwähnt, zur Modellierung des Bevölkerungswachstums ein, allerdings zeigen weder die Daten für die Bundesrepublik Deutschland noch die für die Weltbevölkerung eine gute Übereinstimmung mit diesem Modell. Seit 1940 immer wieder als Beispiel zitiert wird aber die Bevölkerung der Vereinigten Staaten von Amerika, das wir uns deshalb etwas genauer anschauen wollen:

Die Vereinigten Staaten führen seit 1790 in jedem zehnten Jahr einen „Census“ durch, in dessen Rahmen insbesondere auch die Gesamteinwohnerzahl festgestellt wird; sie bieten daher ein ideales Beispiel, für einen über einen langen Zeitraum hinweg dokumentierten Wachtumsprozess. Zwar hat sich das Territorium der USA seit 1790 gewaltig vergrößert, aber da Staaten wie das 1867 für $7 200 000 dazugekaufte Alaska kaum Einwohner haben, ist der Effekt dieser Änderungen auf die Bevölkerungszahlen fast vernachlässigbar – besonders wenn man bedenkt, daß Volkszählungsdaten notorisch unzuverlässig sind. Wir betrachten daher die Daten, wie sie bei den einzelnen Volkszählung für das jeweils aktuelle Territorium ermittelt wurden.

Das U.S. Department of Commerce, Bureau of the Census, veröffentlicht diese Daten ohne jegliche Rundung; in der folgenden Tabelle sind sie zur besseren Übersicht auf volle Hunderttausender gerundet und in Einheiten von einer Million Einwohner angegeben.

Die graphische Darstellung in Abbildung 38 zeigt die Datenpunkte zusammen mit einer logistischen Kurve; wie man sieht, ist die Übereinstimmung recht gut, aber bei weitem nicht perfekt.

Bevölkerungsentwicklung in den USA 1790–2000

Jahr:	1790	1800	1810	1820	1830	1840
Bevölkerung:	3,9	5,3	7,2	9,6	12,9	17,1
Jahr:	1850	1860	1870	1880	1890	1900
Bevölkerung:	23,2	31,4	38,6	50,2	63,0	76,2
Jahr:	1910	1920	1930	1940	1950	1960
Bevölkerung:	92,2	106,0	123,2	132,2	151,3	179,3
Jahr:	1970	1980	1990	2000		
Bevölkerung:	203,3	226,5	248,7	281,4		

Quelle: http://www.census.gov/main/www/cen2000.html

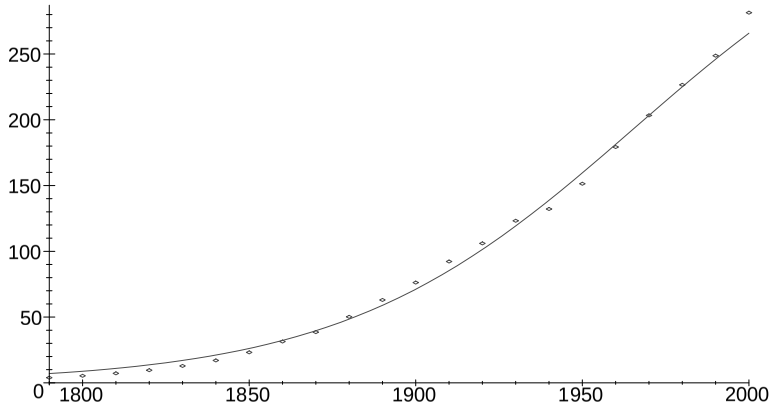


Abb. 38: U.S. Population Census 1790–2000 mit logistischer Kurve

Eine bessere Modellierung mit *zwei* logistischen Kurven zeigt Abbildung 39: Die erste Kurve, die den Datenpunkten bis 1940 angepaßt ist, hat eine Grenzkapazität K von 190 Millionen Einwohner, die zweite, für die Daten ab 1950, hat K = 356 Millionen.

Historisch ist diese Beinahe-Verdoppelung leicht zu erklären: Schließlich gingen die USA aus dem zweiten Weltkrieg als Weltmacht hervor, und eine solche hat die Macht, auch teilweise auf Kosten anderer Staaten

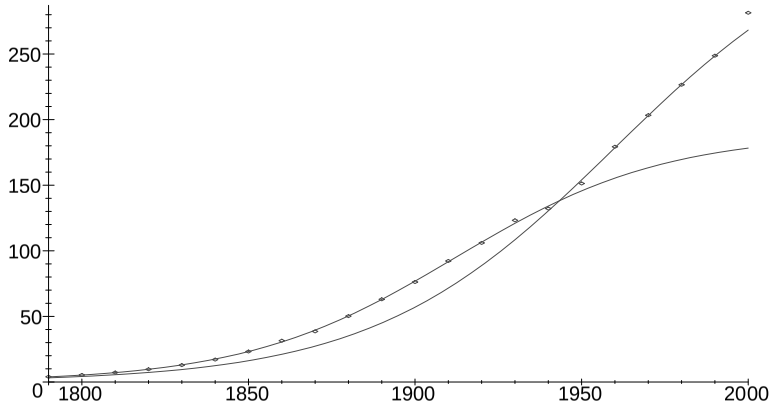


Abb. 39: U.S. Population Census 1790–2000 mit zwei logistischen Kurven

zu leben. Von daher erscheint es durchaus nachvollziehbar, daß sich die Grenzkapazität des Systems USA nach dem zweiten Weltkrieg deutlich vergrößert hat.

Erklärungsbedürftig ist auch der deutlich über der Kurve liegende Wert für 2000. Anhand der vorhandenen Daten läßt sich allerdings nicht beurteilen, ob es sich hier um den Beginn eines neuen Trends handelt, oder ob die ermittelte Bevölkerungszahl für das Jahr 2000 womöglich einfach dadurch zu erklären ist, daß die Vereinigten Staaten im Jahr 2000 Probleme mit dem Zählen hatten. Wenn in fünfzig Jahren fünf weitere Werte vorliegen, wird man mehr sagen können; die Tatsache, daß der dreihundert Millionste Amerikaner nach Schätzungen der Statistiker am 17. Oktober 2006 geboren wurde, ist immerhin ein erstes Indiz für einen neuen Trend.

e) Exakte Differentialgleichungen und integrierende Faktoren

Im letzten Abschnitt ließ sich die Lösung einer Differentialgleichung gelegentlich nur in der impliziten Form $\Phi(y,t,C) = 0$ angeben, wobei C die Integrationskonstante war.

Falls die Lösungsfunktionen, wie dies im allgemeinen der Fall sein wird, wirklich von C abhängen, kann die partielle Ableitung von Φ nach C

nicht überall verschwinden; der Satz über implizite Funktionen (**[HM I]**, *Kap. 2, §1d*) sagt uns, daß die Gleichung $\Phi(y,t,C) = 0$ überall dort nach C aufgelöst werden kann, wo diese partielle Ableitung von Null verschieden ist. Dort lassen sich die Lösungen mit einer geeigneten Funktion F also auch in der Form

$$F(y,t) = C$$

schreiben.

Umgekehrt können wir jede Kurvenschar, die durch eine Gleichung dieser Art implizit gegeben ist, als Lösungsmenge einer linearen Differentialgleichung interpretieren: Setzen wir für y eine Funktion $y(t)$ ein, so ist nach der Kettenregel

$$\frac{d}{dt}F\big(y(t),t\big) = \frac{\partial F}{\partial y}\big(y(t),t\big)\cdot\dot{y}(t) + \frac{\partial F}{\partial y}\big(y(t),t\big)\,.$$

Da $F\big(y(t),t\big) = C$ konstant ist, muß dies verschwinden; die gesuchte Differentialgleichung ist also

$$\frac{\partial F}{\partial y}\big(y(t),t\big)\cdot\dot{y}(t) + \frac{\partial F}{\partial y}\big(y(t),t\big) = 0\,.$$

Andererseits verschwindet die Ableitung einer Funktion genau dann, wenn die Funktion konstant ist; die Lösungen dieser Differentialgleichung sind also genau die Funktionen, die (in impliziter Form) gegeben sind durch

$$F\big(y(t),t\big) = C \quad \text{mit} \quad C \in \mathbb{R}\,.$$

Als Beispiel betrachten wir die Schar von Kreisen

$$y^2 + t^2 = r^2 \quad \text{mit} \quad r \in \mathbb{R}\,.$$

Hier ist $F(y,t) = y^2 + t^2$, also erhalten wir die Differentialgleichung

$$2y(t)\cdot\dot{y}(t) + 2t = 0\,,$$

was sich auch als

$$y(t)\cdot\dot{y}(t) = -t$$

schreiben läßt. In dieser Form haben wir eine Differentialgleichung mit getrennten Veränderlichen; Integration links und rechts liefert

$$\int y\,dy = -\int t\,dt \quad \text{oder} \quad \frac{y^2}{2} = -\frac{t^2}{2} + C\,.$$

Da die Summe zweier Quadrate nicht negativ sein kann, hat diese Gleichung nur für $C \geq 0$ eine Lösung (wobei der Fall $C = 0$ in unserem Zusammenhang nichts Brauchbares liefert), also können wir $C = r^2/2$ setzen und erhalten dann genau die obigen Kreisgleichungen.

Hier ist die implizite Lösung sogar nützlicher als die explizite: Offenbar gibt es durch jeden Punkt (t_0, y_0) genau eine Lösungskurve. Auflösen nach y dagegen liefert die *beiden* Lösungen

$$y(t) = \pm\sqrt{r^2 - t^2}\,,$$

wobei man jeweils anhand der Anfangsbedingung das richtige Vorzeichen finden muß. Für eine Anfangsbedingung $y(t_0) = c_0$ mit $c_0 \neq 0$ ist das einfach das Vorzeichen von c_0; für $c_0 = 0$ aber lösen beide Funktionen

$$y_1(t) = \sqrt{r^2 - t_0^2} \quad \text{und} \quad y_2(t) = -\sqrt{r^2 - t_0^2}$$

das Anfangswertproblem. Dies entspricht natürlich genau dem, was wir nach dem Satz von PICARD-LINDELÖF erwarten, denn die rechte Seite der Differentialgleichung

$$\dot{y}(t) = \frac{-t}{y(t)}$$

erfüllt überall dort eine LIPSCHITZ-Bedingung, wo y nicht verschwindet.

In diesem Abschnitt wollen wir uns allgemein mit Differentialgleichungen beschäftigen, deren Lösungen implizit in der Form

$$F(y,t) = C \quad \text{mit} \quad C \in \mathbb{R}$$

dargestellt werden können.

Wie wir oben gesehen haben, ist

$$\frac{\partial F}{\partial y}\big(y(t),t\big) \cdot \dot{y}(t) + \frac{\partial F}{\partial y}\big(y(t),t\big) = 0$$

eine Differentialgleichung mit diesen Lösungen; eine solche Differentialgleichung bezeichnen wir als *exakt:*

Definition: Die Differentialgleichung

$$a(y,t)\dot{y}(t) + b(y,t) = 0$$

heißt *exakt,* wenn es eine differenzierbare Funktion $F(y,t)$ gibt mit

$$\frac{\partial F}{\partial y}(y,t) = a(y,t) \quad \text{und} \quad \frac{\partial F}{\partial t}(y,t) = b(y,t)\,.$$

Zu diesen exakten Differentialgleichungen zählen insbesondere auch die Differentialgleichungen mit getrennten Veränderlichen aus dem letzten Abschnitt. Diese hatten wir in der Form

$$h\big(y(t)\big)\dot{y}(t) = g(t) \quad \text{oder} \quad h\big(y(t)\big)\dot{y}(t) - g(t) = 0$$

geschrieben; mit

$$F(y,t) = \int h(y)\,dy - \int g(t)\,dt$$

ist hier gerade

$$\frac{\partial F}{\partial y}(y,t) = h(y) \quad \text{und} \quad \frac{\partial F}{\partial t}(y,t) = -g(t)\,.$$

Andererseits ist aber nicht jede Differentialgleichungen der Form

$$a(y,t)\dot{y}(t) + b(y,t) = 0 \qquad (\star)$$

exakt: Das ist sie nur, wenn es eine Funktion $F(y,t)$ gibt mit

$$\frac{\partial F}{\partial y}(y,t) = a(y,t) \quad \text{und} \quad \frac{\partial F}{\partial t}(y,t) = b(y,t)$$

d.h.

$$\nabla F(y,t) = \operatorname{grad} F(y,t) = \begin{pmatrix} a(y,t) \\ b(y,t) \end{pmatrix}.$$

Die Differentialgleichung $(\star)$ ist also genau dann exakt, wenn das Vektorfeld $\left(\begin{smallmatrix} a(y,t) \\ b(y,t) \end{smallmatrix}\right)$ eine Stammfunktion hat.

Dies kann für stetig differenzierbare Funktionen a und b nach dem Lemma von SCHWARZ (**[HM I]**, Kapitel 2, §2) nur dann der Fall sein, wenn

$$\frac{\partial a}{\partial t}(y,t) = \frac{\partial b}{\partial y}(y,t)$$

ist. Umgekehrt reicht diese Bedingung nach **[HM I]**, Kapitel 2, §6*f)* aus für die Existenz einer Stammfunktion, *falls* das Vektorfeld auf einem einfach zusammenhängenden Gebiet definiert ist, falls es also keine Löcher im Definitionsbereich gibt, wie sie etwa durch Nullstellen von Nennern verursacht sein können. Insbesondere reicht die Bedingung also aus, wenn beide Funktionen auf ganz $\mathbb{R}^2$ definiert sind oder in einem Rechteck oder einer Kreisscheibe.

Als Beispiel betrachten wir die Differentialgleichung

$$te^{-ty(t)}\dot{y}(t) + 6t^2 + ye^{-ty(t)} = 0\,.$$

Diese Gleichung ist exakt, denn beide Koeffizientenfunktionen sind auf ganz $\mathbb{R}^2$ definiert und

$$\frac{\partial}{\partial t}te^{-ty(t)} = \frac{\partial}{\partial y}(6t^2 + ye^{-ty(t)}) = e^{-ty(t)} - tye^{-ty(t)}\,.$$

Für die Stammfunktion F des Vektorfelds ist einerseits

$$\frac{\partial F}{\partial y}(y,t) = te^{-ty}, \quad \text{also} \quad F(y,t) = -e^{-ty} + h_1(t)$$

mit irgendeiner nur von t abhängigen Funktion h_1, und andererseits

$$\frac{\partial F}{\partial t}(y,t) = 6t^2 + ye^{-ty(t)}, \quad \text{also} \quad F(y,t) = 2t^3 - e^{-ty} + h_2(y)$$

mit irgendeiner nur von y abhängigen Funktion h_2. Abgesehen von einer additiven Konstanten, die wir nach Belieben addieren können, passen diese beiden Gleichungen für $F(y,t)$ genau dann zusammen, wenn

$$F(y,t) = 2t^3 - e^{-ty}$$

ist; für die Lösungsfunktionen der Differentialgleichung ist also

$$F\big(y(t),t\big) = 2t^3 - e^{-ty(t)} = C \quad \text{oder} \quad e^{-ty(t)} = 2t^3 - C$$

mit einer beliebigen Konstanten $C \in \mathbb{R}$.

In diesem Fall läßt sich die implizite Gleichung unschwer nach $y(t)$ auflösen; wir erhalten

$$y(t) = -\frac{\ln(2t^3 - C)}{t} \quad \text{mit} \quad C \in \mathbb{R}\,.$$

Genau wie nur wenige Vektorfelder Stammfunktionen haben, sind auch nur wenige Differentialgleichungen exakt. Beispiel einer nichtexakten Differentialgleichung der Form ($\star$) ist etwa

$$t^2y(t)\dot{y}(t) + t^3 = 0 \quad \text{mit} \quad a(y,t) = t^2y \ \text{ und } \ b(y,t) = t^3\,,$$

denn hier ist

$$\frac{\partial a}{\partial t}(y,t) = 2ty, \quad \text{aber} \quad \frac{\partial b}{\partial y}(y,t) = 0\,.$$

Trotzdem können wir diese Differentialgleichung lösen: Wenn wir durch t^2 dividieren, erhalten wir die obige Differentialgleichung für konzentrische Kreise um den Nullpunkt, und da eine differenzierbare Funktion durch ihre Werte für $t \neq 0$ auch im Nullpunkt eindeutig festgelegt ist, verlieren wir durch die Division keine Lösungen.

Auch die Differentialgleichung

$$t^2\dot{y}(t) + ty + 1 = 0$$

ist nicht exakt, da

$$\frac{\partial t^2}{\partial t} = 2t \quad \text{und} \quad \frac{\partial(ty+1)}{\partial y} = t$$

voneiander verschieden sind. Multipliziert man die Gleichung aber mit e^{ty}, so wird sie exakt, denn

$$\frac{\partial(t^2e^{ty})}{\partial t} = \frac{\partial(ty+1)e^{ty}}{\partial y} = (t^2y + 2t)e^{ty}\,,$$

und in der Tat sind t^2e^{ty} und $(ty+1)e^{ty}$ die partiellen Ableitungen von $F(y,t) = te^{ty}$ nach y und nach t.

An den Lösungen der Gleichung ändert sich durch die Multiplikation mit der nirgends verschwindenden Funktion e^{ty} nichts, sie sind also also gegeben durch die implizite Gleichung

$$te^{ty(t)} = C$$

oder explizit durch

$$y(t) = \frac{1}{t} \ln \frac{C}{t},$$

wie man sich leicht durch Einsetzen überzeugen kann. Diese Lösung existiert bei positivem C nur für $t > 0$, bei negativem nur für $t < 0$. Für $C = 0$ definiert $te^{ty(t)} = 0$ keine Funktion $y(t)$.

Wenn eine Differentialgleichung, wie in diesen beiden Beispielen, durch Multiplikation mit einer Funktion $\varphi(y,t)$ exakt gemacht werden kann, nennt man $\varphi(y,t)$ einen *integrierenden Faktor.* Ihn zu finden, kann im allgemeinen sehr schwierig sein: Einen einfach zusammenhängenden Definitionsbereich vorausgesetzt, ist die Gleichung

$$\varphi\big(y(t),t\big)a\big(y(t),t\big)\dot{y}(t) + \varphi\big(y(t),t\big)b\big(y(t),t\big) = 0$$

genau dann exakt, wenn die partiellen Ableitungen des ersten Koeffizienten nach t und des zweiten Koeffizienten nach y miteinander übereinstimmen, wenn also

$$\frac{\partial\varphi}{\partial t}(y,t)\cdot a(y,t) + \varphi(y,t)\frac{\partial a(y,t)}{\partial t} = \frac{\partial\varphi}{\partial y}(y,t)\cdot b(y,t) + \varphi(y,t)\frac{\partial b(y,t)}{\partial y}$$

ist. Diese Gleichung für gegebene Funktionen a und b zu lösen ist leider fast aussichtslos; lediglich in speziellen Fällen kann man eine eine Lösung wirklich hinschreiben. Dazu gehört insbesondere der Fall, daß φ nur von einer der beiden Variablen abhängt:

Sucht man ein φ, das nur von t abhängt, vereinfacht sich die obige Gleichung zu

$$\dot{\varphi}(t)\cdot a(y,t) = \varphi(t)\left(\frac{\partial b(y,t)}{\partial y} - \frac{\partial a(y,t)}{\partial t}\right)$$

oder

$$\frac{\dot{\varphi}(t)}{\varphi(t)} = \frac{\frac{\partial b(y,t)}{\partial y} - \frac{\partial a(y,t)}{\partial t}}{a(y,t)}.$$

Die linke Seite dieser Gleichung ist die Ableitung von $\ln\varphi(t)$; falls die rechte Seite unabhängig von y ist, also nur eine Funktion $\psi(t)$, so ist also

$$\varphi(t) = e^{\int \psi(t)\,dt}.$$

Entsprechend kann man argumentieren, wenn φ nur von y abhängen soll; bezeichnen wir die Ableitung nach y durch einen Strich, wird dann die Gleichung zu

$$\varphi'(y)\cdot b(y,t) = \varphi(y)\left(\frac{\partial a(y,t)}{\partial t} - \frac{\partial b(y,t)}{\partial y}\right)$$

oder

$$\frac{\varphi'(y)}{\varphi(y)} = \frac{\frac{\partial a(y,t)}{\partial t} - \frac{\partial b(y,t)}{\partial y}}{b(y,t)}.$$

Falls die rechte Seite nicht von t abhängt, also eine Funktion $\omega(y)$ ist, folgt wie oben

$$\varphi(y) = e^{\int \omega(y)\,dy}.$$

Ist also

$$\frac{\frac{\partial b(y,t)}{\partial y} - \frac{\partial a(y,t)}{\partial t}}{a(y,t)} \quad \text{unabhängig von } y$$

oder

$$\frac{\frac{\partial a(y,t)}{\partial t} - \frac{\partial b(y,t)}{\partial y}}{b(y,t)} \quad \text{unabhängig von } t,$$

läßt sich ein integrierender Faktor auch algorithmisch produzieren; ansonsten kann sich die Suche als schwierig erweisen.

Für die oben betrachtete Differentialgleichung

$$t^2\dot{y}(t) + ty + 1 = 0$$

beispielsweise ist $a(y,t) = t^2$ und $b(y,t) = ty + 1$, also

$$\frac{\frac{\partial a(y,t)}{\partial t} - \frac{\partial b(y,t)}{\partial y}}{b(y,t)} = \frac{t}{ty+1}$$

auch von t abhängig, so daß es keinen integrierenden Faktor gibt, der nur von y abhängt. Da aber

$$\frac{\frac{\partial b(y,t)}{\partial y} - \frac{\partial a(y,t)}{\partial t}}{a(y,t)} = \frac{-t}{t^2} = -\frac{1}{t}$$

nur von t abhängt, gibt es einen integrierenden Faktor, der nur von t abhängt, nämlich

$$\varphi(t) = e^{-\int \frac{dt}{t}} = e^{-\ln t} = \frac{1}{t}\,.$$

In der Tat ist

$$t\dot{y}(t) + y + \frac{1}{t} = 0$$

eine exakte Differentialgleichung, und

$$F(y,t) = yt + \ln t$$

ist eine Stammfunktion, die (abgesehen von der etwas anderen Form der Integrationskonstanten) auf dieselbe Lösung, die wir oben mit einem etwas anderen F hergeleitet haben F.

Integrierende Faktoren, die sowohl von y als auch von t abhängen, können gelegentlich über Symmetriebetrachtungen gefunden werden: Die Differentialgleichung

$$a(y,t)\dot{y}(t) + b(y,t) = 0$$

habe die (unbekannte) Lösung $F(y,t) = C$ mit $C \in \mathbb{R}$. Angenommen, wie kennen eine Transformation der Koordinaten, die in Abhängigkeit von einem Paramter ε eine Lösungskurve $F(y,t) = C$ überführt in eine neue Lösungskurve $F(y_\varepsilon, t_\varepsilon) = C_\varepsilon$, wobei $\varepsilon = 0$ der Ausgangslösung entsprechen soll. Nach dem Satz über implizite Funktionen können wir dann auch eine Transformation finden, für die $C_\varepsilon = C + \varepsilon$ ist. Es könnte schwierig sein, so eine Transformation explizit anzugeben; es reicht uns allerdings, wenn wir sie nur in erster Näherung kennen, d.h. bis auf Terme, die schneller gegen null gehen als ε:

$$y_\varepsilon = y + \varepsilon\eta(y,t) + o(\varepsilon) \quad \text{und} \quad t_\varepsilon = t + \varepsilon\tau(y,t) + o(\varepsilon)\,.$$

Dann ist

$$C + \varepsilon = F(y_\varepsilon, t_\varepsilon) = F(y,t) + \varepsilon\big(F_y(y,t)\eta(y,t) + F_t(y,t)\tau(y,t)\big) + o(\varepsilon)\,,$$

also

$$F_y(y,t)\eta(y,t) + F_t(y,t)\tau(y,t) = 1\,. \qquad (*)$$

Außerdem ist nach dem Satz über implizite Funktionen

$$\dot{y}(t) = -\frac{F_t(y,t)}{F_y(y,t)}\,;$$

da $y(t)$ Lösung der Differentialgleichung ist, folgt

$$-a(y,t)\frac{F_t(y,t)}{F_y(y,t)} + b(y,t) = 0$$

oder

$$-b(y,t)F_y(y,t) + a(y,t)F_t(y,t) = 0\,.$$

Diese Gleichung, zusammen mit Gleichung $(*)$ ist ein lineares Gleichungssystem für $F_y(y,t)$ und $F_t(y,t)$; nach der CRAMERschen Regel hat es die Lösung

$$F_y(y,t) = \frac{\begin{vmatrix} 1 & \tau(y,t) \\ 0 & a(y,t) \end{vmatrix}}{\begin{vmatrix} \eta(y,t) & \tau(y,t) \\ -b(y,t) & a(y,t) \end{vmatrix}} = \frac{a(y,t)}{\eta(y,t)a(y,t) + \tau(y,t)b(y,t)}$$

und

$$F_t(y,t) = \frac{\begin{vmatrix} \eta(y,t) & 1 \\ -b(y,t) & 0 \end{vmatrix}}{\begin{vmatrix} \eta(y,t) & \tau(y,t) \\ -b(y,t) & a(y,t) \end{vmatrix}} = \frac{b(y,t)}{\eta(y,t)a(y,t) + \tau(y,t)b(y,t)}\,.$$

Insbesondere ist

$$\frac{F_y(y,t)}{a(y,t)} = \frac{F_t(y,t)}{b(y,t)} = \frac{1}{\eta(y,t)a(y,t) + \tau(y,t)b(y,t)}$$

ein integrierenden Faktor.

f) Qualitative Theorie

Ein allgemeines System

$$\dot{\vec{y}}(t) = f\left(\vec{y}(t), t\right)$$

wird nur selten eine Lösung haben, die sich in geschlossener Form angeben läßt; meist wird man sich mit (nicht immer unproblematischen)

numerischen Näherungslösungen begnügen müssen. In diesem letzten Abschnitt des Paragraphen über Differentialgleichungen wollen wir uns, hauptsächlich anhand von Beispielen, überlegen, wie man auch bei solchen nicht explizit lösbaren Gleichungen zu Aussagen über das Verhalten der Lösungsfunktionen kommen kann.

Eine wesentliche Rolle spielen dabei die Gleichgewichtslösungen:

Definition: Ein Punkt $\vec{y}_0 \in \mathbb{R}^n$ heißt *Fixpunkt* oder *Gleichgewichtslösung* des Differentialgleichungssystems $\dot{\vec{y}}(t) = f\big(\vec{y}(t), t\big)$, wenn die konstante Funktion $\vec{y}(t) = \vec{y}_0$ eine Lösung ist.

Anschaulich bedeutet dies, daß der Zustand eines Systems, das durch diese Differentialgleichung beschrieben wird, für $\vec{y}(t) = \vec{y}_0$ zeitlich konstant ist, das System befindet sich also im Gleichgewicht.

Da die Ableitung einer konstanten Funktion verschwindet, sind die Fixpunkte des Differentialgleichungssystems $\dot{\vec{y}}(t) = f\big(\vec{y}(t), t\big)$ gerade die Lösungen des Gleichungssystems

$$f(\vec{y}_0, t) = 0 \quad \text{für alle } t \in [t_0, t_1]\,.$$

Bei einem nichtlinearen Differentialgleichungssystem ist das ein nichtlineares Gleichungssystem, man wird sich daher oft mit Näherungslösungen begnügen müssen. (Die Variable t tritt natürlich nur bei nichtautonomen Systemen auf; bei den in Naturwissenschaft und Technik häufigen autonomen Systemen haben wir ein Gleichungssystem, in dem nur die Komponenten von $\vec{y}_0$ vorkommen.)

Ein klassisches Beispiel, bei dem sich die Fixpunkte leicht ausrechnen lassen, ist das Raubtier-Beutetier-Modell, das 1925 von LOTKA und VOLTERRA vorgeschlagen wurde: In einem Gebiet gebe es eine Population von Raubtieren, die sich von genau einer Art von Beutetieren ernähren. Für die Beutetiere sei genügend Nahrung vorhanden, so daß diese sich, falls es keine Raubtiere gäbe, beliebig vermehren könnten. Wenn wir die Populationsstärke zum Zeitpunkt t mit $x(t)$ bezeichnen, können wir also annehmen, daß die Beutetiere bei Abwesenheit der Raubtiere eine konstanter Wachstumsrate hätten, d.h.

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

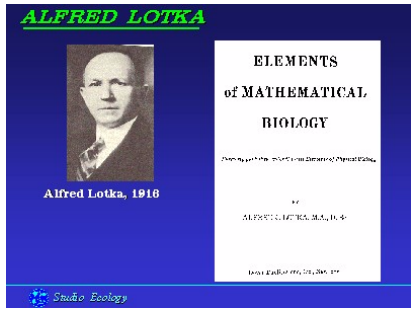
Die Raubtiere, deren Bestand zum Zeitpunkt t wir mit $y(t)$ bezeichnen wollen, würden in Abwesenheit der Beutetiere relativ schnell verhungern und somit aussterben, was wir durch eine negative Wachstumsrate modellieren können:

$$\dot{y}(t) = -\gamma y(t) \quad \text{mit} \quad \gamma > 0\,.$$

Nun sind aber die Raubtiere und die Beutetiere nicht isoliert voneinander, sondern es kommt zu Begegnungen zwischen den beiden Populationen. Deren Häufigkeit ist etwa proportional zum Produkt der beiden Populationsstärken, und die Auswirkung einer solchen Begegnung ist positiv für die Wachstumsrate der Raubtiere, aber negativ für die der Beutetiere. Unser Modell läßt sich somit beschreiben durch das System

$$\begin{aligned}\dot{x}(t) &= \alpha x(t) - \beta x(t) y(t)\\ \dot{y}(t) &= -\gamma y(t) + \delta x(t) y(t)\end{aligned}$$

mit positiven reellen Zahlen $\alpha, \beta, \gamma, \delta$.



Der amerikanische Wissenschaftler ALFRED LOTKA (1880–1949) war von der Ausbildung her ein Mathematiker, interessierte sich aber Zeit seines Lebens stark für Physik, insbesondere Thermodynamik, und war einer der ersten, der die Evolution unter physikalischen Gesichtspunkten betrachtete. Er zählt zu den Pionieren der Selbstorganisation, der Bioenergetik und (auch für eine Versicherung) der Demographie.

VITO VOLTERRA (1860–1940) wurde in Ancona im damaligen Kirchenstaat geboren. Er studierte bereits als Elfjähriger mathematische Literatur, promovierte dann aber in Physik über ein Thema aus der Hydrodynamik. Er hatte Lehrstühle für Mechanik und für mathematische Physik in Pisa, Turin und Rom. Seine wichtigsten Arbeiten beschäftigen sich mit partiellen Differentialgleichungen und vor allem Integralgleichungen. Ab 1922 kämpfte er im italienischen Parlament gegen den Faschismus und verlor deshalb 1931 nach Auflösung des Parlaments seinen Lehrstuhl in Rom. Den Rest seines Lebens verbrachte er größtenteils im Exil.

Zur Bestimmung der Gleichgewichtslösungen müssen wir für $x(t)$ und $y(t)$ Konstanten einsetzen; dies führt auf die Gleichungen

$$0 = \quad \alpha x_0 - \beta x_0 y_0 = x_0(\alpha - \beta y_0)$$
$$0 = -\gamma\, y_0 + \delta x_0 y_0 = y_0(\gamma - \delta x_0)\,.$$

Es gibt somit genau zwei Gleichgewichtslösungen: Einmal die uninteressante Lösung $x(t) = y(t) \equiv 0$, die im wesentlichen besagt, daß ohne Raub- und Beutetiere in diesem System nichts passiert, und dann noch die Lösung

$$x(t) \equiv \frac{\gamma}{\delta} \quad \text{und} \quad y(t) \equiv \frac{\alpha}{\beta}\,.$$

Falls die beiden Populationen diese Stärken haben, fressen also die Raubtiere genau so viele Beutetiere weg, wie nachwachsen; umgekehrt reichen die Beutetiere gerade aus, um die Raubtierpopulation zu ernähren.

Was passiert, wenn die Populationen nicht im Gleichgewicht sind? Wir haben offensichtlich kaum Chancen, das Differentialgleichungssystem explizit zu lösen, aber wir können trotzdem versuchen, etwas über die Lösungs*kurven* in Erfahrung zu bringen.

Wenn wir y als Funktion von x betrachten, ist

$$y'(x) = \frac{dy}{dx} = \frac{\dot y(t)}{\dot x(t)} = \frac{-\gamma y + \delta x y}{\alpha x - \beta x y} = \frac{y}{\alpha - \beta y} \cdot \frac{\delta x - \gamma}{x}\,,$$

wir haben also eine Differentialgleichung mit getrennten Veränderlichen. Trennung der Variablen führt auf

$$\int \left(\frac{\alpha}{y} - \beta\right) dy = \int \left(\delta - \frac{\gamma}{x}\right) dx$$

oder

$$\alpha \ln y - \beta y = \delta x - \gamma \ln x + C\,.$$

Anwendung der Exponentialfunktion macht daraus

$$\frac{y^\alpha}{e^{\beta y}} = \frac{e^{\delta x}}{x^\gamma} \cdot e^C\,.$$

Diese Gleichung können wir zwar weder nach y noch nach x auflösen, aber eine einfache Kurvendiskussion der Funktionen

$$f(y) = \frac{y^\alpha}{e^{\beta y}} \quad \text{und} \quad g(x) = \frac{e^{\delta x}}{x^\gamma}$$

zeigt, daß die Ableitung in beiden Fällen außer im Nullpunkt noch in genau einem weiteren Punkt verschwindet, nämlich dort wo x *bzw.* y gleich der entsprechenden Koordinate des nichttrivialen Gleichgewichtspunkts ist. f hat in diesem Punkt ein Maximum, g ein Minimum, und für $t \to 0$ oder $t \to \infty$ geht f gegen null und g gegen unendlich.

Da beide Funktionen im positiven Bereich der reellen Achse nur positive Werte annehmen, gibt es für eine vorgegebene positive Zahl c somit höchstens zwei Werte, an denen sie von f *bzw.* g angenommen wird; für zu große c, gibt es kein y mehr mit $f(y) = c$, und für zu kleine c kein x mit $g(x) = c$.

Daraus folgt nach kurzer Überlegung, daß die Lösungskurven (abgesehen von den beiden Fixpunkten) auf geschlossenen Kurven um den nichttrivialen Fixpunkt liegen. Da kein Punkt auf einer solchen Kurve Fixpunkt ist, kann keine Lösungskurve für $t \to \infty$ gegen einen Punkt einer solchen Kurve konvergieren, die Lösungskurven müssen also den nichttrivialen Gleichgewichtspunkt permanent umrunden.

Betrachten wir eine konkrete Lösung $\big(x(t), y(t)\big)$ des Differentialgleichungssystems, das wir der Einfachheit halber kurz als

$$\begin{pmatrix}\dot x(t)\\ \dot y(t)\end{pmatrix} = F\big(x(t), y(t)\big)$$

schreiben wollen, und fixieren wir einen Zeitpunkt t_0; für diesen sei $x(t_0) = a$ und $y(t_0) = b$. Dann muß es nach obiger Diskussion ein kleinste Zeitspanne T geben, so daß auch

$$x(t_0 + T) = a \quad \text{und} \quad y(t_0 + T) = b$$

ist. Für die beiden Funktionen

$$u(t) \underset{\text{def}}{=} x(t + T) \quad \text{und} \quad v(t) \underset{\text{def}}{=} y(t + T)$$

ist dann $u(t_0) = a$ und $v(t_0) = b$; außerdem ist

$$\begin{pmatrix}\dot u(t)\\ \dot v(t)\end{pmatrix} = \begin{pmatrix}\dot x(t+T)\\ \dot y(t+T)\end{pmatrix} = F\big(x(t+T), y(t+T)\big) = F\big(u(t), v(t)\big)\,,$$

$\big(x(t), y(t)\big)$ und $\big(u(t), v(t)\big)$ lösen also dasselbe Anfangswertproblem. Falls wir zeigen können, daß F eine LIPSCHITZ-Bedingung erfüllt, müssen die beiden Funktionen also übereinstimmen.

