Wolfgang K. Seiler

Höhere Mathematik I

Vorlesung an der Universität Mannheim
im Sommersemester 1997

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 Ihre Übungsgruppenleiterin. 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.

Literaturhinweise 1

KAPITEL III: DIFFERENTIAL- UND INTEGRALRECHNUNG
MEHRERER VERÄNDERLICHER 4

§1: Skalarwertige Funktionen mehrer Veränderlicher 4

§2: Fehler- und Ausgleichsrechnung 35

a) Fehlerquellen und statistische Kenngrößen 35

b) Das LAPLACEsche Fehlermodell 38

c) Das Fehlerfortpflanzungsgesetz 49

d) Die Standardabweichung des Mittelwerts und die Schätzung der Varianz 51

e) Die *Maximum likelihood* Methode und die Methode der kleinsten Quadrate 53

§3: Maxima und Minima unter Nebenbedingungen 62

a) Eine Kettenregel 63

b) Der Satz über implizite Funktionen 65

c) LAGRANGEsche Multiplikatoren 69

§4: Vektorwertige Funktionen mehrer Veränderlicher 77

a) Der Begriff des Vektorfelds 79

b) Die JACOBI-Matrix 80

c) Die Divergenz eines Vektorfelds 81

d) Die Rotation eines dreidimensionalen Vektorfelds 86

e) Erste Beispiele 91
1) Das elektrische Feld einer Punktladung 91
2) Das Magnetfeld eines stromdurchflossenen Leiters 93
f) Allgemeine Rechenregeln 96
g) Polarkoordinaten, Kugelkoordinaten und Zylinderkoordinaten . 99
1) Polarkoordinaten in $\mathbb{R}^2$ 99
2) Zylinderkoordinaten im $\mathbb{R}^3$ 101
3) Kugelkoordinaten 102
§5: Kurvenintegrale im $\mathbb{R}^n$ 103
a) Kurvenstücke und Kurven 104
b) Die Bogenlänge einer Kurve 106
c) Integration eines Vektorfelds längs einer Kurve 108
d) Zirkulationsfreie und konservative Vektorfelder 113
§6: Mehrdimensionale Integrationstheorie 118
a) Flächeninhalte und Volumina 118
b) Integration über Normalbereiche und der Satz von Fubbini 126

Literaturhinweise

Das Angebot von Lehrbüchern zum Thema „Höhere Mathematik“ ist riesig; die hier zusammengestellte kleine Auswahl besteht aus Büchern, die erstens in der Mathematischen Bibliothek zu finden sind (größtenteils im allgemeinen Bestand, teilweise auch in der Lehrbuchsammlung) und die zweitens von der Stoffauswahl her ungefähr dem Inhalt dieser Vorlesung entsprechen – auch wenn sie meist erheblich mehr Stoff enthalten, als in der Vorlesung behandelt werden kann. Am nützlichsten dürften wohl die beiden erstgenannten Bücher **[D]** und **[MV]** sein; bei Rückverweisen auf den Stoff der Analysis I werde ich außer dem Skriptum

[Bö] S. Böcherer: *Analysis I,* Universität Mannheim, 1996/97

jeweils auch diese Bücher zitieren.

[D] H.J. Dirschmid: *Mathematische Grundlagen der Elektrotechnik,* Vieweg, 1992

Etwa fünf Pfund Mathematik, die nicht nur diese Vorlesung mehr als abdecken, sondern für viele Studenten für deren gesamtes Berufsleben ausreichen dürften. Der Schwerpunkt liegt allerdings eindeutig auf dem Gebiet der Analysis; numerische Mathematik und die Statistik sind so gut wie nicht vorhanden.

[MV] K. Meyberg, P. Vachenauer: *Höhere Mathematik I,* Springer, 1995

Erster Band eines zweibändigen Werks; enthält die für die *Höhere Mathematik* relevanten Teile der *Analysis I* sowie den gesamten Stoff der Vorlesung *Höhere Mathematik I* plus etwas Numerik in recht kompakter Darstellung mit fast vollständigen Beweisen. Für einige Grundalgorithmen sind auch Programme angegeben.

[BDH] BRAUCH/DREYER/HAACKE: *Mathematik für Ingenieure Maschinenbau/Elektrotechnik,* Teubner 1977

Dieses Buch richtet sich an Studenten von Fachhochschulen, deckt aber auch einen großen Teil dieser Vorlesung ab; für uns zu knapp ist vor allem die Darstellung der Vektoranalysis. Dafür gibt es Kapitel, die auch Themen wie Statistik, Fehlerrechnung, *usw.* darstellen.

Dazu seien noch einige „Klassiker" genannt, die seit mehreren Jahrzehnten in immer neuen Auflagen erscheinen und die auch heute noch interessant sind. Es handelt sich um Werke aus meist recht vielen Bänden, wobei selbst der Stoff der Vorlesung *Höhere Mathematik I* je nach Organisation des Gesamtwerks auf bis zu drei Bände verteilt sein kann. Wegen der Vielzahl von Auflagen und Bänden verzichte ich auf die Angabe von Erscheinungsjahren.

[S] W.I. SMIRNOW: *Lehrbuch der Höheren Mathematik,* VEB Deutscher Verlag der Wissenschaften

Ein Klassiker, nach dem Generationen von russischen und (nicht nur ost-)deutschen Naturwissenschaftlern und Ingenieuren ausgebildet wurden; enthält in vier Bänden (von denen die letzten beiden noch in Halbbände unterteilt sind) den gesamten klassischen Stoff der *Höheren Mathematik* und ist auch heute noch sehr gut zu lesen. Die HM I behandelt Stoff aus den Bänden I, II und III/1.

[R] R. ROTHE: *Höhere Mathematik,* Teubner

Sieben (dünne) Bände, zu denen allerdings auch Aufgaben- und Formelsammlung gehören, so daß die Darstellung insgesamt eher knapp ist. Für diese Vorlesung relevant sind die Bände II und III.

[Du] A. DUSCHEK: *Höhere Mathematik,* Springer Wien

Die österreichische Variante, vier recht ausführliche Bände, von denen hier vor allem die ersten beiden von Interesse sind.

[A] G. AUMANN: *Höhere Mathematik,* Bibliographisches Institut Mannheim

Drei relativ dünne Taschenbücher, deren erste beide trotzdem fast alles enthalten, was wir in dieser Vorlesung brauchen. Die Darstellung ist natürlich weniger ausführlich als in den dickleibigen Werken, aber das wird nicht jeder Student als Nachteil empfinden.

[C] R. COURANT: *Vorlesungen über Differential- und Integralrechnung I+II,* Springer

Ein Klassiker eines berühmten Mathematikers, auch heute noch sehr lesenswert. Trotz des Titels beschränkt sich das Buch nicht auf Analysis, sondern behandelt auch beispielsweise die Lineare Algebra in einem Umfang, der für diese Vorlesung völlig ausreicht.

Kapitel 1
Einführung in die Integralrechnung

§1: Heuristische Vorüberlegungen

Die Integralrechnung ist neben der Differentialrechnung die zweite wichtige Säule der Analysis. Seiner großen Bedeutung entsprechend, wollen wir uns dem Integralbegriff zunächst heuristisch nähern, bevor wir ihn – mit beträchtlichem technischem Aufwand – im nächsten Paragraphen formal einführen.

Die Integration dient primär drei Zwecken:
- Sie dient zur Umkehrung der Differentiation.
- Sie dient zur Flächenberechnung.
- Sie dient zur Durchschnittsberechnung.

Wie wollen uns alle drei Aspekte kurz ansehen.

a) Integration als Umkehrung der Differentiation

Das klassische Beispiel zur Einführung der Differentiation ist die Geschwindigkeit als Ableitung des Wegs nach der Zeit. Es liegt daher nahe, zur Einführung des Integrals die Frage nach der Berechnung des Wegs aus der Geschwindigkeit zu betrachten.

Wir nehmen also an, ein Fahrzeug sei zur Zeit $t = 0$ an der Stelle $s = 0$ und beschleunige in fünf Sekunden auf die im Stadtverkehr zulässige Höchstgeschwindigkeit von 13,9 m/sec. Welchen Weg hat es bis dahin zurückgelegt?

Die Geschwindigkeit sei durch folgende Tabelle gegeben:

t [sec]	0	1	2	3	4	5
v [m/sec]	2,5	6,1	9,1	11,5	13,4	13,9

Ist $s(t)$ der bis zum Zeitpunkt t zurückgelegte Weg und $v(t)$ die dann erreichte Geschwindigkeit, so ist bekanntlich

$$v(t) = \frac{ds}{dt} \approx \frac{\Delta s}{\Delta t} = \frac{s(t+1\text{ sec}) - s(t)}{(t+1\text{ sec}) - t} = \frac{s(t+1\text{ sec}) - s(t)}{1\text{ sec}} .$$

Also ist

$$s(t+1\text{ sec}) \approx s(t) + v(t)\cdot 1\text{ sec} .$$

Da wir sowohl $s(0\text{ sec}) = 0\text{ m}$ als auch die Werte $v(t)$ für $t = 0, 1, 2, 3$ und 4 sec kennen, können wir $s(5\text{ sec})$ rekursiv berechnen als

$$\begin{aligned} s(5\text{ sec}) &\approx v(0\text{ sec})\cdot 1\text{ sec} + v(1\text{ sec})\cdot 1\text{ sec} + v(2\text{ sec})\cdot 1\text{ sec} \\ &\quad + v(3\text{ sec})\cdot 1\text{ sec} + v(4\text{ sec})\cdot 1\text{ sec} \\ &= 2{,}5\text{ m} + 6{,}1\text{ m} + 9{,}1\text{ m} + 11{,}5\text{ m} + 13{,}4\text{ m} = 42{,}6\text{ m} . \end{aligned}$$

Also hat das Fahrzeug in fünf Sekunden 42,6 m zurückgelegt?

Zumindest mit dieser Genauigkeit ist das sicherlich falsch, denn bei der Berechnung des Wegs sind wir davon ausgegangen, daß das Fahrzeug während der ersten Sekunde konstant mit einer Geschwindigkeit von 2,5 m/sec gefahren ist, exakt ab deren Ende dann aber plötzlich eine Sekunde lang mit 6,1 m/sec, *usw.* Das ist natürlich unrealistisch; tatsächlich dürfte die Geschwindigkeit etwa so verlaufen sein, wie es die Kurve in Abbildung 1 angibt, wohingegen wir mit dem ebenfalls eingezeichneten stufenförmigen Verlauf gerechnet haben. Wir haben die Geschwindigkeit somit fast durchgängig unterschätzt und damit auch einen zu kleinen Weg berechnet.

Alternativ hätten wir auch einen *zu großen* Weg schätzen können, wenn wir die Geschwindigkeit während eines Sekundenintervalls jeweils auf den bekannten Wert am *Ende* dieses Intervalls gesetzt hätten; das Ergebnis dann wäre gewesen

$$\begin{aligned} s(5\text{ sec}) &\approx v(1\text{ sec})\cdot 1\text{ sec} + v(2\text{ sec})\cdot 1\text{ sec} + v(3\text{ sec})\cdot 1\text{ sec} \\ &\quad + v(4\text{ sec})\cdot 1\text{ sec} + v(5\text{ sec})\cdot 1\text{ sec} \\ &= 6{,}1\text{ m} + 9{,}1\text{ m} + 11{,}5\text{ m} + 13{,}4\text{ m} + 13{,}9\text{ m} = 54{,}0\text{ m} . \end{aligned}$$

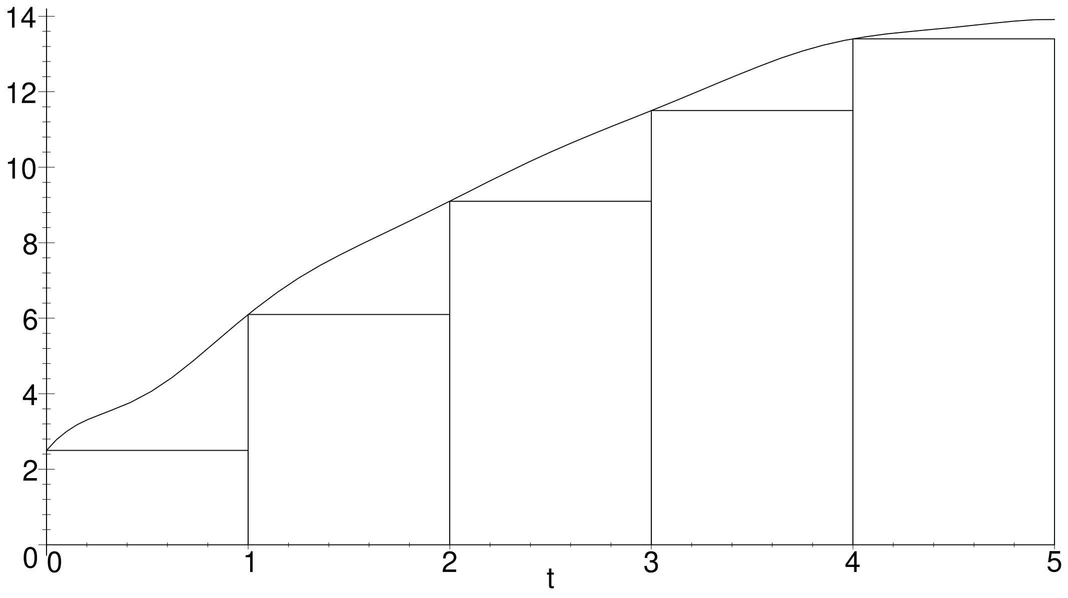


Abb. 1: Tatsächliche und unterschätzte Geschwindigkeit

Tatsächlich wissen wir im Augenblick also nur, daß der zurückgelegte Weg irgendwo zwischen 42,6 m und 54 m liegt.

Eine bessere Schätzung bekommen wir, wenn wir anstelle der Sekundenintervalle Intervalle von nur einer halben Sekunde betrachtet – vorausgesetzt natürlich, wir kennen die Geschwindigkeit auch für halbzahlige Sekundenwerte. Die entsprechenden Werte seien die in der folgenden Tabelle angegebenen:

t [sec]	0	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5
v [m/sec]	2,5	4,4	6,1	7,8	9,1	10,4	11,5	12,6	13,4	13,8	13,9

Jetzt ist die „unterschätzte“ Wegstrecke

$$\sum_{i=0}^{9} v\left(\frac{i}{2}\,\mathrm{sec}\right)\cdot\frac{1}{2}\,\mathrm{sec} = 45{,}6\ \mathrm{m}\,,$$

und die „überschätzte“ ist

$$\sum_{i=1}^{10} v\left(\frac{i}{2}\,\mathrm{sec}\right)\cdot\frac{1}{2}\,\mathrm{sec} = 51{,}3\ \mathrm{m}\,,$$

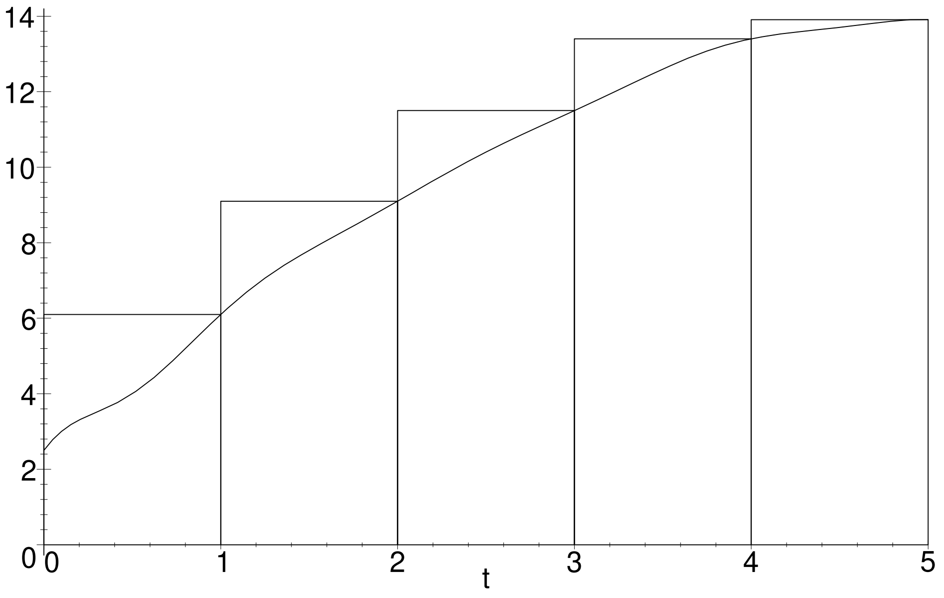


Abb. 2: Tatsächliche und überschätzte Geschwindigkeit

die Unsicherheit hat sich also etwa halbiert. Abbildung 3 zeigt den Verlauf von unterschätzter, tatsächlicher und überschätzter Geschwindigkeit für die Halbsekundenintervalle.

Beide Werte wie auch ihre Differenz lassen sich anhand von Abbildung 3 leicht geometrisch veranschaulichen: Die unterschätzte Geschwindigkeit ist die Fläche der Rechtecke unterhalb der unteren Treppenfunktion, die überschätzte entsprechend die Fläche der Rechtecke unterhalb der oberen Treppenfunktion, wobei die Basis der Rechtecke jeweils auf der Zeitachse liegt. Die Differenz ist somit gleich der Fläche der Differenzrechtecke.

Diese können wir durch weitere Verfeinerung der Abtastung verkleinern. Dazu nehmen wir an, die eingezeichnete Kurve (die tatsächlich einfach eine notdürftig angepaßte Polynomfunktion darstellt) gebe den tatsächlichen Geschwindigkeitsverlauf wieder, und lassen den Computer rechnen:

Wenn wir die Geschwindigkeit viermal pro Sekunde bestimmen und den

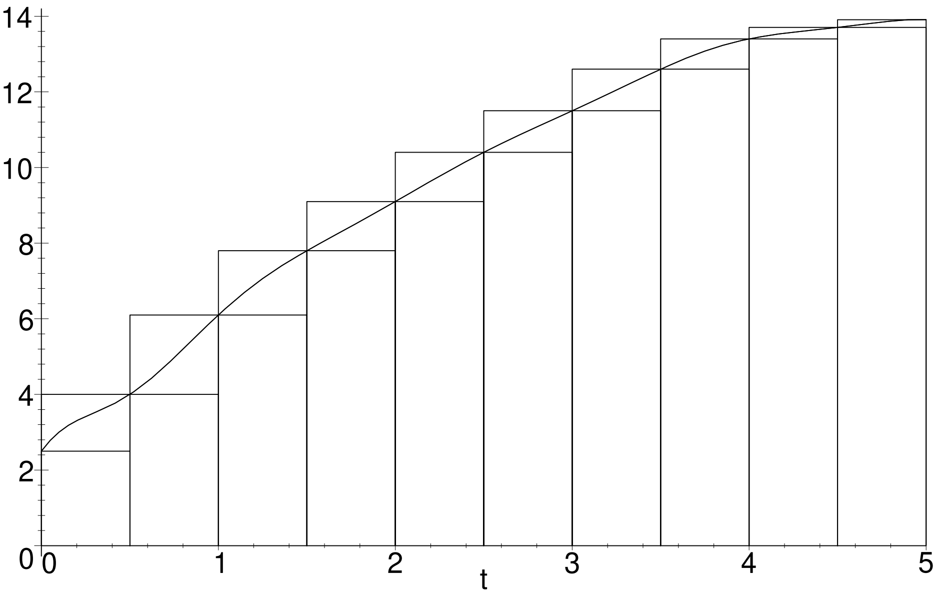


Abb. 3: Abtastung der Geschwindigkeit in Intervallen von einer halben Sekunde

zurückgelegten Weg damit schätzen, erhalten wir als untere Schranke

$$\sum_{i=0}^{19} v\left(\frac{i}{4}\text{ sec}\right)\cdot\frac{1}{4}\text{ sec}\approx 47{,}14\text{ m}$$

und als obere Schranke

$$\sum_{i=1}^{20} v\left(\frac{i}{4}\text{ sec}\right)\cdot\frac{1}{4}\text{ sec}\approx 49{,}99\text{ m}\,.$$

Mit zehn Geschwindigkeitswerten pro Sekunde verbessert sich die untere Schranke auf

$$\sum_{i=0}^{49} v\left(\frac{i}{10}\text{ sec}\right)\cdot\frac{1}{10}\text{ sec}\approx 48{,}02\text{ m}$$

und die obere auf

$$\sum_{i=1}^{50} v\left(\frac{i}{10}\text{ sec}\right)\cdot\frac{1}{10}\text{ sec}\approx 49{,}16\text{ m}\,.$$

Wie Abbildung 4 zeigt, unterscheiden sich nun die Rechtecke der unteren Abschätzung nur noch wenig von denen der oberen, und die Fläche unter der Geschwindigkeitskurve liegt schon recht nahe bei der Fläche der Rechtecke aus jeder der beiden Abschätzungen.

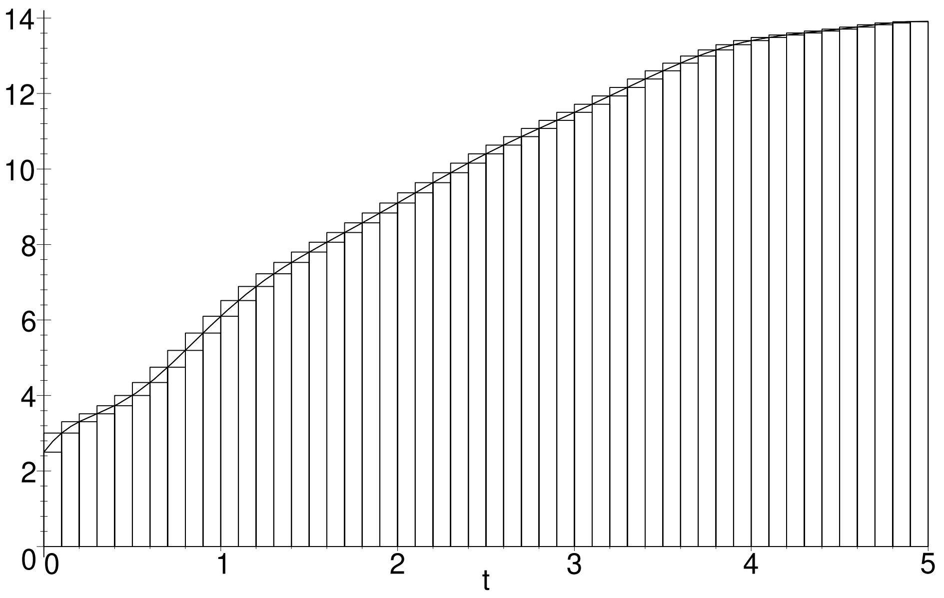


Abb. 4: Abtasten der Geschwindigkeit in Intervallen von einer Zehntelsekunde

Gehen wir von den Zehntelsekunden zu den Hundertsteln, sind die Rechtecke zumindest visuell in Abbildung 5 nicht mehr voneinander und von der Fläche unter der Kurve zu unterscheiden: Daß man hier keine durchweg schwarze Fläche sieht, liegt ausschließlich an sogenannten *alias*-Effekten, d.h. Diskretisierungsfehlern bei der Rastergraphik.