Im betrachteten Beispiel ist

$$F(x,y) = \begin{pmatrix} \alpha x - \beta xy \\ -\gamma y + \delta xy \end{pmatrix},$$

also

$$\|F(x_1,y_1) - F(x_2,y_2)\| = \left\| \begin{pmatrix} \alpha(x_1-x_2) - \beta(x_1y_1 - x_2y_2) \\ -\gamma(y_1-y_2) + \delta(x_1y_1 - x_2y_2) \end{pmatrix} \right\| .$$

Im Quadrat $-R \le x,y \le R$ ist

$$|\alpha(x_1-x_2) - \beta(x_1y_1 - x_2y_2)| \le |\alpha(x_1-x_2)| + |\beta(x_1y_1 - x_2y_2)|$$

und

$$\begin{aligned} |(x_1y_1 - x_2y_2)| &= |x_1y_1 - x_1y_2 + x_1y_2 - x_2y_2| \\ &= |x_1(y_1-y_2) + y_2(x_1-x_2)| \\ &\le |x_1(y_1-y_2)| + |y_2(x_1-x_2)| \\ &\le R\big(|y_1-y_2| + |x_1-x_2|\big) , \end{aligned}$$

also ist

$$\begin{aligned} &|\alpha(x_1-x_2) - \beta(x_1y_1 - x_2y_2)| \\ &\le \alpha\,|(x_1-x_2)| + \beta R\big(|y_1-y_2| + |x_1-x_2|\big) . \end{aligned}$$

Analog folgt die Ungleichung

$$\begin{aligned} &|\gamma(y_1-y_2) + \delta(x_1y_1 - x_2y_2)| \\ &\le \gamma\,|(y_1-y_2)| + \delta R\big(|y_1-y_2| + |x_1-x_2|\big) . \end{aligned}$$

Wir arbeiten hier mit der Maximumsnorm von Vektoren, $\left\| \begin{pmatrix} x \\ y \end{pmatrix} \right\|$ ist also das Maximum von $|x|$ und $|y|$, und entsprechend ist

$$\left\| \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \right\| = \max\big\{|x_1-x_2|\,,|y_1-y_2|\big\} .$$

Mit

$$L = \max\{\alpha + 2\beta R, \gamma + 2\delta R\}$$

ist somit

$$\|F(x_1,y_1) - F(x_2,y_2)\| \le L \left\| \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \right\| ;$$

F erfüllt also eine LIPSCHITZ-Bedingung, so daß wir aus dem Satz von PICARD-LINDELÖF folgern können, daß das Anfangswertproblem in jedem abgeschlossenen Quadrat eindeutig lösbar ist.

Da jede Lösungskurve in einem abgeschlossenen Quadrat liegt (sonst müßte sie irgendwo gegen unendlich gehen), ist also

$$u(t) = x(t) \quad \text{und} \quad v(t) = y(t) \quad \text{für alle } t \ge t_0$$

d.h.

$$x(t+T) = x(t) \quad \text{und} \quad y(t+T) = y(t) \quad \text{für alle } t \ge t_0 .$$

Damit wissen wir, daß alle Lösungsfunktionen periodisch sind.

In der unmittelbaren Umgebung der nichttrivialen Gleichgewichtslösung können wir sogar noch etwas mehr sagen: Durch TAYLOR-Entwicklung der oben betrachteten Funktionen f und g überzeugt man sich leicht davon, daß die Lösungskurven dort näherungsweise als Ellipsen aufgefaßt werden können.

Nach all diesen Vorbereitungen sollten wir uns endlich eine konkrete Lösungskurve anschauen, d.h. wir sollten das Problem in einem Spezialfall numerisch lösen.

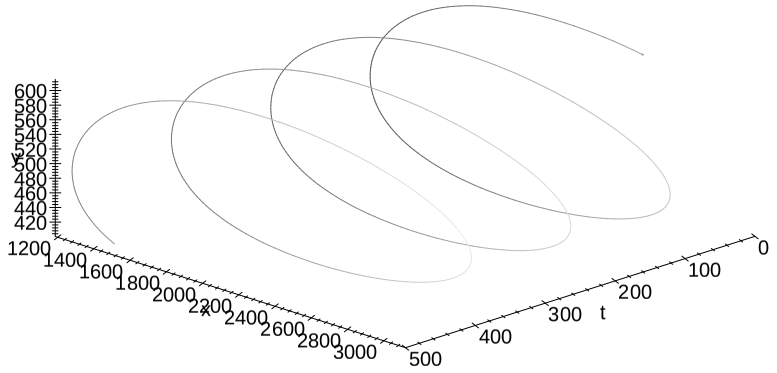


Abb. 40: Numerische Simulation des Raubtier-Beutetier-Systems

Abbildung 40 zeigt das Ergebnis; die spiralförmige Kurve entspricht genau unseren Erwartungen.

Besser können wir diese überprüfen, wenn wir eine Reihe von Lösungskurven in der xy-Ebene betrachten; Abbildung 41 zeigt solche Kurven zu verschiedenen Anfangsbedingungen. Abgesehen von der Gleichgewichtslösung, die einfach ein Punkt ist, sieht man die vorhergesagten geschlossenen Kurven, die das Gleichgewicht umrunden.

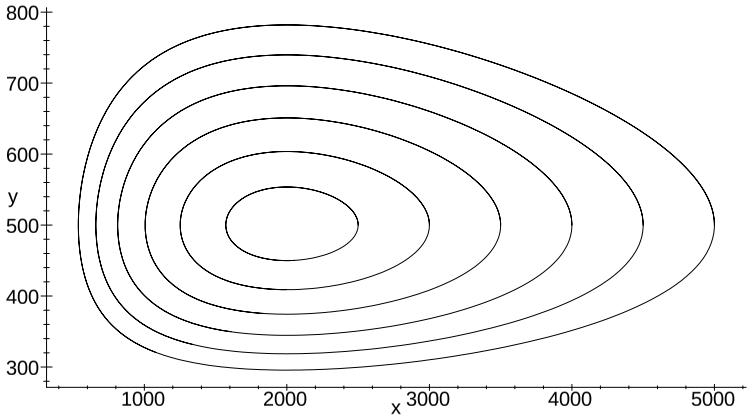


Abb. 41: Lösungskurven in der xy-Ebene

Man kann sich leicht klarmachen, was der Umlauf auf so einer Lösungskurve biologisch bedeutet: Im jeweils untersten Punkt ist die Raubtierpopulation minimal, so daß sich die Beutetiere stark vermehren können; dadurch verbessert sich die Nahrungsgrundlage für die Raubtiere, was nun auch zu deren Vermehrung führt, so daß die Kurve auf ihren am weitesten rechts gelegenen Punkt zusteuert, in dem die Beutetiere ihre maximale Populationsstärke erreichen. Die gestiegene Raubtierpopulation frißt nun, nachdem sie ihren Gleichgewichtswert überschritten hat, mehr Beutetiere als nachwachsen, kann sich aber wegen der großen Anzahl vorhandener Beutetiere weiterhin vermehren auf ein Maximum hin, das am obersten Punkt der Lösungskurve erreicht ist. Danach reicht der bereits gesunkene Bestand an Beutetieren nicht mehr aus als Nahrungsgrundlage für die Raubtiere, ihre Population geht also zurück, reicht

aber immer noch aus, um die Beutetiere weiter zu dezimieren. Im Punkt links außen hat deren Bestand schließlich sein Minimum erreicht; die weiterhin sinkende Raubtierpopulation frißt nun weniger Beutetiere als nachwachsen und leidet trotzdem weiter an Nahrungsmangel. Sobald sie ihr Minimum erreicht hat, schließt sich der Kreis, und der gleiche Zyklus beginnt von vorne.

Auch wenn Raubtiere und Beutetiere in der Technischen Informatik keine große Rolle spielen, sollten wir uns doch zumindest kurz fragen, ob die mathematische Lösung irgendetwas mit der biologischen Realität zu tun hat – der Zusammenhang zwischen idealisierten mathematischen Modellen und realen Systemen ist schließlich auch in der Technischen Informatik von Bedeutung.

Wie in vielen praktischen Anwendungen der Mathematik sind die Annahmen des Modells auch hier viel zu einfach: Es gibt kaum je zwei Arten, die völlig isoliert vom Rest der Welt leben. Trotzdem wurden die vorhergesagten Zykel schon beobachtet: Die Hudson Bay Company sammelte rund hundert Jahre lang Daten über gekaufte Felle von Luchsen (Raubtieren) und Schneehasen (deren Beute); da die Trapper kaum beeinflussen können, was in ihre Fallen läuft, sollten diese Anzahlen ungefähr proportional sein zu den jeweiligen Populationszahlen. Abbildung 42 zeigt ungefähr die vorhergesagten zyklischen Schwankungen – auch wenn die Schwankungen zwischen 1875 und 1905 in die falsche Richtung gehen: Dort wird der Gleichgewichtspunkt nicht gegen den Uhrzeigersinn umrundet, sondern im Uhrzeigersinn.

Über den genauen Grund dafür gibt es immer noch viele Spekulationen; der Grund, warum Abweichungen vom Modell auftreten müssen ist aber einfach zu verstehen: Die Realität deutlich ist komplizierter als das extrem vereinfachte Modell von LOTKA und VOLTERRA.

Das vorliegende Beispiel ist natürlich, wie fast alle Beispiele in einer Anfängervorlesung, viel zu elementar: Im allgemeinen kann man einer Differentialgleichung nicht auf so einfache Weise so viele Eigenschaften der Lösungen ansehen.

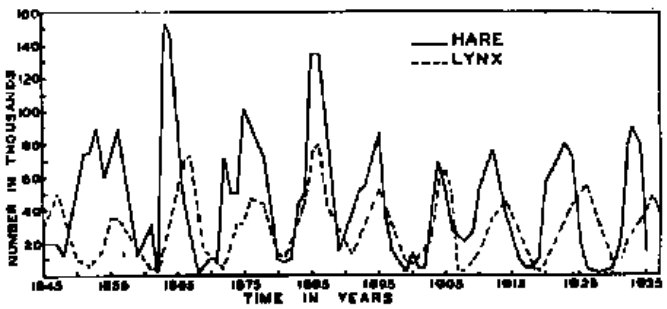


Abb. 42: Luchse und Schneehasen

Die qualitative Theorie der Differentialgleichungen wendet denn auch viele Methoden an, die weit jenseits des Stoffs dieser Vorlesung liegen, und selbst damit kann sie in komplizierteren Fällen nur deutlich weniger Information aus der Differentialgleichung extrahieren als in diesem Beispiel.

Einen ersten Überblick über das Verhalten der Lösungen einer *autonomen* Differentialgleichung in nur zwei Variablen liefert auch bei beliebig komplizierten Systemen ein graphisches Verfahren: Bei einer autonomen Differentialgleichung

$$\dot{\vec{y}}(t) = F\left(\vec{y}(t)\right)$$

definiert die Funktion F ein Vektorfeld, wie wir es aus **[HM1]**, *Kapitel 2*, kennen.

Spätestens an dieser Stelle wird klar, daß wir in diesem Kapitel bei der Unterscheidung von Punkten und Vektoren geschlampt haben: Für die linearen homogenen Differentialgleichungssysteme, die den Hauptinhalt dieses Kapitels bilden, war es völlig in Ordnung, nur von Vektoren zu reden: Dort gibt es einen wohldefinierten Nullpunkt, so daß sich Punkte und Vektoren in kanonischer Weise entsprechen.

Zur geometrischen Interpretation von $\dot{\vec{y}}(t) = F\left(\vec{y}(t)\right)$ ist es aber sinnvoller, das Argument $\vec{y}$ von F als Punkt $\boldsymbol{y}$ aufzufassen und den Funktionswert als Tangenten*vektor* in diesem Punkt zu interpretieren. Eine

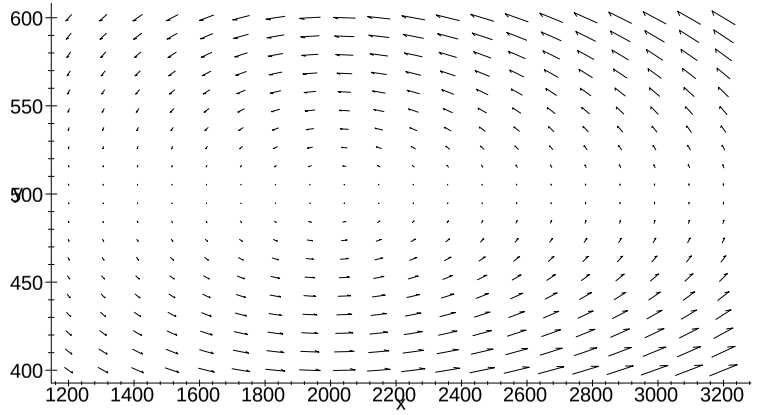


Abb. 43: Das Vektorfeld zur Raubtier-Beutetier-Gleichung

Lösungskurve der Differentialgleichung ist dann eine Kurve, die in jedem Punkt $\boldsymbol{y}$ den Vektor $F(\boldsymbol{y})$ als Tangentenvektor hat.

Abbildung 43 zeigt das Vektorfeld im betrachteten Beispiel; es legt zumindest die Vermutung nahe, daß die Lösungen zyklisch um einen Punkt rotieren. *Genau* können das wir freilich aufgrund der graphischen Information nicht sagen: Eine visuell nicht wahrnehmbare Richtungsänderung der Vektoren gehört zu einer Lösung die sich spiralförmig auf den Gleichgewichtspunkt zusammenzieht oder aber spiralförmig ins Unendliche geht.

Schon bei der Visualisierung von Vektorfeldern haben wir gesehen, daß es gelegentlich übersichtlicher ist, auf die Längeninformation zu verzichten und nur die Richtung zu betrachten. Bei Differentialgleichungen, bei denen es bei einer graphischen Lösung praktisch nur auf die *Richtung* des Vektorfelds in jedem Punkt ankommt, gilt dies umso mehr; oft versucht man daher die Lösungskurve durch ein auf Einheitslänge normiertes Vektorfeld zu führen. Abbildung 44 zeigt, wie dies im vorliegenden Beispiel aussieht.

g) Stabilitätsfragen

Die Beschreibung eines realen Systems durch ein mathematisches Mo-

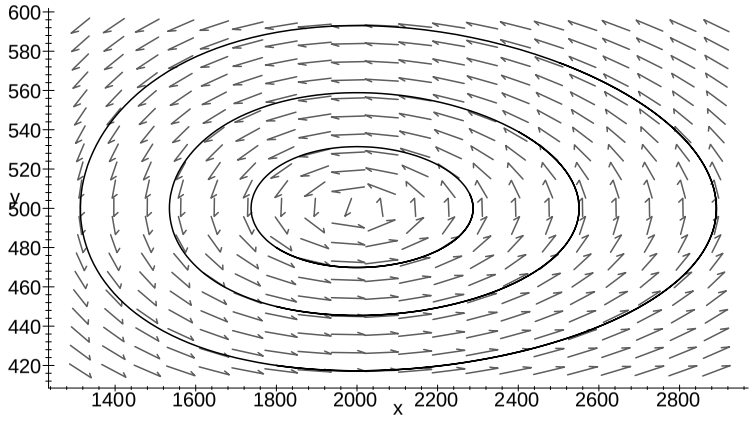


Abb. 44: Anpassung von Lösungskurven an das Vektorfeld

dell wie ein Differentialgleichungssystem ist abgesehen von einigen ganz einfachen Fällen immer mit einer Idealisierung verbunden; das reale System verhält sich daher nicht *exakt* so wie das Modell es vorhersagt. Auch die Anfangsbedingungen des Modells, die dem Zustand des realen Systems zu einem vorgegebenen Anfangszeitpunkt entsprechen, lassen sich nur durch fehlerbehaftete Messungen bestimmen. Hinzu kommt, daß man bei der Auswertung des mathematischen Modells nur selten wirklich mit reellen Zahlen rechnet; meistens rechnet man per Computer und somit (falls man keine sehr spezialisierte Mathematiksoftware benutzt) mit rundungsfehlerbehafteten Gleitkommaoperationen.

Aus einem mathematischen Modell abgeleiteten Aussagen können daher nur dann nützlich für die Vorhersage von realen Systemen sein, wenn sie stabil sind gegenüber kleineren Änderungen von Koeffizienten und Anfangsbedingungen.

Betrachten wir dazu ein Beispiel: Eine Größe $y(t)$ sei beschrieben durch das Anfangswertproblem

$$\dot{y}(t) = y(t) + 2e^{-t} \quad \text{mit} \quad y(0) = 1\,.$$

Wie man sich sofort durch Einsetzen überzeugt, ist $y(t) = e^{-t}$ eine

Lösung. Die rechte Seite

$$F(y,t) = y - e^{-t}$$

genügt offensichtlich auf ganz $\mathbb{R}^2$ einer LIPSCHITZ-Bedingung mit Konstante eins, denn $F_y(y,t) \equiv 1$, und man sieht auch direkt, daß

$$|F(y_1,t) - F(y_2,t)| = |y_1 - y_2| \le 1 \cdot |y_1 - y_2|$$

ist. Somit ist $y(t) = e^{-t}$ die *einzige* Lösung des Anfangswertproblems.

Trotzdem ist diese Lösung für alle praktischen Zwecke völlig wertlos:

$$\dot{y}(t) = y(t) - 2e^{-t}$$

ist eine inhomogene lineare Differentialgleichung, deren zugehörige homogene Gleichung

$$\dot{y}(t) = y(t)$$

die allgemeine Lösung

$$y(t) = \lambda e^{t}$$

hat. Die allgemeine Lösung der inhomogenen Gleichung ist daher

$$y(t) = e^{-t} + \lambda e^{t} \quad \text{und} \quad y(0) = 1 + \lambda\,.$$

Sobald also die Anfangsbedingung auch nur minimal gestört wird, geht die Lösung für große t nicht mehr gegen null, sondern je nach Vorzeichen von λ gegen $\pm\infty$. Bei einem solchen *schlecht gestellten* oder *strukturell instabilen* Problem läßt sich also mathematisch nichts vorhersagen.

Auch eine numerische Lösung des Anfangswertsproblems wird wegen allfälliger Rundungsfehler über kurz oder lang die exakte Lösungskurve $y(t) = e^{-t}$ verlassen und auf eine der zumindest anfänglich benachbarten anderen Kurven überwechseln, so daß auch sie für $t \to \infty$ divergiert. Abbildung 45 zeigt eine mit einem RUNGE-KUTTA-Verfahren der Ordnung vier/fünf berechnete numerische Lösung; wie man sieht, hat sie ab etwa $t = 14$ nichts mehr mit der korrekten Lösung zu tun.

Anders sieht es aus beim Anfangswertproblem

$$\dot{y}(t) = -y(t) + 1 \quad \text{mit} \quad y(0) = 1\,.$$

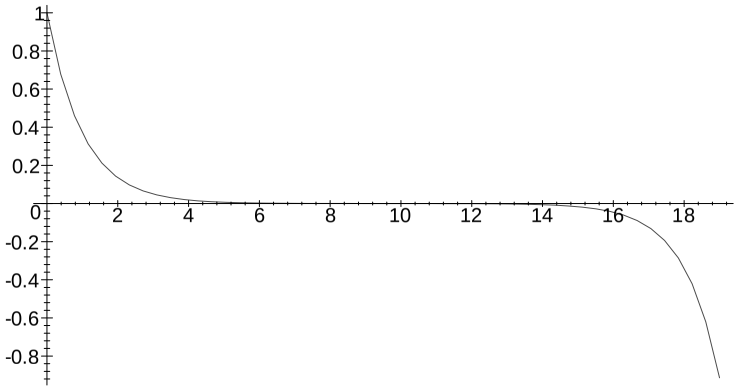


Abb. 45: Divergenz einer numerisch berechneten Lösungskurve

Hier überzeugt man sich leicht, daß $y(t) = 1$ die einzige Lösung ist, aber jetzt hat die zugehörige homogene Differentialgleichung die allgemeine Lösung $y(t) = \lambda e^{-t}$; die allgemeine Lösung der Differentialgleichung

$$\dot{y}(t) = -y(t) + 1$$

ist also

$$y(t) = 1 + \lambda e^{-t} \quad \text{mit} \quad y(0) = 1 + \lambda\,.$$

Kleine Störungen λ des Anfangswerts werden jetzt durch den Faktor e^{-t} weggedämpft; unabhängig von solchen Störungen bleibt also die Lösung $y \equiv 1$ stabil.

Um allgemeiner zu sehen, was in der Umgebung eines Gleichgewichts passieren kann, versuchen wir, die Gleichung in der Umgebung des Gleichbewichts anzunähern durch die einzige Klasse von Differentialgleichungen, die wir wirklich beherrschen, die linearen homogenen Differentialgleichungen mit konstanten Koeffizienten.

Dazu erinnern wir uns an die Definition einer differenzierbaren Funktion mehrerer Veränderlicher: $F\colon \mathbb{R}^n \to \mathbb{R}_n$ ist im Punkt $\boldsymbol{x} \in \mathbb{R}^n$ differenzierbar, wenn in einer Umgebung des Punktex gilt

$$F(\boldsymbol{x} + \vec{h}) = F(\boldsymbol{x}) + J_F(\boldsymbol{x})\vec{h} + o\big(\|\vec{h}\|\big)\,,$$

wobei $J_F(\boldsymbol{x})$ die JACOBI-Matrix von F in $\boldsymbol{x}$ ist.

Wir betrachten ein autonomes Differentialgleichungssystem

$$\dot{\vec{y}}(t) = F\big(\vec{y}(t)\big)$$

mit einer differenzierbaren Funktion F mit Fixpunkt $\vec{y}_0$. In der Umgebung des Fixpunkts ist dann

$$F(\vec{y}_0 + \vec{h}) = F(\vec{y}_0) + J_F(\vec{y}_0)\vec{h} + o\big(\|\vec{h}\|\big) = \vec{y}_0 + J_F(\vec{y}_0)\vec{h} + o\big(\|\vec{h}\|\big)\,;$$

falls wir den Fehlerterm $o\big(\|\vec{h}\|\big)$ vernachlässigen, genügt also die Differenz

$$\vec{u}(t) \underset{\text{def}}{=} \vec{y}(t) - \vec{y}_0$$

zwischen der Gleichgewichtslösung und einer nahe benachbarten Lösung näherungsweise einer linearen homogenen Differentialgleichung

$$\dot{\vec{u}}(t) = J_F(\vec{y}_0)\vec{u}(t)\,.$$

Das Langzeitverhalten von deren Lösungen hängt, wie wir aus §3*c)* wissen, von den Eigenwerten der Matrix $J_F(\vec{y}_0)$ ab: Falls diese allesamt negativen Realteil haben, konvergiert jede Lösung $\vec{u}$ für $t \to \infty$ gegen den Nullpunkt; falls alle Eigenwerte positiven Realteil haben, divergiert jede Lösung außer der Null ins Unendliche. Im ersten Fall sprechen wir von einem *stabilen* oder *anziehenden* Fixpunkt, im zweiten von einem *instabilen* oder *abstoßenden.* Falls einige Eigenwerte positiven und andere negativen Realteil haben, reden wir von einem *Sattelpunkt;* hier hängt es von der Richtung ab, ob eine Störung weggedämpft wird oder nicht, allerdings wird in der Praxis fast jede Störung zur Divergenz führen, denn nur im linearen Unterraum, der von den Eigenvektoren zu den Eigenwerten mit negativem Realteil aufgespannt wird, werden die Störungen weggedämpft. Eine zufällige Störung wird aber meist sehr schnell aus diesem Unterraum herausführen, so daß dann auch Eigenwerte mit positivem Realteil eine Rolle spielen.

Bei Eigenwerten mit Realteil null reicht die Linearisierung nicht aus, um zu Aussagen über das Stabilitätsverhalten zu kommen, da dann die Terme höherer Ordnung das Geschehen dominieren. Es kann sehr schwer sein, in so einem Fall die Dynamik vorherzusagen.

Abbildung 46 zeigt das Vektorfeld in der Nähe eines stabilen Fixpunkts; alle Lösungskurven laufen auf diesen Punkt zu. Bei einem instabilen Fixpunkt hätten wir dasselbe Bild, nur daß dann alle Pfeile in Gegenrichtung zeigen würden.

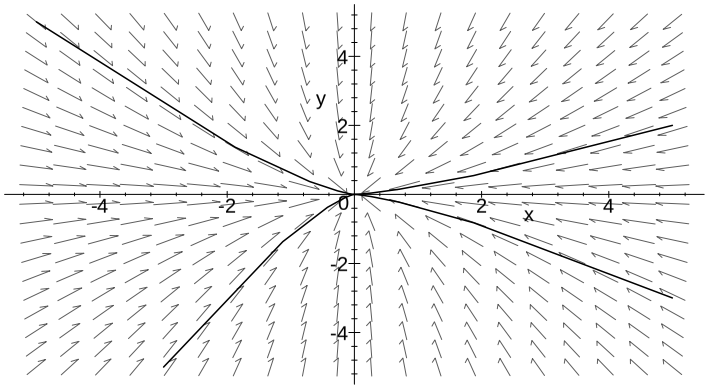


Abb. 46: Die Umgebung eines stabilen Fixpunkts

Auch in Abbildung 47 ist ein stabiler Fixpunkt zu sehen; hier hat aber die JACOBI-Matrix zwei konjugiert komplexe Eigenwerte, so daß sich benachbarte Lösungen spiralförmig auf den Fixpunkt zusammenziehen. Bei einem abstoßenden Fixpunkt hätten wir wieder im wesentlichen dasselbe Bild, aber mit umgedrehten Pfeilen.

Abbildung 48 zeigt die Umgebung eines Sattelpunkts; hier haben wir Lösungskurven, die sich zwar asymptotisch der y-Achse annähern, auf dieser aber gegen plus oder minus unendlich gehen.

Als Beispiel wollen wir die Lösungen der LORENZ-Gleichungen

$$\dot{x}(t) = p\big(y(t) - x(t)\big)$$
$$\dot{y}(t) = rx(t) - y(t) - x(t)z(t)$$
$$\dot{z}(t) = -bz(t) + x(t)y(t)$$

untersuchen. Dieses Differentialgleichungssystem ist eine extreme Vereinfachung der sogenannten NAVIER-STOKES-Gleichung, einer partiellen Differentialgleichung, die Strömungsphänomene beschreibt. Für technische Informatiker interessanter ist wohl, daß dasselbe System nach HAKEN (Phys. Lett. A53 (1975), 77–78) auch das Verhalten von Lasern beschreiben kann.

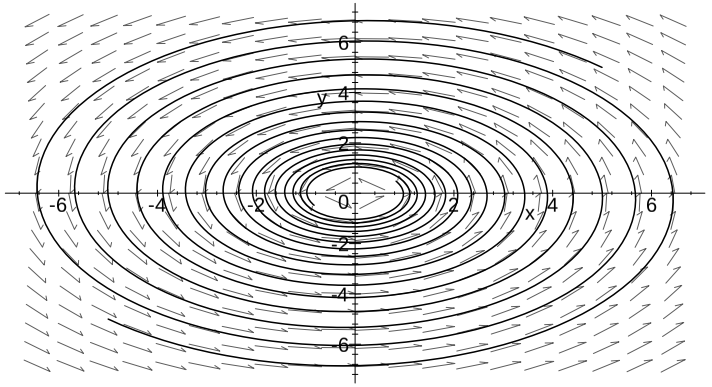


Abb. 47: Zwei konjugiert komplexe Eigenwerte mit negativem Realteil

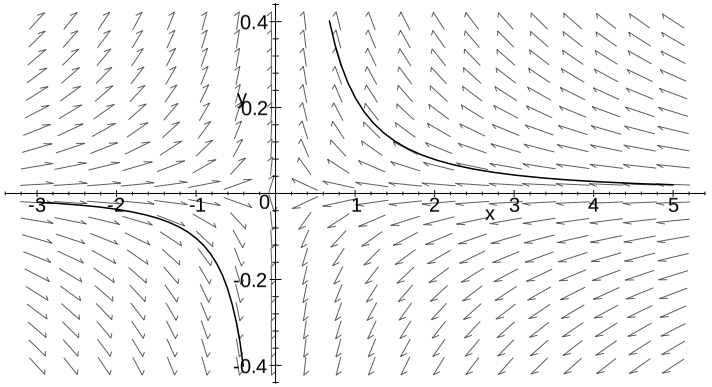


Abb. 48: Umgebung eines Sattelpunkts

Die Funktionen $x(t), y(t)$ und $z(t)$ verlieren im Vereinfachungsprozeß ihre unmittelbare physikalische Bedeutung; die Parameter lassen sich al-

lerdings physikalisch interpretieren: Für die atmosphärische Konvektion ist nach LORENZ

$$p = 10, \quad r = 28 \quad \text{und} \quad b = \frac{8}{3}$$

eine sinnvolle Wahl.

EDWARD NORTON LORENZ wurde 1917 im US-Bundesstaat Connecticut geboren; er studierte Mathematik in Dartmouth College (A.B. 1938) und Harvard (M.A. 1940). Nach seinem Kriegsdienst ging er ans MIT, wo er 1948 über Meteorologie promovierte. Sowohl dem MIT, wo er 1981 als Professor emeritiert wurde, als auch der Meteorologie blieb er fortan treu. Zu seinen vielen Auszeichnungen gehört unter anderem der Kyoto-Preis von 1991, der wohl höchstdotierte Wissenschaftspreis.

Da uns der erste Augenschein nichts über das Verhalten der Lösungen zeigt, empfiehlt es sich, daß wir uns durch numerische Simulation einen ersten Eindruck verschaffen. Abbildung 49 zeigt die Lösung des Anfangswertproblems mit $x(0) = 2$ und $y(0) = z(0) = 10$; die meisten werden ähnliche Bilder wohl schon gesehen haben.

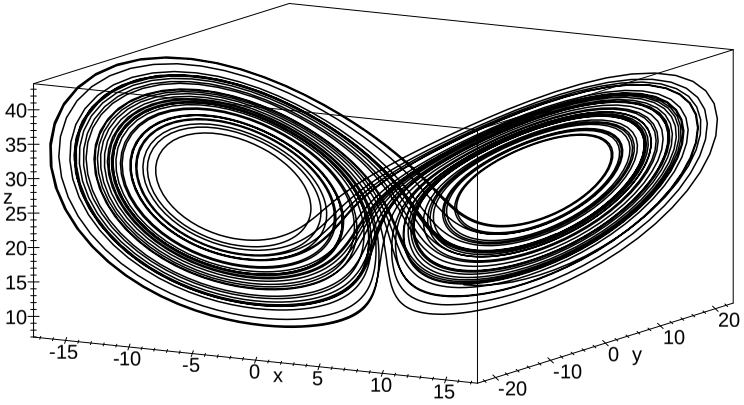


Abb. 49: Eine Bahnkurve des LORENZ-Systems

Leider ist dieses Bild einerseits etwas unübersichtlich, andererseits zeigt es nur eine einzige Lösungskurve. Um besser zu verstehen, was hier passiert, beschränken wir uns auf die Funktion $x(t)$ und betrachten diese für zwei Lösungskurven; Abbildung 50 zeigt die für die Anfangsbedingungen

$$x(0) = 2, \quad y(0) = z(0) = 10 \quad \text{und} \quad x(0) = 2{,}01, \quad y(0) = z(0) = 10\,.$$

Wie man sieht, sind die beiden Lösungskurven bis etwa zum Zeitpunkt $t = 6{,}5$ praktisch ununterscheidbar, danach gehen sie aber recht schnell auseinander und haben ab etwa $t = 10$ nichts mehr miteinander zu tun.

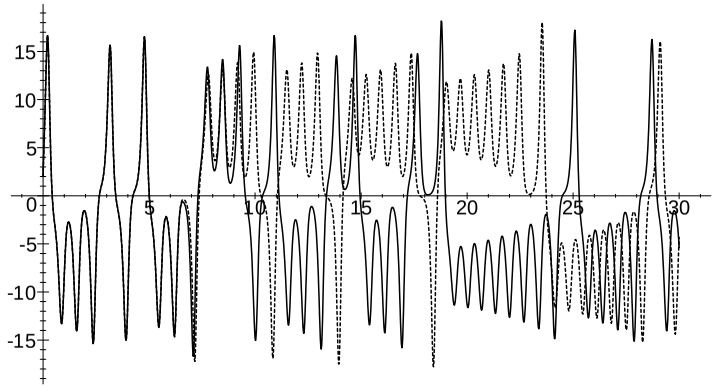


Abb. 50: Die x-Koordinaten zweier Lösungen mit benachbarten Anfangswerten

Fast das gleiche Bild ergibt sich, wenn wir die gestrichelte Kurve nicht mit $x(0) = 2{,}01$ anfangen lassen, sondern – bei sonst unveränderten Werten – bei

$$x(0) = 2{,}000001 = 2 + 10^{-6}\,.$$

Jetzt sind die Kurven zwar bis etwa $t = 16$ praktisch ununterscheidbar, aber spätestens ab etwa $t = 20$ haben sie auch hier nichts mehr miteinander zu tun.

Gerade beim zweiten Fall sollte uns das zu denken geben: Wenn wir Differentialgleichungen zur Vorhersage benutzen, stammen die Anfangsbedingungen im allgemeinen aus einer Messung. Man kann aber nur selten so genau messen, daß sich die beiden Werte 2 und 2,000001 unterscheiden ließen; um eine sinnvolle Voraussage über das Verhalten der Lösung

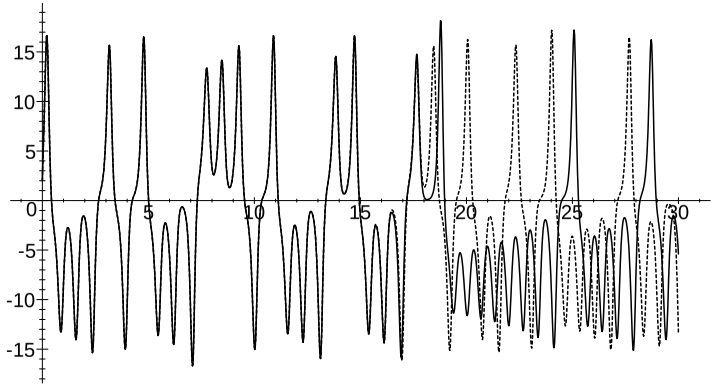


Abb. 51: Effekt einer Störung des Anfangswerts um 10^{-6}

zum Zeitpunkt $t = 20$ zu machen, *muß* man aber nach Abbildung 51 den Wert $x(0)$ mit dieser Genauigkeit kennen.

Es kommt noch schlimmer: In Abbildung 52 ist die dick ausgezogene Kurve wieder eine numerische Simulation der Lösung zu den Anfangsbedingungen $x(0) = 2$ und $y(0) = z(0) = 10$, die gestrichelte Kurve allerdings auch! Die beiden Kurven unterscheiden sich nur dadurch, daß die numerische Simulation bei der dick ausgezogenen Kurve (wie auch bei allen anderen bisherigen Kurven) mit Schrittweite 0,02 arbeitete, wohingegen die Schrittweite für die gestrichelte Kurve mit 0,01 nur halb so groß war. Auch das reicht schon, daß die Kurven ab etwa $t = 10$ nichts mehr miteinander zu tun haben, und damit dürfte wohl auch klar sein, daß keine der bislang betrachteten Kurven für größere Werte von t irgendetwas mit der „wahren“ Lösungsfunktion $x(t)$ zu tun hatte.

LORENZ mußte dieses Verhalten der Lösungen auf die harte Weise lernen: Er fand zu seinem großen Erstaunen, daß sich seine Rechenergebnisse nicht reproduzieren ließen, wenn er Zwischenergebnisse für Kontrollrechnungen nur in gerundeter Form eintippte: Der geringe Rundungsfehler bei der Eingabe des Startwerts reichte bereits aus, um das Langzeitverhalten des Systems grundlegend zu verändern.

Falls das gleiche Phänomen auch in der wirklichen Atmosphäre auftritt,

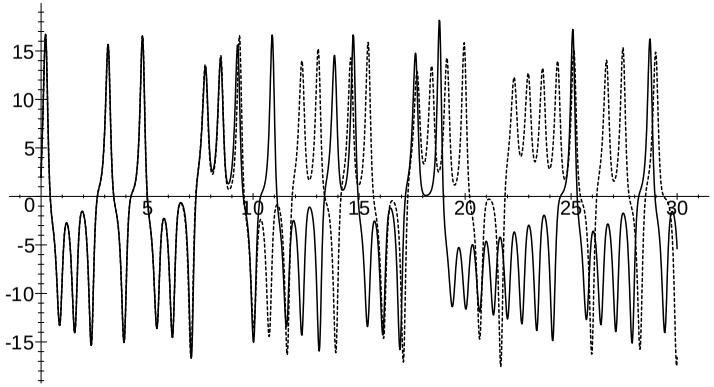


Abb. 52: Effekt einer Schrittweitenhalbierung bei der numerischen Simulation

können also minimale Veränderungen etwa des Luftdrucks oder der Temperatur auf längere Sicht zu einer dramatisch anderen Entwicklung des Wetters führen – eine Idee, die vielen Meteorologen damals als zu phantastisch erschien um ernstgenommen zu werden: Am 22. Januar 1963 berichtete LORENZ vor der New York Academy of Sciences über seine Ergebnisse *(Trans. N.Y. Acad. Sci.* **25** *(1963), 409–432)* und schloß seinen Vortrag mit den Worten:

Als die Instabilität eines gleichförmigen Flusses gegenüber infinitesimalen Störungen erstmals als Erklärung für das Auftreten von Zyklonen und Antizyklonen in der Atmosphäre vorgeschlagen wurde, war diese Idee nicht allgemein akzeptiert. Ein Meteorologe bemerkte, daß, falls die Theorie korrekt wäre, ein Flügelschlag einer Möwe ausreichen würde, um die Entwicklung des Wetters für immer zu verändern. Die Kontroverse ist noch nicht entschieden, aber die neueste Evidenz scheint für die Möwen zu sprechen.

Inzwischen ist der Sieg der Möwen bekanntlich allgemein anerkannt; man fordert sogar nicht einmal mehr den relativ kräftigen Flügelschlag einer Möwe, um das Wetter permanent zu verändern: Im Dezember 1972 hielt LORENZ vor der American Association for the Advancement of Sciences in Washington, DC, einen Vortrag mit dem Titel *Predictability:*

Does the Flap of a Butterfly's Wings in Brazil set off a Tornado in Texas, und seitdem geht das Wort vom *Schmetterlingseffekt* um die Welt.

Auch das Wort *Chaos* wird heute meist auf diese Weise definiert: Kleinste Änderungen bei den Anfangsbedingungen führen zu dramatischen Veränderungen des Langzeitverhaltens. Allerdings muß man hier aufpassen: Bei der Differentialgleichung

$$\dot{y}(t) = y(t)$$

mit Anfangsbedingungen

$$y(0) = 1 \quad \text{und} \quad y(0) = 1 + \varepsilon$$

unterscheiden sich die Lösungen $y(t) = e^t$ und $y(t) = (1+\varepsilon)\cdot e^t$ zur Zeit t um $\varepsilon \cdot e^t$, was auch bei kleinsten ε-Werten sehr schnell eine sehr große Zahl wird: bei $\varepsilon = 10^{-6}$ und $t = 50$ etwa ist die Differenz bereits größer als $5 \cdot 10^{15}$. Trotzdem wird hier niemand von Chaos reden, denn beide Lösungen gehen sehr schnell gegen den „Gleichgewichtspunkt" unendlich. Von „echtem" Chaos verlangt man daher auch noch, daß die Lösungen nicht gegen eine (endliche oder unendliche) Gleichgewichtslösung konvergieren und auch nicht gegen eine periodische Lösung. Chaos in diesem Sinne ist sehr schwer nachzuweisen; für die LORENZ-Gleichung mit den klassischen Parameterwerten wurde erst Ende September 1999 ein *preprint* veröffentlicht, in dem dies (mit großem theoretischen wie auch rechnerischem Aufwand) gezeigt wird; siehe `http://www.math.gatech.edu/~mischaik/papers/lor3.ps` .

Chaos heißt nun allerdings nicht, daß wir dann überhaupt nichts über das Verhalten der Lösungen aussagen können. Beispielsweise können wir bereits mit unseren einfachen Mitteln zeigen, daß das Bild in Abbildung 49 zumindest qualitativ das Verhalten der Lösungskurven korrekt wiedergibt – quantitativ ist natürlich ab spätestens etwa $t = 10$ alles falsch.

Dazu berechnen wir zunächst die Gleichgewichtslösungen: Im Gleichungssystem

$$\begin{aligned} 0 &= p(y-x) \\ 0 &= rx - y - xz \\ 0 &= -bz + xy \end{aligned}$$

zeigt die erste Gleichung, falls wir den uninteressanten Fall $p = 0$ ausschließen, daß die x-Koordinate und die y-Koordinate eines jeden Fixpunkts übereinstimmen müssen.

Falls beide Koordinaten verschwinden, zeigt die dritte Gleichung ($b \neq 0$ vorausgesetzt), daß dann auch die z-Koordinate verschwinden muß; Einsetzen in die Gleichungen zeigt, daß der Nullpunkt in der Tat ein Fixpunkt ist.

Im Fall $x \neq 0$ können wir y in der zweiten Gleichung durch x ersetzen und dann durch x dividieren; dies ergibt die z-Koordinate

$$z = r - 1 \,.$$

Damit zeigt die dritte Gleichung, daß es für $r \neq 1$ noch zwei weitere Fixpunkte gibt mit

$$x = y = \pm\sqrt{b(r-1)} \quad \text{und} \quad z = r - 1 \,.$$

Die Untersuchung des relativ uninteressanten Nullpunkts sei dem Leser als Übungsaufgabe überlassen; hier seien nur die beiden anderen Fixpunkten betrachtet. Für

$$x = y = \pm\sqrt{b(r-1)} \quad \text{und} \quad z = r - 1$$

führen wir, wie oben im allgemeinen Fall, neue Variablen u, v und w ein, die den Abstand zum Fixpunkt beschreiben, d.h.

$$x = \pm\sqrt{b(r-1)} + u, \quad y = \pm\sqrt{b(r-1)} + v \quad \text{und} \quad z = r - 1 + w \,.$$

Zur Linearisierung in der Nähe des Gleichgewichtspunkt vernachlässigen wir alle nichtlinearen Terme in $u(t), v(t)$ und $w(t)$; das entstehende lineare Differentialgleichungssystem hat dann die JACOBI-Matrix im Fixpunkt als Matrix, ist also

$$\begin{pmatrix} \dot{u}(t) \\ \dot{v}(t) \\ \dot{w}(t) \end{pmatrix} = A \begin{pmatrix} u(t) \\ v(t) \\ w(t) \end{pmatrix}$$

mit

$$A = \begin{pmatrix} -p & p & 0 \\ 1 & -1 & \mp\sqrt{b(r-1)} \\ \pm\sqrt{b(r-1)} & \pm\sqrt{b(r-1)} & -b \end{pmatrix} .$$

Das charakteristische Polynom

$$\det(A - \lambda E) = -\lambda^3 - (b + 1 + p)\lambda^2 - b(r - p)\lambda - 2pb(r - 1)$$

ist für beide Fixpunkte dasselbe, verleitet aber nicht dazu, es allgemein lösen zu wollen. Wir setzen daher die von LORENZ vorgeschlagenen speziellen Parameterwerte ein und erhalten

$$-\lambda^3 - \frac{41}{3}\lambda^2 - \frac{304}{3}\lambda - 1440\,,$$

was immer noch so schlimm ist, daß wir es besser numerisch lösen. Die drei Lösungen ergeben sich näherungsweise als

$$-13{,}85457791 \quad \text{und} \quad 0{,}09395562396 \pm 10{,}19450522i\,.$$

Es gibt also einen negativen Eigenwert und zwei konjugiert komplexe Eigenwerte mit positivem Realteil. Damit ist klar, wie Lösungskurven des linearisierten Systems in der Nähe der beiden Fixpunkte aussehen: Der negative Eigenwert sorgt dafür, daß die Lösungen asymptotisch in die Ebene gedrückt werden, die von den Eigenvektoren zu den beiden anderen Eigenwerten aufgespannt wird, und die beiden komplexen Eigenwerte sorgen dafür, daß sie dort spiralförmig nach außen gehen.

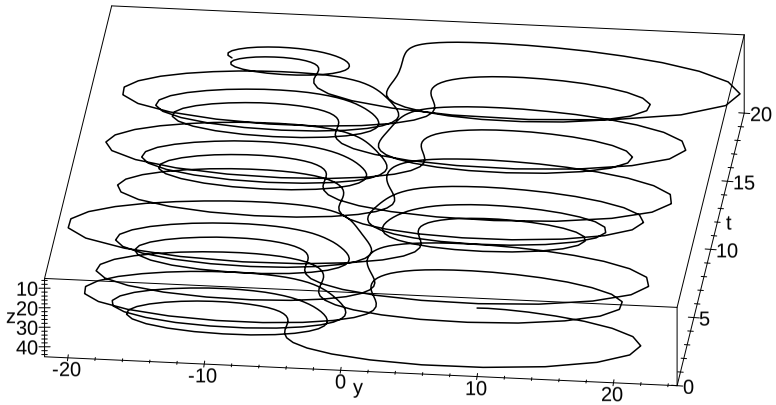


Abb. 53: y- und z-Koordinate als Funktion von t

Damit wird das Verhalten des LORENZ-Systems klar: Wir haben zwei Ebenen, die jeweils einen Sattelpunkt enthalten; kommt eine Lösung in die Nähe eines solchen Sattelpunkts, wird sie von der entsprechenden Ebenen eingefangen und geht dort spiralförmig nach außen. Wenn sie sich hinreichend weit vom Sattelpunkt entfernt hat, sind die Voraussetzungen für die obige Linearisierung nicht mehr gegeben; die Lösung kann daher der Ebenen entkommen, wird aber über kurz oder lang von der Ebenen des anderen Sattelpunkts eingefangen und so weiter. Abbildung 53 zeigt dieses Verhalten etwas klarer als Abbildung 49: Hier sind die die y- und die z-Koordinate der Lösungskurve über der Zeit aufgetragen.

Kapitel 5
Optimierung, Fehlerrechnung und Statistik

In der Schule werden Ableitungen hauptsächlich benutzt, um die Extremwerte einer Funktion zu bestimmen; ein Gesichtspunkt, der im letzten Semester bei der Differentialrechnung mehrerer Veränderlicher keine Rolle spielte. In diesem letzten Kapitel der Vorlesung soll dies nachgeholt werden, wobei insbesondere die Anwendungen auf die Fehler- und Ausgleichsrechnung wichtige Beispiele liefern. Zu deren besseren Verständnis sollen auch einige Grundbegriffe der Statistik erörtert werden.

§ 1: Extrema von Funktionen mehrerer Veränderlicher

a) Der eindimensionale Fall

Erinnern wir uns an die Schule: Wenn die stetig differenzierbare Funktion $f\colon (a, b) \to \mathbb{R}$ im Punkt $x_0 \in (a, b)$ ein Extremum annimmt, verschwindet dort die Ableitung $f'(x_0)$. Der Grund ist klar: Nach Definition der Differenzierbarkeit ist

$$f(x_0 + h) = f(x_0) + hf'(x_0) + o(h)\,;$$

falls $f'(x_0)$ nicht verschwindet, ist $f(x_0 + h)$ für kleine h mit demselben Vorzeichen wie $f'(x_0)$ größer und für solche mit entgegengesetztem Vorzeichen kleiner als $f(x_0)$. In x_0 kann f somit weder ein Maximum noch ein Minimum annehmen.

Die Umkehrung gilt nicht: Standardbeispiel ist die Funktion $f(x) = x^3$, für die $f'(0)$ verschwindet, ohne daß im Nullpunkt ein Maximum oder Minimum wäre.

Für zweimal stetig differenzierbare Funktionen gibt es bekanntlich auch eine hinreichende Bedingung sowie die Möglichkeit, Maxima und Minima voneinander zu unterscheiden: Falls $f'(x_0)$ verschwindet und $f''(x_0)$ negativ ist, hat f im Punkt x_0 ein Maximum; bei positivem $f''(x_0)$ liegt ein Minimum vor. Auch hier folgt alles sofort aus der Definition der zweimaligen Differenzierbarkeit: Wegen

$$\begin{aligned} f(x_0 + h) &= f(x_0) + hf'(x_0) + \frac{h^2}{2} f''(x_0) + o(h^2) \\ &= f(x_0) + \frac{h^2}{2} f''(x_0) + o(h^2) \end{aligned}$$

sieht der Graph von f in diesen Fällen in der unmittelbaren Umgebung von x_0 aus wie eine nach unten *bzw.* oben geöffnete Parabel.

b) Verallgemeinerung aufs Mehrdimensionale

Nun betrachten wir eine stetig differenzierbare Funktion $f\colon D \to \mathbb{R}$ auf einer offenen Teilmenge $D \subset \mathbb{R}^n$. Dann bedeutet Differenzierbarkeit bekanntlich, daß es in jedem Punkt $\boldsymbol{x}_0 \in D$ einen Vektor

$$\nabla f(\boldsymbol{x}_0) = \operatorname{grad} f(\boldsymbol{x}_0) \in \mathbb{R}^n$$

gibt, den Gradienten, so daß für hinreichend kleine Vektoren $\vec{h} \in \mathbb{R}^n$ gilt

$$f(\boldsymbol{x}_0 + \vec{h}) = f(\boldsymbol{x}) + \operatorname{grad} f(\boldsymbol{x}_0) \cdot \vec{h} + o\big(|\vec{h}|\big)\,.$$

Hier muß also für jeden Extremwert $\operatorname{grad} f(\boldsymbol{x}_0)$ gleich dem Nullvektor sein, denn setzt man für $\vec{h}$ ein kleines Vielfaches $t \cdot \operatorname{grad} f(\boldsymbol{x}_0)$ des Gradienten ein, wäre sonst

$$f(\boldsymbol{x}_0 + \vec{h}) = f(\boldsymbol{x}) + t\big(\operatorname{grad} f(\boldsymbol{x}_0) \cdot \operatorname{grad} f(\boldsymbol{x}_0)\big) + o\big(|\vec{h}|\big)$$

für kleine positive t größer als $f(x_0)$ und für kleine negative t kleiner.

Die Frage, welche Nullstellen des Gradienten wirklich Extremwerten entsprechen, ist schwieriger; in der Praxis wird es oft am einfachsten sein, sich die Umgebung des betreffenden Punktes mit irgendwelchen *ad hoc*-Methoden genauer anzusehen und dann zu entscheiden.

Klassisches Beispiel eines Punkts, in dem der Gradient verschwindet, ohne daß ein Extremwert vorliegt, ist der in Abbildung 54 gezeigte

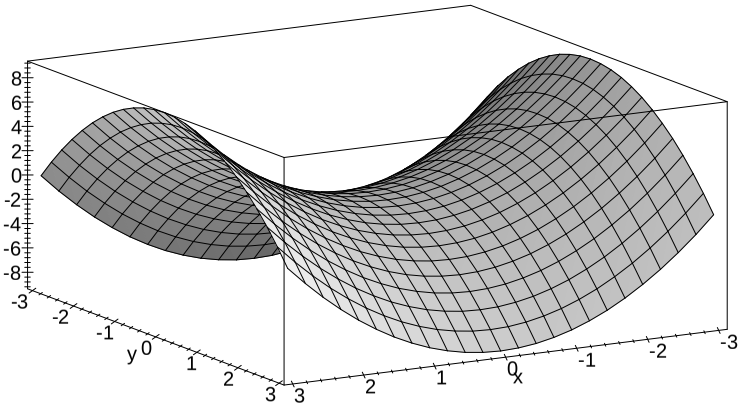


Abb. 54: Graph der Funktion $f(x,y) = x^2 - y^2$

Sattelpunkt, hier dargestellt als Funktionswert über dem Punkt $(0,0)$ für die Funktion $f(x,y) = x^2 - y^2$.

Für zweifach stetig differenzierbare Funktionen kann man genau wie im eindimensionalen Fall ein hinreichendes Kriterium finden, das nur von der zweiten Ableitung im Punkt $\boldsymbol{x}_0$ abhängt:

Die zweite Ableitung von $f \in \mathcal{C}^2(D,\mathbb{R})$ im Punkt $\boldsymbol{x}_0 \in D$ ist bekanntlich gegeben durch die HESSE-Matrix

$$H_f(\boldsymbol{x}_0) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

und zweimalige Differenzierbarkeit bedeutet, daß

$$f(\boldsymbol{x}_0 + \vec{h}) = f(\boldsymbol{x}_0) + \operatorname{grad} f(\boldsymbol{x}_0) \cdot \vec{h} + \tfrac{1}{2}\,{}^t\vec{h}\, H_f(\boldsymbol{x}_0)\, \vec{h} + o\big(|\vec{h}^2|\big)$$

ist für kleine $\vec{h}$.

Wenn grad $f(\boldsymbol{x}_0)$ verschwindet, hängt also das Verhalten von f in der Umgebung von $\boldsymbol{x}_0$ ab von der quadratischen Form

$$\vec{h} \mapsto {}^t\vec{h}\, H_f(\boldsymbol{x}_0)\, \vec{h}\,.$$

Definition: *a)* Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt *positiv definit,* wenn für alle Vektoren $\vec{v} \neq \vec{0}$ aus $\mathbb{R}^n$ gilt:

$${}^t\vec{v} A(\boldsymbol{x}_0)\vec{v} > 0\,.$$

b) A heißt *negativ definit,* wenn für alle $\vec{v} \neq \vec{0}$ aus $\mathbb{R}^n$ gilt:

$${}^t\vec{v} A(\boldsymbol{x}_0)\vec{v} < 0\,.$$

c) A heißt *indefinit,* wenn es Vektoren $\vec{v}, \vec{w} \in \mathbb{R}^n$ gibt mit

$${}^t\vec{v} A(\boldsymbol{x}_0)\vec{v} > 0 \quad \text{und} \quad {}^t\vec{w} A(\boldsymbol{x}_0)\vec{w} < 0\,.$$

Mit dieser Terminologie ist das folgende Lemma klar:

Lemma: Wenn die differenzierbare Funktion $f \in \mathcal{C}^1(D,\mathbb{R})$ im Punkt $\boldsymbol{x}_0 \in D$ ein lokales Extremum hat, ist dort ihr Gradient gleich dem Nullvektor.
Falls umgekehrt für $f \in \mathcal{C}^2(D,\mathbb{R})$ der Gradient im Punkt $\boldsymbol{x} \in D$ verschwindet, gilt:
a) Falls die HESSE-Matrix $H_f(\boldsymbol{x}_0)$ positiv definit ist, hat f im Punkt $\boldsymbol{x}_0$ ein Minimum.
b) Falls $H_f(\boldsymbol{x}_0)$ negativ definit ist, hat f im Punkt $\boldsymbol{x}_0$ ein Maximum.
c) Falls $H_f(\boldsymbol{x}_0)$ indefinit ist, hat f im Punkt $\boldsymbol{x}_0$ kein Extremum. ∎

Damit uns das etwas nützt, brauchen wir jetzt nur noch ein Kriterium, mit dem wir feststellen können, welche Definitheitseigenschaften die HESSE-Matrix hat. Dazu erinnern wir uns daran, daß die HESSE-Matrix symmetrisch ist, und daß nach Kapitel 4, §2*d)* jede symmetrische Matrix diagonalisierbar ist.

Für eine Diagonalmatrix A mit Einträgen $\lambda_1, \ldots, \lambda_n$ und einen Vektor $\vec{v}$ mit Komponenten $v_1, \ldots, v_n$ wird obige quadratische Form zu

$$(v_1, v_2, \ldots, v_n) \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \lambda_1 v_1^2 + \cdots + \lambda_n v_n^2\,;$$

eine Diagonalmatrix ist also genau dann positiv definit, wenn alle Diagonaleinträge positiv sind und genau dann negativ definit, wenn sie alle

negativ sind. Falls es sowohl positive als auch negative Diagonaleinträge gibt, ist die Matrix indefinit.

Nun ist es für den Wertebereich einer Funktion irrelevant, bezüglich welches Koordinatensystems wir die Argumente ausdrücken; wir können eine symmetrische Matrix also bezüglich einer Basis aus Eigenvektoren betrachten, wo sie zur Diagonalmatrix wird mit den Eigenwerten als Einträgen. Daher gilt:

Lemma: Eine symmetrische Matrix ist genau dann positiv definit, wenn alle ihre Eigenwerte positiv sind und genau dann negativ definit, wenn alle ihre Eigenwerte negativ sind. Falls es sowohl positive als auch negative Eigenwerte gibt, ist sie indefinit. ∎

Da die Determinante einer Matrix gleich dem Produkt ihrer Eigenwerte ist, folgt, daß eine Matrix nur dann positiv definit sein kann, wenn ihre Determinante positiv ist; für negativ definite $n \times n$-Matrizen muß die Determinante bei geradem n ebenfalls positiv sein, bei ungeradem negativ.

Für symmetrische 2×2-Matrizen läßt sich daraus leicht ein notwendiges und hinreichendes Kriterium machen: Das charakteristische Polynom von

$$A = \begin{pmatrix} a & b \\ b & d \end{pmatrix}$$

mit Eigenwerten λ_1 und λ_2 ist

$$\lambda^2 - (a+d)\lambda + (ad - b^2) = (\lambda - \lambda_1)(\lambda - \lambda_2)\,;$$

daher ist

$$\lambda_1 + \lambda_2 = a + d\,.$$

(In der Tat rechnet man auf genau die gleiche Weise leicht nach, daß für jede $n \times n$-Matrix die Summe der n Eigenwerte gleich der Summe der n Diagonaleinträge ist, die sogenannte *Spur* der Matrix.)

Wenn $\det A = ad - b^2$ positiv ist, haben nicht nur λ_1 und λ_2, sondern auch a und d dasselbe Vorzeichen, das somit gleich dem von $a + d = \lambda_1 + \lambda_2$ ist. Also ist A genau dann positiv definit, wenn $\det A > 0$ und $a > 0$

ist, negativ definit, wenn $\det A > 0$ und $a < 0$ ist, und indefinit wenn $\det A < 0$ ist. (Anstelle von a könnte hier natürlich überall auch d stehen.)

Beispielsweise ist die Matrix $\begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}$ positiv definit, denn sie hat Determinante eins und positive Diagonaleinträge. Im obigen Beispiel des Sattelpunkts mit $f(x,y) = x^2 - y^2$ ist

$$H_f(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$$

offensichtlich indefinit, was man nicht nur an der negativen Determinanten sieht.

§2: Maxima und Minima unter Nebenbedingungen

Bei einem realen physikalischen oder technischen Prozeß können sich die Variablen selten frei im gesamten $\mathbb{R}^n$ bewegen: Physikalisch sinnvoll ist meist nur eine beschränkte Teilmenge. Im Gegensatz zur Dimension eins, wo diese Teilmenge praktisch immer ein Intervall ist, gibt es aber im Mehrdimensionalen keinen Grund, warum diese Teilmenge offen oder zumindest der Abschluß einer offenen Teilmenge sein sollte: Im $\mathbb{R}^3$ kann man sich beispielsweise auch interessieren für das Maximum oder Minimum der Ladungsdichte auf einer Kugeloberfläche oder die elektrische Feldstärke oder Temperaturverteilung auf der Innenhaut eines Reaktordruckbehälters.

Diese Maxima oder Minima sind im allgemeinen keine lokalen Maxima oder Minima der betrachteten Funktion: Wenn man die jeweilige Fläche verläßt, läßt sich der Funktionswert selbst für einen solchen Extremwert meist noch – je nach Richtung – sowohl vergrößern als auch verkleinern. Dementsprechend können die Methoden, die wir in §1 diskutiert haben, solche Extremwerte üblicherweise nicht finden; wir brauchen weitere Werkzeuge, die in diesem Paragraphen bereitgestellt werden sollen.

Die Situation, um die es hier geht, ist typischerweise die folgende: Gegeben ist eine Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$, möglicherweise auch nur auf einer Teilmenge $D \subset \mathbb{R}^n$ definiert, deren Extremwerte nicht auf $\mathbb{R}^n$ oder D

gesucht werden, sondern nur auf einer Teilmenge, die beispielsweise durch das Verschwinden einer weiteren Funktion $g\colon \mathbb{R}^n \to \mathbb{R}$ gegeben ist. Falls wir uns für Extremwerte auf einer Kugel vom Radius r um den Nullpunkt interessieren, wäre dies etwa die Funktion

$$g\colon \begin{cases} \mathbb{R}^3 & \to \mathbb{R} \\ (x,y,z) & \mapsto x^2+y^2+z^2-r^2 \end{cases} .$$

Eine mögliche Strategie zur Lösung solcher Probleme besteht darin, die Gleichung $g = 0$ nach einer der Variablen aufzulösen, diese dann in f einzusetzen und sodann eine gewöhnliche Extremwertaufgabe zu lösen. Diese Auflösung ist *explizit* nur in sehr einfachen Fällen möglich, aber selbst wenn wir nur wissen, daß eine solche Auflösung *existiert*, können wir doch damit argumentieren und Kriterien ableiten.

Unter Maxima und Minima sollen hier *lokale* Extrema verstanden werden, so daß wir die üblichen Kriterien anwenden können:

Definition: Wir sagen, die Funktion $f\colon D \to \mathbb{R}$ auf einer Teilmenge $D \subseteq \mathbb{R}^n$ habe im Punkt $\boldsymbol{a} \in D$ ein lokales $\left\{\begin{matrix}\text{Maximum}\\ \text{Minimum}\end{matrix}\right\}$ unter der Nebenbedinung $g = 0$, wobei $g\colon D \to \mathbb{R}$ eine weitere Funktion ist, wenn $g(\boldsymbol{a}) = 0$ ist und es eine Umgebung U von $\boldsymbol{a}$ gibt, so daß für alle $\boldsymbol{x} \in U$ gilt: Ist $g(\boldsymbol{x}) = 0$, so ist $f(\boldsymbol{x}) \left\{\begin{matrix}\le\\ \ge\end{matrix}\right\} f(\boldsymbol{a})$.

Als Einstiegsbeispiel betrachten wir eine beliebte Schulbuchaufgabe zur Minimumsbestimmung: Eine Konservendose soll bei einem vorgegebenen Volumen von 100 cm^3 möglichst wenig Blech benötigen, d.h. ihre Oberfläche soll minimal sein.

Die Oberfläche eines Zylinders der Höhe h mit einer Grundfläche vom Radius r ist

$$f(r,h) = 2\pi r^2 + 2\pi r \cdot h\,;$$

die Nebenbedingung für das Volumen $V = \pi r^2 h$ besagt, daß

$$g(r,h) = \pi r^2 h - 100 = 0$$

sein soll.

Hier läßt sich natürlich die Nebenbedingung sofort nach h auflösen:

$$h = \frac{100}{\pi r^2} ,$$

und wir müssen nur noch die Funktion

$$F(r) = f\left(r, \frac{100}{\pi r^2}\right) = 2\pi r^2 + \frac{200}{r}$$

minimieren. Für diese ist

$$F'(r) = 4\pi r - \frac{200}{r^2} ,$$

und dies verschwindet genau dann, wenn

$$4\pi r^3 = 200 \quad \text{oder} \quad r = \sqrt[3]{\frac{50}{\pi}}$$

ist.

In diesem einfachen Fall kann man solche Aufgaben also zurückführen auf gewöhnliche Extremwertaufgaben, indem man die Nebenbedingung nach einer der Variablen auflöst und diese dann in f einsetzt; in anderen Fällen kann man gelegentlich die Nebendedingung durch geeignete Parameterwahl oder Wahl eines angepaßten Koordinatensystems berücksichtigen. Im allgemeinen wird aber beides nicht möglich sein, so daß wir andere Methoden brauchen.

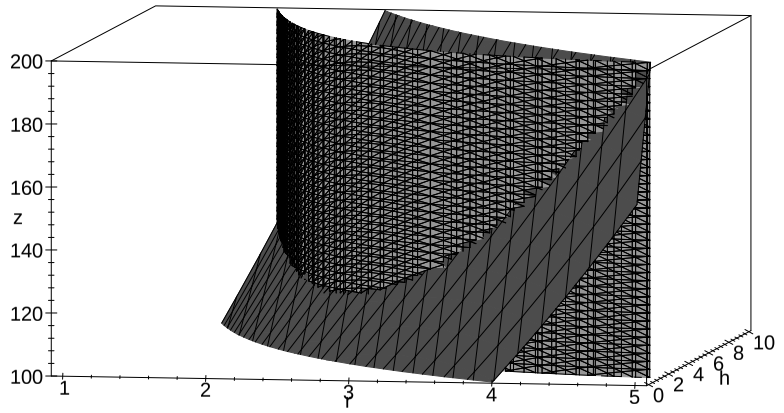


Abb. 55: Oberfläche einer Konservendose mit festem Volumen

Unser bisherige Theorie für lokale Extrema ist in dieser Situation nicht anwendbar, denn die lokalen Extrema von f werden nur in den seltensten Fällen die Nebenbedingung $g = 0$ erfüllen; im obigen Beispiel zeigt Abbildung 55 die Nebenbedingung als eng schraffierte Fläche dargestellt und der Graph von f als weiter schraffierte; wie man sieht, läßt sich der Wert von f problemlos verkleinern, wenn man nur die Fläche $g = 0$ verläßt, und in der Tat ist auch ohne jede Mathematik sofort klar, daß man mit weniger Blech auskommt, wenn man die Konservendose einfach schmäler oder kürzer macht.

Die Grundidee für ein alternatives Verfahren wird klar bei der Betrachtung der Niveaulinien in Abbildung 56: Die Niveaulinie für $g = 0$ ist gestrichelt eingezeichnet, verschiedene Niveaulinien von f als durchgezogene Kurven.

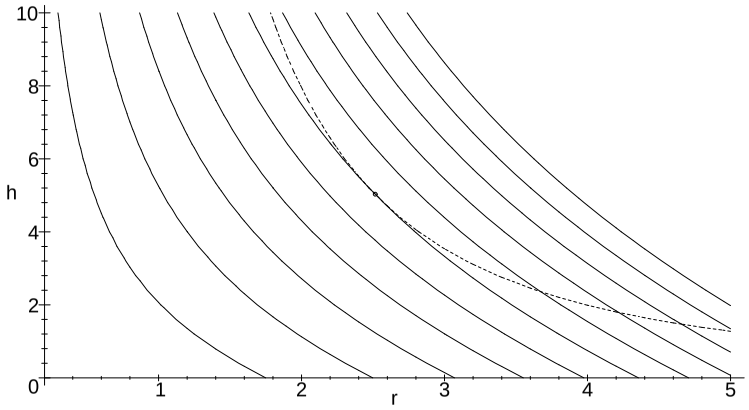


Abb. 56: Niveaulinien für Oberfläche und Volumen

Wie man sieht, schneiden einige dieser Niveaulinien die gestrichelte Kurve überhaupt nicht: Wenn man zu wenig Blech hat, kann man keine Dose mit 100 cm^3 Inhalt zusammenlöten. Wenn es dagegen genug Blech gibt, gibt es gleich zwei Schnittpunkte: Die Dose kann entweder eher höher oder eher breiter gemacht werden. In einem solchen Fall kann man die Niveaulinie durch eine zu einem etwas niedrigeren Niveau ersetzen,

die im allgemeinen auch wieder Schnittpunkte haben wird, so daß das Niveau noch nicht minimal sein kann. Erst wenn man im Minimum ist, fallen die beiden Schnittpunkte zusammen; wenn man nun das Niveau noch weiter erniedrigt, gibt es keine Schnittpunkte mehr.

Da somit im Minimum zwei Schnittpunkte zusammenfallen, berühren sich dort die Niveaulinien von f und von g, d.h. sie haben eine gemeinsame Tangente. Da der Gradient, wie wir wissen, senkrecht auf der Tangenten der Niveaulinien steht (die Richtungsableitung entlang einer Niveaulinie ist schließlich null), sind somit die Gradienten von f und g im Minimum zueinander parallel, d.h. der eine ist ein Vielfaches des anderen.

Dies gilt nicht nur im vorliegenden Beispiel, sondern allgemein:

Satz: $D \subseteq \mathbb{R}^n$ sei eine offene Menge und $f, g \in \mathcal{C}^1(D, \mathbb{R})$ seien stetig differenzierbare Funktionen auf D. Falls f im Punkt $\boldsymbol{a} \in D$ ein Extremum hat unter der Nebenbedinung $g(\boldsymbol{x}) = 0$, so sind grad $f(\boldsymbol{a})$ und grad $g(\boldsymbol{a})$ linear abhängig.

Beweis: Die Grundidee ist einfach: Auch wenn wir die Nebenbedingung nicht *explizit* nach einer der Variablen auflösen können, sagt uns der Satz über implizite Funktionen in vielen Fällen dennoch, daß zumindest lokal eine Auflösung existiert. Diese Auflösung kennen wir zwar nicht, aber wir können mit ihr argumentieren und, zumindest formal, auch rechnen.

Falls grad $g(\boldsymbol{a})$ der Nullvektor ist, gibt es nichts mehr zu beweisen, denn jede Menge, die den Nullvektor enthält, ist linear abhängig.

Wir können daher annehmen, daß grad $g(\boldsymbol{a})$ mindestens eine von Null verschiedene Komponente hat, und durch Umnummerieren der Koordinaten können wir o.B.d.A. annehmen, daß dies die n-te Komponente ist, d.h. $g_{x_n}(\boldsymbol{a}) \neq 0$.

Dann gibt es nach dem Satz über implizite Funktionen (**[HM I]**, *Kap. 2, §3d*) eine Umgebung U von $(a_1, \ldots, a_{n-1})$ und eine Funktion $h\colon U \to \mathbb{R}$ mit $h(a_1, \ldots, a_{n-1}) = a_n$, so daß

$$g\big(x_1, \ldots, x_{n-1}, h(x_1, \ldots, x_{n-1})\big) = 0 \quad \text{für alle } (x_1, \ldots, x_{n-1}) \in U\,.$$

Nachdem f in $\boldsymbol{a}$ ein lokales Extremum unter der Nebenbedinung $g = 0$ hat, nimmt die Funktion

$$F(x_1, \ldots, x_{n-1}) \underset{\text{def}}{=} f\big(x_1, \ldots, x_{n-1}, h(x_1, \ldots, x_{n-1})\big)$$

in $(a_1, \ldots, a_{n-1})$ ein lokales Extremum im üblichen Sinne an, d.h. der Gradient von F verschwindet dort.

Nach der Kettenregel ist für $i = 1, \ldots, n-1$

$$F_{x_i}(a_1, \ldots, a_{n-1}) = f_{x_i}(\boldsymbol{a}) + f_{x_n}(\boldsymbol{a}) \cdot h_{x_i}(a_1, \ldots, a_{n-1}),$$

und nach dem Satz über implizite Funktionen ist $h_{x_i} = -g_{x_i}/g_{x_n}$, d.h.

$$F_{x_i}(a_1, \ldots, a_{n-1}) = f_{x_i}(\boldsymbol{a}) - f_{x_n}(\boldsymbol{a}) \frac{g_{x_i}(\boldsymbol{a})}{g_{x_n}(\boldsymbol{a})}.$$

Da die linke Seite verschwindet, gilt dasselbe auch für die rechte. Die rechte Seite ist im Gegensatz zur linken auch für $i = n$ definiert und verschwindet aus trivialen Gründen; also ist für alle i

$$f_{x_i}(\boldsymbol{a}) - \frac{f_{x_n}(\boldsymbol{a})}{g_{x_n}(\boldsymbol{a})} g_{x_i}(\boldsymbol{a}) = 0$$

oder, anders ausgedrückt,

$$\operatorname{grad} f(\boldsymbol{a}) - \frac{f_{x_n}(\boldsymbol{a})}{g_{x_n}(\boldsymbol{a})} \operatorname{grad} g(\boldsymbol{a}) = \vec{0}.$$

Damit sind die beiden Gradienten in der Tat linear abhängig. ∎

Falls der Gradient von g im Punkt $\boldsymbol{a}$ nicht verschwindet, gibt es somit eine Zahl $\lambda \in \mathbb{R}$, so daß

$$\operatorname{grad} f(\boldsymbol{a}) - \lambda \operatorname{grad} g(\boldsymbol{a}) = \vec{0}$$

ist, nämlich

$$\lambda = \frac{f_{x_n}(\boldsymbol{a})}{g_{x_n}(\boldsymbol{a})}.$$

Diese Zahl bezeichnet man als LAGRANGEschen Multiplikator; mit seiner inhaltlichen Interpretation werden wir uns in Kürze beschäftigen.

JOSEPH-LOUIS LAGRANGE (1736–1813) wurde als GIUSEPPE LODOVICO LAGRANGIA in Turin geboren und studierte dort zunächst Latein. Erst eine alte Arbeit von HALLEY über algebraische Methoden in der Optik weckte sein Interesse an der Mathematik, woraus ein ausgedehnter Briefwechsel mit EULER entstand. In einem Brief vom 12. August 1755 berichtete er diesem unter anderem über seine Methode zur Berechnung von Maxima und Minima; 1756 wurde er, auf EULERs Vorschlag, Mitglied der Berliner Akademie; zehn Jahre später zog er nach Berlin und wurde dort EULERs Nachfolger als mathemascher Direktor der Akademie. 1787 wechselte er an die Pariser Académie des Sciences, wo er bis zu seinem Tod blieb und unter anderem an der Einführung des metrischen Systems beteiligt war. Seine Arbeiten umspannen weite Teile der Analysis, Algebra und Geometrie.

Zur praktischen Bestimmung von Extremwerten unter Nebenbedingungen geht man wie folgt vor: Über die Punkte, in denen der Gradient von g verschwindet, macht obiger Satz keine verwertbare Aussage; diese Punkte müssen also vorab berechnet und untersucht werden.

Danach müssen die Punkte gefunden werden, in denen es ein $\lambda \in \mathbb{R}$ gibt, so daß

$$\begin{aligned} f_{x_1}(\boldsymbol{x}) - \lambda g_{x_1}(\boldsymbol{x}) &= 0 \\ &\vdots \\ f_{x_n}(\boldsymbol{x}) - \lambda g_{x_n}(\boldsymbol{x}) &= 0 \\ g(\boldsymbol{x}) &= 0 \end{aligned}$$

ist. Dies ist ein System von $n + 1$ Gleichungen für die $n + 1$ Unbekannten, allerdings ist dieses Gleichungssystem nur selten linear und damit oft nicht mit bekannten Methoden lösbar. Manchmal kann man das Gleichungssystem durch geeignete Umformungen und Fallunterscheidungen vollständig lösen, in anderen Fällen helfen nur die aus der Numerik bekannten Näherungsverfahren wie etwa die Methode von NEWTON-RAPHSON.

Falls alle Gleichungen Polynomgleichungen sind (oder durch Einführung geeigneter zusätzlicher Variablen auf Polynomgleichungen zurückgeführt werden können), kann man im Falle einer endlichen Lösungsmenge diese auch exakt bestimmen: Genau wie der GAUSS-Algorithmus zur Lösung eines linearen Gleichungssystems dieses auf eine

Treppengestalt bringt, aus der man die Lösungen einfach ermitteln kann, gibt es in der Computeralgebra einen Algorithmus, der dasselbe für beliebige Systeme von Polynomgleichungen versucht; die Gleichungen, die dieser Algorithmus liefert, bezeichnet man als GRÖBNER-Basis oder Standardbasis. Zum Verständnis dieses Algorithmus, den man als eine Art Synthese aus EUKLIDischen Algorithmus und GAUSS-Algorithmus ansehen kann, sind Kenntnisse der kommutativen Algebra erforderlich, für die die Zeit in dieser Vorlesung nicht ausreicht; bei einigen Implementierungen werden zusätzlich auch noch Algorithmen aus der Informatik eingesetzt, die typischerweise nicht in Grundvorlesungen behandelt werden. Deshalb sei hier nur darauf hingewiesen, daß die gängigen universellen Computeralgebrasysteme wie Maple, Mathematica, MuPad allesamt entsprechende Routinen enthalten, mit denen man auch dann experimentieren kann, wenn man die dahinterstehende Theorie nicht versteht.