Rechnerisch gibt es noch einen kleinen Unterschied im Dezimeterbereich:

$$\sum_{i=0}^{499} v\left(\frac{i}{100}\,\text{sec}\right)\cdot\frac{1}{100}\,\text{sec}\approx 48{,}54\text{ m}$$

und

$$\sum_{i=1}^{500} v\left(\frac{i}{100}\,\text{sec}\right)\cdot\frac{1}{100}\,\text{sec}\approx 48{,}66\text{ m}\,.$$

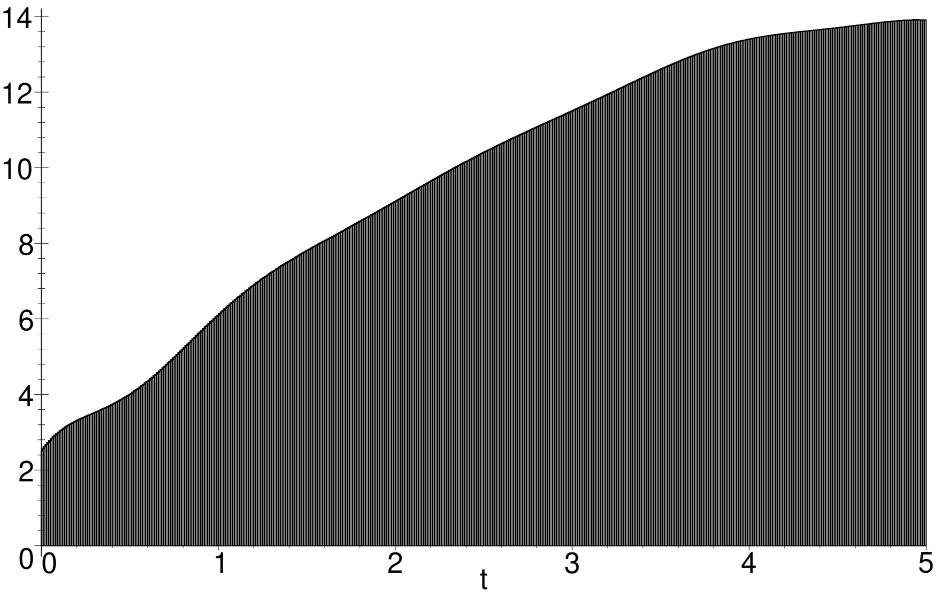


Abb. 5: Abtastung der Geschwindigkeit in Intervallen von einer Hundertstelsekunde

Bei einer nochmaligen Verzehnfachung der Intervallanzahl ist natürlich graphisch nichts neues mehr zu sehen; die Schätzwerte verbessern sich auf

$$\sum_{i=0}^{4999} v\left(\frac{i}{1000}\ \mathrm{sec}\right)\cdot\frac{1}{1000}\ \mathrm{sec}\approx 48{,}595\ \mathrm{m}$$

für die untere Schranke und

$$\sum_{i=1}^{5000} v\left(\frac{i}{1000}\ \mathrm{sec}\right)\cdot\frac{1}{1000}\ \mathrm{sec}\approx 48{,}607\ \mathrm{m}$$

für die obere.

Um die eingangs gestellte Frage nach dem zurückgelegten Weg millimetergenau beantworten zu können (sofern man diese Genauigkeit wirklich als sinnvoll betrachtet), muß die Geschwindigkeitskurve 10 000 Mal pro Sekunde abgetastet werden, und wir erhalten die Schranken

$$\sum_{i=0}^{49999} v\left(\frac{i}{10000}\ \mathrm{sec}\right)\cdot\frac{1}{10000}\ \mathrm{sec}\approx 48{,}6006\ \mathrm{m}$$

und

$$\sum_{i=1}^{50000} v\left(\frac{i}{10000}\right) \sec \cdot \frac{1}{10000} \sec \approx 48{,}6017 \text{ m};$$

das Fahrzeug legte in fünf Sekunden also 48 m und 60,1 cm zurück.

b) Integration als Flächenbestimmung

Die Abbildungen 4 und vor allem 5 zeigen, daß die immer feiner werdenden Rechtecke, die wir für die obigen Abschätzungen wählten, die Fläche unter der Kurve $v = v(t)$ immer besser annähern. Wenn wir die Einheiten vergessen und t wie auch v als Längen ansehen, können wir die oben betrachteten Ausdrücke

$$\sum_{i=0}^{5n-1} v\left(\frac{i}{n}\right) \cdot \frac{1}{n} \quad \text{und} \quad \sum_{i=1}^{5n} v\left(\frac{i}{n}\right) \cdot \frac{1}{n}$$

also auch als Näherungswerte für die Fläche unter der Kurve ansehen und den Flächeninhalt als ihren gemeinsamen Grenzwert für immer größer werdendes n berechnen.

Genau genommen handelt es sich hier allerdings nicht um die *Berechnung* dieser Fläche, sondern um die *Definition* des Flächeninhalts, denn es gibt schließlich *a priori* keinen Begriff des Flächeninhalts einer krummlinig begrenzten Fläche. Die hier gewählte Definition über eine Ausschöpfung mit immer feiner werdenden Rechtecken ist zwar sehr natürlich, aber nicht zwangsläufig. Sie ist übrigens weit älter als jeder Begriff eines Integrals oder auch nur Grenzwerts: Bereits vor über zwei Jahrtausenden, um etwa 370 vor Christus, definierte der griechische Mathematiker EUDOXOS VON CNIDOS (*ca.* 408–*ca.* 355), ein Schüler und späterer Konkurrent PLATOS, Flächeninhalte auf diese Weise; später hat ARCHIMEDES VON SYRAKUS (287–212) diese Methode perfektioniert und angewandt auf Kreise und Parabeln; insbesondere berechnete er damit seine Abschätzung $223/71 < \pi < 22/7$ für π, in Dezimalzahlen ausgedrückt $3{,}1408 < \pi < 3{,}1429$, wobei diese Abschätzung allerdings nicht auf Rechtecken beruht, sondern auf der Konstruktion des regelmäßigen 96-Ecks.

c) Integration als Durchschnittsbestimmung

Kehren wir wieder zurück zum Eingangsbeispiel eines sich beschleunigenden Fahrzeugs, und fragen wir uns, wie hoch die *Durchschnittsgeschwindigkeit* war. Bei endlich vielen Werten ist der Durchschnitt natürlich einfach das arithmetische Mittel, also z.B.

$$\begin{aligned} &\frac{1}{5}\cdot\big(v(0\text{ sec})+v(1\text{ sec})+v(2\text{ sec})+v(3\text{ sec})+v(4\text{ sec})\big)\\ =&\frac{1}{5}\cdot 42{,}6\text{ m/sec} = 8{,}52\text{ m/sec}\,, \end{aligned}$$

wenn wir die Geschwindigkeiten vom *Anfang* jedes Sekundenintervalls nehmen, und

$$\begin{aligned} &\frac{1}{5}\cdot\big(v(1\text{ sec})+v(2\text{ sec})+v(3\text{ sec})+v(4\text{ sec})+v(5\text{ sec})\big)\\ =&\frac{1}{5}\cdot 54{,}0\text{ m/sec} = 10{,}8\text{ m/sec}\,, \end{aligned}$$

wenn wir die vom Ende nehmen.

Natürlich läßt sich auch hier die Abtastung immer weiter verfeinern; die entstehenden Ausdrücke

$$\frac{1}{5n}\cdot\sum_{i=0}^{5n-1} v\left(\frac{i}{n}\text{ sec}\right) \quad\text{und}\quad \frac{1}{5n}\cdot\sum_{i=1}^{5n} v\left(\frac{i}{n}\text{ sec}\right)$$

entsprechen bis auf einen Faktor von $\frac{1}{5\text{ sec}}$ genau denen aus Teil *a),* wo wir den Weg abgeschätzt haben – wie es ja auch in der Tat der Fall sein muß: Die Durchschnittsgeschwindigkeit ist schließlich nichts anderes als der zurückgelegte Weg dividiert durch die zugrundeliegende Zeitspanne von fünf Sekunden.

Trotzdem ist die Interpretation eines Integrals als Durchschnitt gelegentlich von unabhängigem Interesse, da vor allem bei Anwendungen in der Quantenphysik oft zwar Durchschnitte existieren, aber keine „Zähler“ und „Nenner“, als deren Quotienten man sie interpretieren könnte.

d) Integration elementarer Funktionen

Abstrahieren wir vom Beispiel der Wegberechnung anhand einer Geschwindigkeitskurve und betrachten wir das allgemeine Problem, zu einer gegebenen reellwertigen Funktion f eine neue Funktion F zu finden derart, daß $F'(x) = f(x)$ ist.

Da jede konstante Funktion die Ableitung Null hat, ist mit F für jede reelle Zahl c auch $F + c$ eine solche Funktion; um ein konkretes F hinzuschreiben, müssen wir also einen Funktionswert von F festlegen; der Einfachheit halber sei dies, in Analogie zum obigen Beispiel, der Wert $F(0) = 0$.

Damit übernimmt f die Rolle der Geschwindigkeit v, dem Weg entspricht die Funktion F, die Variable ist nun x anstelle der Zeit t, und da uns F an einer beliebigen Stelle x interessiert, nicht nur wie im obigen Beispiel an der Stelle 5, empfiehlt es sich, n jetzt als Gesamtzahl der Intervalle zu nehmen, nicht wie oben als Anzahl der Intervalle pro Einheit (Sekunde). Die Länge eines jeden Intervalls wird dann zu x/n, und die Analoga zu obigen Summen sind die Ausdrücke

$$\sum_{i=0}^{n-1} f\left(\frac{ix}{n}\right) \cdot \frac{x}{n} \quad \text{und} \quad \sum_{i=1}^{n} f\left(\frac{ix}{n}\right) \cdot \frac{x}{n},$$

deren gemeinsamer Grenzwert für immer größer werdendes n der gesuchte Wert $F(x)$ sein sollte.

Wir *definieren* daher

$$F(x) \underset{\text{def}}{=} \lim_{n\to\infty} \sum_{i=0}^{n-1} f\left(\frac{ix}{n}\right) \cdot \frac{x}{n} \tag{$*$}$$

in der Hoffnung, daß dieser Grenzwert existiert und mit dem anderen übereinstimmt – zumindest in hinreichend vielen interessanten Fällen. Bevor wir uns im nächsten Paragraphen dieser Frage zuwenden, wollen wir hier zunächst etwas mit dieser vorläufigen Definition spielen und sehen, was sie bei einfachen Funktionen liefert.

1) Die Funktion** $f(x) = x^2$**: Hier erhalten wir

$$F(x) \underset{\text{def}}{=} \lim_{n\to\infty} \sum_{i=0}^{n-1} f\left(\frac{ix}{n}\right) \cdot \frac{x}{n} = \lim_{n\to\infty} \sum_{i=0}^{n-1} \left(\frac{ix}{n}\right)^2 \cdot \frac{x}{n}$$

$$= x^3 \lim_{n\to\infty} \frac{1}{n^3} \sum_{i=0}^{n-1} i^2 .$$

Wie man in seiner Formelsammlung nachschlägt oder sich von seinem Computeralgebrasystem sagen läßt, ist

$$\sum_{i=0}^{n-1} i^2 = \frac{n(2n-1)(n-1)}{6} ,$$

eine Formel, die dann nachträglich leicht durch vollständige Induktion bewiesen werden kann:

Für $n = 1$ steht links und rechts Null, die Formel ist also korrekt.

Angenommen, wir hätten sie bewiesen für $n - 1$ anstelle von n, d.h. wir wüßten, daß für ein spezielles $n > 1$

$$\sum_{i=0}^{n-2} i^2 = \frac{(n-1)(2n-3)(n-2)}{6}$$

ist. Dann folgt

$$\begin{aligned}\sum_{i=0}^{n-1} i^2 &= \frac{(n-1)(2n-3)(n-2)}{6} + (n-1)^2 \\ &= \frac{n-1}{6} \cdot \big((2n-3)(n-2) + 6(n-1)\big) \\ &= \frac{n-1}{6} \cdot (2n^2 - 7n + 6 + 6n - 6) \\ &= \frac{n-1}{6} \cdot (2n^2 - n) = \frac{(n-1)(2n-1)n}{6} \\ &= \frac{n(2n-1)(n-1)}{6} ,\end{aligned}$$

d.h. die Formel für n selbst.

Nach dem Prinzip der vollständigen Induktion ist diese damit für *alle* Werte von n bewiesen; wir können sie in obige Rechnung einsetzen und erhalten

$$\begin{aligned} F(x) &= x^3 \lim_{n\to\infty} \frac{1}{n^3} \cdot \frac{n(2n-1)(n-1)}{6} = x^3 \lim_{n\to\infty} \frac{n(2n-1)(n-1)}{6n^3} \\ &= \frac{x^3}{6} \lim_{n\to\infty} \left(\frac{n}{n}\right)\left(\frac{2n-1}{n}\right)\left(\frac{n-1}{n}\right) = \frac{x^3}{6} \cdot 2 = \frac{x^3}{3} \, . \end{aligned}$$

Diese Funktion hat tatsächlich die Ableitung $F'(x) = x^2$, zumindest in diesem Beispiel funktioniert also die Definition $(*)$.

2) Die Exponentialfunktion: Versuchen wir dasselbe nochmal für die Funktion $f(x) = e^x$. Hier führt $(*)$ (wieder mit der Festlegung $F(0) = 0$) auf den Ansatz

$$\begin{aligned} F(x) \underset{\text{def}}{=} \lim_{n\to\infty} \sum_{i=0}^{n-1} f\left(\frac{ix}{n}\right) \cdot \frac{x}{n} &= \lim_{n\to\infty} \sum_{i=0}^{n-1} e^{\frac{ix}{n}} \cdot \frac{x}{n} \\ &= x \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \left(e^{\frac{x}{n}}\right)^i \, . \end{aligned}$$

Die Summe ganz rechts ist eine geometrische Reihe und kann leicht ausgerechnet werden: Setzen wir zur Abkürzung $q = e^{x/n}$, so ist

$$\begin{aligned} \sum_{i=0}^{n-1} q^i &= 1 + q + q^2 + q^3 + \cdots + q^{n-1} \\ q \sum_{i=0}^{n-1} q^i &= \phantom{1 +{}} q + q^2 + q^3 + \cdots + q^{n-1} + q^n \\ \hline \Rightarrow (1-q) \sum_{i=0}^{n-1} q^i &= 1 \phantom{{}+ q + q^2 + q^3 + \cdots + q^{n-1}} - q^n \, . \end{aligned}$$

Für $q \neq 1$ können wir durch $1 - q$ dividieren und erhalten die Formel

$$\sum_{i=0}^{n-1} q^i = \frac{1-q^n}{1-q} \, .$$

Da uns nur der Fall $x \neq 0$ interessiert, ist $q = e^{\frac{x}{n}} \neq 1$; wir können die Formel also anwenden und erhalten

$$
\begin{aligned}
F(x) &= x \lim_{n\to\infty} \frac{1}{n} \cdot \frac{1-\left(e^{\frac{x}{n}}\right)^n}{1-e^{\frac{x}{n}}} = x \lim_{n\to\infty} \frac{1}{n} \cdot \frac{1-e^x}{1-e^{\frac{x}{n}}} \\
&= x(1-e^x) \lim_{n\to\infty} \frac{\frac{1}{n}}{1-e^{\frac{x}{n}}} \,.
\end{aligned}
$$

Im Grenzwert ganz rechts gehen für $n \to \infty$ leider sowohl der Zähler als auch der Nenner gegen Null, wir müssen also die Regel von DE L'HOSPITAL (**[Bö]**, *Folgerung 24.7;* **[D]**, *Satz 7.6a;* **[MV]**, *Kap. 3, Satz 2.9*) anwenden. Diese gilt zwar nur für Grenzwerte von Werten stetiger Funktionen, während wir hier den Grenzwert einer diskreten Folge suchen, aber wenn für eine reelle Variable u der Grenzwert

$$\lim_{u\to\infty} \frac{\frac{1}{u}}{1-e^{\frac{x}{u}}}$$

existiert, konvergiert natürlich auch die Folge der Zahlen

$$\frac{\frac{1}{n}}{1-e^{\frac{x}{n}}} \quad \text{für } n = 1, 2, 3, \ldots$$

gegen eben diesen Grenzwert.

Nach DE L'HOSPITAL ist

$$
\begin{aligned}
\lim_{u\to\infty} \frac{\frac{1}{u}}{1-e^{\frac{x}{u}}} &= \lim_{u\to\infty} \frac{\frac{d}{du}\frac{1}{u}}{\frac{d}{du}(1-e^{\frac{x}{u}})} = \lim_{u\to\infty} \frac{\frac{-1}{u^2}}{-e^{\frac{x}{u}}\left(\frac{-x}{u^2}\right)} \\
&= \lim_{u\to\infty} \frac{-1}{xe^{\frac{x}{u}}} = \frac{-1}{x} \,.
\end{aligned}
$$

Eingesetzt in den Ansatz für $F(x)$ führt dies auf

$$F(x) = x(1-e^x) \cdot \frac{-1}{x} = e^x - 1 \,,$$

eine Funktion, deren Ableitung in der Tat e^x ist.

3) Die Dirichletsche Sprungfunktion: Bevor wir zu übermütig werden, möchte ich als letztes Beispiel noch eine Funktion betrachten, die sich im Gegensatz zu Quadrat und Exponentialfunktion alles andere als gut verhält, die DIRICHLETsche Sprungfunktion

$$f(x) = \begin{cases} 1 & \text{für } x \in \mathbb{Q} \\ 0 & \text{für } x \notin \mathbb{Q} \end{cases} .$$

Mit $F(0) \underset{\text{def}}{=} 0$ ist wieder

$$F(x) \underset{\text{def}}{=} \lim_{n\to\infty} \sum_{i=0}^{n-1} f\left(\frac{ix}{n}\right) \cdot \frac{x}{n} .$$

Dabei ist die Zahl ix/n für $i = 0$ rational, da Null; für $i \neq 0$ ist sie genau dann rational, wenn auch x rational ist. In diesem Fall ist also $f(ix/n) = 1$ für alle i, d.h.

$$F(x) = \lim_{n\to\infty} \sum_{i=0}^{n-1} \frac{x}{n} = \lim_{n\to\infty} n \cdot \frac{x}{n} = x$$

für rationales x. Für irrationales x dagegen sind alle ix/n außer der Null irrational, also ist $f(ix/n) = 0$ für alle $i \neq 0$. Damit bleibt von der Summe nur der erste Summand stehen, d.h.

$$F(x) = \lim_{n\to\infty} \frac{x}{n} = 0 .$$

Also ist

$$F(x) = \begin{cases} x & \text{für } x \in \mathbb{Q} \\ 0 & \text{für } x \notin \mathbb{Q} \end{cases} ,$$

eine offensichtlich nicht differenzierbare Funktion, und wenn $F'(x)$ nicht einmal existiert, kann es natürlich auch nicht gleich $f(x)$ sein.

Auch unter dem Gesichtspunkt der Interpretation von $F(x)$ als Fläche unterhalb der Kurve $y = f(x)$ ist das Ergebnis unbefriedigend, denn da $f(x) \geq 0$ für alle x, sollte $F(x)$ eine monoton wachsende (oder zumindest nicht fallende) Funktion von x sein, wohingegen das gerade berechnete F bei jeder rationalen Zahl auf Null zurückfällt.

Der Ansatz $(*)$ hat also auch seine Tücken, und es wird Zeit, diese heuristische Definition durch eine bessere zu ersetzen.

§2: Definition des Riemann-Integrals

Tatsächlich ist der Ansatz ($*$) nicht so schlecht: Für die beiden gutartigen Funktionen und für das Eingangsbeispiel der Streckenbestimmung führte er schließlich zu vernünftigen Ergebnissen, und wie wir bald sehen werden, kann man Integrale über stetige Funktionen *immer* nach diesem Ansatz bestimmen; zumindest als Veranschaulichung des Integralbegriffs sollte man ihn daher durchaus im Hinterkopf behalten. Problematisch wird er erst bei „schlechten“ Funktionen wie der im letzten Beispiel.

a) Warum lohnt sich ein allgemeinerer Ansatz?

Dies legt natürlich die Frage nahe, ob man solche „schlechten“ Funktionen für Anwendungen wirklich braucht, und wenn ja, ob man sie so sehr braucht, daß sich der erhebliche technische Aufwand, den wir in diesem Paragraphen treiben müssen, lohnt.

Die DIRICHLETsche Sprungfunktion ist ein rein theoretisch konstruiertes Gegenbeispiel, das wohl keinerlei praktische Anwendung haben dürfte. Eine ihrer charakteristischen Eigenschaften taucht allerdings auch bei praktisch relevanten Funktionen auf: Auch ein periodisch wiederholter Rechteckimpuls hat abzählbar unendlich viele Sprungstellen, und zumindest als Idealisierung eines realen Signals spielt diese Funktion eine Rolle – sie ist allerdings erheblich harmloser als DIRICHLETS Beispiel, da die Menge ihrer Sprungstellen diskret ist.

Wirklich kompliziert wird die Situation dagegen beispielsweise bei der mathematischen Modellierung des Rauschens in einer elektronischen Schaltung: Hier liegt eine noch einmal deutlich schwierigere Situation vor als im obigen Beispiel, und in der Tat wird auch der in diesem Paragraphen definierte Integralbegriff nicht ausreichen, um mit diesem Problem fertig zu werden. Er ist aber eine unabdingbare Voraussetzung, um die dazu benötigten Techniken zu verstehen.

Obwohl wir im folgenden fast jeden der wichtigeren Sätze aus der Analysis I noch einmal in die Erinnerung zurückrufen und anwenden müssen, ist die nun folgende Konstruktion also kein Luxus, sondern zumindest langfristig notwendig auch für praktische Anwendungen.

b) Wo sollte der bisherige Ansatz modifiziert werden?