Als Beispiel, wie gelegentlich auch ein nichtlineares Gleichungssystem elementar gelöst werden kann, betrachten wir eine Anwendung aus den Wirtschaftswissenschaften: Die Gesamtproduktion eines Unternehmens oder eines Staats in Abhängigkeit von n eingesetzten Ressourcen $x_1, \dots, x_n$ wird oft modelliert durch eine sogenannte COBB-DOUGLAS-Funktion der Form

$$P(x_1, \dots, x_n) = \alpha x_1^{e_1} \dots x_n^{e_n} ,$$

benannt nach den beiden Wissenschaftlern, die dieses Modell 1928 für die amerikanische Gesamtproduktion in Abhängigkeit von Kapital und Arbeit in den Jahren 1899 bis 1922 entwickelten. (Sie fanden $P \approx 1{,}01 A^{3/4} K^{1/4}$ mit A = Anzahl der Beschäftigten und K = Kapitaleinsatz.)

Betrachten wir stattdessen die Produktion eines Wirtschaftsguts aus zwei Resourcen x, y gemäß der Funktion

$$f(x, y) = P(x, y) = x^{1/2} y^{1/4} .$$

Falls wir der Einfachheit halber annehmen, daß die Kosten pro Einheit für x und y gleich sind und die Gesamtkosten höchstens gleich zwölf sein dürfen, müssen wir f maximieren unter der Nebenbedingung

$$x + y \leq 12 .$$

Nun ist aber f eine monoton wachsende Funktion sowohl von x als auch von y, d.h. die maximale Produktion wird sicherlich erreicht in einem Punkt, für den $x + y = 12$ ist, denn für jeden anderen Punkt

(x, y) mit $x + y < 12$ ist $f(x, y) < f(x, 12 - x)$. Daher können wir die Nebenbedingung in der gewohnten Form

$$g(x, y) = x + y - 12 = 0$$

schreiben. Diese Nebenbedingung sowie die zu maximierende Funktion sind in Abbildung 57 dargestellt.

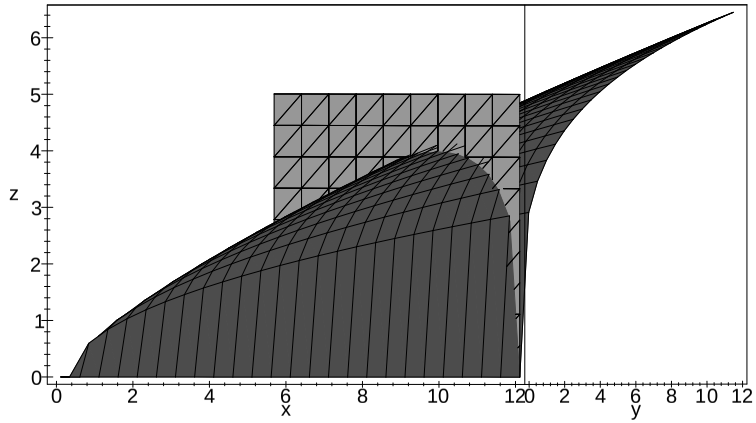


Abb. 57: Maximierung einer Produktionsfunktion bei festem Kapitaleinsatz

Ableitung beider Funktionen zeigt, daß

$$\operatorname{grad} g = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{und} \quad \operatorname{grad} f = \begin{pmatrix} y^{1/4}/2x^{1/2} \\ x^{1/2}/4y^{3/4} \end{pmatrix}$$

ist; das zu lösende Gleichungssystem wird also zu

$$\begin{aligned} \frac{y^{1/4}}{2x^{1/2}} - \lambda &= 0 \\ \frac{x^{1/2}}{4y^{3/4}} - \lambda &= 0 \\ x + y - 12 &= 0 \end{aligned} .$$

(Die Nenner brauchen uns nicht zu stören, denn da $f(0, y) = f(x, 0) = 0$ ist, kommen Lösungen mit $x = 0$ oder $y = 0$ für das Maximum ohnehin nicht in Frage; wir können sie also getrost ausschließen.)

Als Ansatz zu einer möglichen Lösung können wir ausnutzen, daß λ in den beiden ersten Gleichungen isoliert steht; wenn wir danach auflösen und gleichsetzen, erhalten wir die Gleichung

$$\frac{y^{1/4}}{2x^{1/2}} = \frac{x^{1/2}}{4y^{3/4}} .$$

Multiplikation mit dem Hauptnenner macht daraus

$$4y^{1/4}y^{3/4} = 2x^{1/2}x^{1/2} \quad \text{oder} \quad 2y = x .$$

Einsetzen in die dritte Gleichung ergibt $3y = 12$, also ist

$$y = 4 \quad \text{und} \quad x = 8 ;$$

der Maximalwert von f ist

$$f(8,4) = 8^{1/2} \cdot 4^{1/4} = 2\sqrt{2} \cdot \sqrt{2} = 4 .$$

Auch den LAGRANGEschen Multiplikator λ können wir noch ausrechnen:

$$\lambda = \frac{y^{1/4}}{2x^{1/2}} = \frac{4^{1/4}}{2 \cdot 8^{1/2}} = \frac{\sqrt{2}}{2 \cdot 2\sqrt{2}} = \frac{1}{4} .$$

Die Berechnung von λ war für die Bestimmung des Optimums eigentlich überflüssig; λ ist nur eine Hilfsgröße zur Berechnung des Extremums. Wir wollen uns als nächstes überlegen, daß wir λ auch inhaltlich interpretieren können: Dazu betrachten wir eine Nebenbedingung

$$g(x_1, \ldots, x_n) = c$$

mit *variabler* rechter Seite c und ein Extremum der Funktion

$$f(x_1, \ldots, x_n) .$$

Dieses Extremum wird natürlich von c abhängen; wir schreiben es in der Form

$$\big(x_1(c), \ldots, x_n(c)\big)$$

und nehmen an, daß die Funktionen $x_i(c)$ stetig differenzierbar seien. (Ein interessierter Leser kann sich anhand des Satzes über implizite Funktionen überlegen, welche Bedingungen f und g erfüllen müssen,

damit dies garantiert ist.) Der Optimalwert von f in Abhängigkeit von c ist dann

$$F(c) \underset{\text{def}}{=} f\big(x_1(c), \ldots, x_n(c)\big) .$$

Nach der Kettenregel aus **[HM I]**, *Kapitel 2,* §3*c)* ist

$$F'(c) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dx_i(c)}{dc} .$$

Genauso können wir

$$G(c) \underset{\text{def}}{=} g\big(x_1(c), \ldots, x_n(c)\big)$$

betrachten und erhalten

$$G'(c) = \sum_{i=1}^{n} \frac{\partial g}{\partial x_i} \frac{dx_i(c)}{dc} .$$

Da $\big(x_1(c), \ldots, x_n(c)\big)$ ein Optimum ist, sind dort die Gradienten von f und $g - c$ proportional mit Proportionalitätsfaktor λ. Da wir bei der Gradientenbildung nur nach den x_i ableiten, von denen die rechte Seite c nicht abhängt, ist der Gradient von $g - c$ gleich dem von g selbst, d.h.

$$\frac{\partial f}{\partial x_i} = \lambda \frac{\partial g}{\partial x_i} \qquad \text{für alle } i .$$

Somit ist $F'(c) = \lambda G'(c)$. Da der Punkt $\big(x_1(c), \ldots, x_n(c)\big)$ die Nebenbedingung mit rechter Seite c erfüllt, ist aber $G(c) = c$ und damit $G'(c) \equiv 1$. Also ist $\lambda = F'(c)$ die Wachstumsrate für das Optimum bei Änderung der rechten Seite der Nebenbedingung.

Im obigen Beispiel steigt also die Maximalmenge $f(x, y)$, die mit Kapitaleinsatz 12 produziert werden kann, für kleines h ungefähr um $h/4$, wenn wir den Kapitaleinsatz auf $12 + h$ erhöhen. Die Erhöhung des Kapitaleinsatzes lohnt sich, wenn für das fertige Produkt ein Preis pro Einheit erzielt werden kann, der größer ist als vier.

Als letztes wollen wir uns noch überlegen, was passiert, wenn wir nicht nur eine, sondern mehrere Nebenbedingungen erfüllen müssen. Es geht

also wieder darum, eine Funktion $f(x_1, \ldots, x_n)$ zu optimieren, jetzt aber unter den Nebenbedingungen

$$g_1(x_1, \ldots, x_n) \geq 0, \quad \ldots \quad g_r(x_1, \ldots, x_n) \geq 0\,.$$

(Es genügt, Bedingungen mit $\geq$ zu betrachten, denn durch Multiplikation mit minus Eins kann man jede Ungleichung mit $\leq$ in eine mit $\geq$ überführen. Auch Gleichungen $g_i = 0$ kann man zumindest formal durch die beiden Ungleichungen $g_i \geq 0$ und $-g_i \geq 0$ ausdrücken.)

Die wichtigsten Beispiele solcher Optimierungsaufgaben sind die Fälle mit linearen Funktionen f und g_i; hier redet man von *linearen Programmen*. (Das Wort *Programm* in diesem Zusammenhang hat natürlich nichts mit Computerprogrammen zu tun.) Das wichtigste Verfahren zur Lösung solcher Aufgaben, der Simplex-Algorithmus, wird in der Vorlesung *Numerik I* behandelt, so daß wir uns hier auf die *nichtlineare Programmierung* beschränken können.

Man überlegt sich leicht, daß im linearen Fall die Nebenbedingungen ein (endliches oder unendliches) Polyeder im $\mathbb{R}^n$ definieren und eine lineare Funktion, so sie ein endliches Maximum oder Minumum hat, dieses auf dem Rand dieses Polyeders annimmt, und dort sogar in einer Ecke. Man muß daher „nur“ die Ecken dieses Polyeders untersuchen – deren Anzahl allerdings wächst exponentiell mit der Anzahl der Variablen. Trotzdem führt der Simplex-Algorithmus selbst im Fall von Zehntausenden von Variablen in der Regel fast immer sehr schnell ans Ziel; das theoretische Problem der exponentiellen Komplexität im schlimmsten Fall hat also für praktische Anwendungen keine Bedeutung.

Bei nichtlinearen Funktionen ist die Situation komplizierter, denn nun kann es auch im Innern Extrema geben: Die Funktion

$$f(x, y) = e^{-x^2-y^2} \quad \text{mit der Nebenbedingung} \quad x^2 + y^2 \leq 1$$

etwa nimmt ihr Maximum im Punkt $(0, 0)$ an; auf dem Rand des Einheitskreises liegen nur die Minima. Im allgemeinen Fall eines nichtlinearen Programms kann ein Optimum also entweder ganz im Innern liegen oder aber eine beliebige Teilmenge der Nebenbedingungen exakt erfüllen.

Falls wir es mit inneren Punkte zu tun haben, sind diese lokale Maxima oder Minima ohne Nebenbedingungen, und wir haben uns bereits in §1

überlegt, wie man diese bestimmt: In jedem solchen Punkt verschwindet der Gradient der zu optimierenden Funktion.

Im Falle einer einzigen *Gleichung* als Nebenbedingung ist der Gradient von f linear abhängig vom Gradienten der Nebenbedingung; da der Nullvektor von jedem anderen Vektor linear abhängig ist, schließt dies auch den Fall der Optima bei inneren Punkten mit ein. Die naheliegende Verallgemeinerung auf den Fall mehrerer Nebenbedingungen ist der

Satz: Die Funktion $f\colon D \to \mathbb{R}$ auf $D \subseteq \mathbb{R}^n$ habe im Punkt $\boldsymbol{a} \in D$ ein Extremum unter den Nebenbedingungen

$$g_1(\boldsymbol{a}) \geq 0, \quad g_2(\boldsymbol{a}) \geq 0, \quad \ldots, \quad g_r(\boldsymbol{a}) \geq 0\,.$$

Dann sind die $r + 1$ Vektoren

$$\operatorname{grad} f(\boldsymbol{a}), \quad \operatorname{grad} g_1(\boldsymbol{a}), \quad \operatorname{grad} g_2(\boldsymbol{a}), \quad \ldots, \quad \operatorname{grad} g_r(\boldsymbol{a})$$

linear abhängig.

Der *Beweis* erfordert keine wesentlich neuen Ideen gegenüber dem Fall einer einzigen Nebenbedingung und sei daher nur kurz skizziert: Falls die Gradienten der g_i im Punkt $\boldsymbol{a}$ bereits untereinander linear abhängig sind, gibt es nichts mehr zu beweisen; nehmen wir also an, sie seien linear unabhängig. Dann gibt es (mindestens) r verschiedene Variablen x_{j_1} bis x_{j_r}, so daß

$$\frac{\partial g_i}{\partial x_{j_i}}(\boldsymbol{a}) \neq 0$$

ist. Also kann nach dem Satz über implizite Funktionen jede Nebenbedingung zur Elimination einer anderen Variablen benutzt werden, und im wesentlichen dieselbe Rechnung wie im Fall einer Nebenbedingung zeigt die Behauptung. ∎

Die lineare Abhängigkeit der Vektoren

$$\operatorname{grad} f(\boldsymbol{a}), \quad \operatorname{grad} g_1(\boldsymbol{a}), \quad \operatorname{grad} g_2(\boldsymbol{a}), \quad \ldots, \quad \operatorname{grad} g_r(\boldsymbol{a})$$

bezeichnet man als KUHN-TUCKER-Bedingung; sie ist eine offensichtliche Verallgemeinerung der Bedingung von LAGRANGE, ist allerdings deutlich jünger: Sie erschien 1951 in einer gemeinsamen Arbeit von

H.W. KUHN und A.W. TUCKER, vier Jahre, nachdem G. DANTZIG den Simplex-Algorithmus entwickelt hatte, und fast zweihundert Jahre, nachdem LAGRANGE seine Multiplikatoren zur Bestimmung von Extrema unter einer Nebenbedingung eingeführt hatte.

Das Problem bei der praktischen Anwendung des Satzes von KUHN und TUCKER besteht darin, daß in einem Optimum manche Nebenbedingungen als Gleichungen, andere als echte Ungleichungen erfüllt sind; man muß also jede der möglichen Kombinationen untersuchen.

Eine mögliche Abhilfe sind sogenannte *barrier*-Methoden: Man läßt die Nebenbedingungen eine Barriere errichten, indem man (bei der Suche nach einem Maximum) Maxima *ohne* Nebenbedingung der Funktion

$$f(x_1,\ldots,x_n) + \sum_{i=1}^{r} \varepsilon_i \log g_i(x_1,\ldots,x_n)$$

sucht, wobei die ε_i positive Konstanten sind. Da die Logarithmen am Rand gegen $-\infty$ gehen, liegen diese Maxima stets im Innern. Falls man nun alle ε_i in geeigneter Weise gegen Null gehen läßt, kann man in manchen Fällen zeigen, daß diese Maxima gegen Maxima der Funktion *mit* Nebenbedingung konvergiert.

Ein Beispiel dafür ist der 1984 gefundene Algorithmus von KARMAKAR für den Fall linearer Funktionen f, g_i. Er ist eine Alternative zum Simplex-Algorithmus, die stets in polynomialer Zeit zu einer Lösung führt, und war der erste mathematische Algorithmus, der patentiert wurde. In der Praxis ist er jedoch bei fast allen Problemen dem Simplex-Algorithmus unterlegen; lediglich bei einigen wenigen Spezialfällen, bei denen bekannt ist, daß der Simplex-Algorithmus schlecht funktioniert, führt KARMAKAR schneller zu einer Lösung.

§3: Numerische Verfahren

Wie wir gesehen haben, führt die Methode der LAGRANGEschen Multiplikatoren im allgemeinen auf nichtlineare Gleichungssysteme, die nur in einfachen Fällen explizit lösbar sind. In allen anderen Fällen muß man mit numerischen Methoden arbeiten, und da bietet sich an, das Problem

von vornherein ohne den Umweg über LAGRANGEsche Multiplikatoren Extrema numerisch zu bearbeiten.

a) Die Gradientenmethode

Für eine differenzierbare Funktion f auf $D \subseteq \mathbb{R}^n$ ist

$$f(\boldsymbol{x} + \vec{h}) = f(\boldsymbol{x}) + \operatorname{grad} f(\boldsymbol{x}) \cdot \vec{h} + o\big(\big|\vec{h}\big|\big) \,;$$

wenn wir ein Maximum (oder Minimum) von f ansteuern wollen, liegt es daher nahe, $\vec{h}$ so zu wählen, daß sich der Funktionswert möglichst stark vergrößert (oder verkleinert).

Nach der CAUCHY-SCHWARZschen Ungleichung ist

$$\Big|\operatorname{grad} f(\boldsymbol{x}) \cdot \vec{h}\Big| \le \|\operatorname{grad} f(\boldsymbol{x})\| \cdot \Big\|\vec{h}\Big\| \,;$$

wir erhalten also die maximalmögliche Veränderung bei vorgegebener Länge von $\vec{h}$ genau dann, wenn $\vec{h}$ parallel zum Gradienten ist.

Damit bietet sich folgende Strategie an: Wir wählen irgendeinen Ausgangspunkt $\boldsymbol{x}_0$ und berechnen dort den Gradienten $\nabla f(\boldsymbol{x}_0)$. Weiter geben uns eine Länge ℓ_0 für den Vektor $\vec{h}$ vor, die von der Länge des Gradienten abhängen kann oder auch nicht. Dann setzen wir bei der Suche nach einem Maximum

$$\vec{h}_0 = \frac{\ell_0}{\|\nabla f(\boldsymbol{x}_0)\|} \nabla f(\boldsymbol{x}_0) \,;$$

bei der Suche nach Minima nehmen wir das Negative davon.

Als nächstes betrachten wir den Punkt

$$\boldsymbol{x}_1 \underset{\text{def}}{=} \boldsymbol{x}_0 + \vec{h}_0 \,,$$

berechnen dort den Gradienten $\nabla f(\boldsymbol{x}_0)$, setzen mit geeignetem ℓ_1

$$\vec{h}_1 = \pm \frac{\ell_1}{\|\nabla f(\boldsymbol{x}_1)\|} \nabla f(\boldsymbol{x}_1)$$

(+ für Maxima, − für Minima) zur Definition des nächsten Punkts

$$\boldsymbol{x}_2 \underset{\text{def}}{=} \boldsymbol{x}_1 + \vec{h}_1$$

und so weiter. In jedem Schritt erhöhen (oder erniedrigen) wir den Funktionswert soweit wie es mit der vorgegebenen Länge ℓ_i nur möglich ist in der Hoffnung, so irgendwann auf ein Maximum (oder Minimum) zu stoßen. Dieses können wir erreichen, wenn wir am Rand des Definitionsbereichs von f angelangt sind, oder aber wenn wir in einem Punkt sind, in dem der Gradient verschwindet: Von dort aus geht es mit diesem Verfahren nicht mehr weiter.

Da wir mit einem numerischen Verfahren nur ein verschwindend geringe Chance haben, exakt in einem Extremum zu enden, zeigt sich hier auch die Notwendigkeit einer intelligenten Wahl der Schrittweiten ℓ_i: Wenn diese zu groß sind, kann es passieren, daß wir endlos um ein Extremum herumoszillieren.

Theoretisch ist auch möglich, daß wir in einem Sattelpunkt landen, aber wenn man sich überlegt, wie die Gradienten in der Umgebung eines Sattelpunktes aussehen, wird schnell klar, daß dies nur sehr selten passiert.

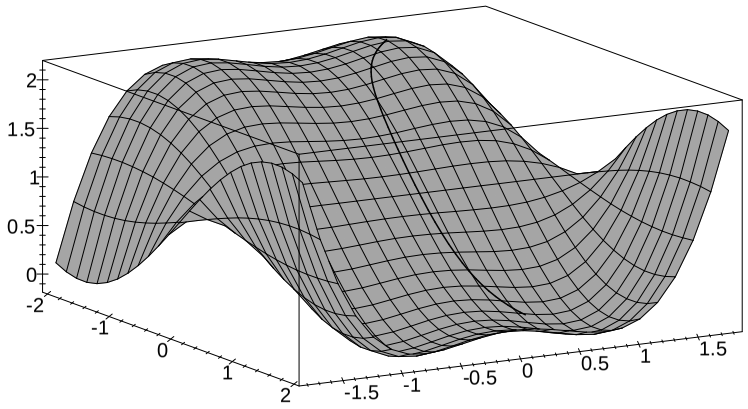


Abb. 58: Eine Anwendung der Gradientenmethode

Abbildung 58 zeigt ein einfaches Beispiel für einen mit der Gradientenmethode zurückgelegten Weg; hier wurde in jedem Schritt

$$\begin{pmatrix} h_i \\ k_i \end{pmatrix} = 0{,}1 \cdot \nabla f(x_i, y_i)$$

gesetzt. Der Weg geht offensichtlich recht zielstrebig auf das Maximum zu.

Abbildung 59 zeigt dasselbe Bild in einen etwas größeren Zusammenhang; hier sehen wir, daß unser Streben nach kurzfristigen Gewinnen langfristig wohl doch nicht so erfolgreich war: Wenn wir vom Startpunkt aus nach rechts in die kleine Mulde abgestiegen wären, hätten wir auf dem gegenüberliegenden Hang deutlich größere Funktionswerte erreicht als im lokalen Maximum, in dem wir schließlich gelandet sind.

Dies ist ein grundsätzliches Problem von Gradientenverfahren: Falls man sie in der Nähe des (absoluten) Optimums starten läßt, führen sie schnell und zuverlässig ans Ziel, ansonsten aber ist die Gefahr sehr groß, daß man in einem nur lokalen Optimum steckenbleibt.

Um von dort wieder weiterzukommen, gibt es verschiedene Strategien. Eine anschaulich recht klare ist die sogenannte „Tunnelung“. Der Name entstand aus der Betrachtung von Minimisierungsproblemen; nehmen wir also an, wir wollen das Minimum der Funktion $f(x, y)$ in einem gewissen Bereich finden und ein Gradientenverfahren hat uns in einen Punkt $\boldsymbol{x}_M$ geführt, von dem aus es nicht mehr weiterkommt. Um zu sehen, ob $z_M = f(\boldsymbol{x}_M)$ wirklich der kleinste Wert ist, den f im betrachteten Bereich annehmen kann, versuchen wir, eine weitere Lösung der Gleichung

$$f(\boldsymbol{x}) = z_M$$

zu finden. Dafür gibt es eine ganze Reihe numerischer Verfahren, z.B. das Verfahren von NEWTON-RAPHSON, mit denen sich zumindest ein solcher Punkt leicht finden läßt. Leider könnte dieser Punkt unser Ausgangspunkt $\boldsymbol{x}_M$ sein; deshalb sucht man tatsächlich nicht nach Lösungen der Gleichung $f(\boldsymbol{x}) = z_M$, sondern nach Lösungen einer leicht abgewandelten Gleichung der Form

$$\widetilde{f}(\boldsymbol{x}) = z_M \,,$$

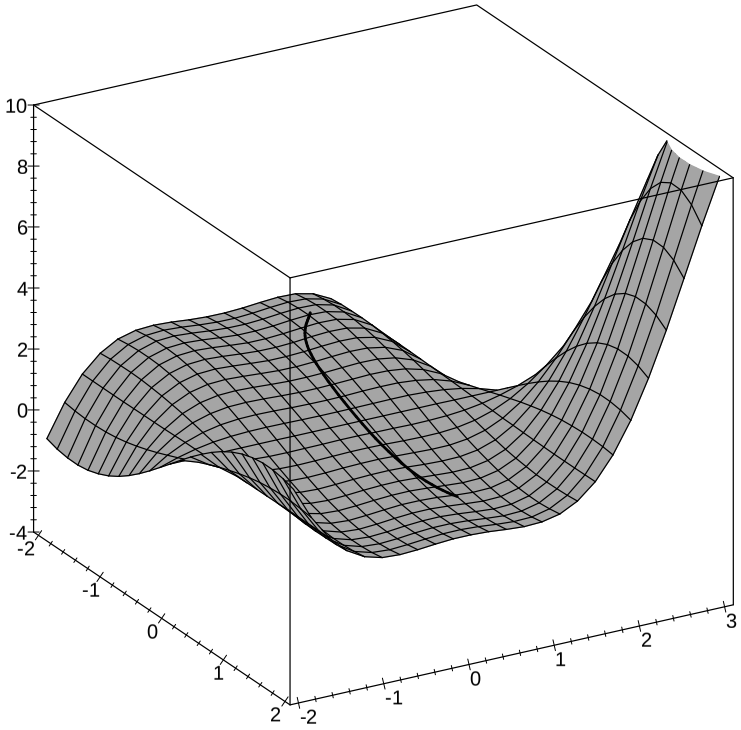


Abb. 59: Der Weg aus Abb. 58 aus einem weiteren Blickwinkel

wobei $\widetilde{f}$ dadurch aus f entsteht, daß man die Funktionswerte in der unmittelbaren Umgebung von (x_M, y_M) starkt anhebt, so daß das dortige Minimum verschwindet. Dazu kann man beispielsweise eine Funktion der Form

$$G(x,y) = ae^{\frac{(x-x_M)^2+(y-y_M)^2}{b}}$$

mit geeigneten Parametern a, b wählen, wie sie in Abbildung 60 zu sehen ist, und

$$\widetilde{f}(x,y) = f(x,y) + G(x,y)$$

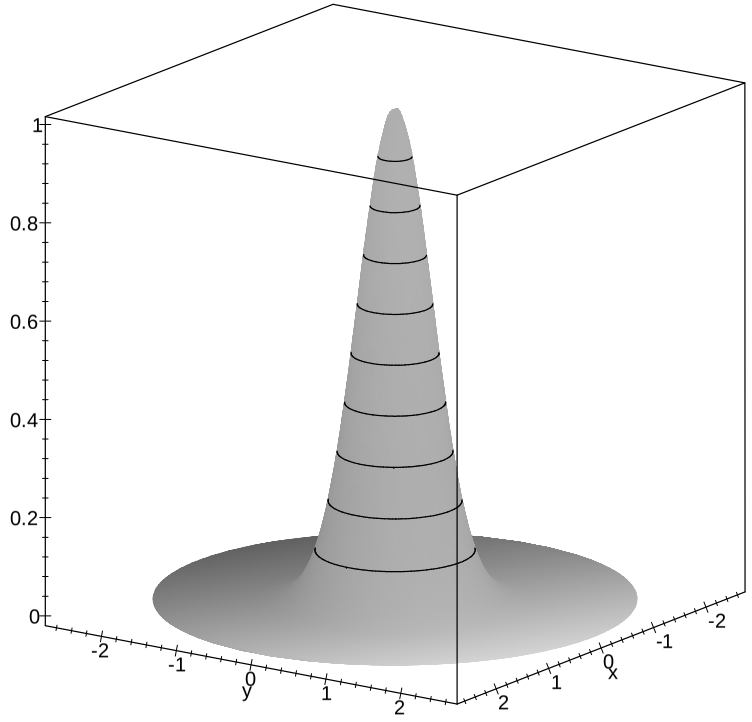


Abb. 60: $G(x,y) = e^{-3(x^2+y^2)}$

setzen.

Dies bringt das Minimum im Punkt $\boldsymbol{x}_M$ zum Verschwinden und verändert die Funktion praktisch nicht, wenn man nur hinreichend weit entfernt ist von $\boldsymbol{x}_M$. (Je kleiner b ist, umso lokalisierter ist die Veränderung.) Eine Lösung der Gleichung

$$\widetilde{f}(\boldsymbol{x}) = z_M\,,$$

so es eine gibt, liegt also nicht in der unmittelbaren Umgebung von $\boldsymbol{x}_M$ und ist daher ein guter Ausgangspunkt, um dort die Gradientenmethode

noch einmal zu starten bis zum nächsten lokalen Minimum und so weiter. Sobald die Gleichung nicht mehr lösbar ist, können wir ziemlich sicher sein, daß z_M das globale Minimum ist – es sei denn, wir hätten die Parameter a und b sehr dumm gewählt.

Tunnelung ist auch ein wichtiges Konzept in der Physik: Dort versucht ein System bekanntlich stets, sein Energieminimum zu erreichen. Dies kann jedoch daran scheitern, daß es sich in einem lokalen Minimum befindet und nicht genügend Energie aufbringen kann, um den Energiewall zu überwinden, der es vom absoluten Minimum trennt. Zumindest im Bereich der Quantentheorie gibt es dann auch den sogenannten *Tunneleffekt*, der es einzelnen Teilchen erlaubt, diesen Wall zu tunneln und auf diese Weise einen Zustand niedrigerer Energie zu erreichen.

Im obigen Beispiel geht es nicht um ein Minimum, sondern um ein Maximum, da die Suche danach graphisch besser darstellbar ist. Also graben wir auch keinen Tunnel, sondern spannen ein Hochseil, das irgendwo auf der eingezeichneten Ebenen liegt und uns vom erreichten Zwischenhoch zur Startposition für einen weiteren Anstieg bringt. (Tatsächlich ist die Ebene etwas zu tief eingezeichnet, damit man das alte Maximum noch erkennen kann; das Seil muß also etwas höher hängen.)

b) Der Metropolis-Algorithmus

Eine weitere Idee zur Vermeidung von Zwischenhochs hat ebenfalls viel mit Physik zu tun: Ein Gas erreicht seinen Zustand minimaler Energie dann, wenn die Bewegungsenergie $\frac{1}{2}mv^2$ eines jeden Teilchens gleich null ist, wenn sich also nichts mehr bewegt. Dies geschieht aber höchstens am absoluten Nullpunkt; bei positiven Temperaturen werden die meisten Teilchen positive kinetische Energie haben. Nach LUDWIG BOLTZMANN ist dabei die Wahrscheinlichkeit dafür, daß ein Teilchen die Energie $E = \frac{1}{2}mv^2$ hat, bei Temperatur T proportional zu

$$e^{-\frac{E}{kT}},$$

mit einer Konstanten $k \approx 1{,}38066 \cdot 10^{-23} J/K$, die heute als BOLTZMANN-Konstante bezeichnet wird.

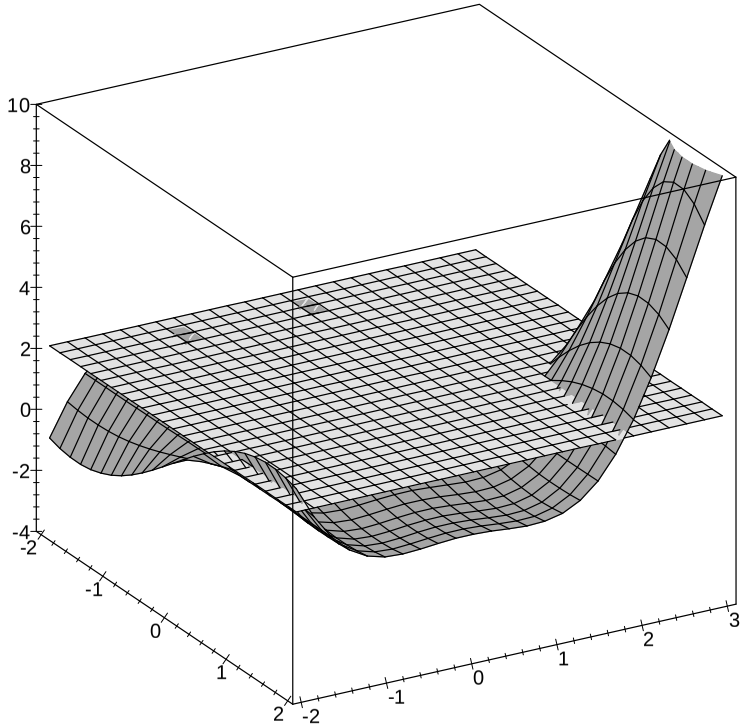


Abb. 61: „Tunnelung“ für Maxima

LUDWIG BOLTZMANN (1844–1906) wuchs auf und studierte in Wien; danach lehrte er in Graz, Heidelberg, Berlin, Graz, Wien, Graz, Wien, Leipzig und Wien. Er war Professor für Theoretische Physik, für Mathematik und für Experimentalphysik. Auf seiner letzten Stelle in Wien hielt er eine so erfolgreiche Philosophievorlesung, daß ihn Kaiser Franz Josef in den Palast einlud. Am bekanntesten ist er für die Begründung der statistischen Mechanik, einer damals sehr umstrittene Theorie. Ob die damit verbundenen Anfeindungen zu seinem Selbstmord führten, ist unbekannt.

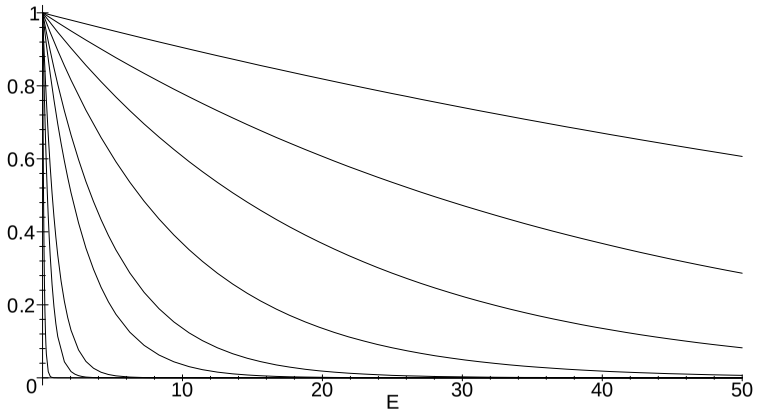


Abb. 62: $e^{-E/kT}$ für $kT = 0{,}1, 0{,}5, 1, 3, 5, 10, 20, 40, 100$

Bei der *simulierten Abkühlung* oder BOLTZMANN-*Maschine* ahmt man dies nach, indem man mit einer hohen Temperatur startet und der Richtung, in der man weitergeht, einer dieser Temperatur entsprechende Freiheit läßt. Man geht also nicht mehr unbedingt in Richtung des Gradienten, sondern geht zufällig in eine von endlich vielen vorgegebenen Richtungen. Die Wahrscheinlichkeit für den Richtungsvektor $\vec{h}_j$ soll dabei analog zur BOLTZMANN-Verteilung festgelegt werden, d.h. wir ordnen ihm eine „Energie“

$$E_j = \pm\big(f(\boldsymbol{x} + \vec{h}_j) - f(x, y)\big)$$

zu (positiv bei der Suche nach einem Minimum, negativ bei der Suche nach einem Maximum) und die Wahrscheinlichkeit dafür, daß wir in Richtung $\vec{h}_j$ gehen, soll proportional sein zu $e^{-E_j/kT}$. Sie ist also, falls N Richtungen zur Verfügung stehen, gleich

$$p_j \underset{\text{def}}{=} \frac{e^{-E_j/kT}}{\sum_{\ell=1}^{N} e^{-E_\ell/kT}}.$$

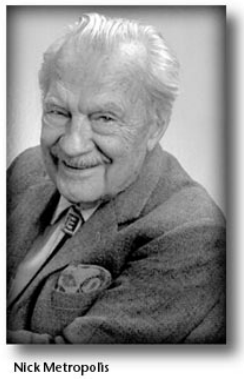
Zur Wahl einer Richtung erzeugen wir uns somit eine Zufallszahl Z zwischen null und eins und gehen in Richtung $\vec{h}_j$, wenn

$$\sum_{\ell=1}^{j-1} p_\ell < Z \le \sum_{\ell=1}^{j} p_\ell$$

ist. (Die Frage, wie lang die Richtungsvektoren im wievielten Schritt sein sollen, wollen wir hier ausklammern.)

Bie hohen Temperaturen ist damit die Richtung fast vollständig zufallsbedingt gewählt, während in der Nähe des absoluten Nullpunkts praktisch nur noch die optimale Richtung eine Chance hat. Falls wir bei hoher Temperatur in einem Zwischenextremum landen, sorgt dies also mit recht hoher Wahrscheinlichkeit dafür, daß wir dort nicht steckenbleiben.

Am Ende wollen wir allerdings im absoluten Optimum steckenbleiben, d.h. wir müssen die Temperatur im Verlauf der Rechnung immer weiter senken – daher der Name *simulated annealing* = simulierte Abkühlung. Bei der Anwendung auf Optimierungsprobleme bezeichnet man diese Vorgehensweise als den METROPOLIS-Algorithmus. In welcher Weise man die Temperatur am besten senkt, ist immer noch ein Gebiet aktiver Forschung. Man kann zeigen, daß man statistisch betrachtet praktisch immer im Optimum landet, wenn man mit einer hinreichend hohen Ausgangstemperatur T_1 startet und im r-ten Schritt mit Temperatur $T_1/\log(r+1)$ arbeitet, aber bei einer derart langsamen Abkühlung braucht der Algorithmus viel zu lange, um ans Ziel zu kommen.