Ein Punkt, bei dem der bisherige Ansatz nur aufgrund der Einfachheit der betrachteten Beispiele so erfolgreich war, ist die *Abschätzung* des Grenzwerts durch die endlichen Approximationen. Beim Beispiel der Geschwindigkeit eines beschleunigenden Fahrzeugs war das problemlos, denn die Geschwindigkeit war eine monoton wachsende Funktion, die für jedes Teilintervall am linken Ende ihr Minimum und am rechten ihr Maximum annimmt.

Auch die Funktion $f(x) = x^2$ ist für $x \geq 0$ monoton wachsend, $f(x) = e^x$ sogar für alle $x \in \mathbb{R}$, so daß wir auch hier leicht untere und obere Schranken für $F(x)$ berechnen können.

Für nichtmonotone Funktionen dagegen ist die Situation völlig anders: Betrachten wir als Beispiel etwa die Sinuslinie zwischen Null und π mit einer Annäherung der Fläche durch sechs Rechtecke. In den Abbildungen 6 und 7 sind diese Rechtecke eingezeichnet, einmal bezogen auf den Funktionswert am linken Ende des Teilintervalls und einmal bezogen auf den am rechten. Wie man sieht, liegen die Rechtecke exakt spiegelsymmetrisch zueinander und haben damit insbesondere die gleiche Summe der Flächeninhalte, nämlich jeweils ungefähr 1,954. Die Fläche unter der Sinuslinie zwischen Null und π ist allerdings, wie wir bald sehen werden,

$$-\cos\pi - (-\cos 0) = -(-1) - (-1) = 1 + 1 = 2\,.$$

Wenn man ein (jedem Wissenschaftler zu empfehlendes) gesundes Mißtrauen gegen Computer und Taschenrechner mitbringt, könnte man versucht sein, den Wert 1,954 als eine „Taschenrechnerzwei“ zu interpretieren; da aber die Sinuswerte an den Stellen 0, $\frac{\pi}{6}$, $\frac{\pi}{3}$, $\frac{\pi}{2}$, $\frac{4\pi}{3}$, $\frac{5\pi}{6}$ und 1 wohlbekannt sind (im Winkelmaß ausgedrückt sind diese Stellen gerade die Vielfachen von 30°), kann man die Summe der Rechteckflächen exakt berechnen mit dem Ergebnis

$$\frac{\pi}{6}(2 + \sqrt{3})\,,$$

das schon wegen der Transzendenz von π nicht gleich zwei sein kann. Somit haben wir definitiv keine obere Schranke für den korrekten Wert,

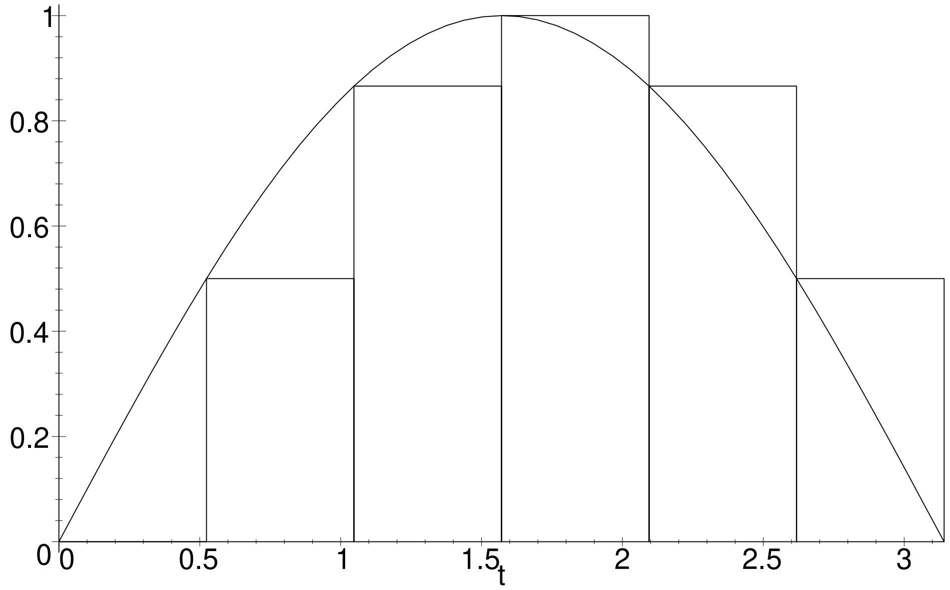


Abb. 6: Sinuslinie mit sechs Rechtecken zum jeweils linken Funktionswert

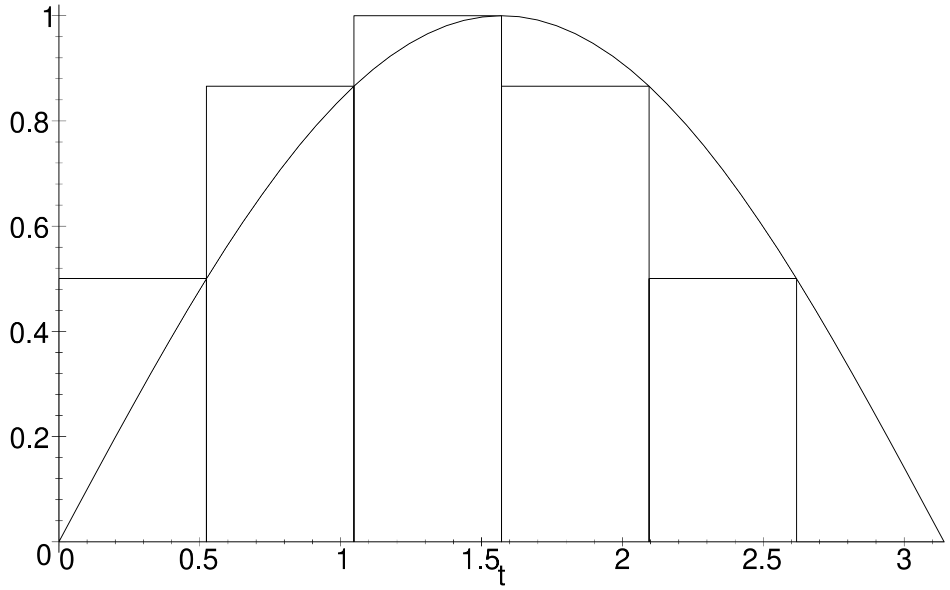


Abb. 7: Sinuslinie mit sechs Rechtecken zum jeweils rechten Funktionswert

und zumindest *a priori* haben wir auch keine untere, denn daß die

Summe der Rechteckflächen kleiner ist als die Fläche unter der Kurve folgte je erst nachträglich durch Vergleich der beiden Werte.

Um wirklich eine untere Schranke für die Fläche zu bekommen, müßte man hier im Beispiel für die ersten drei Rechtecke den Funktionswert am linken Ende des Teilintervalls nehmen und für die letzten drei den vom rechten; für eine obere Schranke müßte man genau umgekehrt vorgehen.

Für eine beliebige Funktion gibt es aber offensichtlich keinen Grund, warum das Minimum oder Maximum überhaupt an einem Intervallende angenommen werden sollte; für einen möglichst allgemeinen Integrallbegriff sollte man also auch Punkte im Intervallinnern zur Ermittlung der Höhe des Rechtecks heranziehen: Anstelle der bislang betrachteten Unterteilung

$$a < a+\delta < a+2\delta < \cdots < a+(n-1)\delta < a+n\delta = b \quad \text{mit} \quad \delta = \frac{b-a}{n}$$

sollte also eine beliebige Unterteilung

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$$

treten.

c) Anwendung des Mittelwertsatzes

Einen weiteren Grund dafür liefert der wohl wichtigste Satz aus der Analysis I, der *Mittelwertsatz der Differentialrechnung* (**[Bö]**, *Satz 24.3;* **[D]**, *Satz 7.3;* **[MV]**, *Kap. 3, Satz 2.2.*): Ist F stetig in einem Intervall $[u,v]$ und differenzierbar in (u,v), so gibt es einen Punkt $\xi \in (u,v)$, so daß

$$F'(\xi) = \frac{f(v)-f(u)}{v-u}$$

ist.

Wir suchen zu einer gegebenen Funktion f eine differenzierbare Funktion F mit $F' = f$. Falls wir annehmen, daß wir diese Funktion schon *hätten,* so wäre $F' = f$, und nach dem Mittelwertsatz gäbe es in jedem Intervall (x_i, x_{i+1}) ein Element ξ_i, so daß

$$\frac{F(x_{i+1}) - F(x_i)}{x_{i+1} - x_i} = F'(\xi_i) = f(\xi_i)$$

wäre, d.h.

$$F(x_{i+1}) = F(x_i) + f(\xi_i)(x_{i+1} - x_i)\,.$$

Rekursiv folgt, daß

$$\begin{aligned} F(x) = F(x_n) &= F(x_{n-1}) + f(\xi_{n-1})(x_n - x_{n-1}) \\ &= F(x_{n-2}) + f(\xi_{n-1})(x_{n-1} - x_{n-2}) + f(\xi_{n-1})(x_n - x_{n-1}) \\ &= \qquad \vdots \\ &= F(x_0) + \sum_{i=0}^{n-1} f(\xi_i)(x_{i+1} - x_i)\,. \end{aligned}$$

Damit ist also die Fläche unter der Kurve $y = f(x)$ *exakt* gleich der Gesamtfläche endlich vieler geeignet gewählter Rechtecke; Abbildung 8 veranschaulicht dies anhand der ganz zu Beginn betrachteten Geschwindigkeitsfunktion. Hier ist nicht nur die Fläche unter der Kurve exakt gleich der Fläche der fünf Rechtecke, sondern auch die Fläche eines jeden Rechtecks gleich der Fläche zwischen Kurve und Grundlinie des Rechtecks, wir kommen also mit endlich vielen Rechtecken aus und brauchen nicht einmal einen Grenzwert zu berechnen.

Der Haken bei der Sache ist natürlich, daß wir die „geeignet gewählten“ Rechtecke nicht kennen: Im obigen Ausdruck ist alles bekannt *außer* den Werten ξ_i, an denen die Funktion f ausgewertet wird.

Die Idee zur Definition des RIEMANN-Integrals ist nun, daß man einfach *beliebige* ξ_i zwischen x_i und x_{i+1} wählt in der Hoffnung, daß bei immer kleiner werdendem Abstand zwischen zwei aufeinanderfolgenden x_i der Grenzwert nicht mehr von der Wahl der ξ_i abhängt.

Diese Hoffnung hat natürlich nur eine Chance auf Erfüllung, wenn die Funktion f hinreichend stetig ist; ansonsten können sich die Werte von f an zwei beliebig nahe beieinanderliegenden Stellen immer noch beliebig stark unterscheiden, indem sie beispielsweise wie oben davon abhängen, ob ξ_i rational ist oder nicht.

d) Gleichmäßige Stetigkeit

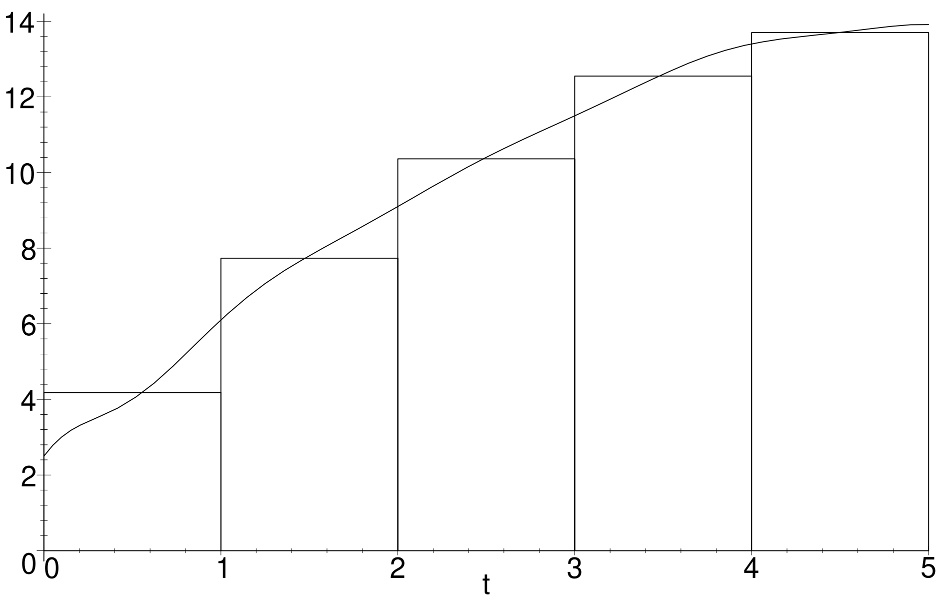


Abb. 8: Die Fläche unter der Kurve ist gleich der Fläche der fünf Rechtecke

Die Stetigkeit allein reicht allerdings immer noch nicht ganz aus:

Erinnern wir uns: f heißt *stetig* auf einer Teilmenge $D \subseteq \mathbb{R}$, wenn es zu jedem $\varepsilon > 0$ und zu jedem $x \in (a, b)$ ein $\delta > 0$ gibt, so daß gilt: Ist $y \in D$ und $|y - x| < \delta$, so folgt, daß $|f(y) - f(x)| < \varepsilon$ ist.

Wir möchten mehr: Wir wollen, daß sich in *jedem* Rechteck der Wert von $f(\xi)$ höchstens um $\varepsilon > 0$ ändert, falls nur die Breite des Rechtecks unter einer gewissen Schranke liegt, d.h. wir möchten, daß δ nicht von x abhängt

Dies führt auf folgende Verschärfung des Stetigkeitsbegriffs:

Definition: $D \subseteq \mathbb{R}$ sei eine Teilmenge von $\mathbb{R}$ und $f\colon D \to \mathbb{R}$ eine Funktion. f heißt *gleichmäßig stetig* in D, wenn es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß für alle $x, y \in D$ gilt:

$$|y - x| < \delta \Longrightarrow |f(y) - f(x)| < \varepsilon\,.$$

Um zu sehen, daß dies mehr ist als die bloße Stetigkeit, betrachten wir die Funktion $f(x) = 1/x$. Diese Funktion ist stetig auf der Menge aller positiver reeller Zahlen, denn für $0 < \delta < x$ und $\varepsilon > 0$ ist

$$\left|\frac{1}{x \pm \delta} - \frac{1}{x}\right| = \left|\frac{\mp\delta}{x(x \pm \delta)}\right| = \frac{\delta}{x(x \pm \delta)} < \varepsilon$$

$$\Leftrightarrow \delta < (x^2 \pm x\delta)\varepsilon \Leftrightarrow \delta \mp x\delta\varepsilon < x^2\varepsilon \Leftrightarrow \delta < \frac{x^2\varepsilon}{1 \mp x\varepsilon}\,.$$

Der Ausdruck ganz rechts ist offensichtlich kleiner, wenn im Nenner das Pluszeichen steht, also gilt: Für gegebenes $x > 0$ und $\varepsilon > 0$ gilt für jede positive reelle Zahl y

$$|y - x| < \delta \underset{\text{def}}{=} \frac{x^2\varepsilon}{1 + x\varepsilon} \Rightarrow \left|\frac{1}{y} - \frac{1}{x}\right| < \varepsilon\,.$$

Damit ist die Stetigkeit der Funktion bewiesen. Sie ist aber nicht gleichmäßig stetig, denn es gibt aber kein von x unabhängiges $\delta > 0$ mit dieser Eigenschaft: Der angegebene Ausdruck für das größtmögliche δ geht wegen des Faktors x^2 im Zähler für $x \to 0$ selbst gegen Null, fällt also unter jede vorgegebene positive Schranke. Abbildung 9 zeigt δ in Abhängigkeit von x für den speziellen Wert $\varepsilon = \frac{1}{10}$.

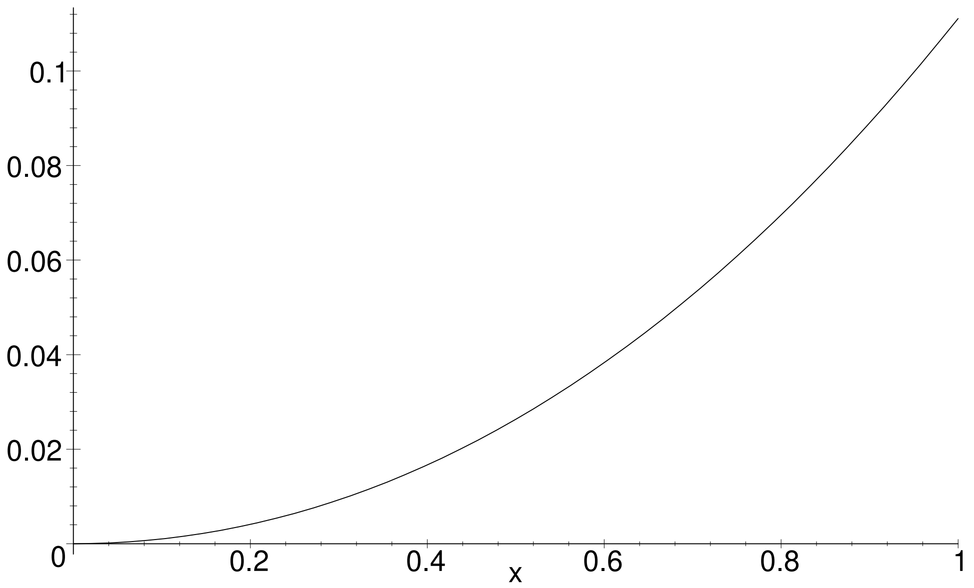


Abb. 9: δ in Abhängigkeit von x für $f(x) = 1/x$ und $\varepsilon = 0{,}1$

In einem abgeschlossenen Intervall sollte so etwas nicht möglich sein, und in der Tat gilt der (durchaus nichttriviale)

Satz: $f: [a, b] \to \mathbb{R}$ sei eine stetige Funktion auf dem abgeschlossenen Intervall $[a, b]$ mit $a, b \in \mathbb{R}$. Dann ist f auf $[a, b]$ gleichmäßig stetig.

Der *Beweis* ist technisch und indirekt:

Angenommen, f wäre *nicht* gleichmäßig stetig. Dann gäbe es für mindestens ein $\varepsilon > 0$ zu jedem $\delta > 0$ Punkte $x, y \in [a, b]$, so daß $|y - x| < \delta$, aber $|f(y) - f(x)| \geq \varepsilon$ wäre.

Aufgrund der Annahme, der Satz sei falsch, können wir ein solches ε fixieren und betrachten die speziellen Werte $\delta = \frac{1}{n}$: Nach dem gerade gesagten gibt es dazu Punkte $x_n, y_n \in [a, b]$, so daß $|y_n - x_n| < \frac{1}{n}$, aber $|f(y_n) - f(x_n)| \geq \varepsilon$ ist.

Nun kommt der nichttriviale Teil: Wir benötigen aus der Analysis I den

Satz von Bolzano-Weierstraß: Jede Folge (x_n) von Punkten aus dem abgeschlossenen Intervall $[a, b]$ hat (mindestens) eine konvergente Teilfolge x_{n_ν}.

Referenz: **[Bö]**, *Satz 13.3* in Verbindung mit *Satz 12.8;* **[D]**, *Satz 2.1* in Verbindung mit *Satz 2.8.*

Insbesondere muß also die hier betrachtete Folge (x_n) eine konvergente Teilfolge $\left(x_{n_\nu}\right)_{\nu\in\mathbb{N}}$ haben; deren Grenzwert sei $c \in [a, b]$.

Da $\left|x_{n_\nu} - y_{n_\nu}\right| < 1/n_\nu$ ist, muß dann auch die Folge der y_{n_ν} gegen c konvergieren, und da f nach Voraussetzung eine stetige Funktion ist, gilt

$$\lim_{\nu\to\infty} f(x_{n_\nu}) = \lim_{\nu\to\infty} f(y_{n_\nu}) = f(c)\,.$$

Andererseits ist aber für jedes ν nach Konstruktion der Folgen (x_n) und (y_n) der Abstand zwischen $f(x_{n_\nu})$ und $f(y_{n_\nu})$ größer oder gleich der festgewählten Zahl ε, so daß die beiden Folgen unmöglich denselben Grenzwert haben können.

Mithin führt die Annahme, f sei *nicht* gleichmäßig stetig, auf einen Widerspruch, und damit ist der Satz bewiesen.

■

e) Definition einer Approximation für das Integral

Nach diesen Vorarbeiten können wir ernsthaft darangehen, das Integral einer Funktion zu definieren. Es gibt in der Mathematik verschiedene Integralbegriffe; für uns reicht die Definition des deutschen Mathematikers GEORG FRIEDRICH BERNHARD RIEMANNN (1826–1866), das inzwischen nach ihm benannte RIEMANN-Integral.

Wir gehen aus von einer auf dem abgeschlossenen Intervall $[a, b]$ definierten Funktion $f\colon [a, b] \to \mathbb{R}$; gesucht ist eine Funktion $F\colon [a, b] \to \mathbb{R}$, für die (idealerweise) $F'(x) = f(x)$ sein sollte für alle x aus dem offenen Intervall (a, b).

Wir wählen eine Unterteilung

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = x$$

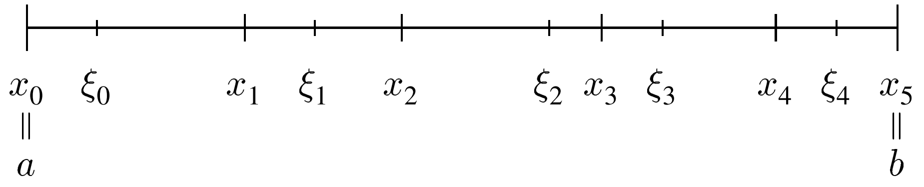
des Intervalls von a bis x. Da wir im folgenden viel mit dieser Unterteilung rechnen werden, führen wir die Abkürzung

$$\underline{x} = (x_0, x_1, \ldots, x_n)$$

dafür ein und verabreden, daß die x_i immer, wenn von $\underline{x}$ die Rede ist, die obigen Gleichungen und Ungleichungen erfüllen sollen, daß sie also tatsächlich eine Unterteilung von $[a, x]$ definieren.

Falls die gesuchte differenzierbare Funktion $F\colon [a, b] \to \mathbb{R}$ existiert, sagt uns – wie wir in Teil *c)* gesehen haben – der Mittelwertsatz der Differentialrechnung, daß für geeignete Elemente ξ_i mit

$$a = x_0 < \xi_0 < x_1 < \xi_1 < x_2 < \cdots < \xi_{n-2} < x_{n-1} < \xi_{n-1} < x_n = x$$



folgt

$$F(x) = F(a) + \sum_{i=0}^{n-1} f(\xi_i)(x_{i+1} - x_i)\,.$$

Natürlich kennen wir die Zahlen ξ_i nicht – selbst wenn wir wissen, daß F und damit die ξ_i überhaupt existieren. Deshalb betrachten wir einfach *beliebige* Zahlen ξ_i mit

$$a = x_0 < \xi_0 < x_1 < \xi_1 < x_2 < \cdots < \xi_{n-2} < x_{n-1} < \xi_{n-1} < x_n = x$$

und führen auch hier wieder die Abkürzung

$$\underline{\xi} = (\xi_0, \xi_1, \ldots, \xi_{n-1})$$

ein mit der Verabredung, daß immer, wenn ein $\underline{\xi}$ in Zusammenhang mit einem $\underline{x}$ auftritt, diese Ungleichungen erfüllt sein sollen.

Definition: Die RIEMANNsche Summe zu einem Paar $(\underline{x}, \underline{\xi})$ ist

$$I(\underline{x}, \underline{\xi}) = \sum_{i=0}^{n-1} f(\xi_i)(x_{i+1} - x_i)\,.$$

Das RIEMANN-Integral soll der Grenzwert einer Folge RIEMANNscher Summen sein, wobei die Unterteilungen $\underline{x}$ immer feiner werden. Dieses „Feinerwerden“ müssen wir natürlich auch noch erklären: Seien $\underline{x}$ und $\underline{y}$ zwei Unterteilungen des Intervalls $[a, x]$; konkret sei

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = x$$

und

$$a = y_0 < y_1 < \cdots < y_{m-1} < y_m = x\,.$$

Dann heißt $\underline{y}$ eine *Verfeinerung* von $\underline{x}$, wenn jedes Teilintervall (y_i, y_{i+1}) von $\underline{y}$ ganz in einem der Teilintervalle (x_j, x_{j+1}) von $\underline{x}$ liegt; $\underline{y}$ entsteht also aus $\underline{x}$, indem einige der Intervalle von $\underline{x}$ noch weiter unterteilt werden. Insbesondere muß dann $m \geq n$ sein.

Nun betrachten wir eine Folge $(\underline{x}^{(\nu)})$ von Unterteilungen derart, daß

- $\underline{x}^{(\nu+1)}$ stets eine Verfeinerung von $\underline{x}^{(\nu)}$ ist *und*
- die maximale Länge der Teilintervalle von $\underline{x}^{(\nu)}$ mit wachsendem ν gegen Null geht, d.h.

$$\lim_{\nu\to\infty} \max_{i=0}^{n_\nu - 1} (x_{i+1}^{(\nu)} - x_i^{(\nu)}) = 0\,.$$

Zu jeder Unterteilung $\underline{x}^{(\nu)}$ wählen wir willkürlich eine (im Sinne obiger Konvention dazu passende) Folge $\underline{\xi}^{(\nu)}$ von Zwischenpunkten und fragen nach Existenz und gegebenenfalls Wert des Grenzwerts

$$\lim_{\nu\to\infty} I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)}) .$$

Dieser Grenzwert muß natürlich im allgemeinen nicht existieren, und wenn er existiert, kann er von der Wahl der Zwischenpunkte $\xi_i^{(\nu)}$ und von der Wahl einer Folge $(\underline{x}^{(\nu)})$ von Unterteilungen abhängen. Wenn er unabhängig von all diesen Wahlen existiert und immer denselben Wert hat, bezeichnen wir diesen gemeinsamen Wert als das RIEMANN-*Integral* von f zwischen a und x, in Zeichen

$$\int_a^x f(\xi)\,d\xi = \lim_{\nu\to\infty} I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)}) .$$

Falls dieses Integral für jedes $x \in [a,b]$ existiert, sagen wir, f sei RIEMANN-*integrierbar* auf $[a,b]$.

ξ heißt *Integrationsvariable* und übernimmt die Rolle des Summationsindex in einer Summe: Genau wie beispielsweise

$$\sum_{i=0}^{n-1} i^2 = \sum_{j=0}^{n-1} j^2 = \sum_{k=0}^{n-1} k^2 = \frac{n(2n-1)(n-1)}{6}$$

davon unabhängig ist, ob der Summationsindex i, j oder k heißt, ist auch

$$\int_a^x f(\xi)\,d\xi = \int_a^x f(u)\,du = \int_a^x f(t)\,dt$$

unabhängig davon, ob die Integrationsvariable mit ξ, u oder t bezeichnet wird.

f) Existenz des Riemann-Integrals für stetige Funktionen

Damit haben wir das RIEMANN-Integral definiert; wir wissen allerdings bisher für keine einzige Funktion, daß es existiert. In diesem Abschnitt wollen wir uns überlegen, daß es zumindest für *stetige* Funktionen immer existiert:

Satz: Jede stetige Funktion $f\colon [a,b] \to \mathbb{R}$ ist RIEMANN-integrierbar auf $[a,b]$.

Zum *Beweis* müssen wir zeigen, daß das RIEMANN-Integral

$$\int\limits_a^x f(\xi)\, d\xi$$

für jedes $x \in [a,b]$ existiert.

Dazu fixieren wir ein beliebig vorgegebenes $x \in [a,b]$ und betrachten Folgen $(\underline{x}^{(\nu)})$ von immer feiner werdenden Unterteilungen des Intervalls $[a,x]$; dazu entsprechende Folgen $(\underline{\xi}^{(\nu)})$ von Zwischenwerten und die Grenzwerte der RIEMANN-Summen $I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)})$.

1. Schritt: Für eine feste Folge $(\underline{x}^{(\nu)})$ von Unterteilungen existiert der Grenzwert und ist unabhängig von der Wahl der Zwischenwerte $\xi_i^{(\nu)}$:

Um das einzusehen, betrachten wir für jede Unterteilung $\underline{x}^{(\nu)}$ die RIEMANNschen Obersummen und Untersummen: Da f nach Voraussetzung stetig ist, nimmt es in jedem abgeschlossenen Intervall sowohl sein Maximum als auch sein Minimum an (**[Bö]**, *Folgerung 21.13;* **[D]**, *Satz 2.15;* **[MV]**, *Kap. 2, Satz 6.5b))*; insbesondere werden also im Intervall $[x_i^{(\nu)}, x_{i+1}^{(\nu)}]$ ein Maximum $M_i^{(\nu)}$ und ein Minimum $m_i^{(\nu)}$ angenommen. Damit gilt unabhängig von der Wahl des Zwischenwerts $\xi_i^{(\nu)}$

$$m_i^{(\nu)} \le f(\xi_i^{(\nu)}) \le M_i^{(\nu)}\,.$$

Nach Definition ist

$$I(\underline{x}^{(\nu)}, \xi^{(\nu)}) = \sum_{i=0}^{n_\nu - 1} f(\xi_i^{(\nu)})\bigl(x_{i+1}^{(\nu)} - x_i^{(\nu)}\bigr)\;;$$

ersetzen wir hierin $\xi_i^{(\nu)}$ jeweils durch $m_i^{(\nu)}$, so erhalten wir eine untere Abschätzung

$$s^{(\nu)} \underset{\text{def}}{=} \sum_{i=0}^{n_\nu - 1} m_i^{(\nu)}\bigl(x_{i+1}^{(\nu)} - x_i^{(\nu)}\bigr)$$

für $I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)})$, die wir als RIEMANN*sche Untersumme* der Unterteilung $\underline{x}^{(\nu)}$ bezeichnen. Entsprechend liefert die Ersetzung durch $M_i^{(\nu)}$ eine obere Abschätzung

$$S^{(\nu)} \underset{\text{def}}{=} \sum_{i=0}^{n_\nu - 1} M_i^{(\nu)} \big(x_{i+1}^{(\nu)} - x_i^{(\nu)}\big)$$

für $I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)})$, die wir als RIEMANN*sche Obersumme* der Unterteilung $\underline{x}^{(\nu)}$ bezeichnen.

Unabhängig von der Wahl der Zwischenwerte $\xi_i^{(\nu)}$ gilt somit

$$s^{(\nu)} \le I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)}) \le S^{(\nu)} ,$$

ein Zusammenhang, der in Abbildung 10 noch einmal graphisch dargestellt ist: Die RIEMANNsche Untersumme ist gleich der Fläche der durchgezogenen Rechtecke, für die Obersumme kommt noch der gestrichelte Anteil dazu, und für die RIEMANNsche Summe $I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)})$ muß man bis zu den punktierten Linien gehen. Die Werte $\xi_i^{(\nu)}$ sind auf der x-Achse durch kleine Kreise gekennzeichnet, ebenso die Punkte $\big(\xi_i^{(\nu)}, f(\xi_i^{(\nu)})\big)$ auf der Kurve.

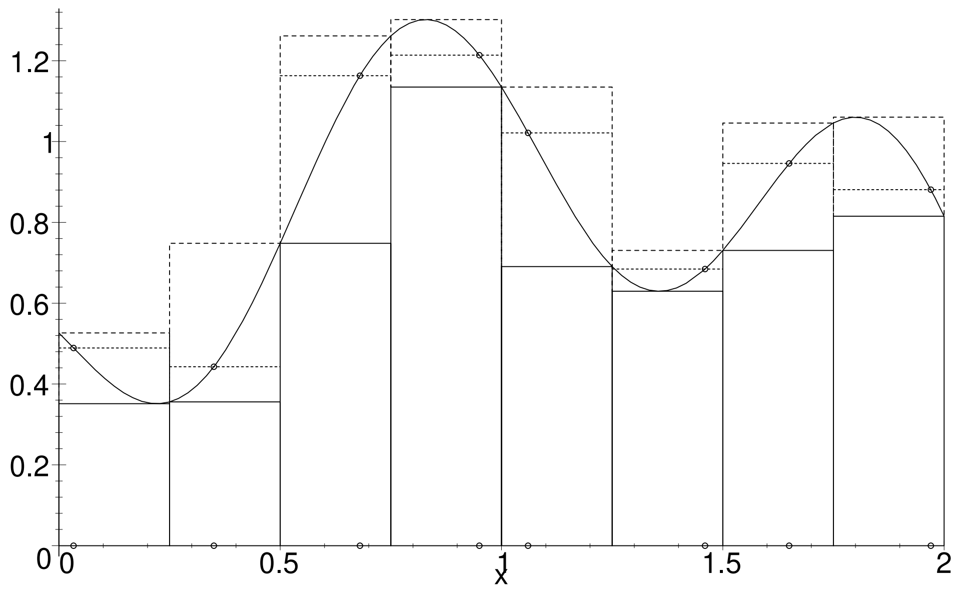


Abb. 10: Riemannsche Summen, Obersummen und Untersummen

Wenn wir jetzt noch zeigen können, daß die Folge der $s^{(\nu)}$ und die der $S^{(\nu)}$ beide gegen denselben Grenzwert konvergieren, muß wegen der Einschließung

$$s^{(\nu)} \leq I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)}) \leq S^{(\nu)}$$

auch $I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)})$ gegen diesen Grenzwert konvergieren, und die Behauptung im ersten Schritt ist bewiesen.

Vergleichen wir dazu zunächst die Untersummen $s^{(\nu)}$ und $s^{(\nu+1)}$: Die Unterteilung $\underline{x}^{(\nu+1)}$ entsteht aus $\underline{x}^{(\nu)}$ dadurch, daß einige der Intervalle weiter unterteilt werden. Ist aber $(x_i^{(\nu+1)}, x_{i+1}^{(\nu+1)})$ Teilintervall von $(x_j^{(\nu)}, x_{j+1}^{(\nu)})$, so kann das Minimum im Teilintervall natürlich nicht kleiner sein als im größeren Intervall, d.h. $s^{(\nu+1)} \geq s^{(\nu)}$. Somit ist die Folge der $s^{(\nu)}$ monoton wachsend.

Genauso folgt, daß die Folge der $S^{(\nu)}$ monoton fallend ist, und da stets $s^{(\nu)} \leq S^{(\nu)}$ ist, folgt weiter, daß

$$s^{(1)} \leq s^{(2)} \leq s^{(3)} \leq \cdots \leq S^{(3)} \leq S^{(2)} \leq S^{(1)}\,.$$

Insbesondere ist also $S^{(1)}$ eine obere Schranke für die Folge der $s^{(\nu)}$ und $s^{(1)}$ eine untere Schranke für die Folge der $S^{(\nu)}$.

Damit ist also die Folge $(s^{(\nu)})$ monoton wachsend und nach oben beschränkt, während $(S^{(\nu)})$ monoton fallend und nach unten beschränkt ist. Nach einem Satz aus der Analysis I (*Jede monotone und beschränkte Folge reeller Zahlen ist konvergent;* siehe **[Bö]**, *Folgerung 15.4;* **[D]**, *Satz 2.15;* **[MV]**, *Kap. 2, Satz 5.4*) folgt daraus die Konvergenz beider Folgen.

Nun fehlt nur noch, daß beide denselben Grenzwert haben.

Dazu betrachten wir die Differenzen

$$\begin{aligned} S^{(\nu)} - s^{(\nu)} &= \sum_{i=0}^{n_\nu - 1} M_i^{(\nu)}(x_{i+1}^{(\nu)} - x_i^{(\nu)}) - \sum_{i=0}^{n_\nu - 1} m_i^{(\nu)}(x_{i+1}^{(\nu)} - x_i^{(\nu)}) \\ &= \sum_{i=0}^{n_\nu - 1} (M_i^{(\nu)} - m_i^{(\nu)})(x_{i+1}^{(\nu)} - x_i^{(\nu)}) \,. \end{aligned}$$

Mit

$$\Delta^{(\nu)} \underset{\text{def}}{=} \max_i \left(M_i^{(\nu)} - m_i^{(\nu)}\right)$$

ist also

$$\begin{aligned} S^{(\nu)} - s^{(\nu)} &\leq \sum_{i=0}^{n_\nu - 1} \Delta^{(\nu)}(x_{i+1}^{(\nu)} - x_i^{(\nu)}) = \Delta^{(\nu)} \sum_{i=0}^{n_\nu - 1} (x_{i+1}^{(\nu)} - x_i^{(\nu)}) \\ &= \Delta^{(\nu)}(x_n^{(\nu)} - x_0^{(\nu)}) = \Delta^{(\nu)}(x - a) \,, \end{aligned}$$

und das ist eine Nullfolge, falls wir zeigen können, daß die Folge der $\Delta^{(\nu)}$ eine ist.

Hier kommt nun die in Abschnitt *c)* eingeführte gleichmäßige Stetigkeit von f zum Tragen: Danach gibt es zu jedem $\varepsilon > 0$ ein $\delta > 0$, so daß für alle Punkte $\xi_1, \xi_2 \in [a, b]$ gilt: Ist $|\xi_1 - \xi_2| < \delta$, so folgt $|f(\xi_1) - f(\xi_2)| < \varepsilon$.

Ist insbesondere für eine Unterteilung $\underline{x}^{(\nu)}$ jede der Differenzen $x_{i+1}^{(\nu)} - x_i^{(\nu)}$ kleiner als δ, so ist auch $M_i^{(\nu)} - m_i^{(\nu)} < \varepsilon$, denn sowohl

$M_i^{(\nu)}$ als auch $m_i^{(\nu)}$ sind Funktionswerte, die irgendwo im Intervall $[x_i^{(\nu)}, x_{i+1}^{(\nu)}]$ angenommen werden.

Nun haben wir aber vorausgesetzt, daß die Unterteilungen $\underline{x}^{(\nu)}$ immer feiner werden, d.h. zu jedem $\delta > 0$ gibt es in der Tat ein ν_0, so daß $x_{i+1}^{(\nu)} - x_i^{(\nu)} < \delta$ für alle $\nu > \nu_0$ und alle i. Somit sind für $\nu > \nu_0$ alle Differenzen $M_i^{(\nu)} - m_i^{(\nu)}$ kleiner als ε, und damit ist auch $\Delta^{(\nu)} < \varepsilon$ für $\nu > \nu_0$. Also sind sowohl $(\Delta^{(\nu)})$ als auch $S^{(\nu)} - s^{(\nu)}$ Nullfolgen, d.h.

$$\lim_{\nu\to\infty} s^{(\nu)} = \lim_{\nu\to\infty} S^{(\nu)} .$$

Wie schon erwähnt, folgt daraus aufgrund der Einschließung

$$s^{(\nu)} \leq I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)}) \leq S^{(\nu)} ,$$

auch die Gleichung

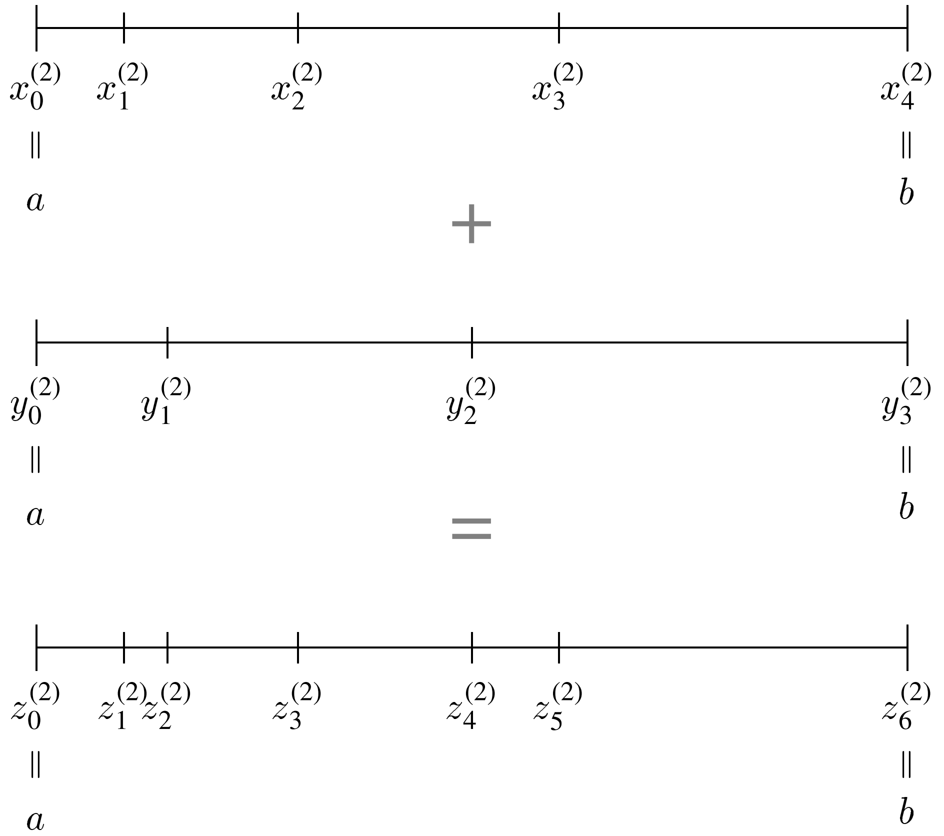
$$\lim_{\nu\to\infty} S^{(\nu)} = \lim_{\nu\to\infty} I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)}) = \lim_{\nu\to\infty} S^{(\nu)} .$$

Da die linke und die rechte Seite obiger Ungleichung nicht von den $\xi_i^{(\nu)}$ abhängen, ist auch $\lim I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)})$ davon unabhängig, und damit ist der erste Schritt des Beweises beendet.

Der zweite ist zum Glück erheblich einfacher:

2. Schritt: $\lim\limits_{\nu\to\infty} I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)})$ ist auch unabhängig von der Folge $(\underline{x}^{(\nu)})$:

Betrachten wir zwei Folgen $(\underline{x}^{(\nu)})$ und $(\underline{y}^{(\nu)})$ von Unterteilungen. Für jeden Index ν können wir zu den Unterteilungen $\underline{x}^{(\nu)}$ und $\underline{y}^{(\nu)}$ eine gemeinsame Verfeinerung $\underline{z}^{(\nu)}$ konstruieren, indem wir einfach die sämtlichen Zahlen $x_i^{(\nu)}$ und $y_j^{(\nu)}$ der Größe nach ordnen als $z_0^{(\nu)} = a, \ldots, z_r^{(\nu)} = x$.



Wie oben seien $s^{(\nu)}$ und $S^{(\nu)}$ die RIEMANNsche Unter- und Obersumme zur Unterteilung $\underline{x}^{(\nu)}$; entsprechend seien $\widetilde{s}^{(\nu)}$ und $\widetilde{S}^{(\nu)}$ die zur Unterteilung $\underline{z}^{(\nu)}$. Da $\underline{z}^{(\nu)}$ eine Verfeinerung von $\underline{x}^{(\nu)}$ ist, wissen wir aus dem ersten Schritt, daß

$$s^{(\nu)} \le \widetilde{s}^{(\nu)} \le \widetilde{S}^{(\nu)} \le S^{(\nu)}$$

ist; da die Folgen $(s^{(\nu)})$ und $(S^{(\nu)})$ denselben Grenzwert haben, müssen auch $(\widetilde{s}^{(\nu)})$ und $(\widetilde{S}^{(\nu)})$ gegen diesen Wert konvergieren und damit auch die Folge der $I(\underline{z}^{(\nu)}, \underline{\tilde{\xi}}^{(\nu)})$, wobei $(\underline{\tilde{\xi}}^{(\nu)})$ irgendeine Folge von Zwischenwerten bezeichnet; wie wir bereits aus dem ersten Schritt wissen, hängt der Grenzwert nicht davon ab. Also ist

$$\lim_{\nu\to\infty} I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)}) = \lim_{\nu\to\infty} I(\underline{z}^{(\nu)}, \underline{\tilde{\xi}}^{(\nu)}) \,.$$

Genauso folgt, daß auch

$$\lim_{\nu\to\infty} I(\underline{y}^{(\nu)}, \underline{\bar{\xi}}^{(\nu)}) = \lim_{\nu\to\infty} I(\underline{z}^{(\nu)}, \underline{\tilde{\xi}}^{(\nu)}) ,$$

wobei $(\underline{\bar{\xi}}^{(\nu)})$ die Folge der Zwischenwerte zu $\underline{y}^{(\nu)}$ bezeichnet, und dies zeigt schließlich die Behauptung

$$\lim_{\nu\to\infty} I(\underline{x}^{(\nu)}, \underline{\xi}^{(\nu)}) = \lim_{\nu\to\infty} I(\underline{y}^{(\nu)}, \underline{\bar{\xi}}^{(\nu)}) .$$

Damit ist der Satz vollständig bewiesen.