Nick Metropolis

NICHOLAS METROPOLIS (1915–1999) wuchs auf in Chicago, wo er Physik studierte und 1941 promovierte. Seit 1943 arbeitete er, unterbrochen durch Professuren an der Universität Chicago von 1945–1948 und 1957–1965, in den Los Alamos Labaratories, die ihn im Nachruf als *giant of mathematics and one of the founders of the Information Age* bezeichneten. Sein Ruhm als Mathematiker beruht vor allem auf den von ihm entwickelten Anwendungen statistischer Verfahren auf eine Vielzahl von mathematischen Problemen; zum Pionier des Informationszeitalter macht ihn u. a., daß er einer der ersten Anwender des ersten elektronischen Computers ENIAC war, dessen Nachfolger MANIAC baute und an der Universität Chicago das Institute for Computer Research gründete und bis 1965 leitete.

In Abbildung 63 sieht man, wie sich der Algorithmus bei einer Abkühlungsregel verhält, die im r-ten Schritt mit Temperatur T_1/r arbeitet: Zumindest im gezeigten Fall funktioniert das recht gut.

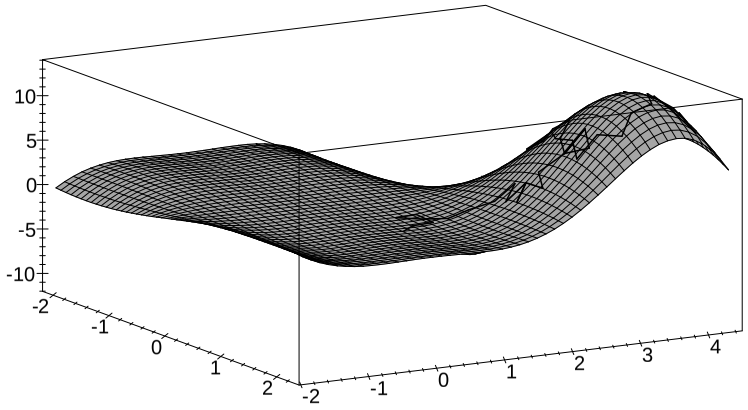


Abb. 63: Der METROPOLIS-Algorithmus für obiges Problem

In anderen Fällen (d.h., wenn andere Zufallszahlen gezogen werden) bleibt man damit aber auch gelegentlich ziemlich lange im Tal hängen; ein Beipspiel dafür zeigt Abbildung 64.

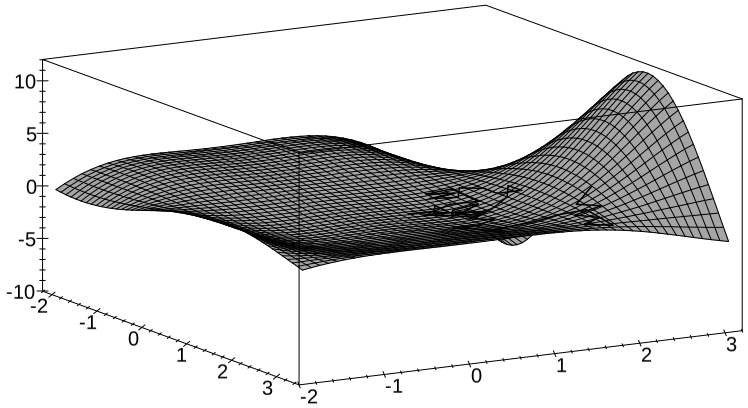


Abb. 64: Ditto mit anderen Zufallszahlen

Auch hier kommt man aber immerhin in eine gute Startposition, und oft wird es am besten sein, nach hinreichend vielen METROPOLIS-Schritten

einfach ein gewöhnliches Gradientenverfahren zu starten.

Zusammenfassend läßt sich sagen, daß der METROPOLIS-Algorithmus und verwandte Verfahren (die sogenannten Monte-Carlo-Methoden) sehr nützliche Hilfsmittel zur Optimierung sind, falls man so gut wie nichts über die zu optimierende Funktion weiß. Sie funktionieren nicht nur bei kontinuierlichen Problemen, wie den hier betrachteten, sondern auch für diskrete und kombinatorische Optimierungsprobleme.

Sie haben allerdings den Nachteil, daß man nie garantieren kann, daß man ein Optimum erreichen wird, und selbst wenn man eines erreicht, kann die Methode dies nicht erkennen. (Es gibt alternative numerische Methoden, die das können.)

c) Zusammenfassung

Die nichtlineare Optimierung ist ein sehr weites Feld, von dem eine Grundvorlesung wie die *Höhere Mathematik* nur einen kleinen Ausschnitt behandeln kann. Dieser Ausschnitt besteht nicht aus den für die Praxis wichtigsten Verfahren, sondern aus denen, die sich am besten in den Stoff der Vorlesung einordnen. Sie sind zwar (in Kombination mit dem aus der *Numerik* bekannten Simplex-Verfahren) die Grundbausteine, aus denen die meisten praktisch relevanten Verfahren zusammengesetzt sind, aber für die vielen kleinen Abwandlungen, die dazu führen, daß man ein Problem wirklich effizient lösen kann, müßte man deutlich mehr Zeit aufwenden, als hier zur Verfügung steht. Interessenten seien auf entsprechende Spezialvorlesung aus dem Bereich der Mathematik oder Operations Research verwiesen.

§4: Grundzüge der Fehler- und Ausgleichsrechnung

Physikalische Gesetze machen meist nur dann eine Aussage über ein reales System, wenn alle Umgebungsbedingungen exakt kontrolliert werden können. Das ist in der Praxis natürlich nie möglich. Insbesondere hat man bei der Anwendung physikalischer Prinzipen zur Messung von Daten keine Chance, den exakten Wert der zu messenden Größe

zu bestimmen; der gemessene Wert wird immer von zahlreichen kleineren Störungen beeinflußt sein, die man bei einem gut durchgeführten Experiment für alle praktischen Zwecke als zufällig betrachten kann.

Zusätzlich kann die Messung noch durch mehr oder weniger große *systematische* Fehler verfälscht sein; diese können hervorgerufen werden durch ein falsch kalibriertes Meßgerät, Ablesen auf der falschen Skala eines Meßinstruments, durch falsche Anwendung von Meßvorschriften *usw.* Mit diesen systematischen Fehlern wollen wir uns hier nicht beschäftigen; in diesem Paragraphen soll es nur um *Zufallsfehler* gehen.

a) Das Laplacesche Fehlermodell

Der französische Mathematiker PIERRE SIMON, MARQUIS DE LAPLACE (1749–1827), dem wir in dieser Vorlesung bereits mehrfach begegnet sind, entwickelte ein extrem vereinfachtes Modell für das Zustandkommen zufälliger Meßfehler. Trotz seiner unrealistischen Annahmen ist es als Einstieg in die Fehlerrechnung noch immer interessant, denn wie wir bald sehen werden, gelten seine Schlußfolgerungen viel allgemeiner und zumindest näherungsweise auch in vielen praktischen Situationen.

Die Grundannahme des LAPLACEschen Fehlermodells können wir uns so vorstellen, daß eine große Anzahl von „Dämonen“ (oder Fehlerquellen) unsere Meßergebnisse verfälschen; jeder einzelne dieser „Dämonen“ verursacht einen Fehler derselben Größe ε in positiver oder negativer Richtung, wobei die Wahrscheinlichkeit für $+\varepsilon$ *bzw.* $-\varepsilon$ für jeden der „Dämonen“ jeweils 50% sein soll und die einzelnen „Dämonen“ unabhängig voneinander handeln sollen.

Im Falle eines einzigen „Dämonen“ wäre der Fehler also mit gleicher Wahrscheinlichkeit $+\varepsilon$ oder $-\varepsilon$, bei zwei „Dämonen“ wäre er in jeweils 25% aller Fälle $+2\varepsilon$ oder -2ε, während sich in 50% der Fälle die beiden Fehler aufheben würden.

Allgemein gibt es bei n „Dämonen“ 2^n gleichwahrscheinliche Möglichkeiten für deren Verhalten; die folgende Tabelle zeigt für $n \le 5$ jeweils die Anzahl der Fälle, die zu dem in der Kopfzeile angegebenen Gesamtfehler führen:

	-5ε	-4ε	-3ε	-2ε	$-\varepsilon$	0	$+\varepsilon$	$+2\varepsilon$	$+3\varepsilon$	$+4\varepsilon$	$+5\varepsilon$
$n=0$						1					
$n=1$					1		1				
$n=2$				1		2		1			
$n=3$			1		3		3		1		
$n=4$		1		4		6		4		1	
$n=5$	1		5		10		10		5		1

Diese dreiecksförmige Anordnung von Zahlen bezeichnet man als PASCAL*sches Dreieck*. Offenbar kann man es dadurch rekursiv zeilenweise berechnen, daß man an jede Stelle die Summe der beiden links und rechts darüberstehenden Zahlen schreibt: Die n-te Störung bringt den Fehler genau dann auf $i \cdot \varepsilon$, wenn sie entweder gleich $+\varepsilon$ ist und die ersten $n-1$ Störungen einen Fehler $(i-1)\cdot\varepsilon$ produziert haben, oder aber wenn sie gleich $-\varepsilon$ ist und die ersten $n-1$ Störungen einen Fehler $(i+1)\cdot\varepsilon$ produziert haben. Entsprechend ist auch klar, daß die Summe aller Zahlen in der n-ten Zeile gleich 2^n ist, denn in der nullten Zeile haben wir Summe eins, und da die jeweils neu hinzukommende Störung genau zwei Möglichkeiten hat, verdoppelt sich die Summe von Zeile zu Zeile. Die Wahrscheinlichkeit dafür, daß sich n Störungen zu $i\cdot\epsilon$ aufsummieren, ist also gerade gleich der Zahl, die in der n-ten Spalte unter $i\cdot\varepsilon$ steht (beziehungsweise Null, wenn dort keine Zahl steht), dividiert durch 2^n.

Bekanntlich kann man die Zahlen in diesem Dreieck auch explizit berechnen: An der n-ten Zeile stehen die $n+1$ Zahlen

$$\binom{n}{i} = \frac{n!}{i!(n-i)!} \quad \text{für} \quad i = 0, \dots, n\,.$$

Wer diese Formel nicht kennt, kann sie leicht durch vollständige Induktion beweisen: Für $n=1$ sowie allgemein für $i=0$ oder $i=n$ ist alles klar; für $n>1$ und $0<i<n$ stehen über $\binom{n}{i}$ die beiden Zahlen $\binom{n-1}{i-1}$ und $\binom{n-1}{i}$, für die in der Tat gilt

$$\binom{n-1}{i-1} + \binom{n-1}{i} = \frac{(n-1)!}{(i-1)!(n-i)!} + \frac{(n-1)!}{i!(n-i-1)!} = \frac{(n-1)!}{i!(n-i)!}\cdot\Big(i+(n-i)\Big)$$
$$= \frac{n!}{i!(n-i)!} = \binom{n}{i}\,.$$

Betrachten wir als nächstes die Größe des Gesamtfehlers. Falls n gerade ist, treten nur Vielfache von 2ε auf und alle diese Vielfachen zwischen $-n\varepsilon$ und $n\varepsilon$ kommen tatsächlich vor; ensprechend sind für ungerades n nur ungeradzahlige Vielfache von ε möglich, und auch hier werden wieder alle solchen Werte zwischen $-n\varepsilon$ und $n\varepsilon$ angenommen. Wir können dies dadurch zusammenfassen, daß in beiden Fällen genau die Werte $(n-2k)\varepsilon$ mit $k = 0, \ldots, n$ angenommen werden, und das PASCALsche Dreieck zeigt, daß der Fehler $(n-2k)\varepsilon$ in

$$\binom{n}{n-k} = \binom{n}{k}$$

Fällen auftritt. Da n „Dämonen“ insgesamt 2^n Möglichkeiten zur Fehlererzeugung haben, ist die Wahrscheinlichkeit für den Gesamtfehler $(n-2k)\varepsilon$ also

$$\binom{n}{k} \cdot 2^{-n}\,.$$

Diese Wahrscheinlichkeit sollte für einen festen Fehlerbetrag im wesentlichen unabhängig von n sein: Da wir nicht wirklich an Dämonen glauben, können wir deren Anzahl schließlich nicht in ein realistisches Fehlermodell einfließen lassen.

Der Formel können wir dies allerdings nicht ansehen, und die Berechnung der Wahrscheinlichkeiten wird für große n auch schnell sehr aufwendig, da die Binomialkoeffizienten schnell sehr groß werden. Um trotzdem einen Eindruck davon zu bekommen, was für größere n passiert, sind auf der nächsten Doppelseite die Wahrscheinlichketen für $n = 5, 10, 50, 100, 500$ und 1000 graphisch dargestellt. (Die Tatsache, daß ab $n = 50$ deutlich weniger als $n+1$ Balken zu sehen sind, erklärt sich daraus, daß die restlichen Wahrscheinlichkeiten zu klein sind, um noch sichtbar zu sein.)

Die Balkendiagramme zeigen, daß sich die Verteilung der Fehlerwahrscheinlichkeiten für große n einer festen Kurve annähern sollte, der in Abbildung 65 (und früher auch auf jedem Zehnmarkschein) zu findenden *Glockenkurve* oder GAUSS-Kurve.

Wenn sich Fehler oder auch beliebige Daten so veteilen, wie es dieser Kurve entspricht, redet man von *normalverteilten* Daten. Damit haben

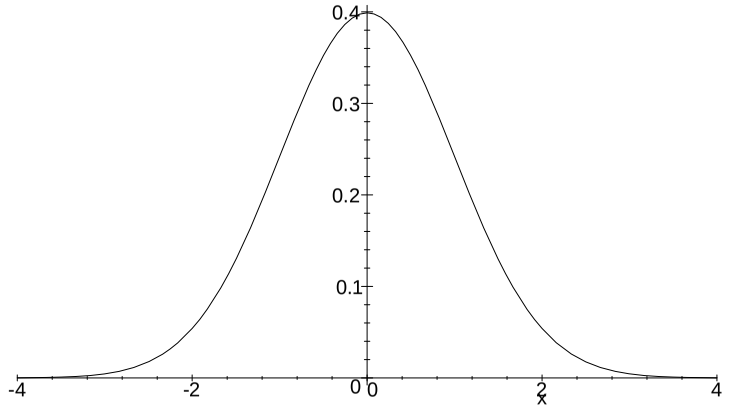


Abb. 65: Die „Glockenkurve“ $y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

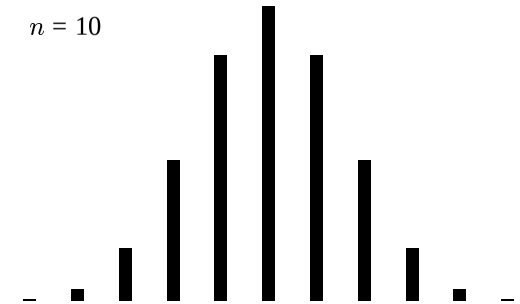
wir also zumindest graphisch gesehen, daß Meßfehler nach dem LAPLACEschen Fehlermodell normalverteilt sind. Das sagt noch nicht unbedingt etwas über die Verteilung realer Meßfehler, da das LAPLACEsche Fehlermodell von unrealistisch einfachen Annahmen ausgeht; nach einem der fundamentale Gesetze der Statistik, dem *zentralen Grenzwertsatz*, führen aber auch realistischere Annahmen zu genau derselben Verteilung: Sind $u_1, \ldots, u_n$ beliebige Quellen von Zufallsfehlern, über deren Verteilung wir (fast) nichts voraussetzen müssen, so ist ihre Summe für hinreichend großes n annähernd normalverteilt; siehe §5. Das eingeklammerte Wort „fast“ ist dabei für praktische Zwecke bedeutungslos, und als „groß“ kann man sich ein n ab etwa dreißig oder vierzig vorstellen.

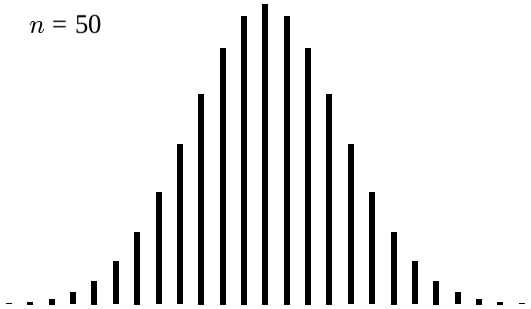
b) Statistische Kenngrößen

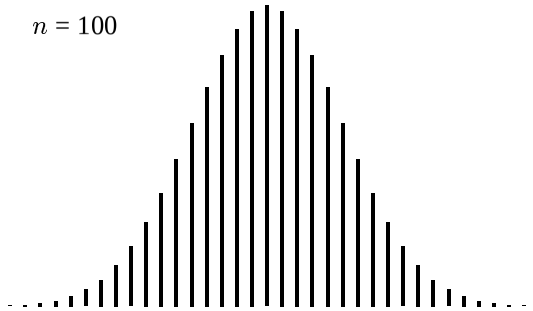
Die übliche Strategie zum Umgang mit Zufallsfehlern ist wohlbekannt: Man begnügt sich nicht mit einer einzigen Messung, sondern mißt mehrfach, so daß man eine ganze Meßreihe $x_1, x_2, \ldots, x_N$ erhält. Dann bildet man das *arithmetische Mittel*

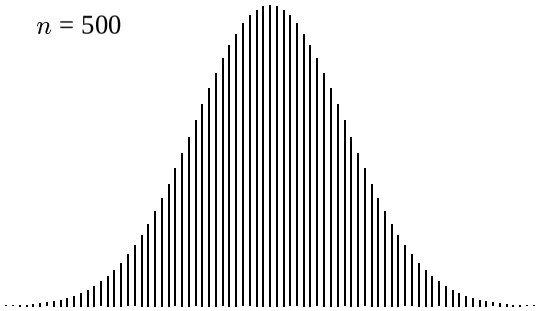
$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

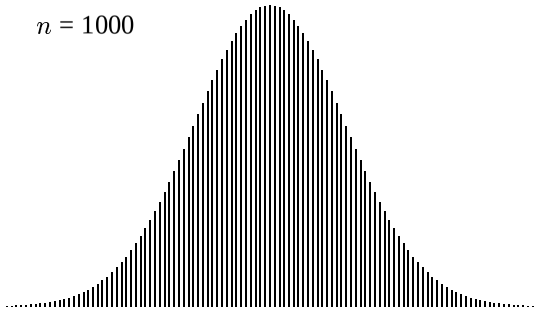

$n = 5$


$n = 10$


$n = 50$


$n = 100$


$n = 500$


$n = 1000$

der Meßreihe in der Hoffnung, daß sich hierbei die Fehler „ausmitteln“, so daß $\overline{x}$ dem theoretisch korrekten Wert $\hat{x}$ nahekommt.

Die Wahl des arithmetischen Mittels läßt sich auch geometrisch begründen: Eine Meßreihe $x_1, \ldots, x_N$ für eine Meßgröße mit exaktem Wert $\hat{x}$ definiert einen Vektor im $\mathbb{R}^n$. Falls es keine Meßfehler gäbe, hätte dieser lauter identische Komponenten $\hat{x}$. Tatsächlich ist dies natürlich nicht der Fall; wir können aber nach einem Vektor mit identischen Komponenten suchen, der möglichst nahe am Vektor der Meßwerte liegt. Für einen Vektor, dessen sämtliche Komponenten gleich x sind, ist der EUKLIDische Abstands zum Vektor der Meßwerte gleich

$$d(x) = \sqrt{\sum_{i=1}^{N}(x - x_i)^2} = \sqrt{Nx^2 - 2x\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} x_i^2}\,.$$

Die quadratische Funktion $d(x)^2$ hat ein eindeutig bestimmtes Minimum bei der Nullstelle ihrer Ableitung

$$2Nx - 2\sum_{i=1}^{N} x_i\,,$$

also beim arithmetischen Mittel $\overline{x}$, und dieses ist auch das einzige Minimum von $d(x)$. Wir nehmen daher das arithmetische Mittel $\overline{x}$ als besten verfügbaren Schätzwert für den unbekannten korrekten Wert $\hat{x}$.

Als Maß für die Schwankungen innerhalb der Meßreihe und damit für die Meßfehler könnte man versucht sein, den Abstand $d(\hat{x})$ zu nehmen; er hat aber den Nachteil, daß er mit steigendem N immer größer wird, d.h. die Schwankungen würden umso größer, je mehr man mißt. Das ist natürlich absurd; daher dividieren wir das Abstandsquadrat noch durch N und definieren die *mittlere quadratische Abweichung* oder *Varianz* der Meßreihe als

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\hat{x} - x_i)^2\,.$$

Die (nichtnegative) Quadratwurzel σ hieraus heißt *Standardabweichung* der Meßreihe.

Das Ergebnis einer Messung wird meist angegeben in der Form

$$x = \overline{x} \pm \sigma\,,$$

man betrachtet also die Standardabweichung der Meßreihe als Maß für den Meßfehler. Da deren Definition allerdings vom (im allgemeinen unbekannten) korrekten Wert $\hat{x}$ abhängt, können wir sie nicht berechnen, sondern müssen im folgenden sehen, wie wir sie zumindest schätzen können.

Als einfachste Möglichkeit bietet sich an, σ^2 durch

$$\frac{1}{N}\sum_{i=1}^{N}(\overline{x} - x_i)^2$$

zu schätzen, aber das führt sicherlich zu einem zu kleinen Ergebnis: Schließlich ist $d(\overline{x})$ das eindeutig bestimmte Minimum der Abstandsfunktion d, so daß der korrekte Wert $d(\hat{x})$ für $\hat{x} \neq \overline{x}$ notwendigerweise größer sein muß.

In Abschnitt *d)* werden wir aus dem Fehlerfortpflanzungsgesetz einen besseren Schätzwert für σ herleiten.

Warum betrachten wir eigentlich quadratische Abweichungen und nicht die einfacheren linearen Abweichungen? Nun, der Mittelwert aller Abweichungen ist

$$\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x}) = \frac{1}{N}\sum_{i=1}^{N} x_i - \frac{1}{N}N\overline{x} = \overline{x} - \overline{x} = 0\,,$$

also ist dies keine geeignete Maßzahl. Möglich wäre die mittlere *betragsmäßige* Abweichung

$$\frac{1}{N}\sum_{i=1}^{N}|x_i - \overline{x}|\,,$$

allerdings wird die im allgemeinen nicht für das arithmetische Mittel $\overline{x}$ minimal, sondern, wie man sich leicht überlegen kann, für jede Zahl $\overline{x}$ mit der Eigenschaft, daß gleichviele Meßwerte größer und kleiner als $\overline{x}$ sind; einen solchen Wert $\overline{x}$ bezeichnet man als *Median* der Meßreihe. Für die Beschreibung wirtschafts- und sozialwissenschaftlicher Daten ist dieser Median meist eine aussagekräftigere Kennzahl als das arithmetische Mittel; in den Naturwissenschaften und der Technik spielt er allerdings keine große Rolle. Im nächsten Paragraphen werden wir sehen, daß auch das LAPLACEsche Fehlermodell in natürlicher Weise auf quadratische Abweichungen führt.

c) Das Fehlerfortpflanzungsgesetz

Gegeben seien zwei Größen

$$x = \hat{x} \pm \sigma_x \quad \text{und} \quad y = \hat{y} \pm \sigma_y$$

(die Verallgemeinerung auf mehr als zwei Größen erfordert, wie man sich bei der folgenden Rechnung leicht klarmacht, nur etwas mehr Schreibaufwand; sie ist nicht prinzipiell schwieriger), und eine Größe

$$w = f(x, y)\,,$$

die von diesen beiden abhängt. Um vernünftige Aussagen machen zu können, setzen wir dabei f als stetig differenzierbar voraus.

Für x seien N Meßwerte $x_1, \ldots, x_N$ gegeben, und für y entsprechend M Werte $y_1, \ldots, y_M$. Wenn wir echte Zufallsfehler haben, können wir davon ausgehen, daß die Fehler der x-Werte und die der y-Werte voneinander unabhängig sind, und das wollen wir im folgenden auch annehmen.

Für w haben wir dann NM Werte $w_{ij} = f(x_i, y_j)$, deren Mittelwert die beste Schätzung für den „wahren“ Wert $\widehat{w} = f(\hat{x}, \hat{y})$ ist. Dieser Mittelwert ist für komplizierte Funktionen f und/oder große Werte von n und m umständlich auszurechnen; günstiger wäre es, einfach den Mittelwert $\overline{x}$ der x_i und den Mittelwert $\overline{y}$ der y_j zu berechnen, um dann $f(\overline{x}, \overline{y})$ als Schätzung für $\widehat{w}$ zu benutzen. Zur Abschätzung des dadurch bedingten Fehler setzen wir

$$x_i = \overline{x} + h_i \quad \text{und} \quad y_j = \overline{y}_j + k_j\,;$$

dann ist wegen der Differenzierbarkeit von f

$$\begin{aligned} w_{ij} &= f(x_i, y_j) = f(\overline{x} + h_i, \overline{y} + k_j) \\ &= f(\overline{x}, \overline{y}) + f_x(\overline{x}, \overline{y})h_i + f_y(\overline{x}, \overline{y})k_j + o\left(\sqrt{h_i^2 + k_j^2}\right)\,, \end{aligned}$$

wobei

$$f_x = \frac{\partial f}{\partial x} \quad \text{und} \quad f_y = \frac{\partial f}{\partial y}$$

die partiellen Ableitungen von f bezeichnen. Da die h_i und die k_j als Abweichungen vom Mittelwert die Summe null haben, ist also der Mittelwert der w_{ij} bis auf einen Fehler der Größenordnung $o\left(\sqrt{h^2 + k^2}\right)$ gleich $f(\overline{x}, \overline{y})$, wobei h, k die Batragsmaxima der h_i, k_j sind.

Als nächstes müssen wir den Fehler von $\widehat{w}$ berechnen, also den Mittelwert der $(w_{ij} - \widehat{w})^2$. Dazu schreiben wir zunächst

$$x_i = \hat{x} + u_i \quad \text{und} \quad y_j = \hat{y}_j + v_j\,,$$

betrachten also anstelle der Abweichungen vom Mittelwert die echten Meßfehler, und erhalten genau wie eben

$$w_{ij} - \widehat{w} = f(x_i, y_j) - f(\hat{x}, \hat{y}) \approx u_i \cdot f_x(\hat{x}, \hat{y}) + v_j \cdot f_y(\hat{x}, \hat{y})$$

mit Quadrat

$$\begin{aligned} (w_{ij} - \widehat{w})^2 &\approx \left(u_i \cdot f_x(\hat{x}, \hat{y}) + v_j \cdot f_y(\hat{x}, \hat{y})\right)^2 \\ &= u_i^2 \cdot f_x(\hat{x}, \hat{y})^2 + v_j^2 \cdot f_y(\hat{x}, \hat{y})^2 + 2u_i \cdot v_j \cdot f_x(\hat{x}, \hat{y}) \cdot f_y(\hat{x}, \hat{y})\,. \end{aligned}$$

Hier sind die Werte u_i^2, v_j^2 und $u_i v_j$ jeweils Zufallsgrößen, über deren Werte wir nichts sagen können. Wir haben aber gewisse Erwartungen darüber, wie sie sich *im Mittel* verhalten: u_i^2 sollte, da σ_x^2 die mittlere quadratische Abweichung von $\hat{x}$ ist, im Mittel gleich σ_x^2 sein und v_j^2 entsprechend σ_y^2. Genauso sollten u_i und v_j im Mittel gleich null sein, und wenn wir annehmen, daß die Fehler u_i und v_j voneinander unabhängig sind, sollte auch ihr Produkt im Mittel verschwinden. Diese sogenannten *Erwartungswerte* sind offensichtlich die bestmöglichen Schätzwerte für die jeweiligen Größen; als beste Schätzung für σ_w^2 erhalten wir damit das GAUSS*sche Fehlerfortpflanzungsgesetz*

$$\sigma_w^2 = f_x(\hat{x}, \hat{y})^2 \cdot \sigma_x^2 + f_y(\hat{x}, \hat{y})^2 \cdot \sigma_y^2$$

oder

$$\sigma_w = \sqrt{f_x(\hat{x}, \hat{y})^2 \cdot \sigma_x^2 + f_y(\hat{x}, \hat{y})^2 \cdot \sigma_y^2}\,.$$

Genauso gilt dieses Gesetz auch für Funktionen von mehr als zwei Größen; für $w = f(x_1, \ldots, x_n)$ ist

$$\sigma_w = \sqrt{f_{x_1}(\hat{x}_1, \ldots \hat{x}_n)^2 \cdot \sigma_{x_1}^2 + \cdots + f_{x_n}(\hat{x}_1, \ldots \hat{x}_n)^2 \cdot \sigma_{x_n}^2}\,.$$

d) Die Standardabweichung des Mittelwerts und die Schätzung der Varianz

Als einfache Anwendung des Fehlerfortpflanzungsgesetzes betrachten wir die Funktion

$$\bar{x} = f(x_1, \ldots, x_N) = \frac{x_1 + \cdots + x_N}{N} ,$$

also den Mittelwert der x_i. Jede Messung x_i sei mit demselben erwarteten Fehler σ behaftet; da alle partiellen Ableitungen von f gleich $1/N$ sind, folgt für den Fehler des Mittelwerts

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} .$$

Dies bestätigt die implizit stets angewandte Regel, daß man durch mehrfaches Messen ein zuverlässigeres Ergebnis erhält; durch 25 Messungen beispielsweise läßt sich der Fehler auf ein Fünftel reduzieren, und für $N \to \infty$ geht er gegen Null *(Gesetz der großen Zahl)*.

Damit wissen wir, wie man aus den Meßwerten auf den Fehler des Mittelwerts schließen kann – sofern man die Fehler der Meßwerte kennt. Wie lassen sich diese schätzen?

Zunächst ist

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\hat{x} - x_i)^2 = \frac{1}{N}\sum_{i=1}^{N}\big((\hat{x} - \bar{x}) + (\bar{x} - x_i)\big)^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\hat{x} - \bar{x})^2 + \frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2 + \frac{2}{N}\sum_{i=1}^{N}(\hat{x} - \bar{x}) \cdot (\bar{x} - x_i) .$$

Die letzte dieser drei Summen ist

$$2 \cdot \frac{(\hat{x} - \bar{x})}{N}\sum_{i=1}^{N}(\bar{x} - x_i) = 0 ,$$

da $\bar{x}$ der Mittelwert der x_i ist. Die zweite Summe ist der Mittelwert der $(\bar{x} - x_i)^2$, also die Varianz der Meßreihe, und von der ersten schließlich wissen wir, daß $(\hat{x} - \bar{x})^2$, das Quadrat des Fehlers des Mittelwerts, den Erwartungswert σ^2/N hat. Die gesamte erste Summe ist somit

$$\frac{1}{N} \cdot N \cdot \frac{\sigma^2}{N} = \frac{\sigma^2}{N} ,$$

und obige Formel wird zu

$$\sigma^2 = \frac{\sigma^2}{N} + \frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2 .$$

Bringt man hier noch den Term σ^2/N auf die linke Seite, so folgt

$$\frac{N-1}{N} \cdot \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2$$

oder

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{x} - x_i)^2 ,$$

also

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(\bar{x} - x_i)^2}{N-1}} .$$

Somit läßt sich auch σ aus den Meßdaten berechnen, der Meßfehler kann also ohne Kenntnis des „wahren“ Werts anhand der gemessenen Werte geschätzt werden.

§5: Zufallsvariablen und ihre Verteilungen

Zufallsvariablen sind, anschaulich ausgedrückt, Funktionen, die bei jeder Anwendung einen zufälligen Wert liefern derart, daß die Verteilung dieser Werte gewissen Gesetzmäßigkeiten unterliegt.

Eine exakte Definition müßte mit Wahrscheinlichkeitsräumen und meßbaren Funktionen arbeiten; dafür fehlt im Rahmen dieser Vorlesung erstens die Zeit, und zweitens wäre der dafür notwendige Aufwand bei den wenigen hier betrachteten anwendungen auch übertrieben.

Wir begnügen uns daher im nächsten Abschnitt mit nicht wirklich präzisen *ad hoc* Definitionen der beiden wichtigsten Arten von Zufallsvariablen, den diskreten mit endlichem Wertebereich und den kontinuierlichen mit stetiger Verteilungsfunktion.

a) Zufallsvariablen

Definition: Eine *diskrete Zufallsvariablen* ist ein Prozeß, der zufällig einen Wert aus einer vorgegebenen endlichen Menge

$$\{x_0, \ldots, x_m\}$$

liefert; eine *kontinuierliche Zufallsvariablen* liefert entsprechend einen zufälligen Wert aus $\mathbb{R}$.

Dieser „Zufall“ muß natürlich, falls er mathematisch faßbar sein soll, irgendwelchen Regeln genügen.

Im diskreten Fall nehmen wir dazu an, daß für jeden der möglichen Werte x_i feststeht, mit welcher *Wahrscheinlichkeit* p_i er angenommen wird. Diese „Wahrscheinlichkeit“ definieren wir informell so, daß bei einer großen Anzahl m von Versuchen *ungefähr* $p_i m$-mal der Wert x_i geliefert wird. Das „Gesetz der großen Zahlen“, das wir aus dem letzten Paragraphen kennen, sagt, daß diese Definition sinnvoll ist und man die Wahrscheinlichkeiten p_i damit in wohldefinierter Weise mit beliebiger Genauigkeit bestimmen kann.

Für eine Zufallsvariable, die kontinuierliche Werte annimmt, ist die Wahrscheinlichkeit dafür, daß ein konkreter Wert angenommen wird, praktisch immer gleich Null; hier können wir sinnvollerweise nur fragen, mit welcher Wahrscheinlichkeit die Werte in einem gegebenen Intervall $[a, b]$ liegen. Wie sagen, die Zufallsvariable X habe die *Wahrscheinlichkeitsdichte* f, wenn diese Wahrscheinlichkeit gleich

$$\int_a^b f(x)\,dx$$

ist.

Zwei formale Konsequenzen der Definition sind offensichtlich: Im diskreten Fall ist

$$0 \le p_i \le 1 \quad \text{für alle } i \text{ und} \quad \sum_{i=0}^{m} p_i = 1\,,$$

denn bei m Versuchen muß die Anzahl der auftretenden x_i für jedes i zwischen null und m liegen, und die Summe aller dieser Anzahlen ist m. Im kontinuierliche Fall ist entsprechend

$$f(x) \ge 0 \quad \text{für alle } x \in \mathbb{R} \quad \text{und} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1\,.$$

(Da $f(x)$ eine Wahrscheinlichkeits*dichte* ist, können wir dafür keine allgemein gültige obere Schranke angeben.)

b) Statistische Kenngrößen von Zufallsvariablen

Statistische Kenngrößen sind Zahlen, die Informationen über Zufallsvariablen liefern. Die wichtigste davon ist der *Erwartungswert;* anschaulich betrachtet ist das der erwartete Durchschnitt aus einer großen Anzahl von Werten.

Definition: Der *Erwartungswert* $\mathbb{E}(X)$ einer diskreten Zufallsvariablen X ist

$$\mathbb{E}(X) = \sum_{i=0}^{m} p_i x_i\,;$$

der einer kontinuierlichen Zufallsvariablen ist

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\,dx\,.$$

Manche Zufallsvariablen nehmen nur Werte an, die in der Nähe dieses Erwartungswerts liegen; bei anderen streuen die Werte in einem weiten Bereich. Ein erstes Maß dafür ist die Varianz, die wir analog zur mittleren quadratischen Abweichung bei Meßreihen als Erwartungswert der quadratischen Abweichung definieren:

Definition: Die *Varianz* einer Zufallsvariablen X mit Erwartungswert $\mathbb{E}(X)$ ist $\sigma_X^2 = \mathbb{E}\big((X - \mathbb{E}(X))^2\big)$; im diskreten Fall ist also

$$\sigma_X^2 = \sum_{i=0}^{m} p_i \big(x_i - \mathbb{E}(X)\big)^2 = \sum_{i=0}^{m} p_i x_i^2 - \mathbb{E}(X)^2\,;$$

im kontinuierlichen Fall ist

$$\sigma_X^2 = \int_{-\infty}^{\infty} \big(x - \mathbb{E}(X)\big)^2 \cdot f(x)\,dx = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mathbb{E}(X)^2 .$$

Die Quadratwurzel $\sigma_X = \sqrt{\sigma_X^2}$ heißt *Standardabweichung* von X.