∎

g) Stückweise stetige Funktionen

Gelegentlich möchte man die Bedingung der Stetigkeit wenigstens ein bißchen lockern, um beispielweise auch einen Rechteckimpuls, der periodisch zwischen 0 und 1 (oder -1 und 1) wechselt, behandeln zu können. Die Existenz des RIEMANN-Integral wird durch solche kleineren Abweichungen nicht beeinträchtigt:

Definition: Eine Funktion $f\colon [a, b] \to \mathbb{R}$ heißt *stückweise stetig,* wenn es eine Unterteilung

$$a = a_0 < a_1 < \cdots < a_{r-1} < a_r = b$$

des Intervalls $[a, b]$ gibt, so daß f auf jedem der offenen Intervalle (a_i, a_{i+1}) stetig ist.

f ist also überall stetig außer eventuell in endlich vielen Punkten $a_0, \ldots, a_r$. Der Wert in diesen Punkten kann, aber muß nicht mit dem linksseitigen oder rechtsseitigen Grenzwert der Funktion übereinstimmen; beispielsweise definiert man Rechteckimpulse gelegentlich auch durch die Funktion

$$f(x) = \begin{cases} 1 & \text{falls } 2n-1 < x < 2n \text{ für ein } n \in \mathbb{Z} \\ 0 & \text{falls } x \in \mathbb{Z} \\ -1 & \text{falls } 2n < x < 2n+1 \text{ für ein } n \in \mathbb{Z} \end{cases} .$$

Satz: Eine stückweise stetige Funktion $f\colon [a, b] \to \mathbb{R}$ ist RIEMANN-integrierbar.

Beweis: f sei stetig in den offenen Intervallen (a_i, a_{i+1}) mit

$$a = a_0 < a_1 < \cdots < a_{r-1} < a_r = b\,.$$

Wiederholt man den Beweis des Satzes aus §*2f)* für die Funktion f, so ist alles problemlos außer den Intervallen einer Unterteilung $\underline{x}$, die einen der Punkte a_i enthalten. In diesen Intervallen ist f nicht notwendigerweise stetig, so daß die Differenz zwischen dem Supremum und dem Infimum von f dort nicht mit Verkleinerung des Intervalls gegen Null gehen muß, sondern auch gegen einen von Null verschiedenen Wert, nämlich die „Sprunghöhe“ bei a_i, konvergieren kann.

Nun gibt es aber in jeder Unterteilung nur endlich viele Intervalle, die ein a_i enthalten, und auch die Sprunghöhen sind begrenzt; gehen also alle Intervallängen gegen Null, so geht auch wie im Beweis des Satzes die Differenz zwischen RIEMANNschen Ober- und Untersummen gegen Null.

■

h) Noch einmal die Dirichletsche Sprungfunktion

In §1 waren wir nicht zufrieden mit dem Verhalten des dort heuristisch eingeführten Integrals für die DIRICHLETsche Sprungfunktion; schauen wir, was das nun eingeführte RIEMANN-Integral daraus macht. Sei also

$$f(x) = \begin{cases} 1 & \text{für } x \in \mathbb{Q} \\ 0 & \text{für } x \notin \mathbb{Q} \end{cases};$$

wir interessieren uns für $\int\limits_0^1 f(x)\,dx$.

Dazu sei $\underline{x}$ eine Unterteilung des Intervalls $[0, 1]$. Unabhängig von der Wahl dieser Unterteilung können wir stets eine Folge $\underline{\xi}$ von rationalen Zwischenwerten finden, aber auch eine Folge $\underline{\widetilde{\xi}}$ von irrationalen. Dann ist, unabhängig von der Unterteilung $\underline{x}$,

$$I(\underline{x}, \underline{\xi}) = \sum_{i=0}^{n-1} f(\xi_i)(x_{i+1} - x_i) = \sum_{i=0}^{n-1}(x_{i+1} - x_i) = 1 - 0 = 1$$

und

$$I(\underline{x}, \underline{\widetilde{\xi}}) = \sum_{i=0}^{n-1} f(\widetilde{\xi}_i)(x_{i+1} - x_i) = 0\,.$$

Damit kann kein gemeinsamer Grenzwert und somit auch kein RIEMANN-Integral existieren – ein sehr viel überzeugenderes Ergebnis als das aus §*1d3)*.

i) Ausblick: Das Lebesgue-Integral

Wir könnten allerdings auch argumentieren, daß es „nur" abzählbar unendlich viele rationale Zahlen, aber überabzählbar viele irrationale zwischen null und eins gibt; daher sollten letztere das Geschehen dominieren, und $\int_0^1 f(x)\,dx$ sollte verschwinden.

In der Tat kann man eine Verallgemeinerung des RIEMANN-Integrals definieren, das LEBESGUE-Integral, das für stückweise stetige Funktionen mit dem RIEMANN-Integral übereinstimmt und für die DIRICHLETsche Sprungfunktion den Wert Null liefert. Dazu geht man nach dem französischen Mathematiker HENRI LÉON LEBESGUE (1875-1941) bei den Unter- und Obersummen nicht wie bei RIEMANN von *endlich* vielen Rechtecken aus, sondern von *abzählbar unendlich* vielen. (Wie man dies im einzelnen macht, braucht uns hier nicht zu interessieren.) Dieses LEBESGUE-Integral existiert dann fast immer: Man kann zwar die Existenz von Funktionen, für die es nicht existiert, beweisen, aber explizite Beispiele solcher Funktionen sind nicht bekannt.

j) Anwendung auf Flächeninhalte

Nach der heuristischen Diskussion von Flächeninhalten in §1 und der Einführung des RIEMANN-Integrals in diesem Paragraphen bietet sich nun an, Flächen über dieses Integral zu *definieren:*

Definition: Ist $f\colon [a,b] \to \mathbb{R}$ eine nichtnegative Funktion, so bezeichnen wir die Zahl

$$\int_a^b f(x)\,dx$$

als *Fläche* zwischen der Kurve $y = f(x)$ und der x-Achse zwischen den Geraden $x = a$ und $x = b$.

Was passiert, wenn f auch negative Werte annimmt? Für $f(x) = x$ etwa ist $\int_{-1}^{1} f(x)\,dx = 0$, wie man sich leicht überlegt anhand von Unterteilungen, die symmetrisch zur Null liegen. Auch dies läßt sich als Aussage über Flächeninhalte interpretieren – wenn man davon ausgeht, daß die Formel

$$\textit{Fläche} = \textit{Länge} \times \textit{Breite}$$

für die Rechteckfläche auch bei *negativer* Länge und/oder Breite gelten soll. Da Integrale nicht nur zur Flächenbestimmung, sondern auch etwa zur Bestimmung der Ladung in einem Kondensator verwendet werden, ist dies die Interpretation, die man in der Mathematik in den meisten Fällen vorzieht; für die klassische Fläche im Sinne des Tapezierens muß man mit

$$\int_a^b |f(x)|\,dx$$

arbeiten.

Im Sinne dieser negativen Längen und Breiten ist es auch sinnvoll, Integrale für $b < a$ zu definieren durch

$$\int_a^b f(x)\,dx \underset{\text{def}}{=} -\int_b^a f(x)\,dx\,.$$

Insbesondere ist dann natürlich $\int_a^a f(x)\,dx = 0$.

§3: Erste Integrationsregeln

§2 stellte die heuristischen Überlegungen aus §1 insofern auf eine exaktere mathematische Grundlage, als wir nun einen Integralbegriff haben, der auch bei extrem ungewöhnlichen Funktionen wie der DIRICHLETschen Sprungfunktion keine unerwarteten Ergebnisse liefert. Was allerdings den Ausgangspunkt von §1 betrifft, die Umkehrung der Differentiation, wissen wir über das neue Integral noch gar nichts, und auch sonst kennen wir noch nicht viele Regeln über den Umgang damit und insbesondere auch über seine Berechnung *ohne* die umständlichen und sehr langsam konvergierenden RIEMANN-Summen.

In diesem Paragraphen sollen die ersten (und einfachsten) solchen Regeln zusammengestellt werden; nachdem wir im nächsten Paragraphen unser Instrumentarium erweitert haben, werden dann in §5 weiterführende Techniken entwickelt werden.

a) Monotonieregel

Eine der einfachsten, aber trotzdem oft nützlichen Regeln für den Umgang mit Integralen übersetzt Größerbeziehungen zwischen Integranden in Größenbeziehungen zwischen Integralen:

Satz: f und g seien stückweise stetige Funktionen auf dem Intervall $[a, b]$, und für alle $x \in [a, b]$ sei $f(x) \leq g(x)$. Dann ist auch

$$\int_a^b f(x)\,dx \leq \int_a^b g(x)\,dx\,.$$

Insbesondere ist

$$\left|\int_a^b f(x)\,dx\right| \leq \int_a^b |f(x)|\;dx$$

Beweis: Wir betrachten eine Unterteilung $\underline{x}$ des Intervalls $[a, b]$ und eine dazu kompatible Sequenz $\underline{\xi}$ von Zwischenwerten; $I(\underline{x}, \underline{\xi})$ sei die zugehörige RIEMANN-Summe für f und $J(\underline{x}, \underline{\xi})$ die für g. Da für jedes ξ_i nach Voraussetzung $f(\xi_i) \leq g(\xi_i)$ ist, folgt unmittelbar aus der Definition der RIEMANN-Summen, daß

$$I(\underline{x}, \underline{\xi}) \leq J(\underline{x}, \underline{\xi})$$

ist. Da sich eine solche Kleinergleichbeziehung auch auf Grenzwerte überträgt, ist daher

$$\int_a^b f(x)\,dx \leq \int_a^b g(x)\,dx\,,$$

wie behauptet.

Insbesondere können wir dies auch anwenden auf die Ungleichungskette

$$-|f(x)| \leq f(x) \leq |f(x)|$$

und erhalten

$$-\int_a^b |f(x)|\, dx \leq \left|\int_a^b f(x)\, dx\right| \leq \int_a^b |f(x)|\, dx\,,$$

d.h.

$$\left|\int_a^b f(x)\, dx\right| \leq \int_a^b |f(x)|\, dx\,.$$

■

b) Linearität und Zusammensetzung

Genauso einfach ist die Linearitätsregel:

Satz: Für zwei stückweise stetige Funktionen $f, g\colon [a,b] \to \mathbb{R}$ und zwei reelle Zahlen α, β ist

$$\int_a^b \big(\alpha f(x) + \beta g(x)\big)\, dx = \alpha \int_a^b f(x)\, dx + \beta \int_a^b g(x)\, dx\,.$$

Beweis: Wie beim letzten Satz: Für jede RIEMANN-Summe zu einer festen Unterteilung $\underline{x}$ und jede damit kompatible Zwischenwertsequenz $\underline{\xi}$ gilt die zur Behauptung des Satzes analoge Gleichung, und damit gilt im Limes auch der Satz selbst.

■

Für die Zusammensetzung von Integrationsintervallen gilt erwartungsgemäß

Satz: Für $a \leq b \leq c$ und eine stückweise stetige Funktion $f\colon [a,c] \to \mathbb{R}$ ist

$$\int_a^c f(x)\, dx = \int_a^b f(x)\, dx + \int_b^c f(x)\, dx\,.$$

Beweis: $\int_a^c f(x)\,dx$ kann als Grenzwert von RIEMANN-Summen zu einer beliebigen Folge sich verfeinernder Unterteilungen von $[a,c]$ berechnet werden; insbesondere können also Unterteilungen gewählt werden, die den Zwischenpunkt b enthalten, und dafür folgt die Behauptung unmittelbar. ■

c) Der Mittelwertsatz der Integralrechnung

Der Mittelwertsatz der *Differentialrechnung* ist der wohl wichtigste Satz aus der Analysis I. Mit dieser überragenden Bedeutung kann der Mittelwertsatz der *Integralrechnung* nicht konkurrieren; trotzdem ist auch er eine sehr nützliche Aussage, die insbesondere auch die in §1 postulierte Interpretation von Integralen als Mittelwerten für das RIEMANN-Integral auf eine solide Grundlage stellt.

Im Hinblick auf spätere Anwendungen sei der Satz zunächst etwas allgemeiner formuliert, als wir ihn im Augenblick benötigen; an der Schwierigkeit des Beweises ändert diese Verallgemeinerung nichts.

Satz: $f\colon [a,b] \to \mathbb{R}$ sei eine stetige Funktion; $g\colon [a,b] \to \mathbb{R}$ sei stückweise stetig und nirgends negativ. Dann gibt es ein $\xi \in [a,b]$, so daß

$$\int_a^b f(x)g(x)\,dx = f(\xi)\int_a^b g(x)\,dx\,.$$

Die Funktion g sollte man sich dabei als eine „Gewichtsfunktion“ vorstellen, die die verschiedenen x-Werte verschieden stark gewichtet. Am anschaulichsten ist der Spezialfall $g \equiv 1$, der deshalb noch einmal getrennt angegeben sei:

Korollar: Für eine stetige Funktion $f\colon [a,b] \to \mathbb{R}$ gibt es ein $\xi \in [a,b]$ mit

$$f(\xi) = \frac{1}{b-a}\int_a^b f(x)\,dx\,.$$

Mit anderen Worten: Der „Mittelwert“ der Funktion f im Intervall $[a, b]$ existiert nicht nur, sondern wird auch an (mindestens) einer Stelle im Intervall angenommen.

Zum *Beweis* des Satzes betrachten wir das Minimum m und das Maximum M von f auf $[a, b]$. Da g nirgends negativ wird, gilt dann auch für jedes $x \in [a, b]$

$$m\, g(x) \leq f(x)g(x) \leq M\, g(x)$$

und damit nach der Monotonieregel

$$m \int_a^b g(x)\, dx \leq \int_a^b f(x)g(x)\, dx \leq M \int_a^b g(x)\, dx\, .$$

Es gibt daher eine reelle Zahl μ zwischen m und M, den Mittelwert, so daß

$$\int_a^b f(x)g(x)\, dx = \mu \int_a^b g(x)\, dx$$

ist, und da f stetig ist, gibt es nach dem Zwischenwertsatz (**[Bö]**, *Spezialfall 22.4;* **[D]**, *Satz 5.7;* **[MV]**, *Kap. 2, Satz 5.6c*) mindestens ein $\xi \in [a, b]$, für das $f(\xi) = \mu$ ist.

Damit ist der Satz bewiesen, und das Korollar ist natürlich einfach der Spezialfall $g \equiv 1$, für den das Integral über g gleich $b - a$ ist.

■

Man beachte, daß es hier nicht genügt, wenn f nur eine stückweise stetige Funktion ist: Für

$$f(x) = \begin{cases} 0 & \text{für } x < 0 \\ 1 & \text{für } x \geq 0 \end{cases}$$

ist

$$\frac{1}{1-(-1)} \int_{-1}^{1} f(x)\, dx = \frac{1}{2} \int_0^1 1\, dx = \frac{1}{2}$$

nicht in der Form $f(\xi)$ darstellbar.

d) Der Hauptsatz der Differential- und Integralrechnung

Die Motivation zur Einführung eines Integrals in §1 war die Suche nach einer Umkehroperation zur Differentiation: Genau wie die Differentiation aus einem zeitabhängigen Weg eine Geschwindigkeit macht, wollten wir ausgehend von einer zeitabhängigen Geschwindigkeit den Weg zurückbekommen. Der Hauptsatz der Differential- und Integralrechnung sagt uns nun endlich, daß wir dieses Ziel erreicht haben:

Hauptsatz der Differential- und Integralrechung: $f\colon [a,b] \to \mathbb{R}$ sei eine stetige Funktion und

$$F_a(x) = \int_a^b f(\xi)\,d\xi\,.$$

Dann ist $F_a'(x) = f(x)$ für alle $x \in (a,b)$.

Ist umgekehrt F eine differenzierbare Funktion, deren Ableitung auf (a,b) mit f übereinstimmt, so gibt es eine reelle Zahl C derart, daß $F(x) = F_a(x) + C$ ist.

Beweis: Die Ableitung von F_a ist

$$F_a'(x) = \lim_{h\to 0} \frac{F_a(x+h) - F_a(x)}{h}\,;$$

für positives h ist

$$F_a(x+h) - F_a(x) = \int_a^{x+h} f(x)\,dx - \int_a^x f(x)\,dx = \int_x^{x+h} f(x)\,dx$$

(s. Abschnitt *b)*); für negatives h ist entsprechend

$$F_a(x+h) - F_a(x) = -\int_{x+h}^{x} f(x)\,dx\,.$$

Nach dem Mittelwertsatz der Integralrechnung aus dem vorigen Abschnitt gibt es in jedem der beiden Fälle ein ξ zwischen x und $x+h$, so daß das Integral gleich $|h| \cdot f(\xi)$ ist, d.h.

$$\frac{F_a(x+h) - F_a(x)}{h} = f(\xi)\,.$$

Geht nun h gegen Null, so muß die zwischen x und $x+h$ liegende Zahl ξ gegen x gehen; wegen der Stetigkeit von f ist also $F_a'(x) = f(x)$

Ist F eine weitere Funktion mit Ableitung f, so hat die Differenz $h = F - F_a$ die Ableitung Null. Nun erinnern wir uns an die Analysis I: Wenn die Ableitung einer Funktion $h(x)$ verschwindet, muß die Funktion konstant sein. In der Tat: Gäbe es im Definitionsbereich von h zwei Werte $x_1 \neq x_2$ mit $h(x_1) \neq h(x_2)$, so gäbe es nach dem Mittelwertsatz der Differentialrechnung ein $\xi \in (x_1, x_2)$, so daß

$$\frac{h(x_2) - h(x_1)}{x_2 - x_1} = h'(\xi)$$

wäre. Hier steht links eine von Null verschiedene Zahl, rechts aber Null, ein Widerspruch.

Also gibt es eine Konstante C, so daß $F(x) - F_a(x) = C$ oder $F(x) = F_a(x) + C$ ist, wie behauptet.

■

Als Anwendung für die Berechnung bestimmter Integrale ergibt sich das folgende

Korollar: Ist F irgendeine differenzierbare Funktion mit der Eigenschaft, daß $F'(x) = f(x)$ ist im Intervall $[a, b]$, so gilt

$$\int_a^b f(x)\,dx = F(b) - F(a)\,.$$

Beweis: Mit obigen Bezeichnungen ist das Integral gleich $F_a(b)$. Zur Funktion F gibt es nach dem Hauptsatz eine Konstante C, so daß $F(x) = F_a(x) + C$ ist. Damit ist $F(b) - F(a) = F_a(b)$, wie gewünscht.

■

Für „einfache" Integranden ist dies im allgemeinen die beste Möglichkeit zur Berechnung eines Integrals; die Funktionen F haben daher einen Namen verdient:

Definition: Eine differenzierbare Funktion F heißt *Stammfunktion* von f, wenn $F'(x) = f(x)$ ist; wir schreiben

$$F(x) = \int f(\xi)\,d\xi + C\,,$$

wobei die *Integrationskonstante* C die Nichteindeutigkeit der Stammfunktion ausdrückt.

Diese Nichteindeutigkeit sorgt auch dafür, daß zwar nach obigem Hauptsatz die Differentiation die Integration rückgängig macht, die Integration umgekehrt aber die Differentiation nur bis auf auf eine Konstante:

Korollar: $\int f'(x)\,dx = f(x) + C$

Beweis: Klar, denn f ist eine Stammfunktion von f'. ∎

Falls wir eine Stammfunktion von f kennen, ist die Berechnung des Integrals

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

also ganz einfach. Eine erste Auswahl von Stammfunktionen erhalten wir durch Rückwärtslesen von Differentiationsregeln, zum Beispiel

$$\frac{d}{dx}x^n = nx^{n-1} \Rightarrow \int x^n\,dx = \frac{x^{n+1}}{n+1} \quad \text{falls } n \neq -1$$
$$\frac{d}{dx}e^{ax} = ae^{ax} \Rightarrow \int e^{ax}\,dx = \frac{e^{ax}}{a} \quad \text{falls } a \neq 0\,.$$

e) Partielle Integration

Eine wichtige Regel der Differentialrechnung ist die *Produktregel*

$$\frac{d}{dx}\big(f(x)g(x)\big) = f'(x)g(x) + f(x)g'(x)\,.$$

Für die Zwecke der Integralrechnung schreiben wir sie besser als

$$f'(x)g(x) = \frac{d}{dx}\big(f(x)g(x)\big) - f(x)g'(x)\,,$$

was integriert auf

$$\int f'(x)g(x)\,dx = f(x)g(x) - \int f(x)g'(x)\,dx$$

führt – die Regel der partiellen Integration. Sie ist dann nützlich, wenn sich der Integrand als Produkt schreiben läßt, wobei einer der Faktoren eine bekannte Stammfunktion hat; gelegentlich ist dann das Produkt $f(x)g'(x)$ leichter integrierbar als $f'(x)g(x)$.

Mit den wenigen Funktionen, die wir bislang kennen, lassen sich nur wenige interessante Beispiele konstruieren; die volle Kraft dieser Regel werden wir erst später kennenlernen, wenn wir unser Repertoire an Funktionen erweitert haben. Hier sei nur $\int xe^x\,dx$ betrachtet: Da die Exponentialfunktion ihre eigene Ableitung und Stammfunktion ist, empfiehlt sie sich für die Rolle der Funktion f, während $g(x) = x$ mit $g'(x) = 1$ auf Vereinfachungen auf der rechten Seite hoffen läßt. In der Tat führt partielle Integration mit

$$\int xe^x\,dx = xe^x - \int e^x\,dx = xe^x - e^x + C = (x-1)e^x + C$$

auf eine Stammfunktion. Genauso lassen sich rekursiv auch die Integrale $\int x^n e^x\,dx$ berechnen:

$$\int x^2e^x\,dx = x^2e^x - \int 2xe^x\,dx = (x^2 - 2x + 2)e^x + C$$

und

$$\int x^n e^x\,dx = x^n e^x - n\int x^{n-1}e^x\,dx\,.$$

f) Substitutionsregel

Auch die Kettenregel

$$\frac{d}{dx}f\big(g(x)\big) = f'\big(g(x)\big)\cdot g'(x)$$

läßt sich umschreiben zu einer Integrationsregel, der *Substitutionsregel*

$$\int f\big(g(x)\big)\cdot g'(x)\,dx = F\big(g(x)\big) + C\,,$$

wobei F eine Stammfunktion von f ist.

Diese Regel dürfte wohl meist der erfolgversprechendste Versuch zum Auffinden einer Stammfunktion sein – vor allem, wenn man sie von rechts nach links liest, um für eine bekannte Funktion f die unbekannte Stammfunktion F zu berechnen. Als Substitution g wählt man hier eine „geeignete“ bijektive Funktion, die zu einer Vereinfachung auf der linken Seite führt. Wir werden in §5 zahlreiche Beispiele dazu kennenlernen; im Augenblick fehlt uns allerdings noch ein hinreichender Vorrat an „geeigneten“ Funktionen.

§4: Komplexe Zahlen und die elementaren transzendenten Funktionen

Bevor ein weiterer Ausbau der Integralrechnung möglich ist, müssen wir also unbedingt mehr Funktionen kennenlernen, insbesondere die trigonometrischen Funktionen und ihre Umkehrfunktionen. Sowohl in der Elektrotechnik als auch in der Wellenoptik treten diese meist in komplexer Form auf; es bietet sich daher an, in diesem Zusammenhang gleich auch noch die komplexen Zahlen einzuführen.

a) Der Körper der komplexen Zahlen

Rekapitulieren wir kurz, wie die Mathematik zu den reellen Zahlen kam.

Im Anfang waren die natürlichen Zahlen $1, 2, 3, \ldots$. Man kann sie addieren und multiplizieren, aber man kann dort weder die Gleichung $5 + x = 2$ noch die Gleichung $5 \cdot x = 2$ lösen. Ersteres Problem führt auf die Erweiterung der Menge $\mathbb{N}$ der natürlichen Zahlen zur Menge $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$ der ganzen Zahlen, in der auch die Subtraktion uneingeschränkt ausführbar ist; letzteres führte – zumindest für positive Zahlen schon erheblich früher – zur Einführung von Brüchen und letztlich zum Körper $\mathbb{Q}$ der rationalen Zahlen, in dem *alle* Grundrechenarten uneingeschränkt ausführbar sind – ausgenommen natürlich die Division durch Null.

Daß auch Brüche nicht ausreichen, bemerkten die Pythagoräer, als sie zu ihrem Entsetzen feststellen mußten, daß die Länge der Diagonalen eines Quadrats in keinem rationalen Verhältnis zur Seitenlänge steht, mit anderen Worten, daß $\sqrt{2} \notin \mathbb{Q}$. Kombiniert mit dem Interesse an Grenzwerten bei EUDOXOS, ARCHIMEDES und vor allem rund 2 000 Jahre später bei EULER, LEIBNIZ, NEWTON und vielen anderen, führte dies schließlich zum Körper $\mathbb{R}$ der reellen Zahlen – wenn auch der Begriff des *Körpers* erst viel später in den Zwanzigerjahren unseres Jahrhunderts auftauchte.

Auch hier ist noch nicht alles möglich; beispielsweise hat zwar die Gleichung $x^2 = 3$ eine reelle Lösung (genauer gesagt zwei), nicht aber die Gleichung $x^2 = -3$. Diese könnte gelöst werden, wenn man die Quadratwurzel aus -1 ziehen könnte, denn dann wäre $\sqrt{-3} = \sqrt{-1} \cdot \sqrt{3}$. Genauso überlegt man sich leicht, daß dann jede quadratische Gleichung eine Lösung hätte.

Es liegt also nahe, auch die reellen Zahlen zu erweitern, indem man ein neues Element hinzufügt, dessen Quadrat -1 ist. Dieses Element wird in der Mathematik und Physik traditionellerweise mit i wie *imaginär* bezeichnet, da es zur Zeit seiner Einführung als imaginäre, d.h. nur in der Vorstellung vorhandene, Zahl angesehen wurde. In der Elektrotechnik, wo I und i eine feste andere Bedeutung haben, verwendet man stattdessen den Buchstaben j.

Natürlich genügt es nicht, nur die Menge $\mathbb{R} \cup \{i\}$ zu betrachten, denn wir möchten mit den neuen Zahlen auch rechnen und dabei möglichst wenig von den bewährten Rechenregeln für reelle Zahlen aufgeben.

Insbesondere sollten also alle Zahlen der Form $x+iy$ mit $x, y \in \mathbb{R}$ in der neuen Menge liegen, und es sollte möglich sein, mit ihnen wie gewohnt zu rechnen. Für zwei Zahlen $x + iy$ und $u + iv$ müßte also gelten

$$(x + iy) + (u + iv) = (x + u) + (iy + dv) = (x + u) + i(y + v)$$

und

$$\begin{aligned}(x + iy)(u + iv) &= xu + iyu + xiv + (iy)(iv) = xu + iyu + ixv + i^2 yv \\ &= (xu - yv) + (yu + xv)i\,.\end{aligned}$$

Damit lassen sich Zahlen dieser Form insbesondere addieren und multiplizieren, ohne daß neue Zahlen entstehen; wir können hoffen, daß sie vielleicht sogar schon für die geplante Erweiterung ausreichen.

Um dies zu überprüfen, nehmen wir die oben heuristisch abgeleiteten Regeln als *Definition* von Addition und Multiplikation und untersuchen die entstehende Struktur:

Definition: *a)* Die Menge $\mathbb{C}$ der komplexen Zahlen ist die Menge aller formaler Ausdrücke der Form $x + iy$ mit $x, y \in \mathbb{R}$.
b) Auf $\mathbb{C}$ wird eine Verknüpfung „+“ definiert durch die Vorschrift

$$(x + iy) + (u + iv) = (x + u) + i(y + v)\,.$$

c) Dazu kommt eine Verknüpfung „·“, definiert durch

$$(x + iy) \cdot (u + iv) = (xu - yv) + i(xv + yu)\,.$$

d) Für $z = x + iy \in \mathbb{C}$ nennen wir x den *Realteil* und y den *Imaginärteil* von z; in Zeichen

$$x = \mathfrak{Re}\, z \quad \text{und} \quad y = \mathfrak{Im}\, z\,.$$

(Eine komplexe Zahl heißt *komplex,* weil sie aus einem Realteil und einem Imaginärteil *zusammengesetzt* ist.)

Satz: Die Menge $\mathbb{C}$ mit den so definierten Verknüpfungen ist ein Körper.

Beweis: Eigentlich müßten wir alle Körperaxiome einzeln überprüfen; um aber nicht garzu viel Papier zu produzieren, möchte ich mich hier im Sinne des Umweltschutzes auf die interessantesten beschränken.

Völlig uninteressant sind die Axiome, die festlegen, daß $(\mathbb{C}, +)$ eine abelsche Gruppe sein soll: Da die Addition komponentenweise für Realteil und Imaginärteil definiert ist, folgen alle Axiome sofort aus den entsprechenden Axiomen für $\mathbb{R}$. Das Neutralelement bezüglich der Addition ist natürlich $0 + i0$, und $-(x + iy) = (-x) + i(-y)$.

Die Gruppenstruktur von $\mathbb{C} \setminus \{0\}$ bezüglich der Multiplikation ist weniger offensichtlich. Unmittelbar einsichtig ist das Kommutativitätsgesetz; das Assoziativitätsgesetz ist eine sture Nachrechnerei, die jeder einmal, aber nur einmal in seinem Leben wirklich ausführen sollte. (Ich habe sie glücklicherweise schon hinter mir.)

Neutralelement bezüglich der Multiplikation kann, wenn alles Sinn haben soll, nur $1 + i0$ sein, und in der Tat sieht man sofort, daß jedes Element $x + iy$ sowohl bei Links- wie auch bei Rechtsmultiplikation hiermit sich selbst liefert.

Bleibt die Existenz eines multiplikativen Inversen, und hier hilft nur ein Trick: Für $x + iy \neq 0 + i0$ ist

$$(x + iy) \cdot (x - iy) = x^2 + y^2 \in \mathbb{R}_{>0}$$

eine positive reelle Zahl; falls also ein Inverses existiert und die üblichen Regeln der Bruchrechnung gelten, ist

$$\frac{1}{x + iy} = \frac{x - iy}{(x + iy)(x - iy)} = \frac{x - iy}{x^2 + y^2} = \frac{x}{x^2 + y^2} - i \cdot \frac{y}{x^2 + y^2} .$$

Eine leichte Rechnung zeigt, daß dieses ganz rechts stehende Element von $\mathbb{C}$ in der Tat sowohl bei Links- als auch bei Rechtsmultiplikation mit $x + iy$ das Neutralelement $1 + i0$ liefert.

Die noch verbleibenden Distributivgesetze sind wieder Rechnerei zum Abhaken, die man genau einmal in seinem Leben durchführen sollte. ■

Damit sind also alle von der reellen Zahlen gewohnte Rechenregeln für die Grundrechenarten auch für komplexe Zahlen gültig; nicht zu retten sind allerdings die Regeln über die *Ordnungsbeziehung:* Falls es in $\mathbb{C}$ eine mit der algebraischen Struktur kompatible Ordnungsrelation gäbe, müßte $i^2 \geq 0$ sein, was nicht im Sinne des Erfinders ist.

Es ist nun klar, daß die Abbildung

$$\begin{cases} \mathbb{R} & \hookrightarrow \mathbb{C} \\ x & \mapsto x + i0 \end{cases}$$

eine Einbettung des Körpers der reellen in den der komplexen Zahlen definiert; da wir ersteren erweitern wollen, betrachten wir diese Einbettung als Identifikation, d.h. wir identifizieren den „formalen Ausdruck“ $x+i0$ mit der reellen Zahl x. Insbesondere sind jetzt also 0 und 1 das additive und das multiplikative Neutralelement. Außerdem schreiben wir kurz i anstelle von $0 + i1$; nach den Rechenregeln, die wir inzwischen kennen, ist der „formale Ausdruck $x + iy$“ dann nichts anderes als die mit den Rechenoperationen von $\mathbb{C}$ berechnete komplexe Zahl $x + i \cdot y$.

Der Erfolg, den wir bei der Herleitung des multiplikativen Inversen durch Erweiterung mit $x - iy$ hatten, verdient es, genauer untersucht zu werden:

Definition: Für $z = x + iy \in \mathbb{C}$ heißt $\overline{z} = x - iy$ die zu z konjugiert komplexe Zahl.

(In einigen Büchern wird die konjugiert komplexe Zahl zu z auch mit z^* bezeichnet.)

Offensichtlich ist

$$\overline{z + w} = \overline{z} + \overline{w}\,;$$

auch das Nachrechnen von

$$\overline{zw} = \overline{z} \cdot \overline{w}$$

bereitet keinerlei Schwierigkeiten.

Für die Herleitung des Inversen wesentlich war die Tatsache, daß

$$(x + iy)(x - iy) = x^2 + y^2$$

eine reelle Zahl ist, die genau dann verschwindet, wenn sowohl x als auch y und damit auch $x+iy$ verschwinden; wir bezeichnen die Quadratwurzel aus dieser nichtnegativen reellen Zahl als *Betrag* der komplexen Zahl:

Definition: Der Betrag einer komplexen Zahl $z = x + iy$ ist

$$|z| = \sqrt{z \cdot \overline{z}} = \sqrt{x^2 + y^2}\,.$$

Für reelles z stimmt das natürlich genau mit dem gewohnten Betrag einer reellen Zahl überein.

Offensichtlich ist $|z| = 0 \Leftrightarrow z = 0$ und

$$\frac{1}{z} = \frac{\overline{z}}{z \cdot \overline{z}} = \frac{\overline{z}}{|z|^2}\,.$$

Der Betrag hat noch eine weitere nützliche Interpretation: Für zwei komplexe Zahlen $z = x + iy$ und $w = u + iv$ ist

$$|z - w| = \sqrt{(x - u)^2 + (y - v)^2}$$

gerade der EUKLIDische Abstand zwischen den Punkten (x, y) und (u, v) der EUKLIDischen Ebenen $\mathbb{R}^2$. Da die komplexen Zahlen natürlich über die Abbildung

$$\begin{cases} \mathbb{C} & \to \mathbb{R}^2 \\ x + iy & \mapsto (x, y) \end{cases}$$

in Bijektion mit den Punkten der EUKLIDischen Ebenen stehen, können wir die komplexen Zahlen also auch identifizieren mit den Punkten der EUKLIDischen Ebenen, wobei der Betrag der Differenz zwischen zwei Zahlen gerade dem Abstand entspricht. Man bezeichnet den Körper der komplexen Zahlen in diesem Zusammenhang auch als GAUSS*sche Zahlenebene* nach dem großen deutschen Mathematiker CARL FRIEDRICH GAUSS (1777–1855). Sie ist zwar nicht zusammen mit GAUSS auf dem Zehnmarkschein abgebildet, war aber das Thema der Sondermarke zu seinem zweihundertsten Geburtstag.

b) Polarkoordinaten und trigonometrische Funktionen

Wenn wir den Körper der komplexen Zahlen als GAUSSsche Zahlenebene auffassen, sind Real- und Imaginärteil die kartesischen Koordinaten. In manchen Anwendungen nützlicher sind die *Polarkoordinaten,* die einen Punkt z festlegen durch seinen Abstand r vom Nullpunkt und den Winkel φ, den der Strahl von Null nach z mit der reellen Achse bildet. Diesen Winkel messen wir im sogenannten *Bogenmaß:* Die Größe des Winkels ist demnach die Länge des Kreisbogens auf dem Kreis mit Radius eins, den der Winkel ausschneidet. Das Vorzeichen ist dabei so gewählt, daß die Richtung von der positiven reellen zur positiven imaginären Achse positiv, die Gegenrichtung negativ angegeben wird. Der mathematisch positive Drehsinn ist demnach der *Gegenuhrzeigersinn;* der Uhrzeigersinn wird negativ gerechnet.

Um Winkel konkret angeben zu können, müssen wir noch den Umfang des Einheitskreises kennen. In Übereinstimmung mit der klassischen Definition der griechischen Geometer, wonach π das Verhältnis zwischen Umfang und *Durchmesser* eines Kreises ist, definieren wir eine reelle Zahl π dadurch, daß der Einheitskreis Umfang 2π haben soll. (Der Einheitskreis hat *Radius* eins, also *Durchmesser* zwei.)

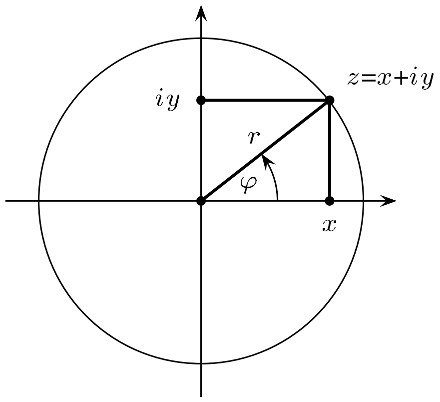


Abb. 10: Polarkoordinaten einer komplexen Zahl

Strenggenommen müßten wir hierzu erst definieren, was wir unter der Länge eines Kurvenstücks verstehen wollen: Genau wie wir beliebige Flächen als Grenzwerte von Summen von Rechteckflächen definiert haben, können wir Längen von Kurvenstücken als Grenzwerte von Summen von Streckenlängen definieren. Da wir aber gerade das relativ trockene Kapitel über die Präzisierung des Flächenbegriffs durch das RIEMANN-Integral hinter uns haben, sei die exakte Definition bis zum Abschnitt über Linienintegrale zurückgestellt; wir nehmen einfach an, daß wir bei einer so einfachen Kurve wie dem Kreis schon wissen, was gemeint ist.

Insbesondere ist anschaulich ziemlich klar, daß sich die 360° der klassischen Gradeinteilung gleichmäßig auf den Kreisumfang verteilen, so daß ein Winkel von α Grad das Bogenmaß

$$\alpha \cdot \frac{2\pi}{360^\circ} = \alpha \cdot \frac{\pi}{180^\circ}$$

hat. Einige wichtige Werte sind

0°	30°	45°	60°	90°	120°	135°	150°	180°
0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$	$\frac{2\pi}{3}$	$\frac{3\pi}{4}$	$\frac{5\pi}{6}$	π

Der Abstand r vom Nullpunkt läßt sich leicht aus den kartesischen Koordinaten berechnen: Für $z = x + iy$ ist nach dem Satz des PYTHAGORAS, angewandt auf eines der Dreiecke in Abbildung 10, $r^2 = x^2 + y^2$, also $r = |z|$. Für die Berechnung de Winkels φ fehlen uns im Augenblick noch die notwendigen Funktionen. Auf dem Weg zu ihrer Einführung starten wir besser mit dem umgekehrten Problem:

Gegeben seien die Polarkoordinaten (r, φ) einer komplexen Zahl $z = x + iy \neq 0$. Dann ist z/r eine komplexe Zahl vom Betrag eins, die somit nur vom Winkel φ abhängt.

Definition: Wir bezeichnen den *Realteil* der komplexen Zahl z mit Betrag eins und Argument φ als *Cosinus* von φ und ihren *Imaginärteil* als *Sinus* von φ, in Zeichen

$$\sin\varphi \underset{\text{def}}{=} \mathfrak{Im}\, z \quad \text{und} \quad \cos\varphi \underset{\text{def}}{=} \mathfrak{Re}\, z\,.$$

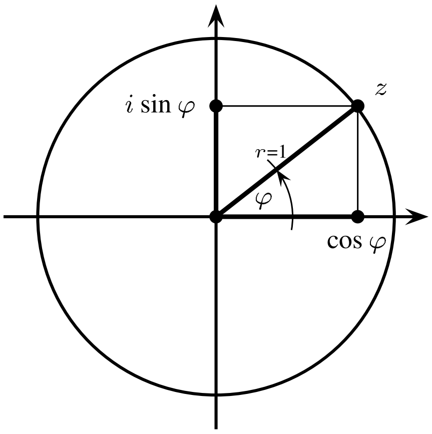


Abb. 11: Sinus und Cosinus eines Winkels

Diese Definition stimmt mit der (vielleicht) aus der Schule bekannten Definition der Winkelfunktionen überein: Dort geht man bekanntlich aus von einem rechtwinkligen Dreieck, in dem φ ein Winkel ist, und setzt

$$\cos\varphi = \frac{\text{Ankathete}}{\text{Hypothenuse}} \quad \text{und} \quad \sin\varphi = \frac{\text{Gegenkathete}}{\text{Hypothenuse}}\,.$$

Zeichnet man dieses Dreieck in der GAUSSschen Zahlenebenen so ein, daß der Scheitelpunkt des Winkels im Nullpunkt liegt und die Ankathete auf der reellen Achse und betrachtet man die Schnittpunkte der beiden Strahlen des Winkels mit dem Einheitskreis, so zeigt eine einfache Anwendung des Strahlensatzes auf die beiden Dreiecke in Abbildung 12

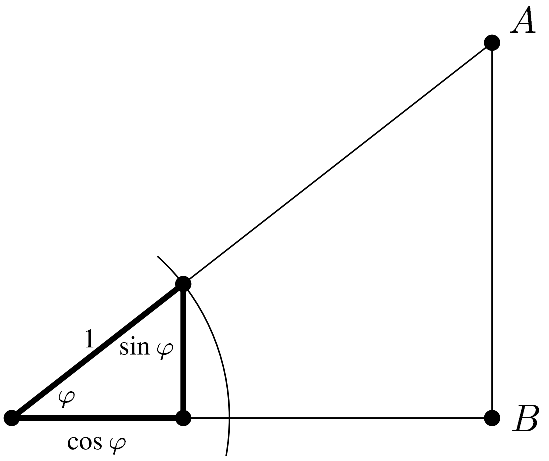


Abb. 12: Winkelfunktionen und Längenverhältnisse im rechtwinkligen Dreieck

sofort, daß die angegebenen Verhältnisse tatsächlich gleich dem hier definierten Sinus *bzw.* Cosinus sind.

Wendet man diese Regeln an auf das gleichschenklige rechtwinklige Dreieck, das entlang einer Mittelsenkrechten zerteilte gleichseitige Dreieck und das zu einer Strecke entartete rechtwinklige Dreieck, kommt man leicht auf die speziellen Werte

$$\begin{array}{llllll} \varphi & 0 & \frac{\pi}{6} & \frac{\pi}{4} & \frac{\pi}{3} & \frac{\pi}{2} \\ \sin\varphi & 0 & \frac{1}{2} & \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{3} & 1 \\ \cos\varphi & 1 & \frac{1}{2}\sqrt{3} & \frac{1}{2}\sqrt{2} & \frac{1}{2} & 0\,. \end{array}$$

Als Eselsbrücke kann man sich dies auch in der Form

$$\begin{array}{llllll} \varphi & 0 & \frac{\pi}{6} & \frac{\pi}{4} & \frac{\pi}{3} & \frac{\pi}{2} \\ \sin\varphi & \frac{1}{2}\sqrt{0} & \frac{1}{2}\sqrt{1} & \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{3} & \frac{1}{2}\sqrt{4} \\ \cos\varphi & \frac{1}{2}\sqrt{4} & \frac{1}{2}\sqrt{3} & \frac{1}{2}\sqrt{2} & \frac{1}{2}\sqrt{1} & \frac{1}{2}\sqrt{0} \end{array}$$

merken.

Abbildung 11 zeigt einige erste Eigenschaften der gerade definierten trigonometrischen Funktionen; beispielsweise ist nach dem Satz des PYTHAGORAS

$$\sin^2\varphi + \cos^2\varphi = 1\,,$$

wobei hier wie auch im folgenden die bei trigonometrischen Funktionen allgemein übliche Konvention

$$\sin^n \varphi \underset{\text{def}}{=} (\sin \varphi)^n, \quad \cos^n \varphi \underset{\text{def}}{=} (\cos \varphi)^n$$

verwendet wird.