Als erstes Beispiel wollen wir eine diskrete Zufallsvariable betrachten, die das Würfeln beschreibt. Bei einem idealen Würfel wird jede Augenzahl i mit derselben Wahrscheinlichkeit $p_i = \frac{1}{6}$ angenommen; der Erwartungswert ist also

$$\mathbb{E}(X) = \frac{1+2+3+4+5+6}{6} = 3\frac{1}{2} .$$

Damit läßt sich auch die Varianz einfach berechnen:

$$\sigma_X^2 = \frac{\big(-2\frac{1}{2}\big)^2 + \big(-1\frac{1}{2}\big)^2 + \big(-\frac{1}{2}\big)^2 + \big(\frac{1}{2}\big)^2 + \big(1\frac{1}{2}\big)^2 + \big(2\frac{1}{2}\big)^2}{6} = \frac{35}{12}$$

und

$$\sigma_X = \sqrt{\frac{35}{12}} \approx 1{,}7078 .$$

§6: Erste Beispiele von Verteilungen

a) Die Gleichverteilung

Das einfachste Beispiel einer kontinuierlichen Verteilung ist die *Gleichverteilung:* Hier kann die Zufallsvariable X nur Werte x annehmen, die zwischen zwei vorgegebenen Werten a und b liegen, und jeder dieser Werte ist gleich wahrscheinlich, d.h., exakt ausgedrückt, die *Wahrscheinlichkeitsdichte* dieser Verteilung ist gleich einer Konstanten γ im Intervall zwischen a und b und ist null außerhalb dieses Intervalls. Da

$$p(a \le x \le b) = \int_a^b f(x)\,dx = \int_a^b \gamma\,dx = \gamma \cdot (b-a) = 1$$

sein muß, ist also

$$f(x) = \gamma = \frac{1}{b-a} \quad \text{für } a \le x \le b .$$

Der *Erwartungswert* einer gleichverteilten Zufallsvariablen ist

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \int_a^b x \cdot f(x)\,dx \\ &= \int_a^b x \cdot \gamma\,dx = \gamma \int_a^b x\,dx \\ &= \gamma \cdot \left(\frac{1}{2}b^2 - \frac{1}{2}a^2\right) = \frac{\frac{1}{2}b^2 - \frac{1}{2}a^2}{b-a} \\ &= \frac{a+b}{2} ; \end{aligned}$$

wie zu erwarten war, ist es also der Mittelpunkt des Intervalls.

Die Varianz errechnet sich demnach als

$$\sigma^2 = \int_a^b \gamma \cdot \left(x - \frac{a+b}{2}\right)^2 dx ;$$

da $\frac{1}{3}(x-c)^3$ eine Stammfunktion von $(x-c)^2$ ist, folgt

$$\begin{aligned} \sigma^2 &= \frac{\gamma}{3} \cdot \left(\left(b - \frac{a+b}{2}\right)^3 - \left(a - \frac{a+b}{2}\right)^3\right) \\ &= \frac{\gamma}{3} \cdot \left(\left(\frac{b-a}{2}\right)^3 - \left(\frac{a-b}{2}\right)^3\right) \\ &= \frac{\gamma}{3 \cdot 8} \cdot 2 \cdot (b-a)^3 = \frac{2 \cdot (b-a)^3}{3 \cdot 8 \cdot (b-a)} \\ &= \frac{a^2 + ab + b^2}{12} . \end{aligned}$$

b) Die Binomialverteilung

Als nächstes Beispiel einer konkreten Verteilung wollen wir eine Verteilung für *diskrete* Zufallsvariablen betrachten, die einerseits für sich

selbst interessant ist, andererseits aber auch zwei der wichtigsten kontinuierlichen Verteilungen als Grenzfälle liefert.

Ausgangspunkt ist das n-malige Werfen einer Münze; diese falle jeweils mit Wahrscheinlichkeit p so, daß *Kopf* oben liegt und dementsprechend mit Wahrscheinlichkeit $q = 1 - p$ so, daß die *Zahl* zu sehen ist. Wir wollen die Wahrscheinlichkeit dafür berechnen, daß beim n-maligen Werfen k-mal *Kopf* und somit $(n - k)$-mal *Zahl* erscheint.

Anstelle des Münzwurfs kann man sich natürlich genausogut jedes andere Ereignis vorstellen, das genau zwei mögliche Ausgänge hat; lediglich der Anschaulichkeit halber soll vorläufig vom Werfen einer Münze die Rede sei – auch wenn für ein stark von $\frac{1}{2}$ abweichendes p die Anschaulichkeit vielleicht nicht allzu groß erscheint.

Wir definieren eine Zufallsvariable X durch

$$X = \text{Anzahl der Würfe mit Ergebnis } \textit{Kopf}$$

und suchen die Wahrscheinlichkeiten

$$p_k \underset{\text{def}}{=} p(X = k)$$

für $k = 0, \ldots, n$. Die Verteilung dieser Zufallsvariablen heißt *Binomialverteilung* oder BERNOULLI-*Verteilung* mit Parametern n und p. (Der Parameter $q = 1 - p$ muß nicht eigens erwähnt werden, da er durch p eindeutig bestimmt ist.)

Im Fall $n = 1$ ist alles klar: Die Wahrscheinlichkeit für *Kopf* ist p, die für *Zahl* entsprechend $q = 1 - p$, d.h.

$$p_0 = q \quad \text{und} \quad p_1 = p\,.$$

Auch für $n > 1$ wird den meisten klar sein, wie man die Wahrscheinlichkeiten berechnet: Es gibt $\binom{n}{k}$ Möglichkeiten, aus den n Versuchen diejenigen k auszuwählen, die das Ergebnis *Kopf* haben, und für jede einzelne dieser Möglichkeiten haben wir die Wahrscheinlichkeit $p^k q^{n-k}$ für k-mal *Kopf* und $(n - k)$-mal *Zahl;* somit ist

$$p_k = \binom{n}{k} p^k q^{n-k}\,.$$

Für diejenigen, die das nicht aus der Schule wissen, sei diese Formel kurz hergeleitet.

Für $n = 2$ gibt es drei mögliche Ausgänge des Zufallsexperiments: zweimal *Kopf,* zweimal *Zahl* oder je einmal *Kopf* und *Zahl.* Zweimal *Kopf* ist nur möglich, wenn beim ersten wie beim zweiten Wurf *Kopf* oben liegt. Beide Ereignisse haben, jeweils für sich betrachtet, die Wahrscheinlichkeit p; da sie voneinander unabhängig sind, ist die Wahrscheinlichkeit ihres gemeinsamen Auftretens gleich dem Produkt der beiden Einzelwahrsheinlichkeiten, also $p \cdot p = p^2$. Entsprechend ist die Wahrscheinlichkeit für zweimal *Zahl* natürlich q^2.

Das Ereignis „einmal *Kopf* und einmal *Zahl*“ kann auf zweierlei Weise zustande kommen: Entweder fällt beim ersten Wurf *Kopf* und beim zweiten Wurf *Zahl,* oder umgekehrt. Die Wahrscheinlichkeit für die erste Möglichkeit berechnet sich analog zu oben als $p \cdot q$, die für die zweite als $q \cdot p$, was natürlich genau der gleiche Wert ist. Da beide Möglichkeiten offensichtlich nie gleichzeitig auftreten können, ist die Wahrscheinlichkeit für das Auftreten von einer der beiden gerade die Summe der beiden Einzelwahrscheinlichkeiten, also $2pq$. Somit ist hier

$$p_0 = q^2,\ p_1 = 2pq \quad \text{und} \quad p_2 = p^2\,.$$

Die Wahrscheinlichkeiten für die drei möglichen Ausgänge des Zufallsexperiments addieren sich, wie es sich gehört, zu *eins*, denn

$$p^2 + q^2 + 2pq = (p + q)^2 = 1\,,$$

da sich p und q zu eins ergänzen.

Bei *dreimaligem* Werfen der Münze gibt es *vier* Möglichkeiten: Dreimal *Kopf*, dreimal *Zahl* sowie zweimal *Kopf* und einmal *Zahl* oder einmal *Kopf* und zweimal *Zahl.*

Anstatt diese Wahrscheinlichkeiten wieder einzeln zu berechnen, können wir vom Fall der zwei Würfe ausgehen und für jeden möglichen Ausgang den Effekt des dritten Wurfs betrachten: Für das Ereignis „zweimal *Kopf* und einmal *Zahl*“ müssen dann nur zwei Fälle betrachtet werden: Entweder war der dritte Wurf *Zahl* und damit die ersten beiden Würfe „zweimal *Kopf*“, oder der dritte Wurf lieferte *Kopf,* und die ersten beiden hatten das Ergebnis „einmal *Kopf* und einmal *Zahl*“. Die Wahrscheinlichkeit ist also

$$q \cdot p^2 + p \cdot pq = 3p^2 q\,.$$

Entsprechend ist die Wahrscheinlichkeit des Ereignisses „zweimal *Zahl* und einmal *Kopf*“ gleich $3pq^2$, und die Wahrscheinlichkeiten von „dreimal *Kopf*“ beziehungsweise „dreimal *Zahl*“ sind natürlich p^3 beziehungsweise q^3. Somit ist

$$p_0 = q^3,\ p_1 = 3pq^2,\ p_2 = 3p^2 q \quad \text{und} \quad p_3 = p^3\,,$$

und wieder ist die Summe aller Wahrscheinlichkeiten eins, denn

$$1 = (p + q)^3 = p^3 + 3p^2 q + 3pq^2 + q^3\,.$$

Entsprechend können wir argumentieren bei einer beliebigen Anzahl n von Würfen: X kann die $n + 1$ Werte $0 \le k \le n$ annehmen. Wie oben setzt sich das Ereignis $X = k$ zusammen aus verschiedenen Einzelereignissen wie etwa „zuerst k mal *Kopf,* dann $(n - k)$

mal Zahl“ und entsprechend den Ereignissen, bei denen k andere Wurfnummern für *Kopf* verlangt werden. Jedes dieser Ereignisse hat die Wahrscheinlichkeit $p^k q^{n-k}$, so daß p_k ein ganzzahliges Vielfaches von $p^k q^{n-k}$ ist. Wir schreiben dieses Vielfache als

$$p_k = \binom{n}{k} p^k q^{n-k}$$

und nennen $\binom{n}{k}$ den *Binomialkoeffizienten „n über k“*.

Zu seiner Berechnung können wir vorgehen wie im Fall $n = 3$, indem wir alles auf das Werfen von $n-1$ Münzen zurückführen und somit versuchen, $\binom{n}{k}$ durch geeignete Binomialkoeffizienten der Form $\binom{n-1}{j}$ auszudrücken:

Das Ereignis $X = k$ kann zustandekommen

- *entweder* dadurch, daß bei den ersten $n-1$ Würfen schon k mal *Kopf* gefallen ist und daß dann beim n-ten Wurf *Zahl* fällt
- *oder* dadurch, daß bei den ersten $n-1$ Würfen nur $(k-1)$ mal *Kopf* gefallen ist und daß dann beim n-ten Wurf noch einmal *Kopf* fällt.

Definieren wir also eine neue Zufallsvariable Y dadurch, daß Y die Häufigkeit des Ereignisses *Kopf* bei den ersten $n-1$ Würfen zählt, und eine Zufallsvariable Z als 0 oder 1, je nachdem, ob im n-ten Wurf *Zahl* oder *Kopf* fällt, so ist

$$\begin{aligned} p(X=k) &= p(Y=k-1)\cdot p(Z=1) + p(Y=k)\cdot p(Z=0) \\ &= \binom{n-1}{k} p^k q^{n-1-k}\cdot q \\ &\quad + \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)}\cdot p \\ &= \binom{n-1}{k} p^k q^{n-k} + \binom{n-1}{k-1} p^k q^{n-k}\,. \end{aligned}$$

Andererseits ist

$$p(X=k) = \binom{n}{k} p^k q^{n-k}\,,$$

also erhalten wir die Beziehung

$$\binom{n}{k} = \binom{n-k}{k-1} + \binom{n-1}{k}\,,$$

aus der sich $\binom{n}{k}$ rekursiv berechnen läßt.

Dies geschieht am besten über das PASCALsche Zahlendreieck: Wenn wir die Binomialkoeffizienten in der Form

$$\begin{array}{lccccccc} n=0 & & & & \binom{0}{0} & & & \\ n=1 & & & \binom{1}{0} & & \binom{1}{1} & & \\ n=2 & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & \\ n=3 & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} \end{array}$$

$$\begin{array}{lccccccccc} n=4 & & \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4} \\ n=5 & \binom{5}{0} & & \binom{5}{1} & & \binom{5}{2} & & \binom{5}{3} & & \binom{5}{4} & & \binom{5}{5} \end{array}$$

anordnen, besagt die gerade hergeleitete Beziehung, daß jeder Binomialkoeffizient die *Summe* der beiden rechts und links über ihm stehenden ist. Außen muß offenbar immer 1 stehen, denn die Wahrscheinlichkeit für 0 mal *Kopf*, d.h. n mal *Zahl*, ist $1\cdot q^n$, und die für n mal *Kopf* ist $1\cdot p^n$. Damit läßt sich das Dreieck leicht hinschreiben:

$$\begin{array}{lccccccccc} n=0 & & & & & 0 & & & & \\ n=1 & & & & 1 & & 1 & & & \\ n=2 & & & 1 & & 2 & & 1 & & \\ n=3 & & 1 & & 3 & & 3 & & 1 & \\ n=4 & 1 & & 4 & & 6 & & 4 & & 1 \\ n=5 & 1 & & 5 & & 10 & & 10 & & 5 & & 1 \end{array}$$

Für kleine Werte von n läßt sich $\binom{n}{k}$ auf diese Weise einfach bestimmen; für großes n ist dieses Verfahren allerdings zu umständlich. Hier verwendet man besser die explizite Formel

$$\binom{n}{k} = \frac{n!}{k!\cdot(n-k)!}\,,$$

in der $n!$ (gesprochen n *Fakultät*) für das Produkt

$$n! = 1\cdot 2\cdot 3\cdot\,\cdots\,\cdot(n-1)\cdot n$$

der Zahlen von 1 bis n steht und verabredungsgemäß das „leere“ Produkt $0! = 1$ sein soll. (Wer sich noch an seine Schulmathematik erinnert, kann diese Formel leicht mit elementarer Bruchrechnung durch Induktion nach n beweisen.) Für das praktische Rechnen ist es oft von Vorteil, durch $(n-k)!$ zu kürzen; dadurch wird obige Formel zu

$$\binom{n}{k} = \frac{n\cdot(n-1)\cdots(n-k+1)}{k!}\,.$$

Für große Werte von n wird die Binomialverteilung nur selten benutzt, da man sie dann meist durch handlichere Verteilungen annähern kann. Trotzdem sei der Vollständigkeit halber auf die in §7*c)* bewiesene STIRLINGsche Formel verwiesen, über die man $n!$ für große Werte von n näherungsweise berechnen kann:

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}\,,$$

wobei

$$e \approx 2{,}71828\,18284\,59045\,23536$$

die EULERsche Zahl ist und π die Kreiszahl. Der Schätzwert ist stets zu klein mit einem Fehler von höchstens $1/12n$.

Die Berechnung von Erwartungswert und Standardabweichung unmittelbar aus der Definition ist wegen der dazu auszuwertenden Summen mit vielen Binomialkoeffizienten eher unangenehm; glücklicherweise kann man darauf aber verzichten und alles auf den Fall $n = 1$ zurückführen: Für $n = 1$ ist der Erwartungswert

$$E(X) = p \cdot 1 + 1 \cdot 0 = p$$

und die Varianz

$$\begin{aligned} E\big((X-p)^2\big) &= p \cdot (1-p)^2 + q \cdot (0-p)^2 \\ &= p \cdot q^2 + q \cdot p^2 = pq(q+p) = pq\,, \end{aligned}$$

die *Standardabweichung* ist also $\sigma = \sqrt{pq}$.

Um daraus die entsprechenden Werte für beliebiges n zu berechnen, können wir uns zunutze machen, daß die verschiedenen Würfe der Münze voneinander *unabhängig* sind, so daß sich die Erwartungswerte einfach addieren, d.h.

$$E(X) = n \cdot p \quad \text{bei } n \text{ Würfen.}$$

Nicht ganz so klar ist, daß sich auch die Varianzen addieren; hierzu müssen wir uns an die Rechnung erinnern, in der wir den Fehler des arithmetischen Mittels aus n Werten einer Zufallsvariablen berechnet haben, indem wir aus dem Verschwinden der Erwartungswerte der gemischten Produkte genau dies geschlossen haben. Somit ist die Varianz im Falle von n Würfen gleich npq und die Standardabweichung entsprechend $\sqrt{npq}$.

c) Die Poisson-Verteilung

Auch hier handelt es sich um eine diskrete Verteilung, nämlich den Grenzfall der Binomialverteilung für großes n und kleines p, wobei der Erwartungswert $\lambda = np$ konstant gehalten wird. Dieses Produkt λ ist demgemäß der einzige Parameter der POISSON-Verteilung.

In Alltagssprache übersetzt bedeuten großes n und kleines p, daß wir die Häufigkeit eines seltenen Ereignisses betrachten; insbesondere interessieren also nur die kleinen Werte von k, da die großen ohnehin extrem unwahrscheinlich sind.

Für die Binomialverteilung mit Parametern n und p ist

$$\begin{aligned} p_k &= \binom{n}{k} p^k q^{n-k} = \frac{n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{k!} p^k q^{n-k} \\ &= \frac{n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{k!} \cdot \left(\frac{\lambda}{n}\right)^k \cdot (1-p)^{n-k} \\ &= \frac{n \cdot (n-1) \cdot \cdots \cdot (n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n}{n} \cdot \frac{n-1}{n} \cdot \cdots \cdot \frac{n-k}{n} \cdot \frac{\lambda^k}{k!} \cdot \frac{(1-\frac{\lambda}{n})^n}{\big(1-\frac{\lambda}{n}\big)^k}\,, \end{aligned}$$

da $q = 1 - p$ und $p = \lambda/n$ ist. Wie aus der Schule bekannt (sein sollte), ist

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}\,,$$

und da k klein, n aber groß sein soll, machen wir keinen großen Fehler, wenn wir

$$\frac{n}{n} \approx \frac{n-1}{n} \approx \cdots \approx \frac{n-k}{n} \approx 1 \quad \text{und} \quad \left(1 - \frac{\lambda}{n}\right)^k \approx 1$$

setzen. Damit ist

$$p_k \approx \frac{e^{-\lambda} \cdot \lambda^k}{k!}\,.$$

Führt man den Grenzübergang explizit aus, so erhält man die Wahrscheinlichkeiten

$$p_k = \frac{e^{-\lambda} \cdot \lambda^k}{k!}\,;$$

wir bezeichnen eine Zufallsvariable demnach als POISSON-verteilt, wenn sie diese Wahrscheinlichkeiten hat.

Die Kennzahlen der POISSON-Verteilung bestimmt man am einfachsten über die der Binomialverteilung: Da die POISSON-Verteilung Grenzwert von Binomialverteilungen mit Erwartungswert $\lambda = np$ ist, hat sie

natürlich auch selbst den Erwartungswert λ. Die *Varianz* einer Binomialverteilung mit Parametern n und p ist

$$n \cdot p \cdot q = n \cdot p \cdot (1 - p) = np - np^2 = \lambda - p \cdot \lambda \,.$$

Da die POISSON-Verteilung der Grenzwert für $p \to 0$ ist, hat sie also die Varianz λ und damit die Standardabweichung $\sqrt{\lambda}$.

Praktisch kann man davon ausgehen, daß man für Werte von n ab etwa $n = 100$ und für Wahrscheinlichkeiten bis etwa 5%, d.h. $p \leq 0{,}05$, die Binomialverteilung durch die erheblich einfacher zu berechnende POISSON-Verteilung ersetzen kann, ohne einen nennenswerten Fehler zu machen. Auch für etwas kleinere Werte von n ist die Übereinstimmung im allgemeinen schon recht gut.

Beispiele für POISSON-verteilte Zufallsvariablen gibt es viele: Häufigkeiten von Naturkatastrophen wie Erdbeben, Überschwemmungen *usw.*, die Anzahl radioaktiver Zerfälle pro Minute eines schwach radioaktiven Materials wie etwa C^{14} oder Co^{60} und viele mehr.

Da die POISSON-Verteilung nur einen Parameter hat, ist ihre Standardabweichung durch ihren Erwartungswert bestimmt; in der Tat ist sie einfach die Quadratwurzel davon. Daher steigt die Standardabweichung mit zunehmendem λ nur sehr viel langsamer als der Erwartungswert. Interpretiert man die Standardabweichung als Abweichung vom Mittelwert, so wird daher die *relative* Abweichung für große λ immer geringer und für sehr große λ praktisch vernachlässigbar. Dies ist einer der Gründe dafür, daß zahlreiche Vorgänge, die auf der mikroskopischen Ebene eigentlich nur statistisch beschrieben werden können, auf der makroskopischen Ebene im Rahmen der erzielbaren Meßgenauigkeit ein deterministisches Verhalten zeigen.

Ein zwar etwas exotisches, wegen der ungewöhnlich guten Übereinstimmung von Theorie und Praxis aber häufig in Statistiklehrbüchern anzutreffendes Beispiel für POISSON-verteilte Daten sind die Anzahlen der jährlich durch Hufschlag getöteten Offiziere preußischer Kavallerieregimenter: Eine Untersuchung von zehn Regimentern über zwanzig Jahre hinweg ergab insgesamt 122 Todesfälle, also im Mittel

$$\lambda = \frac{122}{10 \cdot 20} = 0{,}61$$

pro Jahr und Regiment.

Die Wahrschenlichkeit dafür, daß es in einem Regiment genau k Todesfälle in einem Jahr gab, sollte damit unter der Annahme einer POISSON-Verteilung mit dem Parameter (= Erwartungswert) 0,61 ungefähr gleich

$$\frac{e^{-0{,}61} \cdot 0{,}61^k}{k!} \approx \frac{0{,}331444 \cdot 0{,}61^k}{k!}$$

sein; wir erwarten also, daß dies in etwa

$$200 \cdot \frac{0{,}331444 \cdot 0{,}61^k}{k!} = \frac{66{,}2888 \cdot 0{,}61^k}{k!}$$

der zweihundert Fälle vorkommt. Die folgende Tabelle zeigt die tatsächlichen und die (gerundeten) berechneten Werte:

k	*tatsächliche Fallzahl*	*berechnete Fallzahl*
0	109	108,67
1	65	66,29
2	22	20,22
3	3	4,11
4	1	0,63
≥ 5	0	0,08

Wie man sieht, ist die Übereinstimmung in der Tat erstaunlich gut.

§ 7: Die Normalverteilung

a) Die Normalverteilung als Grenzfall der Binomialverteilung

Hinter dem LAPLACEschen Fehlermodell steckt offenbar eine Bionomialverteilung mit Parametern $p = \frac{1}{2}$ und n = Anzahl der „Dämonen“. Wenn wir hier einen vernünftigen Grenzübergang $n \to \infty$ machen wollen, müssen wir den Parameter ε offensichtlich so von n abhängen lassen, daß der Erwartungswert und die Varianz des Fehlers von n unabhängig sind.

Beim Erwartungswert ist das kein Problem: Er ist ohnehin immer null.

Zur Berechnung der Varianz müssen wir die Fehlerquadrate über die 2^n möglichen Verhaltensweisen der „Dämonen“ summieren, d.h. wir müssen

$$\sigma^2 = \frac{1}{2^n} \sum (\varepsilon_1 + \ldots + \varepsilon_n)^2$$

berechnen, wobei sich die Summation über alle n-tupel

$$(\varepsilon_1, \cdots, \varepsilon_n) \quad \text{mit} \quad \varepsilon_i = \pm\varepsilon$$

erstreckt. Beim Ausmultiplizieren heben sich alle gemischten Terme der Form $\varepsilon_i\varepsilon_j$ gegenseitig weg, denn das Tupel, bei dem nur an der i-ten Stelle das Vorzeichen geändert wurde, liefert einen Summanden $-\varepsilon_i\varepsilon_j$. Also bleiben nur die Quadrate; diese sind alle gleich ε^2, und es sind pro Summand n Stück. Da die Anzahl der Summanden gleich dem Nenner des Vorfaktors ist, berechnet sich die Varianz daher zu

$$\sigma^2 = n\varepsilon^2$$

Somit müssen wir für $n \to \infty$ den Einfluß ε jedes einzelnen „Dämonen“ so gegen null gehen lassen, daß $n\varepsilon^2$ konstant bleibt, d.h. wir setzen

$$\varepsilon = \frac{\sigma}{\sqrt{n}}$$

für eine geeignet zu wählende Konstante $\sigma > 0$, die Standardabweichung.

Für festes n kann der Fehler einen der $n + 1$ Werte

$$-n\varepsilon, \quad -(n-2)\varepsilon, \quad \ldots, \quad (n-2)\varepsilon, \quad n\varepsilon$$

annehmen, was wir in der Form

$$u = (n-2k)\varepsilon = \frac{(n-2k)\sigma}{\sqrt{n}} \quad \text{mit} \quad k = -n, \ldots, n$$

schreiben wollen. Dieser Fehler tritt genau dann auf, wenn k der Dämonen den Fehler ε erzeugen und die restlichen $n - k$ den Fehler $-\varepsilon$. Dies geschieht in $\binom{n}{k}$ der 2^n möglichen Fälle; die Wahrscheinlichkeit dafür ist also $\binom{n}{k}2^{-n}$.

Für $n \to \infty$ geht dieser Ausdruck gegen null, denn mit n geht schließlich auch die Anzahl der zu betrachtenden Fälle gegen unendlich. Falls es ein Intervall gäbe, in dem die Wahrscheinlichkeit für jeden darin liegenden Fehler größer als irgendein $\alpha > 0$ wäre, ginge allein schon die Summe der Wahrscheinlichkeiten für Fehler aus diesem Teilintervall mit n gegen unendlich, da die Anzahl der dort liegenden möglichen u-Werte wegen der $\sqrt{n}$ im Nenner von u gegen unendlich geht. Da die Summe aller Wahrscheinlichkeiten aber nicht größer als eins werden kann, muß die Wahrscheinlichkeit also in jedem einzelnen Punkt für $n \to \infty$ gegen null gehen.

Wenn wir n variieren lassen, ist es allerdings ohnehin sinnlos, einen genauen Wert des Fehlers zu betrachten: Für jedes n gibt es nur $n + 1$ mögliche Werte, und bei den meisten größeren Werten von n werden diese Zahlen – abgesehen von der Null – nicht auftreten. Wenn wir etwas von n unabhängiges definieren möchten (und sofern wir nicht an Dämonen glauben, bleibt uns kaum etwas anderes übrig), dürfen wir also nicht den genauen *Wert* des Fehlers festlegen, sondern müssen ein *Fehlerintervall* betrachten.

Nun wollen wir aber natürlich als Ergebnis keine Funktion, die von einem Intervall abhängt, sondern eine gewöhnliche Funktion von u, d.h. wir wollen, wie bei kontinuierlichen Verteilungen üblich, eine *Wahrscheinlichkeitsdichte* $\varphi(u)$.

Da die Wahrscheinlichkeit dafür, daß u im Intervall $[a, b]$ liegt, gleich dem Integral von f über dieses Intervall ist, ist für jeden Punkt u_0

$$\varphi(u_0) = \lim_{\varepsilon \to 0} \frac{\text{Wahrscheinlichkeit für } u_0 - \varepsilon \le u \le u_0 - \varepsilon}{2\varepsilon},$$

d.h. also als Wahrscheinlichkeit dividiert durch die Intervallbreite.

Für festes n haben die möglichen Fehlerwerte den Abstand 2ε, wir betrachten daher Intervalle der Länge 2ε mit den Werten $(n - 2k)\varepsilon$ als Mittelpunkten; diese überdecken den möglichen Fehlerbereich lückenlos, und die Wahrscheinlichkeit für einen Fehler in diesem Intervall ist $\binom{n}{k}2^{-n}$.

Wir interessieren uns daher für $\lim_{n\to\infty} \dfrac{\binom{n}{k}2^{-n}}{2\varepsilon} = \dfrac{\binom{n}{k}2^{-n}}{2\sigma/\sqrt{n}} = \binom{n}{k}2^{-n-1}\dfrac{\sqrt{n}}{\sigma}$.

b) Die Eulersche Summenformel

Das Problem bei der Berechnung dieses Grenzwerts ist der Binomialkoeffizient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!};$$

um diesen abzuschätzen brauchen wir einen handhabbaren Ausdruck für $n!$.

Dazu schreiben wir

$$\ln n! = \sum_{k=1}^{n} \ln k$$

und berechnen dies nach einer Methode von EULER, die nicht nur für Summen von Logarithmen anwendbar ist.

Wir betrachten allgemeiner irgendeine reellwertige differenzierbare Funktion f, deren Definitionsbereich das Intervall $[1,\, n]$ enthält.

Für eine reelle Zahl x bezeichnen wir wie üblich mit $[x]$ die größte ganze Zahl kleiner oder gleich x und mit $\{x\} \underset{\text{def}}{=} x-[x]$ den gebrochenen Anteil von x; ist k eine ganze Zahl, ist somit $\{x\} = x - k$ für $x \in [k,\, k+1)$.

Partielle Integration führt auf die Gleichung

$$\int_k^{k+1} \left(\{x\} - \tfrac{1}{2}\right) f'(x)\,dx = \left(x - k - \tfrac{1}{2}\right) f(x) \Big|_k^{k+1} - \int_k^{k+1} f(x)\,dx$$

$$= \frac{f(k+1) + f(k)}{2} - \int_k^{k+1} f(x)\,dx\,.$$

Addition aller solcher Gleichungen von $k = 1$ bis $k = n-1$ liefert

$$\int_1^n \left(\{x\} - \tfrac{1}{2}\right) f'(x)\,dx = \frac{f(1)}{2} + \sum_{k=2}^{n-1} f(k) + \frac{f(n)}{2} - \int_1^n f(x)\,dx\,,$$

womit man die Summe der $f(k)$ berechnen kann:

Satz (EULERsche Summenformel): Für eine differenzierbare Funktion $f\colon D \to \mathbb{R}$, deren Definitionsbereich das Intervall $[1,\, n]$ umfaßt, ist

$$\sum_{k=1}^{n} f(k) = \int_1^n f(x)\,dx + \frac{f(1)+f(n)}{2} + \int_1^n \left(\{x\} - \tfrac{1}{2}\right) f'(x)\,dx\,.$$

■

Für die Abschätzung der Binomialkoeffizienten und Fakultäten interessiert uns speziell der Fall $f(x) = \ln x$; hierfür wird die EULERsche Summenformel zu

$$\ln n! = \int_1^n \ln x\,dx + \frac{\ln n}{2} + \int_1^n \frac{\{x\} - \frac{1}{2}}{x}\,dx$$

$$= x(\ln x - 1)\Big|_1^n + \frac{\ln n}{2} + \int_1^n \frac{\{x\} - \frac{1}{2}}{x}\,dx$$

$$= n(\ln n - 1) + 1 + \frac{\ln n}{2} + \int_1^n \frac{\{x\} - \frac{1}{2}}{x}\,dx\,.$$

In dieser Formel stört noch das rechte Integral; dieses können wir wie folgt abschätzen: Für eine natürliche Zahl k ist

$$\int_k^{k+1} \frac{\{x\} - \frac{1}{2}}{x}\,dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{x}{k + \frac{1}{2} + x}\,dx$$

$$= \int_0^{\frac{1}{2}} \left(\frac{x}{k+\frac{1}{2}+x} - \frac{x}{k+\frac{1}{2}-x}\right) dx = \int_0^{\frac{1}{2}} \frac{-2x^2}{\left(k+\frac{1}{2}\right)^2 - x^2}\,dx\,.$$

Im Intervall von 0 bis $\frac{1}{2}$ ist der Integrand monoton fallend, d.h.

$$0 \geq \frac{-2x^2}{\left(k+\frac{1}{2}\right)^2 - x^2} \geq \frac{-\frac{1}{2}}{\left(k+\frac{1}{2}\right)^2 - \frac{1}{4}} = \frac{-2}{(2k+1)^2 - 1} \geq -\frac{1}{4k^2}\,,$$

und damit ist

$$0 \geq \int_k^{k+1} \frac{\{x\} - \frac{1}{2}}{x}\,dx = \int_0^{\frac{1}{2}} \frac{-2x^2}{\left(k+\frac{1}{2}\right)^2 - x^2}\,dx \geq -\frac{1}{8k^2}\,,$$

denn wir können das Integral abschätzen durch das Produkt aus der Länge des Integrationsintervalls und dem Minimum des Integranden. Summation von $k = 1$ bis $n-1$ schließlich gibt die Abschätzung

$$0 \geq \int_1^n \frac{\{x\} - \frac{1}{2}}{x}\,dx \geq -\sum_{k=1}^{n-1} \frac{1}{4k^2}$$

für das störende Integral aus der obigen Formel.

Wie wohl jeder schon einmal in einer Analysis I Übungsaufgabe zeigen mußte, konvergiert die rechtsstehende Summe (egal ob mit oder ohne acht im Nenner) für $n \to \infty$; aus Kapitel III, §3*f)* wissen wir sogar, daß der Grenzwert $\pi^2/48$ ist. Auf jeden Fall können wir folgern, daß das uneigentliche Integral

$$\int_1^\infty \frac{\{x\} - \frac{1}{2}}{x}\,dx$$

konvergiert; den uns bislang noch unbekannten Grenzwert wollen wir mit I bezeichnen. Damit ist

$$\ln n! = n(\ln n - 1) + \frac{\ln n}{2} + C + o(1) \quad \text{mit} \quad C = I + 1$$

oder $n! \approx e^C \cdot n^n e^{-n} \sqrt{n}$.

Für Binomialkoeffizienten folgt, daß

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \approx \frac{e^C n^n e^{-n}\sqrt{n}}{e^C k^k e^{-k}\sqrt{k}\, e^C (n-k)^{n-k} e^{-(n-k)}\sqrt{(n-k)}} = \frac{1}{e^C}\,\frac{n^n}{k^k\,(n-k)^{n-k}}\sqrt{\frac{n}{k\cdot(n-k)}}\,.$$

c) Die Stirlingsche Formel und die Normalverteilung

Ausgedrückt durch $u = (n-2k)\varepsilon = (n-2k)\sigma/\sqrt{n}$ ist

$$k = \frac{n}{2} - \frac{u\sqrt{n}}{2\sigma}\,,$$

und setzen wir zur Vereinfachung der Schreibweise

$$m = \frac{n}{2}, \quad v = \frac{u}{2\sigma} \quad \text{und} \quad \ell = \sqrt{2m}\cdot v\,,$$

so ergeben sich die Formeln

$$k = m - \sqrt{2m}\cdot v = m - \ell \quad \text{und} \quad n - k = m + \sqrt{2m}\cdot v = m + \ell\,.$$

Setzen wir dies alles in die Formel für die Wahrscheinlichkeitsdichte ein, erhalten wir

$$\begin{aligned}
&\binom{n}{k} 2^{-n-1}\frac{\sqrt{n}}{\sigma} \\
&\approx \frac{1}{e^C}\,\frac{n^n}{k^k\,(n-k)^{n-k}}\sqrt{\frac{n}{k\cdot(n-k)}}\,2^{-n-1}\frac{\sqrt{n}}{\sigma} \\
&= \frac{1}{e^C\sigma}\,\frac{n^n\cdot 2^{-n}}{k^k\,(n-k)^{n-k}}\,\frac{n\cdot 2^{-1}}{\sqrt{k\cdot(n-k)}} \\
&= \frac{1}{e^C\sigma}\,\frac{(2m)^{2m}\cdot 2^{-2m}}{(m-\ell)^{m-\ell}(m+\ell)^{m+\ell}}\,\frac{2m\cdot 2^{-1}}{\sqrt{(m-\ell)(m+\ell)}} \\
&= \frac{1}{e^C\sigma}\,\frac{m^{2m}}{(m-\ell)^m(m+\ell)^m}\left(\frac{m-\ell}{m+\ell}\right)^{\ell}\frac{m}{\sqrt{m^2-\ell^2}} \\
&= \frac{1}{e^C\sigma}\,\frac{m^{2m}}{(m^2-\ell^2)^m}\left(\frac{m-\ell}{m+\ell}\right)^{\ell}\frac{m}{\sqrt{m^2-\ell^2}} \\
&= \frac{1}{e^C\sigma}\,\frac{1}{\left(1-\frac{\ell^2}{m^2}\right)^m}\left(\frac{1-\ell/m}{1+\ell/m}\right)^{\ell}\frac{1}{\sqrt{1-\ell^2/m^2}} \\
&= \frac{1}{e^C\sigma}\,\frac{1}{\left(1-\frac{2v^2}{m}\right)^m}\left(\frac{1-\sqrt{2/m}\cdot v}{1+\sqrt{2/m}\cdot v}\right)^{\sqrt{2m}\cdot v}\frac{1}{\sqrt{1-2v^2/m}}\,.
\end{aligned}$$

Nun können wir langsam daran denken, n (und damit auch m) gegen unendlich gehen zu lassen; wir verwenden dazu die aus der Analysis I und wahrscheinlich auch aus der Schule bekannte Beziehung

$$e^x = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n .$$

Danach ist insbesondere

$$\lim_{m\to\infty} \left(1 - \frac{2v^2}{m}\right)^m = e^{-2v^2}$$

und

$$\lim_{m\to\infty} \left(1 \pm \sqrt{\frac{2}{m}} \cdot v\right)^{\sqrt{2m}\cdot v} = \lim_{m\to\infty} \left(1 \pm \frac{\sqrt{2}\cdot v}{\sqrt{m}}\right)^{\sqrt{m}\cdot\sqrt{2}\cdot v}$$

$$= \lim_{q\to\infty} \left(1 \pm \frac{\sqrt{2}\cdot v}{q}\right)^{q\cdot\sqrt{2}\cdot v} = \left(e^{\pm\sqrt{2}\cdot v}\right)^{\sqrt{2}\cdot v} = e^{\pm 2v^2} ,$$

denn es bleibt sich natürlich gleich, ob m oder $q = \sqrt{m}$ gegen unendlich geht. Da der Term v^2/m gegen null geht, erhalten wir somit als Grenzwert des gesamten obigen Ausdrucks

$$\frac{1}{e^C\sigma} \cdot \frac{1}{e^{-2v^2}} \cdot \frac{e^{-2v^2}}{e^{+2v^2}} \cdot 1 = \frac{1}{e^C\sigma} e^{-2v^2} .$$

Beachten wir nun noch, daß $v = u/2\sigma$ war, erhalten wir

$$\frac{1}{e^C\sigma} e^{-u^2/2\sigma^2} .$$

Damit sind wir fast am Ziel; das einzige, was noch fehlt, ist die Konstante C. Diese können wir bestimmen, indem wir ausnutzen, daß jeder Fehler mit Wahrscheinlichkeit eins zwischen $-\infty$ und ∞ liegt, d.h.

$$\frac{1}{e^C\sigma} \int_{-\infty}^{\infty} e^{-u^2/2\sigma^2}\, du = 1 .$$

Aus **[HM1]**, *Kap.* 2, §6*c)*, wissen wir, daß

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}$$

ist; mit der Substitution $x = u/\sqrt{2}\sigma$ folgt, daß dann

$$\int_{-\infty}^{\infty} e^{-u^2/2\sigma^2}\, du = \sqrt{2\pi}\sigma$$

ist und

$$e^C = \sqrt{2\pi} \quad \text{oder} \quad C = \tfrac{1}{2}\ln(2\pi) .$$

Damit haben wir die Wahrscheinlichkeitsdichte endlich vollständig berechnet; das Endergebnis ist

$$\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-u^2}{2\sigma^2}} .$$

Auch die Formel für $n!$ können wir nach der Bestimmung von C nun vollständig hinschreiben:

$$\ln n! = n(\ln n - 1) + \ln\sqrt{2\pi n} + o(1) \quad \text{oder} \quad n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n} .$$

Beides bezeichnet man als STIRLINGsche Formel.