Durch Spiegelung an der ersten Winkelhalbierenden erhält man auch sofort die Beziehungen

$$\sin\left(\frac{\pi}{2} - \varphi\right) = \cos \varphi \quad \text{und} \quad \cos\left(\frac{\pi}{2} - \varphi\right) = \sin \varphi\,;$$

entsprechend führt Spiegelung an der reellen Achse auf

$$\sin(-\varphi) = -\sin \varphi \quad \text{und} \quad \cos(-\varphi) = \cos \varphi$$

und Spiegelung am Nullpunkt auf

$$\sin(\varphi + \pi) = -\sin \varphi \quad \text{und} \quad \cos(\varphi + \pi) = -\cos \varphi\,.$$

Kombiniert man mehrere dieser Regeln miteinander, folgt weiter, daß beispielsweise

$$\sin\left(\frac{\pi}{2} + \varphi\right) = \cos \varphi \quad \text{und} \quad \cos\left(\frac{\pi}{2} + \varphi\right) = -\sin \varphi$$

und

$$\sin(\pi - \varphi) = \sin \varphi \quad \text{und} \quad \cos(\pi - \varphi) = -\cos \varphi$$

ist. Da Winkel nur bis auf Addition eines Vielfachen des Kreisumfangs 2π bestimmt sind, folgt auch sofort die Periodizität der Winkelfunktion mit Periode 2π; somit gilt auch

$$\sin(\varphi + 2k\pi) = \sin \varphi \quad \text{und} \quad \cos(\varphi + 2k\pi) = \cos \varphi$$

für jede ganze Zahl $k \in \mathbb{Z}$.

c) Die Ableitungen der trigonometrischen Funktionen

Da $\cos x = \sin\left(x + \frac{\pi}{2}\right)$ ist, genügt es, den Sinus abzuleiten. Wie üblich ist

$$\sin'(x) = \lim_{h \to 0} \frac{\sin(x + h) - \sin x}{h}\,.$$

Betrachten wir dies geometrisch in der komplexen Ebenen: Wir tragen x als Winkel zur reellen Achse ab dem Nullpunkt auf; in Abbildung 14

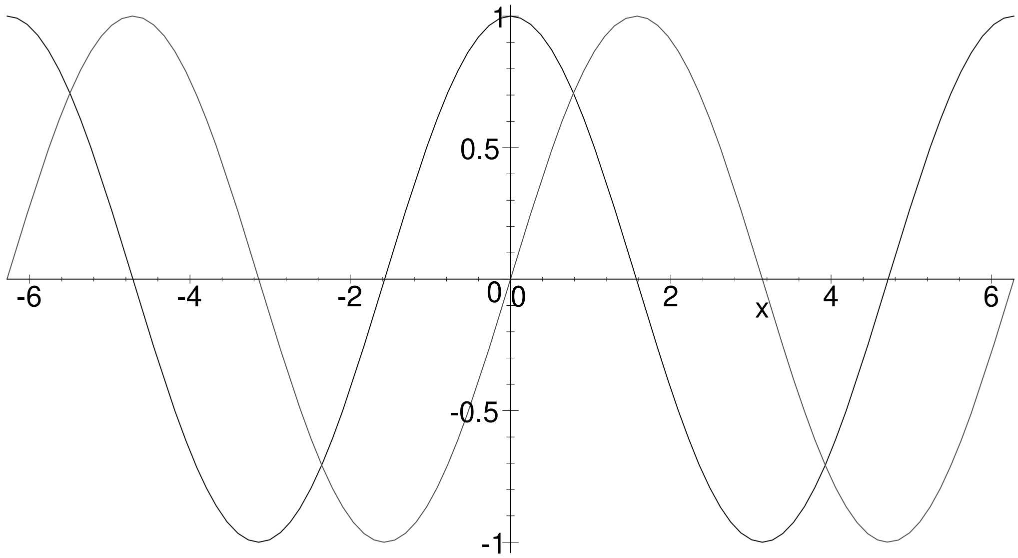


Abb. 13: Sinus und Cosinus zwischen -2π und 2π

ist x also gleich der Länge des Bogens von 1 nach A. Entsprechend ist $x + h$ gleich der Länge des Bogens von 1 bis B, und h selbst ist die Länge des Bogens von A nach B.

Nach Definition ist $\sin x$ der Imaginärteil von A und $\sin(x+h)$ der von B; obiger Zähler $\sin(x + h) - \sin x$ ist also die Länge der Strecke $\overline{CB}$. Der Nenner h ist die Länge des Bogens von A nach B. Da wir Bogenlängen, wie oben angedeutet, über immer feiner unterteilte Streckenzüge definieren, ist dieser Bogen für immer kleiner werdendes h beliebig wenig von der Strecke $\overline{AB}$ unterscheidbar, d.h.

$$\lim_{h\to 0} \frac{\sin(x + h) - \sin x}{h} = \lim_{h\to 0} \frac{\overline{CB}}{\overline{AB}} .$$

Da $\triangle ABC$ rechtwinklig ist, ist dieses Streckenverhältnis gerade der Cosinus des Winkels β bei B, und wir müssen diesen berechnen.

Dazu betrachten wir zunächst das Dreieck $\triangle 0AB$. Da A und B auf dem Einheitskreis liegen, ist es gleichschenklig und hat somit bei A und B denselben Winkel. Da die Winkelsumme im Dreieck gleich π und der Winkel bei 0 gleich h ist, errechnet sich dieser zu $\frac{\pi-h}{2}$.

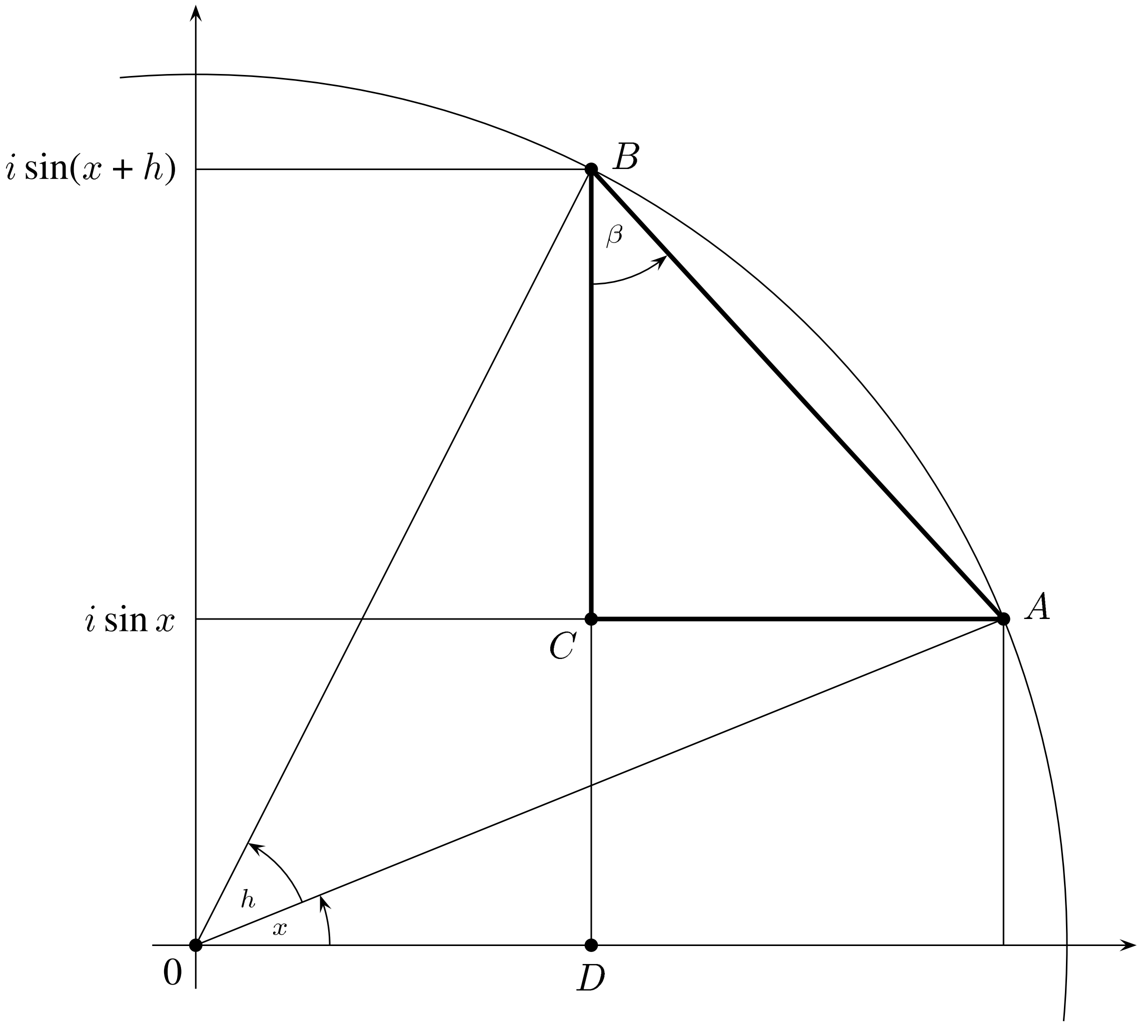


Abb. 14: Zur Ableitung des Sinus

Als nächstes betrachten wir das rechtwinklige Dreieck $\triangle 0BD$. Da es bei 0 den Winkel $x+h$ hat, muß der Winkel bei B gleich

$$\pi - \frac{\pi}{2} - (x+h) = \frac{\pi}{2} - (x+h)$$

sein, und β als Differenz dieser beiden Winkel ergibt sich zu

$$\beta = \frac{\pi - h}{2} - \frac{\pi}{2} + x + h = x + \frac{h}{2}\,.$$

Somit ist

$$\sin'(x) = \lim_{h\to 0} \cos\left(x + \frac{h}{2}\right) = \cos x\,,$$

die Ableitung des Sinus ist also der Cosinus.

Daraus folgt nun sofort, daß

$$\cos'(x) = \frac{d}{dx} \sin\left(x + \frac{\pi}{2}\right) = \cos\left(x + \frac{\pi}{2}\right) = -\sin x\,.$$

ist.

d) Lösungen der Differentialgleichung $f''(x) = -\omega^2 f(x)$

Werden Sinus oder Cosinus zweimal abgeleitet, so wird je einmal ein Sinus und ein Cosinus abgeleitet; im letzteren Fall kehrt sich das Vorzeichen um, im ersteren bleibt es gleich. Daher ist die zweite Ableitung von Sinus *bzw.* Cosinus gerade die negativ genommene jeweilige Funktion.

Allgemein gilt für jedes $\omega \in \mathbb{R}$, daß sowohl für $f(x) = \cos\omega x$ als auch für $f(x) = \sin\omega x$

$$f''(x) = -\omega^2 f(x)$$

ist, oder, wie man auch sagt: Beide Funktionen genügen der *Differentialgleichung* $f''(x) = -\omega^2 f(x)$.

Diese Gleichung sollte, vor allem wenn man x durch die Zeit t ersetzt, aus der Physik bekannt sein: Sie beschreibt die Schwingungen eines Gewichts an einer Feder, deren Rückstellkraft sich aus dem HOOKEschen Gesetz ergibt, die Schwingungen eines Pendels bei kleiner Auslenkung, den Strom in einem ungedämpften elektrischen Schwingkreis *usw.;* Grund genug also, sie etwas genauer zu betrachten.

Da $\sin\omega x$ und $\cos\omega x$ Lösungen sind, ist auch jede Linearkombination

$$f(x) = a\cos\omega x + b\sin\omega x \quad \text{mit} \quad a, b \in \mathbb{R}$$

eine Lösung. Das Hauptergebnis dieses Paragraphen besagt, daß dies im Fall $\omega \neq 0$ bereits *alle* Lösungen sind: (Was gilt für $\omega = 0$?)

Satz: Die in einem offenen Intervall (r, s) mit $r < 0 < s$ definierte reelle Funktion f genüge der Differentialgleichung $f''(x) = -\omega^2 f(x)$ mit einem nichtverschwindenden ω. Dann gibt es Konstanten $a, b \in \mathbb{R}$, so daß $f(x) = a\cos\omega x + b\sin\omega x$ ist.

Beweis: Wir multiplizieren die Differentialgleichung mit $2f'(x)$ und erhalten

$$2f'(x)f''(x) = -2\omega^2 f(x)f'(x)\,.$$

Nach der Produktregel können wir dies auch als

$$\frac{d}{dx}f'(x)^2 = -\omega^2\frac{d}{dx}f(x)^2 \quad \text{oder} \quad \frac{d}{dx}\left(f'(x)^2 + \omega^2 f(x)^2\right) = 0$$

schreiben.

Wie schon in §3*d)* folgt damit aus dem Mittelwertsatz der Differentialrechnung, daß die Funktion $f'(x)^2 + \omega^2 f(x)^2$ konstant sein muß, es gibt also eine Konstante $c \in \mathbb{R}$, so daß

$$f'(x)^2 + \omega^2 f(x)^2 = c$$

für alle x. Wir unterscheiden zwei Fälle:

Erster Fall: $c = 0$. Da Quadrate reeller Zahlen nie negativ sein können, muß dann für jedes x gelten $f(x) = f'(x) = 0$, und mit $a = b = 0$ läßt sich die Nullfunktion in der Tat als Linearkombination von $\cos\omega x$ und $\sin\omega x$ schreiben, so daß der Satz für diesen Fall bewiesen ist.

Zweiter Fall: $c \neq 0$. Dieser Fall soll auf den ersten zurückgeführt werden.

Falls wir annehmen, der Satz sei richtig, gibt es $a, b \in \mathbb{R}$ mit $f(x) = a\cos\omega x + b\sin\omega x$. Dann ist

$$f(0) = a \quad \text{und} \quad f'(0) = b\omega\,,$$

also

$$a = f(0) \quad \text{und} \quad b = \frac{f'(0)}{\omega}\,.$$

Für eine beliebige Lösung f der Differentialgleichung $f''(x) = -\omega f(x)$ *definieren* wir reelle Zahlen a, b über diese Gleichungen und betrachten die Funktion

$$h(x) = f(x) - f(0)\cos\omega x - \frac{f'(0)}{\omega}\sin\omega x\,.$$

Auch h ist, wie man leicht nachrechnet, eine Lösung der Differentialgleichung, und nach Konstruktion ist $h(0) = h'(0) = 0$. Daher ist für h die oben definierte Konstante c, die man ja über jeden beliebigen x-Wert berechnen kann, Null; nach dem, was wir im ersten Fall gezeigt haben, ist h somit die Nullfunktion. Damit ist aber

$$f(x) = f(0)\cos\omega x + \frac{f'(0)}{\omega}\sin\omega x\,,$$

wie behauptet. ∎

e) Potenzreihenentwicklung der trigonometrischen Funktionen

Nachdem wir aus Abschnitt *c)* die ersten Ableitungen von Sinus und Cosinus kennen, bereiten auch die höheren Ableitungen keine Schwierigkeiten mehr: Die Ableitung von Sinus ist Cosinus, dessen Ableitung wiederum −Sinus, die Ableitung davon ist −Cosinus, und die Ableitung davon wieder der Sinus, womit alles wieder von vorne losgeht. Formal aufgeschrieben heißt das

$$\frac{d^{(n)}}{dx^n}\sin x = \begin{cases} \sin x & \text{falls } n \equiv 0 \bmod 4 \\ \cos x & \text{falls } n \equiv 1 \bmod 4 \\ -\sin x & \text{falls } n \equiv 2 \bmod 4 \\ -\cos x & \text{falls } n \equiv 3 \bmod 4\,, \end{cases}$$

wobei

$$a \equiv b \bmod m$$

bedeuten soll, daß $a - b$ ohne Rest durch m teilbar ist. Entsprechend ist

$$\frac{d^{(n)}}{dx^n}\cos x = \begin{cases} \cos x & \text{falls } n \equiv 0 \bmod 4 \\ -\sin x & \text{falls } n \equiv 1 \bmod 4 \\ -\cos x & \text{falls } n \equiv 2 \bmod 4 \\ \sin x & \text{falls } n \equiv 3 \bmod 4\,. \end{cases}$$

Da $\sin 0 = 0$ und $\cos 0 = 1$ ist, sind damit auch die Werte aller höherer Ableitungen an der Stelle $x = 0$ bekannt, und die TAYLOR-Polynome

$$\begin{aligned} P_{2n+1}(x) = P_{2n+2}(x) &= \sum_{\nu=0}^{n} (-1)^\nu \frac{x^{2\nu+1}}{(2\nu+1)!} \\ &= x - \frac{x^3}{6} + \frac{x^5}{120} - \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} \end{aligned}$$

für Sinus und

$$\begin{aligned} Q_{2n}(x) = Q_{2n+1}(x) &= \sum_{\nu=0}^{n} (-1)^\nu \frac{x^{2\nu}}{(2\nu)!} \\ &= 1 - \frac{x^2}{2} + \frac{x^4}{24} - \cdots + (-1)^n \frac{x^{2n}}{(2n)!} \end{aligned}$$

für Cosinus bereiten keine Schwierigkeiten mehr.

Noch nicht ganz klar ist allerdings, ob die als Limes dieser Polynome erhaltenen TAYLOR-Reihen die Funktionen Sinus und Cosinus tatsächlich darstellen; bei einigen Funktionen wie etwa $f(x) = e^{-1/x^2}$ kann es ja durchaus vorkommen, daß die TAYLOR-Reihe gegen eine andere Funktion (im Beispiel die Nullfunktion) konvergiert als die Ausgangsfunktion.

Hier ist das glücklicherweise nicht der Fall: Zunächst einmal zeigt fast jedes der gängigen Konvergenzkriterien für Reihen, unter anderem etwa das Quotientenkriterium, fast sofort, daß die beiden TAYLOR-Reihen für alle $x \in \mathbb{R}$ konvergieren und damit dort zumindest irgendwelche Funktionen darstellen. Gliedweises Differenzieren zeigt, daß auch die Ableitungen existieren und konvergieren und daß beide TAYLOR-Reihen die Differentialgleichung $f''(x) = -f(x)$ erfüllen.

Nach dem Ergebnis des letzten Abschnitts stellen daher beide TAYLOR-Reihen Funktionen der Form $a \cos x + b \sin x$ dar. Vergleicht man die Funktionswerte und die Werte der ersten Ableitung an der Stelle Null, folgt sofort, daß in der Tat

$$\sin x = \sum_{\nu=0}^{\infty} (-1)^\nu \frac{x^{2\nu+1}}{(2\nu+1)!} \quad \text{und} \quad \cos x = \sum_{\nu=0}^{\infty} (-1)^\nu \frac{x^{2\nu}}{(2\nu)!}$$

ist.

f) Die Eulerschen Formeln und ihre Anwendung

Die TAYLOR-Reihen für Sinus und Cosinus erinnern von ihrer Struktur her an die der Exponentialfunktion: Der einzige Unterschied besteht darin, daß jeweils nur „die Hälfte“ der Terme auftritt und daß das Vorzeichen alterniert.

Um den Zusammenhang mit der Exponentialfunktion zu verstehen, gehen wir zurück zum Beginn dieses Paragraphen, wo wir das Rechnen mit komplexen Zahlen lernten. Wenn wir die Potenzen von i betrachten, haben wir wie bei den Ableitungen von Sinus und Cosinus eine Periodizität mit Periode vier: Wir erhalten nämlich abwechselnd die Werte $i, -1, -i$

und 1. Dies können wir in die TAYLOR-Reihe der Exponentialfunktion einsetzen für ein rein imaginäres Argument $ix,\ x \in \mathbb{R}$:

$$\begin{aligned} e^{ix} &= \sum_{\mu=0}^{\infty} \frac{(ix)^\mu}{\mu!} = \sum_{\nu=0}^{\infty} (-1)^\nu \frac{x^{2\nu}}{(2\nu)!} + i \sum_{\nu=0}^{\infty} (-1)^\nu \frac{x^{2\nu+1}}{(2\nu+1)!} \\ &= \cos x + i \sin x\,. \end{aligned}$$

Entsprechend ist

$$e^{-ix} = \cos x - i \sin x\,,$$

also

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

und

$$\sin x = \mathfrak{Im}\, e^{ix} = \frac{e^{ix} - e^{-ix}}{2i}\,.$$

Auch die Exponentialfunktion für ein beliebiges komplexes Argument läßt sich leicht ausrechnen: Für $z = x + iy$ ist

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

Alle diese Formeln werden nach dem Schweizer Mathematiker LEONHARD EULER (1707–1783) als EULERsche Formeln bezeichnet. Sie erlauben es, die trigonometrischen Funktionen auf die erheblich handlichere Exponentialfunktion zurückzuführen; sowohl in der Wechselstromtechnik als auch in der Wellenoptik macht man hiervon regen Gebrauch.

Als Beispiel betrachten wir die eher unhandlichen Additionsformeln für die trigonometrischen Funktionen: Für die Exponentialfunktion ist natürlich

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

Nach den EULERschen Formeln ist die linke Seite gleich

$$\cos(x+y) + i \sin(x+y)\,,$$

und die rechte ist

$$\begin{aligned} &(\cos x + i \sin x)(\cos y + i \sin y) \\ &= \cos x \cos y - \sin x \sin y + i(\sin x \cos y + \cos x \sin y)\,. \end{aligned}$$

Daher ist

$$\cos(x + y) = \cos x \cos y - \sin x \sin y$$

und

$$\sin(x + y) = \sin x \cos y + \cos x \sin y \, .$$

Entsprechend lassen sich auch $\cos nx$ und $\sin nx$ berechnen.

Die EULERschen Formeln sagen uns auch, welche Werte die Exponentialfunktion im Komplexen annimmt: Für $z = x + iy$ ist $e^z = e^x(\cos y + i \sin y)$ eine komplexe Zahl vom Betrag e^x mit Argument y. Wenn z die komplexen und damit x die reellen Zahlen durchläuft, nimmt e^x jeden positiven Wert an und y jeden reellen Wert, e^z durchläuft also alle komplexen Zahlen außer der Null, d.h. im Komplexen bildet die Exponentialfunktion $\mathbb{C}$ surjektiv ab auf $\mathbb{C} \setminus \{0\}$.

Schließlich erlauben die EULERschen Formeln auch eine kompakte Darstellung der komplexen Zahlen in Polarkoordinaten: Die Zahl mit Betrag r und Argument φ ist $z = re^{i\varphi}$.

g) Logarithmen, Wurzeln und allgemeine Potenzen

Die nächsten Funktionen, die wir dringend kennenlernen müssen, hängen alle in irgendeiner Weise mit der Umkehrung der Exponentialfunktion zusammen. Dier wichtigste davon ist der natürliche Logarithmus, die Umkehrfunktion der Exponentialfunktion.

1) Der natürliche Logarithmus im Reellen: Hier ist seine Definition problemlos: Die Exponentialfunktion

$$\mathbb{R} \to \mathbb{R}_{>0}; \quad x \mapsto e^x$$

ist monoton wachsend, hat also eine eindeutig bestimmte Umkehrfunktion

$$\mathbb{R}_{>0} \to \mathbb{R}; \quad x \mapsto \ln x \, ,$$

die als *natürlicher Logarithmus* bezeichnet wird.