Der schottische Mathematiker JAMES STIRLING (1692–1770) war Anhänger des gestürzten Königs Jakob II Stuart und hatte deshalb große politische Probleme bei seinem Studium; unter anderem wurde er deshalb von der Universität Oxford ausgeschlossen. 1717–1722 lebte er in Venedig und hatte auch gute Kontakte zu NICOLAUS BERNOULLI an der Universität von Padua; außerdem brachte er aus Venedig die Produktionsgeheimnisse der dortigen Glasbläser mit. Ab 1724 arbeitete er zehn Jahre lang als Mathematiklehrer in London, wo er viel mit NEWTON zusammentraf; 1735 wurde er Direktor einer schottischen Bergbaugesellschaft. In seine Londoner Zeit fällt die Veröffentlichung seines bedeutendsten Werks *Methodus Differentialis sive Tractatus de Summatione er Interpolatione Serierum Infinitarum* im Jahre 1730, das die obige Formel als Beispiel zwei zu Proposition 28 enthält. Ebenfalls ziemlich bekannt wurde seine 1735 veröffentlichte Arbeit über die Gestalt der Erde.

d) Der zentrale Grenzwertsatz

Betrachten wir die Summe zweier unabhängiger Zufallsvariablen X und Y mit Verteilungsfunktionen f und g, so nimmt $X+Y$ offensichtlich genau dann einen Wert zwischen a und b an, wenn Y einen Wert aus dem Intervall $[a-x, b-x]$ annimmt, wobei x der von der Zufallsvariablen X

gelieferte Wert ist. Diese Wahrscheinlichkeit ist für einen festen Wert von x gleich

$$\int_{a-x}^{b-x} g(y)\,dy = \int_a^b g(u)\,du \quad \text{mit} \quad u = x + y\,.$$

Die Wahrscheinlichkeit dafür, daß $X+Y$ einen Wert aus $[a,b]$ annimmt ist daher nach dem Satz von FUBBINI und wegen der Unabhängigkeit der beiden Variablen

$$\int_{-\infty}^{\infty} f(x) \left(\int_a^b g(u-x)\,du \right) dx = \int_a^b \left(\int_{-\infty}^{\infty} f(x) g(u-x)\,dx \right) du$$
$$= \int_a^b (f \star g)(u)\,du\,.$$

Die Wahrscheinlichkeitsdichte der Summe $X + Y$ ist daher gleich der Faltung $f \star g$ der Wahrscheinlichkeitsdichten der Summanden.

Falls wir eine Summe aus N unabhängigen Zufallsvariablen X_i mit Verteilungsfunktionen f_i betrachten, ist deren Verteilungsfunktion somit gleich der Faltung $s = f_1 \star f_2 \star \cdots \star f_N$.

Die Verteilungsfunktion f des Mittelwerts der X_i genügt offensichtlich der Bedingung

$$\int_a^b f(x)\,dx = \int_{Na}^{Nb} s(x)\,dx\,.$$

Die Substitution $u = x/N$ macht das rechte Integral zu

$$\int_{Na}^{Nb} s(x)\,dx = \int_a^b s(Nu)\,N\,du = N \int_a^b s(Nu)\,du\,,$$

also ist $f(x) = N \cdot s(Nx)$.

Faltungen, insbesondere solche mit vielen Faktoren, sind eher unangenehm zu berechnen; durch FOURIER-Transformation werden sie zu

harmlosen Produkten. Also nehmen wir an, daß für alle betrachten Verteilungsfunktionen FOURIER-Transformierten existieren (was bei den üblicherweise vorkommenden kontinuierlichen Verteilungsfunktionen keine nennenswerte Einschränkung bedeutet), und rechnen mit diesen:

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(x) e^{-i\omega x}\,dx = N \cdot \int_{-\infty}^{\infty} s(Nx) e^{-i\omega x}\,dx$$
$$= N \cdot \int_{-\infty}^{\infty} s(u) e^{-\frac{i\omega u}{N}} \frac{du}{N} = \int_{-\infty}^{\infty} s(u) e^{-i\frac{\omega}{N}u}\,du = \widehat{s}\left(\frac{\omega}{N}\right).$$

Da außerdem $\widehat{s}(\omega) = \widehat{f}_1(\omega) \cdots \widehat{f}_N(\omega)$ ist, folgt

$$\widehat{f}(\omega) = \prod_{k=1}^{N} \widehat{f}_i\left(\frac{\omega}{N}\right).$$

Um zu sehen, was die FOURIER-Transformierte einer Verteilungsfunktion ist, schreiben wir den Exponentialfaktor im FOURIER-Integral als Potenzreihe:

$$\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(x) e^{-i\omega x}\,dx = \int_{-\infty}^{\infty} f(x) \sum_{k=0}^{\infty} \frac{(-i\omega x)^n}{n!}\,dx$$
$$= \sum_{k=0}^{\infty} \frac{(-i\omega)^n}{n!} \int_{-\infty}^{\infty} f(x) x^n\,dx = \sum_{k=0}^{\infty} \frac{(-i\omega)^n}{n!} \mathbb{E}(X^n)\,.$$

$\mathbb{E}(X^0)$ ist natürlich die Konstante Eins, und $\mathbb{E}(X^1) = \mathbb{E}(X)$ ist der Erwartungswert von X. Falls dieser verschwindet, ist $\mathbb{E}(X^2)$ der Erwartungswert der mittleren quadratischen Abweichung vom Mittelwert, also die Varianz.

Betrachten wir der Einfachheit halber zunächst den Fall, daß die Erwartungswerte aller X_i verschwinden. Dann ist $\widehat{f}_k(\omega) = 1 - \frac{\sigma_k^2 \omega^2}{2} + \cdots$ und

$$\widehat{f}(\omega) = \prod_{k=1}^{N} \widehat{f}_k\left(\frac{\omega}{N}\right) = \prod_{k=1}^{N} \left(1 - \frac{\sigma_k^2}{2}\left(\frac{\omega}{N}\right)^2 + \cdots\right),$$

wobei σ_k^2 die Varianz von X_k ist. Die weggelassenen und nur durch Punkte angedeuteten Terme enthalten Potenzen von ω/N mit Exponent mindestens drei; für große N können diese Terme gegenüber dem Quadrat von ω/N vernachlässigt werden. Also ist

$$\widehat{f}(\omega) = \prod_{k=1}^{N} \widehat{f}_k\Big(\frac{\omega}{N}\Big) \approx \prod_{k=1}^{N}\Big(1 - \frac{\sigma_k^2}{2}\Big(\frac{\omega}{N}\Big)^2\Big)\,.$$

Falls alle σ_k einen gemeinsamen Wert σ_0 haben, hat der Mittelwert der X_k nach §4*d)* die Varianz $\sigma^2 = \frac{1}{N}\sigma_0^2$ und

$$\widehat{f}(\omega) = \Big(1 - \frac{\sigma_0^2}{2}\Big(\frac{\omega}{N}\Big)^2\Big)^N = \Big(1 - \frac{\sigma^2\omega^2}{2N}\Big)^N\,,$$

was bekanntlich für $N \to \infty$ gegen $e^{-\frac{\sigma^2\omega^2}{2}}$ konvergiert.

Auch wenn die σ_k verschieden sind, können wir den Grenzwert leicht ausrechnen: Für große Werte von N ist ω/N klein, und für kleine Werte von x ist

$$\ln(1-x) = \ln 1 - \frac{d\,\ln x}{dx}(1)\cdot x + o(x) = -x + o(x)\,,$$

also hier

$$\begin{aligned}\ln \widehat{f}(\omega) &= \sum_{k=1}^{N} \ln \widehat{f}_k\Big(\frac{\omega}{N}\Big) = \sum_{k=1}^{N} \ln\Big(1 - \frac{\sigma_k^2}{2}\Big(\frac{\omega}{N}\Big)^2 + \cdots\Big)\\ &\approx \sum_{k=1}^{N} \frac{\sigma_k^2}{2}\Big(\frac{\omega}{N}\Big)^2 = \frac{1}{N^2}\sum_{k=1}^{N}\sigma_k^2\frac{\omega^2}{2}\,.\end{aligned}$$

Nach dem Fehlerfortpflanzungsgesetz ist die von $\frac{1}{N}\sum X_k$ gleich

$$\sigma^2 = \sum \frac{\sigma_k^2}{N^2}, \quad \text{also ist} \quad \widehat{f}(\omega) \approx e^{\frac{\sigma^2\omega^2}{2}}\,.$$

Wie wir aus Kapitel 3, §7*b)* wissen, ist dies die FOURIER-Transformierte von

$$f(t) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{t^2}{2\sigma^2}}\,,$$

die Verteilung des Mittelwerts ist also eine Normalverteilung mit Erwartungswert null und Varianz σ^2.

Falls die Zufallsvariablen X_k von Null verschiedene Erwartungswerte haben, etwa $\mathbb{E}(X_k) = \mu_k$, haben die Zufallsvariablen $Y_k \underset{\text{def}}{=} X_k - \mu_k$ Erwartungswert null und dieselbe Varianz σ_k^2 wie die X_k. Das arithmetische Mittel der X_k unterscheidet sich um

$$\mu = \frac{1}{N}\sum_{k=1}^{N}\mu_k$$

vom arithmetischen Mittel null der Y_k, also genügt es einer Normalverteilung mit Mittelwert μ und Varianz σ^2.

e) Eigenschaften der Normalverteilung

Oft interessiert nicht so sehr die Verteilung der Fehler, sondern die der Meßwerte selbst. Ist $\hat{x}$ der korrekte Wert und x_i der i-te Meßwert dafür, der gemäß $x_i = \hat{x} + u_i$ mit dem Fehler u_i behaftet ist, so können wir mit $x = \hat{x} + u$ die obigen Wahrscheinlichkeitsdichte auch als

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\hat{x})^2/2\sigma^2}$$

schreiben.

Als *Normalverteilung mit Mittelwert a und Standardabweichung σ* bezeichnen wir daher die Verteilung mit Wahrscheinlichkeitsdichte

$$\varphi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}\,.$$

Diese Wahrscheinlichkeitsdichte hängt offensichtlich nur von der *normierten* Variablen

$$z = \frac{x-a}{\sigma}$$

ab; diese hat Mittelwert null und Standardabweichung eins. Daher gibt es für die Normalverteilung nicht – wie für viele andere statistische Verteilungen – je nach Parameterwerten verschiedene Tabellen, sondern man findet in allen Tabellenwerken nur die Normalverteilung mit Mittelwert null und Standardabweichung eins, man findet also die Wahrscheinlichkeitsdichte

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

und deren Integral

$$F(z) = \int_{-\infty}^{z} f(u)\,du = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}}\,du\,.$$

Dieses Integral läßt sich nicht weiter vereinfachen, da sich die Stammfunktion von $e^{-u^2/2}$ nicht durch elementare Funktionen ausdrücken läßt. Für die Bestimmung von $F(z)$ ist man daher auf Tabellen oder Computerprogramme angewiesen; eine graphische Darstellung von $F(z)$ ist in Abbildung 68 zu sehen. Mit dieser Funktion läßt sich die Wahrscheinlichkeit dafür, daß

$$c \le z = \frac{x-a}{\sigma} \le d$$

ist berechnen als $F(d) - F(c)$, und damit läßt sich auch leicht die Wahrscheinlichkeit berechnen, daß x selbst zwischen zwei gegebenen Schranken liegt.

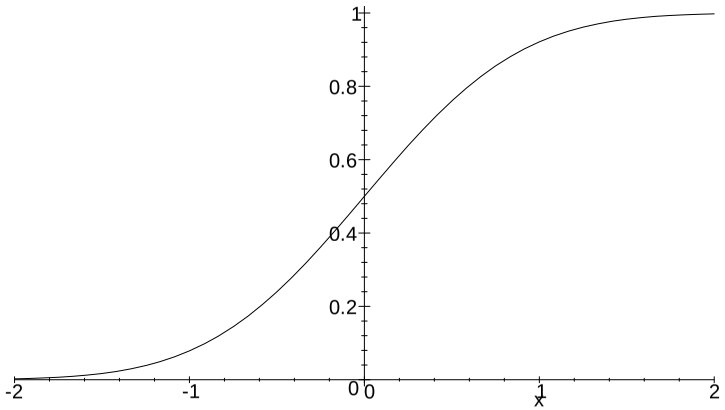


Abb. 68: Das Integral $F(z)$ über die „Glockenkurve“

Mißt man beispielsweise die Temperatur eines Wasserbads eine Viertelstunde lang jede Minute und erhält dabei 15 Meßwerte mit Mittelwert 20,1°C und Standardabweichung 0,2°C, so ist die Standardabweichung des Mittelwerts

$$\sigma_{\overline{\vartheta}} = \frac{0,2^\circ\mathrm{C}}{\sqrt{14}} \approx 0,053^\circ\mathrm{C}\,.$$

Wenn wird dann beispielsweise wissen wollen, mit welcher Wahrscheinlichkeit die „tatsächliche“ mittlere Temperatur zwischen 20,0°C und 20,2°C liegt, müssen wir dazu zunächst die normalisierten Werte berechnen:

$$z_1 = \frac{20,0-20,1}{0,053} \approx -1,89 \quad \text{und} \quad z_2 = \frac{20,2-20,1}{0,053} \approx 1,89\,.$$

Die Wahrscheinlichkeit ist also

$$F(1,89) - F(-1,89) \approx 0,94\,;$$

oder rund 94%.

Schaut man in einer Tabelle nach, wird man dort allerdings im allgemeinen nur den Wert $F(1,89)$ finden, nicht aber $F(-1,89)$. Der Grund dafür liegt in der Symmetrie des Graphen von F bezüglich des Punktes $(0, \frac{1}{2})$. Was dahinter steckt, sieht man am besten, wenn man die Dichtefunktion der Normalverteilung betrachtet, also die Glockenkurve: Für $z > 0$ ist $F(-z)$ die in Abbildung 69 links eingezeichnete schraffierte Fläche. Diese Fläche ist wegen der Symmetrie der Glockenkurve zur senkrechten Achse gleich der rechts eingezeichneten schraffierten Fläche, und deren Komplement ist $F(z)$. Also ist

$$F(-z) = 1 - F(z)\,,$$

und es reicht, wenn wir die Werte von F im positiven Bereich kennen.

Oft interessiert auch die Wahrscheinlichkeit dafür, daß der Betrag des Fehlers unterhalb einer bestimmten Schranke liegt, etwa $z \cdot \sigma$; in Abbildung 69 wäre dies der nichtschraffierte Bereich unter der Glockenkurve.

Wie man sich anhand der Abbildung leicht klarmacht, ist diese Wahrscheinlichkeit gleich

$$F(z) - F(-z) = 2F(z) - 1\,;$$

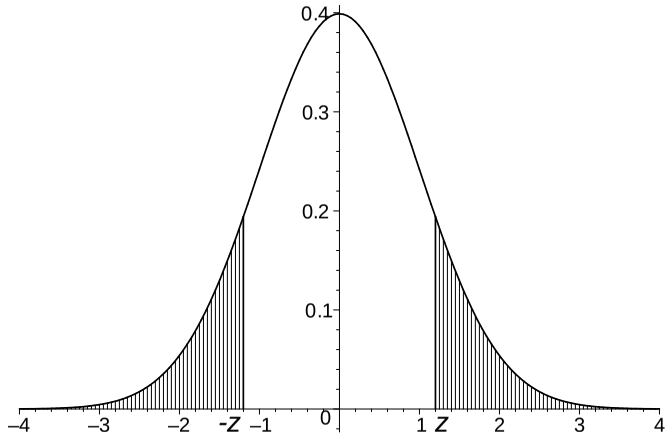


Abb. 69: Zusammenhang zwischen $F(z)$ und $F(-z)$

die Wahrscheinlichkeit, daß wir im obigen Beispiel die mittlere Temperatur mit einem Fehler von höchstens $0{,}05^\circ$ gemessen haben, ist also

$$2F\left(\frac{0{,}05}{0{,}053}\right) \approx F(0{,}94) \approx 0{,}83\,.$$

Ein Wasserbad hat üblicherweise den Sinn, ein Experiment unter kontrollierten Temperaturbedingungen durchzuführen; daher interessiert vor allem, inwieweit es gelingt, die Temperatur innerhalb gewisser Schranken zu halten. Die Wahrscheinlichkeit dafür können wir mit denselben Methoden berechnen, allerdings müssen wir dazu mit der Standardabweichung der Meßreihe selbst arbeiten.

Wenn wir etwa wollen, daß die Temperatur immer zwischen 19,5 und $20{,}5^\circ$C liegt, so ist die Wahrscheinlichkeit, daß wir dies mit dem oben ausgemessenen Versuchsaufbau erreichen, gleich

$$F\left(\frac{20{,}5-20{,}1}{0{,}2}\right) - F\left(\frac{29{,}5-20{,}1}{0{,}2}\right) = F(2) - F(-3) \approx 0{,}976\,.$$

In knapp zweieinhalb Prozent aller Fälle, im Schnitt also alle vierzig Minuten, müssen wir also damit rechnen, daß die Toleranzgrenzen überschritten werden.

Wie Abbildung 68 zeigt, liegt $F(-2)$ sehr nahe bei null und $F(2)$ sehr nahe bei eins. In der Tat ist die Wahrscheinlichkeit dafür, daß ein Wert z Betrag größer z hat, nach obiger Diskussion gleich

$$1 - (2F(z) - 1) = 2F(z) - 2\,,$$

was für $z = 2$ zu -0,0455 wird; die Wahrscheinlichkeit ist also kleiner als 5%. Allgemein gilt für eine beliebige Normalverteilung, daß der Wert der Variablen mit folgenden Wahrscheinlichkeiten um höchstens $i\sigma$ vom Mittelwert abweicht:

$i =$	1	2	3	4
Wahrscheinlichkeit:	0,683	0,954	0,9973	0,99994

Damit liegen also etwa zwei Drittel aller Fehler zwischen $-\sigma$ und σ, 95% liegen zwischen -2σ und 2σ und 99,7% zwischen -3σ und 3σ; die Wahrscheinlichkeit dafür, daß der Fehler größer als 3σ ist, beträgt nur etwa 0,27%. Da Ereignisse mit einer so geringen Wahrscheinlichkeit seltener als in einem von 300 Fällen auftreten, betrachtet man Fehler, die außerhalb des 3σ-Bereichs liegen, oft als „Ausreißer“, d.h. als grobe Meßfehler, die bei der Bestimmung des Ergebnisses nicht berücksichtigt werden. Sehr vorsichtige Leute reden allerdings erst ab einer Abweichung von 4σ von Ausreißern; solche Fehler treten zufällig weniger als einmal pro 15 000 Messungen auf.

Für Leser, die ihren Computer selbst programmieren und keine spezielle Statistiksoftware haben, sei hier eine Näherungsformel für $F(z)$ angegeben: Mit einem Fehler von höchstens $7{,}5 \cdot 10^{-8}$ ist

$$F(z) = 1 - \varphi(z) \cdot \left(a_1 t + a_2 t^2 + a_3 a^3 + a_4 t^4 + a_5 t^5\right)$$

mit $t = \dfrac{1}{1+pz}$ und $\varphi(z) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ sowie

$a_1 = \;\; 0{,}319\,381\,530 \quad a_2 = -0{,}356\,563\,782 \quad a_3 = 1{,}781\,477\,973$
$a_4 = -1{,}821\,255\,978 \quad a_5 = \;\; 1{,}330\,274\,429 \quad p = 0{,}231\,641\,9$.

Beim Rechnen mit dem Taschenrechner kann man sich auch mit einer vereinfachten Version begnügen, bei der $a_4 = a_5 = 0$ ist und

$a_1 = 0{,}436\,1836 \quad a_2 = -0{,}120\,1676 \quad a_3 = 0{,}937\,2980 \quad p = 0{,}332\,67$;

hier kann der Fehler bis zu 10^{-5} betragen.

f) Die Maximum Likelihood Methode

GAUSS gab im Laufe seines Lebens mehrere Begründungen für die Methode der kleinsten Quadrate (die er bei sowohl bei seinen astronomischen Arbeiten wie auch bei der von ihm geleiteten Vermessung des Königreichs Hannover zwischen 1818 und 1832 ständig benutzte); die unter dem Gesichtspunkt einer in sich geschlossenen Fehlertheorie interessanteste beruht auf dem LAPLACEschen Fehlermodell:

Danach sollte der Wert u_i für die korrekten Parameterwerte $a, b, \dots$ aus einer Normalverteilung mit Mittelwert $f(a, b, \dots; t_i)$ kommen, deren Standardabweichung σ_i von der Genauigkeit abhängt, mit der u_i bestimmt werden kann. Die Wahrscheinlichkeit dafür, daß u_i zwischen zwei Werten a und b liegt, ist damit

$$\int_a^b e^{-\big(u - f(a,b,\dots;\,t_i)\big)^2/2\sigma_i^2}\,.$$

Von der Wahrscheinlichkeit, daß u_i gleich einem Wert c ist, können wir natürlich nicht reden, da diese nach obiger Formel ein Integral von c nach c wäre, also Null. Aber die Wahrscheinlichkeit dafür, daß u_i in einem kleinen Intervall der Länge ε_i um einen Wert c_i liegt, ist ungefähr proportional zum ε_i-fachen Wert des Integranden an der Stelle c_i, also zu

$$\varepsilon_i \cdot e^{-\big(c_i - f(a,b,\dots;\,t_i)\big)^2/2\sigma_i^2}\,.$$

Entsprechend ist

$$\varepsilon_j \cdot e^{-\big(c_j - f(a,b,\dots;\,t_j)\big)^2/2\sigma_j^2}$$

ungefähr gleich der Wahrscheinlichkeit dafür, daß u_j in einem Intervall der Breite ε_j um c_j liegt.

Wenn wir wie üblich davon ausgehen, daß keine systematischen Fehler auftreten, sind die Fehler von u_i und u_j voneinander unabhängig, die Wahrscheinlichkeit dafür, daß (u_i, u_j) in einem Rechteck mit Seiten ε_i und ε_j um (c_i, c_j) liegt, ist also proportional zum Produkt der beiden

obigen Einzelwahrscheinlichkeiten, d.h. zu

$$\varepsilon_i\varepsilon_j\, e^{-\big(c_i - f(a,b,\dots;\,t_i)\big)^2/2\sigma_i^2 \,-\, \big(c_j - f(a,b,\dots;\,t_j)\big)^2/2\sigma_j^2}\,.$$

Entsprechend kann auch die Wahrscheinlichkeit dafür berechnet werden, daß der Punkt $(u_1, \dots, u_n)$ in einem kleinen gegebenen Quader mit Kantenlängen $\varepsilon_1, \dots, \varepsilon_n$ liegt; sie ergibt sich zu

$$L(a, b, \dots) \cdot \prod_{i=1}^{n} \varepsilon_i$$

mit

$$L(a, b, \dots) \underset{\text{def}}{=} e^{\sum_{i=1}^{n}\big(u_i - f(a,b,\dots;\,t_i)\big)^2/2\sigma_i^2}\,.$$

Diese Größe ist selbst keine Wahrscheinlichkeit, sondern der Quotient aus einer Wahrscheinlichkeit und einem Volumen; man spricht daher von einer *Wahrscheinlichkeitsdichte.*

Wenn wir diese Wahrscheinlichkeitsdichte als Funktion von $a, b, \dots$ betrachten, macht sie eine Aussage über die Güte der Parameter: Schließlich wird man einem Modell, das dem beobachteten Ausgang eines Experiments eine hohe Wahrscheinlichkeit zuweist, eher glauben als einem alternativen Modell, das die beobachteten Daten zu Ausreißern erklärt. Aus diesem Grund kann die Funktion L auch als Maß dafür betrachtet werden, wie „wahrscheinlich“ in irgendeinem umgangssprachlichen (und schwer präzisierbaren) Sinne die Parameter $a, b, \dots$ sind.

Im englischen gibt es zwei Wörter für Wahrscheinlichkeit: Das romanische Wort *probability* und das germanische Wort *likelihood.* Für den mathematisch exakten Wahrscheinlichkeitsbegriff verwendet man probability, für „Wahrscheinlichkeit“ im Sinne der Funktion L likelihood. Da es im deutschen kein zweites Wort für Wahrscheinlichkeit gibt, spricht man hier in Anlehnung an das Englische von einer *Likelihoodfunktion.*

Die Maximum Likelihood Methode besteht nun genau in dem, was ihr Name besagt: *Man wähle die Parameter $a, b, \dots$ so, daß die Likelihoodfunktion maximal wird.*

Da $L(a, b, \ldots)$ durch eine Exponentialfunktion beschrieben wird, wird die Likelihoodfunktion genau dann maximal, wenn ihr Exponent maximal wird. Dieser Exponent ist eine negative Zahl, wird also genau dann maximal, wenn sein Betrag *minimal* wird, das heißt, wenn die Quadratsumme

$$\sum_{i=1}^{n} \frac{\big(u_i - f(a, b, \ldots; t_i)\big)^2}{2\sigma_i^2}$$

minimal wird.

In vielen Fällen wird die Zuverlässigkeit der einzelnen Paare (t_i, u_i) miteinander vergleichbar sein, so daß alle σ_i gleich sind; in diesem Fall kann man die σ_i ignorieren und einfach die Quadratsumme

$$\sum_{i=1}^{n} \big(u_i - f(a, b, \ldots; t_i)\big)^2$$

minimieren, d.h. wir kommen wieder zur klassischen Methode der kleinsten Quadrate. Es gibt aber auch Anwendungen, wie etwa oben beim überexponentiellen Bevölkerungswachstum, bei denen die Verschiedenheit des σ_i sehr wesentlich ist: Sicherlich wird man etwa der auf Volkszählungen beruhenden Weltbevölkerungszahl, die die Vereinten Nationen für 1995 veröffentlichten, mehr Vertrauen entgegenbringen als der Schätzung eines Historikers für die Weltbevölkerung des Jahres Null, und selbst bei ein und derselben Meßreihe im Labor kommt es gelegentlich vor, daß (beispielsweise aufgrund unterschiedlicher Genauigkeit eines Meßinstruments in verschiedenen Bereichen) manche Daten zuverlässiger sind als andere.

§ 8: Kompression von Bild- und Audiodaten

Zum Abschluß der Vorlesung wollen wir wenigstens kurz eine praktische Anwendung kennenlernen, in der mit Eigenwerten und Eigenvektoren symmetrischer Matrizen, FOURIER-Transformationen und Statistik gleich mehrere der Methoden aus diesem Semester gleichzeitig benötigt werden: die Komprimierung von Bild- und Audiodaten.

a) Datenkompression

Ziel der Datenkompression ist es, eine Datei für Zwecke der Speicherung oder Übertragung möglichst stark zu verkleinern, das aber in einer solchen Weise, daß sich die ursprüngliche Datei aus der verkleinerten wieder exakt rekonstruieren läßt.

Es ist klar, daß es keinen universellen Algorithmus zur Datenkompression geben kann: Gäbe es nämlich ein Verfahren, das für beliebige Dateien einen Kompressionsfaktor $\alpha < 1$ garantieren würde, so könnte man dieses Verfahren iterativ anwenden und nach n Anwendungen eine Kompressionsrate von α^n erreichen. Wenn man n nur hinreichend groß wählt, könnte man daher jede Datei auf weniger als ein Bit komprimieren, was natürlich absurd ist.

Ein Kompressionsverfahren kann also nur auf Dateien mit spezieller Struktur erfolgreich angewandt werden und muß die spezielle Redundanz in diesen Dateien ausnutzen. In Textdateien beispielsweise ist dies die Redundanz der Sprache, die schon bei bloßer Beachtung der höchst unterschiedlichen Buchstabenhäufigkeiten Kompressionen von rund 50% gestattet.

Bilddaten werden typischerweise als Matrizen aus ganzen Zahlen zwischen 0 und 255 digitalisiert; bei Audiodaten nimmt man Vektoren von ganzen Zahlen zwischen 0 und $65\,535 = 2^{16} - 1$ oder $16\,777\,215 = 2^{24} - 1$. (Der Unterschied zwischen den Wertenbereichen liegt darin begründet, daß unser Auge selbst bei gedruckten Bildern mit nur 64 Graustufen praktisch keine Artefakte mehr erkennen kann, wohingegen unser Gehör noch auf sehr feine Unterschiede reagiert.)

Bei einer Musik-CD etwa wird das Signal 44 100-mal pro Sekunde abgetastet (dies bedeutet nach dem Abtasttheorem von NYQUIST, daß ein auf den Bereich von 0 bis 22,05kHz bandbegrenztes Signal fehlerfrei rekonstruiert werden kann), und das Ergebnis wird dann so skaliert und quantisiert (d.h. gerundet), daß eine Zahl zwischen 0 und 65535 entsteht. Bei Bilddaten werden je nach Auflösung und Seitenverhältnis zwischen etwa 256×256 und 1024×1024 Bildpunkte abgetastet, für Schwarzweißbilder nur nach Helligkeit, für Farbbildern nach insgesamt

drei Größen, die vom jeweiligen Farbmodell abhängen. Das Ergebnis dieser Abtastungen wird dann entsprechend skaliert und quantisiert.

Typische Komprimierungsverfahren arbeiten daher mit Vektoren oder Matrizen aus Zahlen zwischen 0 und einer geeigneten Zahl M, die aus praktischen Gründen meist von der Form $2^{8r}-1$ ist, wobei die Zahl r der Empfindlichkeit unserer Sinne angepaßt zwischen eins und drei liegt.

Da man zur eindeutigen Festlegung von N beliebigen Zahlen zwischen 0 und 2^{8r} nicht mit weniger als den $8Nr$ Bit auskommen kann, die man zum Hinschreiben der Zahlen braucht, sehen wir auch hier wieder, daß kein Verfahren *alle* solchen Vektoren komprimieren kann; wir müssen also eine Teilmenge auszeichnen.

Die ideale solche Teilmenge wäre hier natürlich die Menge aller möglicher Bilder (oder Audiosequenzen), aber diese Menge dürfte mathematisch kaum definierbar sein: Schließlich hängt es sehr vom Betrachter ab, welches Pixelmuster er noch als „Bild“ gelten läßt und welches nicht. Sinnvoll läßt sich eine solche Menge daher höchstens definieren, wenn von vornherein feststeht, welche Bilder berücksichtigt werden sollen – und dann ist wohl ein Verfahren, das statt vom Bildinhalt von einer Bildnummer ausgeht, unschlagbar.

Die meisten klassischen Verfahren, die beliebige, aber realistische Bilder komprimieren sollen, gehen aus von einem *statistischen Modell,* das zwar auch viele Matrizen produziert, die niemand als „Bilder“ anerkennen würde, das aber dennoch genügend viele Eigenschaften realer Bilder reproduziert, um eine große Anzahl von „Nichtbildern“ auszuschließen.

Ausgangspunkt ist die Beobachtung, daß es in einem Bild oder Musikstück nur wenige aprupte Übergänge gibt. Zwar gibt es natürlich immer wieder ein plötzliches *fortissimo,* das auf eine leise Stelle folgt, aber da das Signal 44 100-mal pro Sekunde abgetastet wird und solche Übergänge selbst bei der schrägsten Musik deutlich seltener als im Sekundenrythmus erfolgen, sind diese Sprünge innerhalb des zu behandelnden Datenstroms in der Tat sehr seltene Ereignisse. Wir können daher davon ausgehen, daß sich die unmittelbaren Nachbarn eines Datums *im Mittel* nur wenig vom gegebenen Datum unterscheiden.

Dasselbe gilt auch für Bilddaten: Falls das Bild digital hinreichend fein dargestellt wird, so daß keine Rastereffekte erkennbar sind, kommen große Sprünge in den Helligkeitswerten nur selten vor.

Bei diesem engen Zusammenhang zwischen benachbarten Werten setzen viele gängige Komprimierungsalgorithmen an: Wenn zwei Größen typischerweise sehr ähnlich sind, wird bei der Übertragung oder Speicherung *beider* Werte ein großer Teil der Information doppelt betrachtet; die Informationsdichte kann also deutlich erhöht werden, wenn man nur Informationen betrachtet, die weitgehend unabhängig voneinander sind.

Aus dem letzten Semester kennen wir ein Maß für die gegenseitige Abhängigkeit von Daten: Als wir dort untersuchten, wie das Klausurergebnis eines Studenten von seiner Arbeit bei den wöchentlichen Übungen abhängt oder die Korruption eines Staats vom Bruttosozialprodukt pro Einwohner, überprüften wir die Qualität unserer Modelle mit Hilfe des Korrelationskoeffizienten: Dieser lag bei ± 1 bei perfekter Übereinstimmung, und nahe Null, wenn das Modell keinen Zusammenhang zwischen den Daten lieferte.

Dieselbe Technik können wir auch anwenden, um Abhängigkeiten innerhalb einer Folge zu finden; bevor wir diese sogenannte Autokorrelation verstehen können, brauchen wir aber zunächst noch einige Vorbereitungen aus der Stochastik.

b) Korrelation von Zufallsvariablen

Ein guter Komprimierungsalgorithmus muß auch für Bilder funktionieren, die wir erst in ein paar Jahren photographieren. Für den Grundalgorithmus zur Datenkompression müssen wir daher Daten zulassen, über die wir noch nichts konkretes wissen – abgesehen von gewissen vagen Gesetzmäßigkeiten, durch die sich „echte“ Bilddaten von beliebigen Matrizen unterscheiden. Es bietet sich daher an, die Helligkeits- und/oder Farbwerte der einzelnen Pixel *bzw.* die Schalldruckwerte bei Tonaufnahmen durch Zufallsvariablen zu beschreiben. Da wir an digitalen Bild- und Audiodaten interessiert sind, verwenden wir dazu diskrete Zufallsvariablen.

So, wie wir sie bislang definiert haben, ist jede Zufallsvariable ein eigenständiger Prozeß, und zwei verschiedene Zufallsvariablen haben nichts miteinander zu tun. Das ist natürlich nicht das, was wir hier brauchen; wir müssen wir davon ausgehen, daß ein einziger Prozeß gleichzeitig einen ganzen Vektor *bzw.* eine ganze Matrix von Zufallswerten erzeugt, wobei deren einzelne Komponenten dann sehr wohl voneinander abhängig sein können.

Für zwei solche Komponenten X und Y mit jeweiligen Wertebereichen $\{x_0,\dots,x_m\}$ und $\{y_0,\dots,y_n\}$ sowie Wahrscheinlichkeiten p_i für x_i und q_j für y_j ist die Wahrscheinlichkeit dafür, daß X den Wert x_i liefert und Y den Wert y_j dann nicht p_iq_j, wie das bei unabhängigen Variablen der Fall wäre, sondern irgendeine Wahrscheinlichkeit π_{ij}, von der wir nur wissen, daß aus offensichtlichen Gründen etwa

$$\sum_{j=0}^{m}\pi_{ij} = p_i \quad \text{und} \quad \sum_{i=0}^{m}\pi_{ij} = q_j$$

sein muß. Für so ein Paar definieren wir

Definition: *a)* Die Kovarianz eines solchen Paars (X,Y) ist

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathbb{E}\big((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\big)\\ &= \sum_{i=0}^{m}\sum_{j=0}^{n}\pi_{ij}\big((x_i-\mathbb{E}(X)\big)\big((y_j-\mathbb{E}(Y)\big)\,.\end{aligned}$$

b) Die Korrelation von (X,Y) ist $\rho(X,Y) = \dfrac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}$.

Der Vollständigkeit halber sei auch die entsprechende Definition für koninuierliche Zufallsvariablen angegeben: Haben X und Y die zweidimensionale Wahrscheinlichkeitsdichte f, ist also die Wahrscheinlichkeit dafür, daß das Paar (X,Y) einen Wert in einer Teilmenge $B\subseteq\mathbb{R}^2$ liefert, gleich

$$\iint_B f(x,y)\,dx\,dy\,,$$

so definieren wir die Kovarianz des Paars als

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathbb{E}\big((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\big)\\ &= \iint_{\mathbb{R}^2}\big(x-\mathbb{E}(X)\big)\big(y-\mathbb{E}(Y)\big)f(x,y)\,dx\,dy\,;\end{aligned}$$

die Korrelation wird dann über dieselbe Formel wie im diskreten Fall definiert.

Wir bezeichnen zwei diskrete Zufallsvariablen X und Y entsprechend der üblichen Definition für Ereignisse als voneinander unabhängig, falls für alle i,j gilt: $\pi_{ij} = p_iq_j$; im kontinuierliche Fall verlangen wir entsprechend, daß die zweidimensionale Wahrscheinlichkeitsdichte f das Produkt der Wahrscheinlichkeitsdichten von X und von Y ist. Alsdann ist im diskreten Fall

$$\begin{aligned}\operatorname{cov}(X,Y) &= \sum_{i=0}^{m}\sum_{j=0}^{n}\pi_{ij}\big(x_i-E(X)\big)\big((y_j-E(Y)\big)\\ &= \sum_{i=0}^{m}\sum_{j=0}^{m}\Big[p_i(x_i-E(X))\Big]\Big[q_j\big(y_j-E(Y)\big)\Big]\\ &= \left(\sum_{i=0}^{m}p_i\big(x_i-E(X)\big)\right)\left(\sum_{j=0}^{n}q_j\big((y_j-E(Y))\big)\right) = 0\,,\end{aligned}$$

denn

$$\sum_{i=0}^{m}p_i\big((x_i-E(X)\big) = \sum_{i=0}^{m}p_ix_i - \sum_{i=0}^{m}p_iE(X) = \sum_{i=0}^{m}p_ix_i - E(X)$$

verschwindet nach Definition des Erwartungswerts.

Damit haben zwei voneinander unabhängige diskrete Zufallsvariablen also Kovarianz und Korrelation null; man rechnet leicht nach, daß dies auch im kontinuierlichen Fall gilt. Solche Zufallsvariablen heißen unkorreliert.

Bei Bilddaten wird das im allgemeinen nicht der Fall sein; hier wird man ims Gegenteil davon ausgehen, daß die Zufallsvariablen zu benachbarten Pixeln sehr stark miteinander korrelieren. Wir können beispielsweise

annehmen, daß $Y = \rho X + Z$ ist mit einer von X unabhängigen Zufallsvariablen Z und einer positiven reellen Zahl $\rho < 1$. Dann ist

$$E(Y) = E(\rho X + Z) = \rho E(X) + E(Z)\,;$$

falls wir annehmen, daß X und Y denselben Erwartungswert haben, ist daher

$$E(Z) = (1-\rho)E(X)\,,$$

was für ein ρ nahe eins deutlich kleiner ist als $E(Z)$.

In dieser Situation ist

$$\begin{aligned}
\operatorname{cov}(X,Y) &= \sum_{i=0}^{m}\sum_{j=0}^{n} \pi_{ij}\big(x_i - E(X)\big)\big((y_j - E(Y)\big) \\
&= \sum_{i=0}^{m}\sum_{j=0}^{n} \pi_{ij}\big(x_i - E(X)\big)\big(\rho x_i + z_j - \rho E(X) - E(Z)\big) \\
&= \sum_{i=0}^{m}\sum_{j=0}^{n} \pi_{ij}\big(x_i - E(X)\big)\big[\rho\big(x_i - E(X)\big) + \big(z_j - E(Z)\big)\big] \\
&= \sum_{i=0}^{m}\sum_{j=0}^{n} \pi_{ij}\rho\big(x_i - E(X)\big)^2 \\
&\qquad + \sum_{i=0}^{m}\sum_{j=0}^{n} \pi_{ij}\big(x_i - E(X)\big)\big(z_j - E(Z)\big) \\
&= \rho\sum_{i=0}^{m} p_i\big(x_i - E(X)\big)^2 + \operatorname{cov}(X,Z) = \rho\sigma_X^2\,,
\end{aligned}$$

da X und Z voneinander unabhängige Zufallsvariablen sind.

Wenn wir jetzt noch annehmen, daß $\sigma_X = \sigma_Y$ ist, folgt

$$\rho(X,Y) = \frac{\rho\sigma_X^2}{\sigma_X\sigma_Y} = \rho\,,$$

wir können auf diese Weise also für beliebiges $\rho \in [0,1]$ ein Paar voneinander abhängiger Zufallsvariablen mir Korrelation ρ erzeugen.

Besser noch: Wann immer zwei Zufallsvariablen mit gleicher Standardabweichung Korrelation ρ haben, sind wir immer im obigen Fall, denn definieren wir eine neue Zufallsvariable Z durch $Z = Y - \rho X$, so ist $E(Z) = E(Y) - \rho E(X)$, und das Paar (X, Z) ist unkorreliert, da

$$\begin{aligned}
\operatorname{cov}(X,Z) &= \sum_{i=0}^{m}\sum_{j=0}^{n} \big(x_i - E(X)\big)\big(z_j - E(Z)\big) \\
&= \sum_{i=0}^{m}\sum_{j=0}^{n} \big(x_i - E(X)\big)\Big[\big(y_j - E(Y)\big) - \rho\big(x_i - E(X)\big)\Big] \\
&= \sum_{i=0}^{m}\sum_{j=0}^{n} \big(x_i - E(X)\big)\big(y_j - E(Y)\big) \\
&\qquad - \rho\sum_{i=0}^{m}\sum_{j=0}^{n} \big(x_i - E(X)\big)^2 \\
&= \operatorname{cov}(X,Y) - \rho\sigma_X^2 = \rho\sigma_X\sigma_Y - \rho\sigma_X^2 = 0\,.
\end{aligned}$$

c) Das Datenmodell

Wir modellieren Bild- und Audiodaten durch eine Folge (X_i) von Zufallsvariablen, die allesamt denselben Erwartungswert μ und dieselbe Varianz σ^2 haben. Außerdem nehmen wir noch an, daß die Korrelation zwischen X_i und X_{i+1} stets denselben Wert κ haben soll, die sogenannte *Autokorrelation* der Folge.

Bei Bildern haben wir es natürlich tatsächlich mit einer zweifach indizierten Folge (X_{ij}) zu tun; hier verlangen wir, daß sowohl die Korrelation zwischen X_{ij} und $X_{i+1,j}$ als auch die zwischen X_{ij} und $X_{i,j+1}$ gleich κ sein soll.

Um die Bedeutung der Kenngrößen μ, σ^2 und κ in der Bildverarbeitung zu veranschaulichen, sind auf der nächsten Doppelseit sechs beliebte Testbilder zusammen mit den Werten dieser Kenngrößen abgedruckt. Die Werte sind entnommen aus

P.M. FARELLE: Recursive Block Coding for Image Data Compression, *Springer,* 1990 ;

Peppers

μ = 115,6
σ^2 = 5632
σ = 75,0
ρ = 0,98
$x_{\min}$ = 0
$x_{\max}$ = 237

Lenna

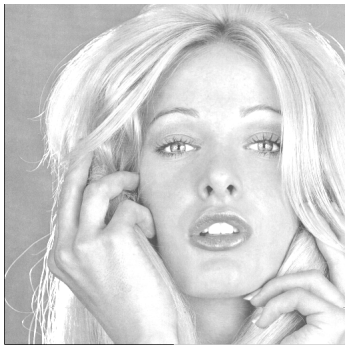
μ = 99,1
σ^2 = 2796
σ = 52,9
ρ = 0,97
$x_{\min}$ = 3
$x_{\max}$ = 248

Sailboat

μ = 124,3
σ^2 = 6027
σ = 77,6
ρ = 0,97
$x_{\min}$ = 0
$x_{\max}$ = 249

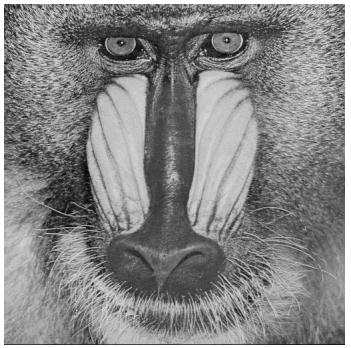
Stream

μ = 113,8
σ^2 = 2996
σ = 54,7
ρ = 0,94
$x_{\min}$ = 0
$x_{\max}$ = 255



Tiffany

μ = 208,6
σ^2 = 1126
σ = 33,6
ρ = 0,87
$x_{\min}$ = 3
$x_{\max}$ = 255



Baboon

μ = 128,9
σ^2 = 2282
σ = 47,8
ρ = 0,86
$x_{\min}$ = 0
$x_{\max}$ = 236

sie beziehen sich natürlich auf die Originalbilder und nicht auf das, was der Druckvorgang hier im Skriptum daraus gemacht hat. Trotzdem sollte der Vergleich von Bildern und Daten einen einigermaßen korrekten Eindruck zumindest der relativen Situation vermitteln, da hoffentlich alle hier abgedruckte Bilder in derselben Weise verunstaltet sind.

Die mittlere Helligkeit eines Bildes, dessen (viele) Pixel durch je eine Zufallsvariable mit Erwartungswert μ produziert werden, sollte ziemlich nahe bei μ liegen; der beste Schätzwert für den gemeinsamen Erwartungswert der Zufallsvariablen ist also die mittlere Helligkeit des Bildes. Typischerweise werden Helligkeiten durch Zahlen zwischen 0 und 255 kodiert, wobei schwarz der Zahl Null entspricht und weiß der 255. Dies sieht man gut an den Beispielbildern, wo das mit Abstand hellste Bild „Tiffany“ auch den mit Abstand größten Mittelwert μ hat; den kleinsten Wert hat das auch visuell dunkelste Bild „Lenna“.

Die nächste wichtige Kenngröße ist die Varianz, welche angibt, wie stark eine Zufallsvariable um ihren Erwartungswert streut. Auch hier wollen wir wieder davon ausgehen, daß alle Zufallsvariablen zu einem gegebenen Bild *bzw.* einer gegebenen Audiosequenz aus N dieselbe Varianz haben. Wir schätzen diese gemeinsame Varianz aufgrund der vorliegenden Daten als

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2\,,$$

wobei $y_1,\ldots,y_N$ die Helligkeits- *bzw.* Lautstärkewerte sind. (Wer sich wundert, daß vor dieser Summe mit N Summanden nur $N-1$ im Nenner steht, sollte zu §4*d)* zurückblättern.)

Was die Varianz und die Standardabweichung bedeuten, sieht man wieder deutlich an den Beispielbildern: Bilder mit geringem Kontrast wie „Tiffany“ oder „Lenna“ haben deutlich geringere Werte als die kontrastreicheren Bilder „Peppers“ und „Sailboat“.

Da der Begriff der Autokorrelation das wohl am schwersten verständliche der hier eingeführten statistischen Konzepte ist, sind die sechs Testbilder in Richtung fallender Autokorrelation geordnet: Die höchste

Autokorrelation hat mit $\rho = 0{,}98$ das Bild „Peppers“, wo die recht homogenen Flächen der Paprikaschoten dafür sorgen, daß sich ein Pixel nur selten von seinen Nachbarn unterscheidet; auch „Lenna“ und „Sailboat“ werden von flächigen Strukturen dominiert. Bei „Stream“ kommen in stärkerem Maße feine Verästelungen von Bäumen und Büschen ins Spiel, so daß die Autokorrelation auf 0,94 absinkt, und bei „Tifffany“ und „Baboon“ schließlich sorgen die vielen Haare für feine Details, die die Autokorrelation auf 0,87 *bzw.* 0,86 herunterdrücken.

d) Komprimierung durch Dekorrelation

Um zu sehen, wie sich die Autokorrelation zur Komprimierung der Daten ausnutzen läßt, betrachten wir der Einfachheit halber zunächst nur eine Folge $X_1,\ldots,X_n$ von Zufallsvariablen; der gemeinsame Erwartungswert sei μ, die gemeinsame Standardabweichung σ, und die Korrelation zwischen X_i und X_{i+1} sei jeweils ρ.

Nach dem oben Gesagten gibt es dann für jede der Zufallsvariablen X_i mit $i<n$ eine davon unabhängige Zufallsvariable Z_i, so daß

$$X_{i+1} = \rho X_i + Z_i$$

ist; entsprechend ist für $i<n-1$

$$X_{i+2} = \rho X_{i+1} + Z_{i+1} = \rho^2 X_i + \rho Z_i + Z_{i+1}$$

usw.; wenn wir zusätzlich annehmen, daß alle Z_i voneinander unabhängig sind, ist also $\rho(X_i,X_{i+2}) = \rho^2$ und allgemein

$$\rho(X_i,X_j) = \rho^{|i-j|}\,.$$

In der Signalverarbeitung spricht man bei einer solchen Folge von Zufallsvariablen von einem *autoregressive Prozeß;* der hier betrachtete allereinfachste Fall, bei dem alle Korrelationen nur von der Korrelation zwischen zwei benachbarten Zufallsvariablen abhängen, wird als AR(1)-Modell bezeichnet; in der Sprechweise der Wahrscheinlichkeitstheorie handelt es sich hier um spezielle sogenannte MARKOV-Ketten.

Der russische Mathematiker ANDREĬ ANDREEVICH MARKOV (1856–1922) studierte in Sankt Petersburg, wo er später auch Professor wurde. Er beschäftigte sich zunächst hauptsächlich mit Zahlentheorie und Analysis; erst später kommen die wahrscheinlichkeitstheoretischen Arbeiten, für die er heute vor allem bekannt ist. MARKOV-Ketten sind Prozesse ohne Erinnerung, in denen das zukünftige Verhalten nur vom augenblicklichen Zustand abhängt, nicht aber von der Geschichte des Systems. Damit sind sie gerade hier bei Bilddaten nur eine unvollkommene Approximation an die Realität, aber dennoch sehr nützlich.

Falls wir bei einer solchen Folge von Zufallsvariablen die Werte von $X_1, \ldots, X_n$ nacheinander übertragen, übertragen wir zuerst den Wert von X_1, dann mit X_2 noch einmal zu $100 \times \rho$ % denselben Wert, mit X_3 dasselbe noch einmal zu $100 \times \rho^2$ %, *usw.*

Eine offensichtliche Alternative hierzu wäre, nur den Wert von X_1 zu übertragen und ansonsten nur die Werte der Z_i. Eine ähnliche Vorgehensweise wird tatsächlich gelegentlich angewandt, allerdings macht man es sich dann noch einfacher und überträgt nur die *Differenzen*, also

$$X_1, X_2 - X_1, \ldots, X_n - X_{n-1}\,.$$

Der Nachteil dieses Verfahrens ist, daß sowohl diese Differenzen als auch die Z_i von Zeit zu Zeit sehr groß werden *müssen*, da es in fast jedem Bild oder Musikstück gelegentliche aprupte Veränderungen gibt.

Die Idee hinter allen Komprimierungsverfahren, die auf Transformationen beruhen, ist es, anstelle der Zufallsvariablen X_i geeignete Linearkombinationen

$$Y_i = \sum_{j=1}^{n} \alpha_{ij} X_j$$

zu betrachten, wobei (α_{ij}) eine *invertierbare* $n \times n$-Matrix ist, so daß sich auch umgekehrt die X_i wieder aus den Y_j rekonstruieren lassen. Diese Matrix wird so gewählt, daß die neuen Variablen möglichst unkorreliert sind und daß man die Größe der neuen Variablen möglichst gut abschätzen kann.

Letzterer Aspekt erfordert statistische Betrachtungen, auf die wir hier verzichten wollen; die Dekorrelation der Zufallsvariablen aber führt uns geradewegs zu Eigenvektoren symmetrischer Matrizen:

Wir definieren für eine Folge von Zufallsvariablen deren *Korrelationsmatrix*

$$\mathrm{Kor}(X_1, \ldots, X_n) \in \mathbb{R}^{n \times n}$$

dadurch, daß der Eintrag an der Stelle ij dieser Matrizen jeweils die Korrelation $\rho(X_i, X_j)$ sein soll.

Das Ideal, auf das wir hinarbeiten, sind Zufallsvariablen, deren Korrelationsmatrix eine Diagonalmatrix ist, denn dann sind je zwei verschiedene Variablen unkorreliert.

Da die Korrelationsmatrix eine symmetrische Matrix ist, gibt es eine Orthonormalbasis des $\mathbb{R}^n$ aus reellen Eigenvektoren, bezüglich derer sie Diagonalgestalt hat; die Vektoren dieser Orthonormalbasis seien

$$\vec{b}_1 = \begin{pmatrix} \alpha_{11} \\ \vdots \\ \alpha_{1n} \end{pmatrix}, \ldots, \vec{b}_n = \begin{pmatrix} \alpha_{n1} \\ \vdots \\ \alpha_{nn} \end{pmatrix}\,.$$

Wir definieren die neuen Zufallsvariablen durch

$$Y_i = \sum_{j=1}^{n} \alpha_{ij} X_j\,;$$

dann ist

$$\begin{aligned}
\rho(Y_i, Y_k) = \vec{v}_{Y_i} \cdot \vec{v}_{Y_k} &= \left(\sum_{j=1}^{n} \alpha_{ij} \vec{v}_{X_j} \right) \cdot \left(\sum_{\ell=1}^{n} \alpha_{k\ell} \vec{v}_{X_\ell} \right) \\
&= \sum_{j=1}^{n} \sum_{\ell=1}^{n} \alpha_{ij}\ \vec{v}_{X_j} \cdot \vec{v}_{X_\ell}\ \alpha_{k\ell} = \sum_{j=1}^{n} \sum_{\ell=1}^{n} \alpha_{ij}\ \rho(X_j, X_\ell)\ \alpha_{k\ell} \\
&= \sum_{\ell=1}^{n} \left(\sum_{j=1}^{n} \alpha_{ij}\ \rho(X_j, X_\ell) \right)\ \alpha_{k\ell}\,.
\end{aligned}$$

Der Inhalt der großen Klammer ist offensichtlich der Eintrag an der Stelle $i\ell$ der Produktmatrix $A \cdot \mathrm{Kor}(X_1, \ldots, X_n)$, wobei $A = (\alpha_{ij})$ die

Matrix der Koeffizienten α_{ij} ist, und die Summation über ℓ macht daraus den Eintrag an der Stelle ik des Produkts mit ${}^tA$. Insgesamt haben wir also gezeigt, daß

$$\mathrm{Kor}(Y_1, \ldots, Y_n) = A \cdot \mathrm{Kor}(X_1, \ldots, X_n) \cdot {}^tA$$

ist. Nun müssen wir nur noch beachten, daß die Spaltenvektoren der Matrix A als die Vektoren einer Orthonormalbasis des $\mathbb{R}^n$ gewählt waren; der Eintrag an der Stelle ij der Matrix $A\,{}^tA$ ist also das Standardskalarprodukt des i-ten und des j-ten Vektors aus einer Orthonormalbasis und somit null für $i \neq j$ und eins für $i = j$. Daher ist $A \cdot {}^tA = E$, also ${}^tA^{-1}$ und somit auch

$$\mathrm{Kor}(Y_1, \ldots, Y_n) = A \cdot \mathrm{Kor}(X_1, \ldots, X_n) \cdot A^{-1}\,.$$

Damit ist $\mathrm{Kor}(Y_1, \ldots, Y_n)$ eine Diagonalmatrix, denn für jede Matrix $B \in \mathbb{R}^{n\times n}$ ist ABA^{-1} die Matrix B bezüglich der Basis aus den Spaltenvektoren vom A. Diese Basis besteht hier aber aus lauter Eigenvektoren der Korrelationsmatrix, die transformierte Matrix ist also eine Diagonalmatrix.

Unter den Annahmen unseres statistischen Modells können wir also jede Folge von Zufallsvariablen durch eine lineare Transformation in eine Folge unkorrelierter Zufallsvariablen überführen. Diese Transformation bezeichnet man, obwohl sie zuerst von HOTELLING vorgeschlagen wurde, als KARHUNEN-LOÈVE-Transformation.

HAROLD HOTELLING (1895–1973) war ein amerikanischer Statistiker und Ökonom; er lehrte an der Columbia University und der University of North Carolina. In einer 1933 veröffentlichten Arbeit im *Journal of Educational Psychology* schlug er erstmalig diese Transformation vor, die von Statistikern heute in Anlehnung an den Titel seiner Arbeit meist als *Hauptkomponentenanalyse* bezeichnet wird. In Europa erschien die Transformation fast gleichzeitig um 1947 *bzw.* 1948 in wahrscheinlichkeitstheoretischen Arbeiten des Finnen KARI KARHUNEN (*1915) und des Franzosen MICHEL LOÈVE (1907–1979), nach denen sie in der technischen Literatur benannt wird.

Die Matrix A der linearen Transformation hängt nur von ρ ab und kann daher für gängige Werte von ρ vorberechnet werden; die KARHUNEN-LOÈVE-Transformation ist also einfach die Multiplikation mit einer bekannten Matrix.

e) Die diskrete Cosinus-Transformation

Für die Multiplikation zweier $n \times n$-Matrizen benötigt man allerdings n^3 Multiplikationen und noch einmal $n^2(n-1)$ Additionen; der Aufwand steigt mit großem n also sehr stark an. In der Praxis gibt man sich daher mit einen Kompromiß zufrieden und zerlegt eine Folge von Zufallszahlen in kurze Teilsequenzen, die bei eindimensionalen Folgen typischerweise die Länge 8 haben; dies ist beispielsweise der Standard bei Musik-CDs.

Allerdings wird weder bei Musik-CDs noch sonstwo die KARHUNEN-LOÈVE-Transformation wirklich angewandt. Der Grund liegt an der Struktur der Eigenvektoren der Korrelationsmatrix: Betrachten wir etwa als typisches Beispiel den $n = 8$; dann haben wir die Matrix

$$\mathrm{Cov}(X_1, \ldots, X_8) = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 & \rho^7 \\ \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^7 & \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}.$$

Ihre Eigenwerte für $\rho = 0{,}95$ können zumindest näherungsweise berechnet werden, und auch die Eigenvektoren lassen sich bestimmen. Diese sollen hier jedoch nicht numerisch angegeben werden: Eine Folge von acht reellen Zahlen ist schließlich im allgemeinen eher unanschaulich. Stattdessen sind in den Abbildungen 70 bis 77 die Eigenvektoren *graphisch* dargestellt, wobei einem Vektor

$$(a_1, \ldots, a_8) \in \mathbb{R}^8$$

die acht Striche vom Punkt $(i, 0)$ bis (i, a_i) in der Ebenen entsprechen sollen. Zusätzlich ist in jedes dieser Diagramme noch eine der Kurven

$$y = \cos\left(\frac{(2x-1)(j-1)\pi}{16}\right)$$

für $j = 1, \ldots, 8$ eingezeichnet; wie man sieht, lassen sich die Komponenten der Eigenvektoren sehr gut durch diese Cosinuswerte annähern. Dies gilt nicht nur für den speziellen Wert $\rho = 0{,}95$, sondern für jeden Wert von ρ, der hinreichend nahe bei eins liegt.

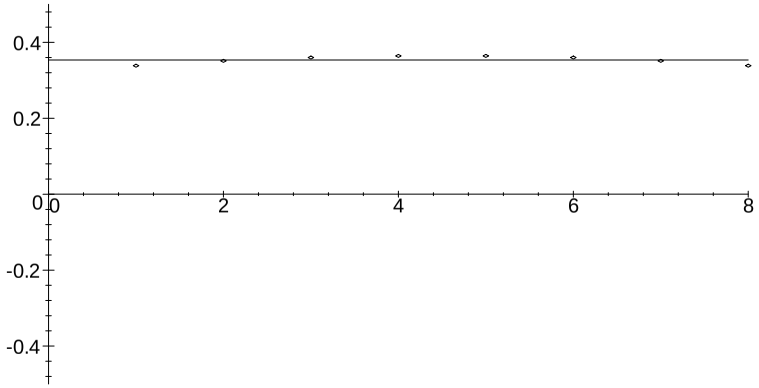


Abb. 70: Der erste Eigenvektor der Korrelationsmatrix

Aus diesem Grund arbeitet man in der Praxis lieber mit den Cosinuswerten; der Basiswechsel hin zur Basis aus den Cosinusvektoren bezeichnet man als *diskrete Cosinustransformation.* Ihr Hauptvorteil gegenüber der KARHUNEN-LOÈVE-Transformation ist, daß sie durch einen schnellen Algorithmus berechnet werden kann, der anstelle des Aufwands n^3 für eine Matrixmultiplikation nur den Aufwand $n^2 \log n$ hat. Für Einzelheiten sei auf die Vorlesung *Numerik I* verwiesen.

Die diskrete Cosinustransformation ist Teil fast aller gängiger Normen zur Bildkomprimierung: Sowohl der JPEG-Standard für Photographien, die Standards MPEG 1 und 2 für digitale (Unterhaltungs-)Videos als

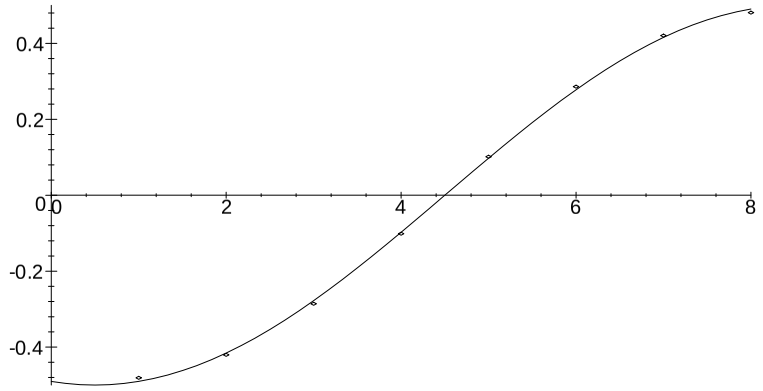


Abb. 71: Der zweite Eigenvektor der Korrelationsmatrix

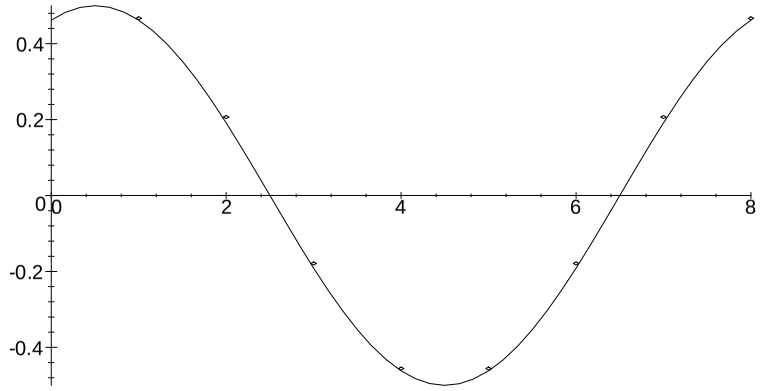


Abb. 72: Der dritte Eigenvektor der Korrelationsmatrix

auch der Standard CCITT H.261 für Videokonferenzen enthalten (neben anderen Bestandteilen) jeweils eine diskrete Cosinustransformation. Auch bei Audio-CDs ist sie ein Teil der Codierung.

Die Transformation allein ist natürlich noch keine Komprimierung: Schließlich haben wir nur einen Vektor in einer anderen Basis hingeschrieben, und die Anzahl der reellen Zahlen, die man zur Beschrei-

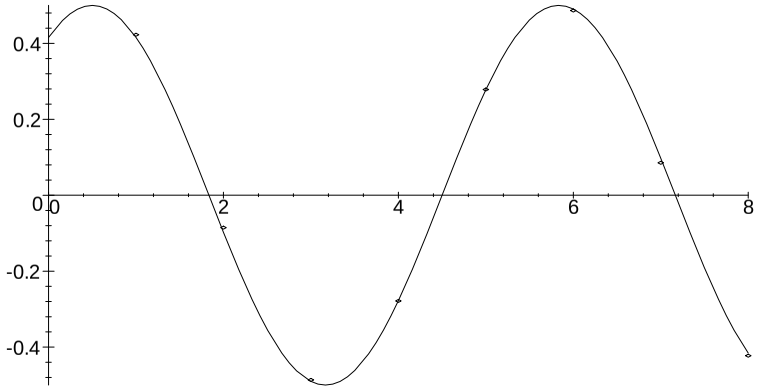


Abb. 73: Der vierte Eigenvektor der Korrelationsmatrix

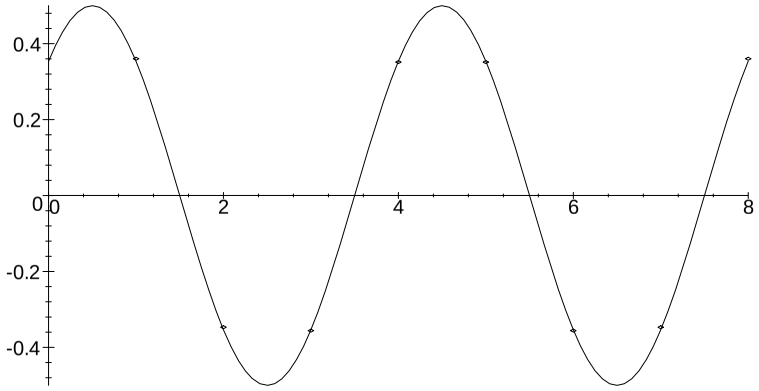


Abb. 74: Der fünfte Eigenvektor der Korrelationsmatrix

bung eines solchen Vektors benötigt, ist unabhängig von der Basis. Der wesentliche Vorteil der neuen Basis ist, daß man statistisch recht gute Aussagen über die Größe der Komponenten machen können. Hier wollen wir auf exakte statistische Berechnungen verzichten und stattdessen informell diskutieren, warum dies der Fall sein könnte.

Wie die Abbildungen der Basisvektoren zur KARHUNEN-LOÈVE-Trans-

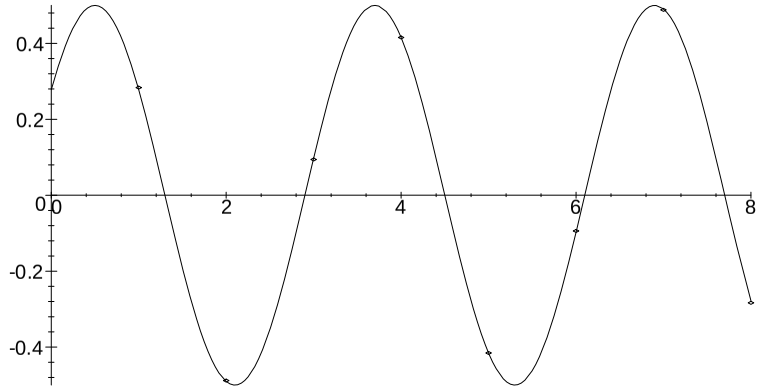


Abb. 75: Der sechste Eigenvektor der Korrelationsmatrix

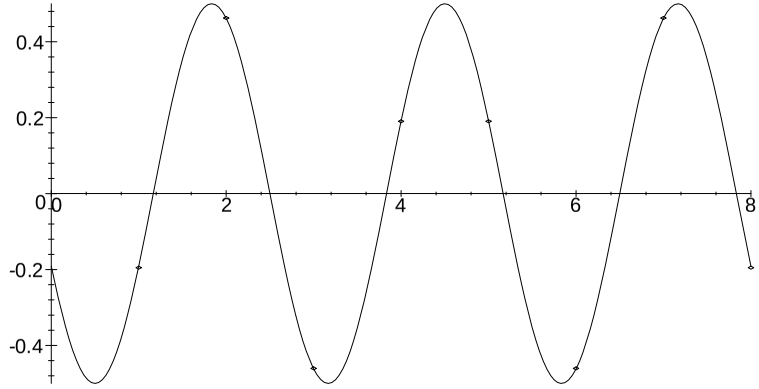


Abb. 76: Der siebte Eigenvektor der Korrelationsmatrix

formation und die Formeln für die Basisvektoren zur diskreten Cosinustransformation zeigen, werden die Basisvektoren, wenn man sie in der hier angegebenen Reihenfolge betrachtet, immer hochfrequenter. In einem hinreichend fein abgetasteten Bild oder Audiosignal erwarten wir, daß hochfrequente Schwankungen keine große Rolle spielen und somit die entsprechenden Basisvektoren nur kleine Koeffizienten haben

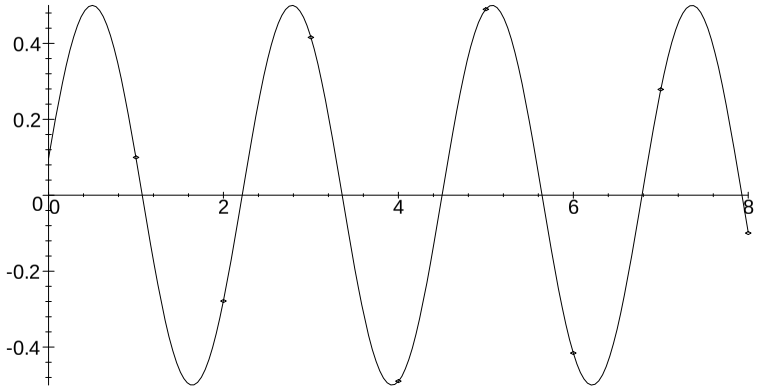


Abb. 77: Der achte Eigenvektor der Korrelationsmatrix

oder in vielen Fällen sogar gleich gar nicht auftreten. Dementsprechend genügt es, für die Übertragung dieser Koeffizienten nur wenige Bits bereitzustellen; bei nur geringen Abstrichen an die Qualität kann man auf gewisse Koeffizienten sogar ganz verzichten.

Ein Kompressionsverfahren wird daher, je nach Anspruch an die Qualität, entweder alle Koeffizienten des Signals in der neuen Basis übertragen und durch eine geeignete Darstellung der Daten dafür sorgen, daß Folgen von Nullen nur wenig Platz benötigen, oder aber es wird nur eine Auswahl der Koeffizienten übertragen und auch für diese jeweils festlegen, wie viele Bit dafür in Anspruch genommen werden. Diese Anzahl wird umso geringer sein, je höher die Frequenz des jeweiligen Basisvektors ist; bei einigen Verfahren wie etwa JPEG können die Anzahlen auch variablel in Abhängigkeit von einer Qualitätszahl gewählt werden.

Zum Schluß sei noch ganz kurz erwähnt, daß die KARHUNEN-LOÈVE-Transformation und damit (mit ganz geringen Abstrichen) auch die diskrete Cosinustransformation zwar die Korrelationsmatrix in optimaler Weise diagonalisieren, daß aber daraus nicht folgt, daß sie auch optimale Kompressionsverfahren liefern: Ausßer der Kovarianz gibt es noch weitere Quellen für Redundanz eines Bildes.

Ein gewisser Nachteil der Cosinustransformation ist außerdem, daß man für aprupte Übergänge, wie sie etwa bei Kanten immer wieder einmal auftauchen, die hochfrequenten Basisvektoren braucht, die dann aber nicht nur die Kante selbst beeinflussen, sondern das gesamte Quadrat, auf das die Transformation angewandt wird.

Eine bessere Möglichkeit wäre es daher, wenn man anstelle von Cosinusfunktionen Funktionen verwenden könnte, die sowohl im Zeit- als auch im Frequenzbereich lokalisiert sind. Solche Funktionen gibt es in der Tat, etwa die sogenannten *Wavelets*. Hierbei handelt es sich um schnell abklingende Wellen, und neuere Arbeiten deuten darauf hin, daß diese für gewisse Bildmodelle (die im Gegensatz zum hier betrachteten nicht mit Wahrscheinlichkeiten arbeiten) nicht zu weit vom Optimum entfernt sein sollten. Im Rahmen dieser Vorlesung ist es jedoch zeitlich weder möglich, auf diese Modelle einzugehen, noch ist an eine genauere Behandlung von Wavelets zu denken.

Einen allgemein verständlichen Überblick über Wavelets findet man etwa bei

BARBARA BURKE HUBBARD: Wavelets: Die Mathematik der kleinen Wellen, *Birkhäuser* 1997;

das zitierte Optimalitätsresultat ist beschrieben im Vortrag

STÉPHANE MALLAT: Applied Mathematics meets signal processing

auf dem Internationalen Mathematikerkongress 1998 in Berlin, nachzulesen in Band I der Proceedings, S. 319–338, oder unter http://www.mathematik.uni-bielefeld.de/documenta/xvol-icm/00/Mallat.MAN.html.

Eine für Technische Informatiker gut geeignete fundierte Einführung in diesen Themenkreis ist etwa

STÉPHANE MALLAT: A wavelet tour of signal processing, *Academic Press, 1998*.

$\mathcal{E}\ \mathcal{N}\ \mathcal{D}\ \mathcal{E}$

$\mathcal{S}\ \mathcal{C}\ \mathcal{H}\ \ddot{\mathcal{O}}\ \mathcal{N}\ \mathcal{E}\quad \mathcal{F}\ \mathcal{E}\ \mathcal{R}\ \mathcal{I}\ \mathcal{E}\ \mathcal{N}\ !$