Da die Exponentialfunktion für positive Argumente sehr schnell wächst, ist der natürliche Logarithmus für Argumente ab eins eine sehr langsam wachsende Funktion; beispielsweise erreicht er erst für $x \approx 2{,}688 \cdot 10^{43}$

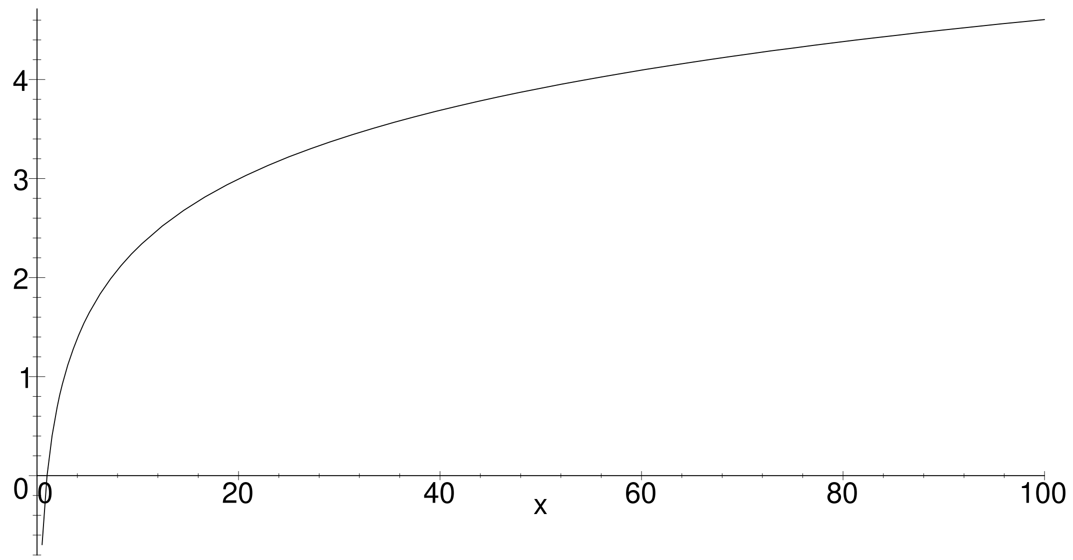


Abb. 15: Der natürliche Logarithmus

den Wert 100. Zwischen eins und null dagegen geht er natürlich sehr schnell gegen minus Unendlich.

Seine Ableitung läßt sich leicht berechnen: Ist g eine Umkehrfunktion von f, d.h. $f\big(g(x)\big) = x$ für alle x aus dem Definitionsbereich von g, so ist nach der Kettenregel

$$\frac{d}{dx} f\big(g(x)\big) = f'\big(g(x)\big)\, g'(x) = \frac{dx}{dx} = 1 \,,$$

also

$$g'(x) = \frac{1}{f'\big(g(x)\big)} \,.$$

Im vorliegenden Fall ist daher

$$\ln'(x) = \frac{1}{e^{\ln x}} = \frac{1}{x} \,.$$

Dies löst auch das einzige noch verbleibenden Integrationsprobleme für Potenzen $f(x) = x^n$: Für $n = -1$ ist

$$\int \frac{dx}{x} = \ln x + C \,.$$

Das Integral des Logarithmus läßt sich, wenn man erst einmal auf diese Idee kommt, sehr einfach durch partielle Integration bestimmen:

$$\int \ln x \, dx = \int 1 \cdot \ln x = x \ln x - \int x \cdot \frac{1}{x} \, dx = x(\ln x - 1) + C \, .$$

2) Abschätzung von $\ln n!$***:*** Als Anwendung dieser Formel können wir den Logarithmus von $n!$ abschätzen, indem wir nicht wie bisher immer ein Integral durch seine RIEMANN-Summen abschätzen, sondern umgekehrt:

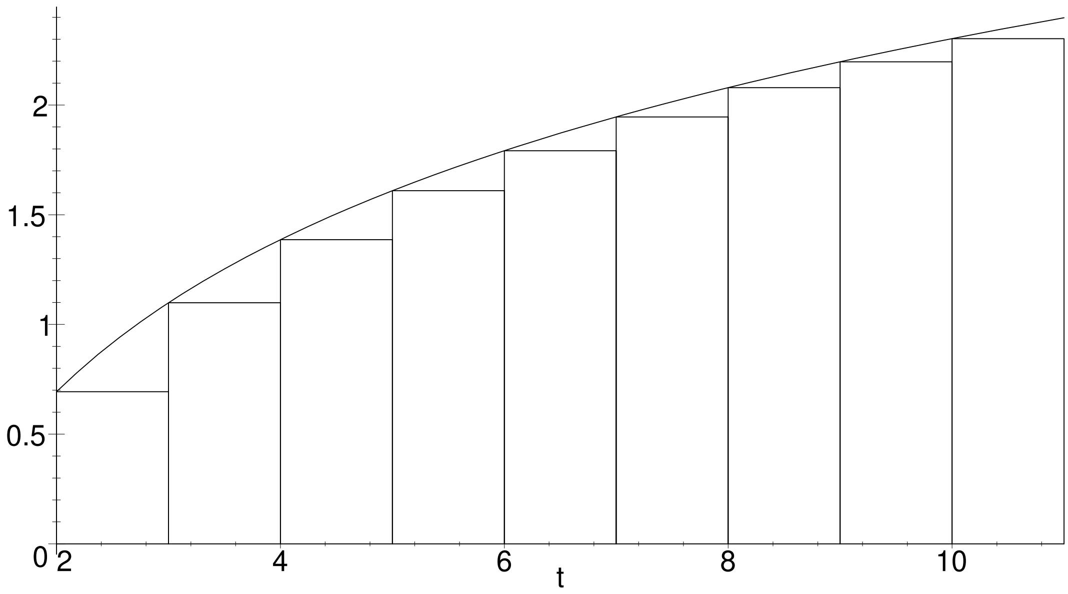


Abb. 16: Eine RIEMANNsche Untersumme für den Logarithmus

$$\ln n! = \sum_{\nu=1}^{n} \ln \nu = \sum_{\nu=2}^{n} \ln \nu = \sum_{\nu=2}^{n} \ln \nu \cdot \big((\nu + 1) - \nu\big)$$

ist wegen der Monotonie des Logarithmus eine RIEMANNsche Untersumme für den Logarithmus im Bereich zwischen 2 und $n + 1$, d.h.

$$\ln n! \leq \int_{2}^{n+1} \ln x \, dx = (n + 1)\big(\ln(n + 1) - 1\big) - 2(\ln 2 - 1) \, .$$

Genauso ist

$$\ln n! = \sum_{\nu=1}^{n} \ln \nu = \sum_{\nu=1}^{n-1} \ln(\nu+1)\big((\nu+1)-\nu\big)$$

wegen der Monotonie des Logarithmus eine RIEMANNsche Obersumme für den Logarithmus im Bereich zwischen 1 und n, d.h.

$$\ln n! \geq \int_1^n \ln x \, dx = n(\ln n - 1) + 1 \; .$$

Insgesamt erhalten wir die Abschätzung

$$n(\ln n - 1) + 1 \leq \ln n! \leq (n+1)\big(\ln(n+1) - 1\big) - 2(\ln 2 - 1) \; .$$

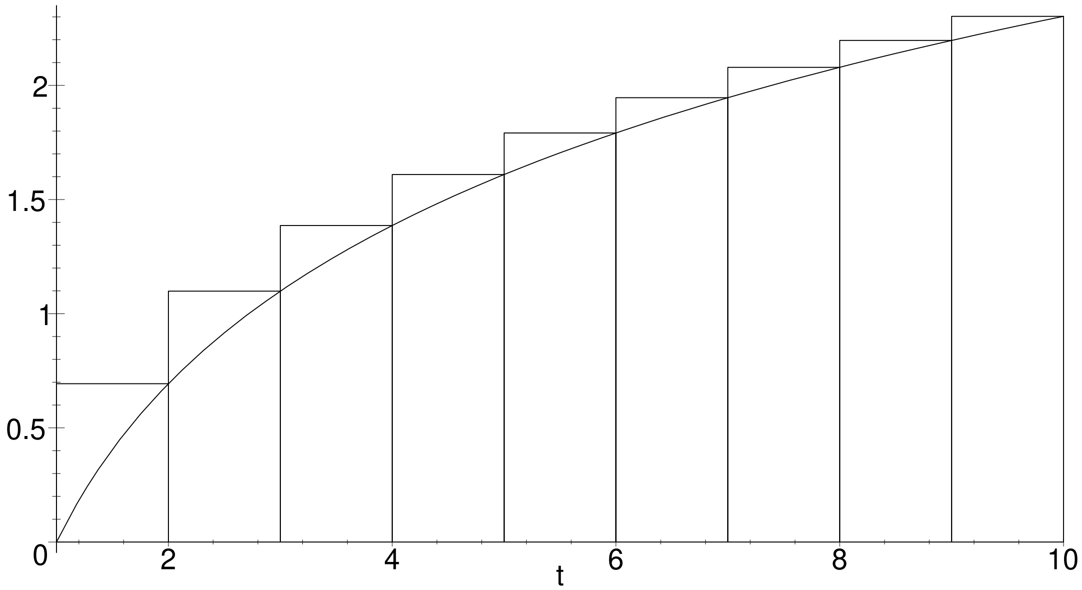


Abb. 17: Eine RIEMANNsche Obersumme für den Logarithmus

Numerische Rechnungen zeigen, daß diese Abschätzung nach beiden Seiten hin ziemlich gut ist; für $n = 10$ etwa erhalten wir

$$14{,}0259 \leq \ln 10! \leq 15{,}9906$$

bei einem tatsächlichen Wert von etwa 15,1044, für $n = 100$

$$361{,}517 \leq \ln 100! \leq 365{,}741$$

bei einem tatsächlichen Wert von etwa 363,739 und für $n = 1000$

$$5908{,}76 \leq \ln n! \leq 5915{,}27\,,$$

während $\ln 1000! \approx 5912{,}13$ ist. Für $n = 10\,000$ schließlich als letztes Beispiel ergibt sich

$$82\,104{,}4 \leq \ln 10\,000! \leq 82\,113{,}2$$

verglichen mit $\ln 10\,000! \approx 82\,108{,}9$.

3) Der Logarithmus im Komplexen: Schwieriger wird die Situation, wenn wir Logarithmen auch im Komplexen betrachten: Dort ist die Exponentialfunktion nicht mehr injektiv, denn nach den EULERschen Formeln ist

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

und die trigonometrischen Funktionen sind periodisch mit Periode 2π. Dies ist allerdings auch die einzige Möglichkeit, wie zwei verschiedene komplexe Zahlen auf denselben Wert der Exponentialfunktion führen können: Ist für $z = x + iy$ und $w = u + iv$

$$e^z = e^{x+iy} = e^{u+iv} = e^w\,,$$

so ist

$$1 = \frac{e^{x+iy}}{e^{u+iv}} = e^{(x-u)+i(y-v)} = e^{x-u}\bigl(cos(y-v) + i\sin(y-v)\bigr)\,.$$

Die rechts stehende Zahl hat den Betrag e^{x-u}; da die Zahl gleich Eins ist, muß also $e^{x-u} = 1$ und damit $x = u$ sein, denn im Reellen ist die Exponentialfunktion injektiv. Damit muß auch die Klammer rechts den Wert eins haben, d.h.

$$\cos(y-v) = 1 \quad \text{und} \quad \sin(y-v) = 0\,.$$

Die Nullstellen des Sinus sind gerade die ganzahligen Vielfachen von π, also muß $y - v = n\pi$ sein mit einem $n \in \mathbb{Z}$. Für solche Zahlen ist $\cos(y - v) = (-1)^n$, also muß $n = 2k$ eine gerade Zahl sein, d.h.

$$e^z = e^w \Leftrightarrow z - w = 2k\pi i \quad \text{mit} \quad k \in \mathbb{Z}\,.$$

Damit ist die Exponentialfunktion injektiv in jedem Bereich, in dem keine zwei komplexen Zahlen liegen, deren Imaginärteil sich um ein Vielfaches von 2π unterscheidet, beispielsweise also in jedem Streifen $a < \mathfrak{Im}\, z \leq a + 2\pi$. In jedem solchen Bereich hat die Exponentialfunktion daher eine eindeutig bestimmte Umkehrfunktion, und diese Umkehrfunktion bezeichnen wir als einen *Zweig* des natürlichen Logarithmus.

Die Nichteindeutigkeit des Logarithmus entspricht exakt der Nichteindeutigkeit des Arguments einer komplexen Zahl: In der Tat ist für $z = re^{i\varphi}$

$$\ln z = \ln r + i\varphi = \ln |z| + i \arg z\,.$$

Der wichtigste Zweig ist der für den Bereich $-\pi < \mathfrak{Im}\, z \leq \pi$; die Umkehrfunktion der Exponentialfunktion für diesen Bereich nennen wir den *Hauptwert* des natürlichen Logarithmus. Dieser Hauptwert ist unstetig auf der negativen reellen Achse, da sein Imaginärteil bei deren Überschreiten von $-\pi$ auf π oder π auf $-\pi$ springt. Auch die Beziehung

$$\ln zw = \ln z + \ln w$$

gilt nicht für den Hauptwert: Beispielsweise ist wegen

$$e^{\pi i} = \cos \pi + i \sin \pi = -1$$

zwar $\ln(-1) = \pi i$, aber

$$\ln(-1) + \ln(-1) = 2\pi i$$

ist nicht der Hauptwert des Logarithmus von $(-1) \cdot (-1) = 1$, denn dieser ist Null.

Die mit der Vielzahl von Zweigen verbundene Mehrdeutigkeit des komplexen Logarithmus ist einerseits sehr störend, andererseits aber der unvermeidliche Preis, den wir dafür bezahlen müssen, daß der Logarithmus nicht mehr nur für positive reelle Zahlen erklärt ist, sondern für beliebige komplexe Zahlen mit Ausnahme der Null. In Anwendungen, beispielsweise in der Wellenoptik, wird man im allgemeinen nach stetigen Umkehrfunktionen suchen und dabei in den seltensten Fällen mit dem Hauptwert auskommen.

4) Die allgemeine Potenzfunktion: Über den Logarithmus und die Exponentialfunktion lassen sich beliebige Potenzen definieren: Betrachten wir zunächst einmal wieder nur reelle Zahlen und Funktionen, so folgt durch Iteration der Regel

$$\ln ab = \ln a + \ln b \,,$$

daß $\ln na = n \ln a$ ist für jede natürliche Zahl $n \in \mathbb{N}$ und jede positive reelle Zahl a. Rückwärts gelesen heißt dies, daß auch

$$\ln \sqrt[n]{a} = \ln a^{1/n} = \tfrac{1}{n} \ln a$$

ist und damit für jede rationale Zahl $\frac{p}{q} \in \mathbb{Q}$

$$\ln \left(\sqrt[q]{a}\right)^p = \ln a^{p/q} = \tfrac{p}{q} \ln a.$$

Also ist für jede rationale Zahl x und jede positive reelle Zahl a

$$a^x = e^{\ln a^x} = e^{x \ln a} \,.$$

Rechts steht für festes a und variables x eine stetige Funktion von x; der Ausdruck links ist nur für rationales x erklärt, aber natürlich spricht nichts dagegen, ihn über diese Beziehung für beliebiges reelles x zu definieren:

Definition: Für eine positive reelle Zahl a und eine beliebige reelle Zahl x ist

$$a^x \underset{\text{def}}{=} e^{x \ln a} \,.$$

5) Logarithmen zu beliebigen Basen: Obige Definition erklärt für jedes $a > 0$ eine stetige Funktion $x \mapsto a^x$; diese ist offensichtlich sogar differenzierbar mit Ableitung $x \mapsto (\ln a) \cdot a^x$. Da diese Ableitung überall dasselbe Vorzeichen hat wie $\ln a$, bildet die Funktion $a \mapsto a^x$ für $a \neq 1$ die reellen Zahlen monoton ab auf die positiven reellen Zahlen und hat somit für $a \neq 1$ eine Umkehrfunktion

$$\mathbb{R}_{>0} \to \mathbb{R}; \quad x \mapsto \log_a x \,,$$

die wir als Logarithmus zur Basis a bezeichnen.

Neben dem natürlichen Logarithmus, der im Sinne dieser Sprechweise der Logarithmus zur Basis e ist, sind vor allem noch zwei weitere Logarithmen wichtig: der *dekadische Logarithmus*

$$\lg x = \log_{10} x$$

zur Basis zehn und der *binäre Logarithmus*

$$\operatorname{lb} x = \log_2 x$$

zur Basis zwei. Letzterer tritt vor allem in der Informationstheorie auf, wo man etwa die Information, die in der Auswahl von einer aus n gleichwahrscheinlichen Alternativen liegt, als $\operatorname{lb} n$ Bit definiert; ersterer war früher zum praktischen Rechnen sehr wichtig und wird noch heute vielfach zur Angabe von Verhältnissen benutzt:

Nach dem WEBER-FECHNERschen Gesetz ist die Stärke einer Sinnesempfindung wie etwa einer Lautstärke nicht proportional zur entsprechenden physikalischen Größe, etwa dem Schalldruck, sondern zu dessen Logarithmus. Ein Lautstärkenunterschied wird daher gemessen über den Logarithmus des Verhältnisses der Schalldruckwerte, und obwohl dies physikalisch eigentlich Unsinn ist, wurde hierfür sogar eine Einheit definiert, das Bel. Da diese Einheit allerdings sehr groß ist, verwendet man in der Praxis die um den Faktor zehn kleinere Einheit, das deziBel oder kurz dB. Falls Musik oder Sprache also um 40 dB verstärkt wird, heißt dies, daß der erzeugte Schalldruck $10^{\frac{40}{10}} = 10\,000$ mal so hoch ist wie der ursprüngliche. Entsprechendes gilt auch für Rauschabstände *usw.*

6) Komplexe Wurzeln: Im Komplexen können wir formal genauso vorgehen wie im Reellen und für beliebige komplexe Zahlen $a \neq 0$ und z definieren, daß $a^z = e^{z \ln a}$ sein soll. Dabei stellt sich nun aber das Problem, daß $\ln a$ kein eindeutig bestimmter Wert ist und somit auch der Wert a^z vom gewählten Zweig des Logarithmus abhängt. Betrachten wir etwa die komplexe Potenz

$$\left(\frac{1+i}{\sqrt{2}}\right)^{1997i} .$$

Da Realteil und Imaginärteil von $a = \frac{1+i}{\sqrt{2}}$ übereinstimmen und positiv sind, ist $\arg a = \frac{\pi}{4}$; da man außerdem leicht nachrechnet, daß $|a| = 1$ ist, ergibt sich der *Hauptwert* von $\ln a$ zu $\frac{\pi i}{4}$; der Hauptwert von a^{1997i} ist also

$$e^{1997i\cdot\frac{\pi i}{4}} = e^{\frac{-1997\pi}{4}},$$

eine extrem kleine positive reelle Zahl. Andererseits ist beispielsweise auch $\ln a = \frac{\pi i}{4} - 100\pi i$ ein Wert des Logarithmus von a, und wenn man diesen Wert des Logarithmus zugrundelegt, ist

$$a^z = e^{1997i\cdot\left(\frac{\pi i}{4}-100i\right)} = e^{\frac{-1997\pi}{4}+199700}$$

eine sehr große reelle Zahl. Wie man sieht, ist der Potenzbegriff in dieser extremen Allgemeinheit offensichtlich nicht sehr nützlich.

In Spezialfällen ist diese Nichteindeutigkeit der Potenz allerdings genau das, was wir brauchen. Betrachten wir etwa als ein zunächst völlig absurd erscheinendes Beispiel die Potenz $1^{1/n}$ für eine natürliche Zahl $n \geq 2$. Hier erhalten wir, komplex gerechnet, für den Wert $2k\pi i$ von $\ln 1$

$$1^{1/n} = e^{\frac{1}{n}\cdot 2k\pi i} = e^{\frac{2k\pi i}{n}}.$$

Für $k = 0, 1, \ldots, n-1$ liefert dies jeweils verschiedene Werte, davor und danach wiederholt sich alles. Wir haben also n verschiedene Lösungen, die allesamt die Eigenschaft haben, daß ihre n-te Potenz gleich 1 ist, d.h. wir bekommen die sämtlichen n-ten Einheitswurzeln. Für $n = 2$ sind dies beispielsweise

$$e^0 = 1 \quad \text{und} \quad e^{\pi i} = -1,$$

für $n = 3$ erhalten wir außer $e^0 = 1$ noch

$$e^{\frac{2\pi i}{3}} = \cos\left(\frac{2\pi}{3}\right) + i\sin\left(\frac{2\pi}{3}\right) = -\cos\left(\frac{\pi}{3}\right) + i\sin\left(\frac{\pi}{3}\right) = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$$

und die dazu konjugiert komplexe Zahl

$$e^{\frac{4\pi i}{3}} = -\frac{1}{2} - i\frac{\sqrt{3}}{2}.$$

$n = 4$ führt auf die bekannten Lösungen $e^0 = 1$, $e^{\frac{\pi i}{2}} = i$, $e^{\pi i} = -1$ und $e^{\frac{3\pi i}{2}} = -i$, während wir für $n = 5$ nicht mehr mit der obigen Tabelle der Sinus- und Cosinuswerte auskommen. Wer solche Dinge mag, kann nachrechnen, daß außer der Eins noch die vier Zahlen

$$\pm\frac{\sqrt{5}}{4} - \frac{1}{4} \pm \frac{i}{4}\sqrt{2}\sqrt{5 \pm \sqrt{5}}$$

auftreten, wobei das erste und das letzte $\pm$ dasselbe Vorzeichen sein sollen, während das mittlere unabhängig davon frei gewählt werden kann. Der Fall $n = 6$ ist wieder eine einfache Übungsaufgabe für interessierte Leser; für $n = 7$ gibt es keine geschlossene Darstellung der von eins verschiedenen Werte durch Wurzelausdrücke, Brüche und ganze Zahlen.

Allgemein überlegt man sich leicht, daß es für jede Gleichung der Form $z^n = w \neq 0$ genau n verschiedene Werte von $w^{1/n}$ gibt, die gerade die sämtlichen Lösungen z dieser Gleichung sind.

h) Die Umkehrfunktionen der trigonometrischen Funktionen

Bevor wir die trigonometrischen Funktionen umkehren, sollten wir sie zunächst alle einführen. Neben dem bereits bekannten Sinus und dem Cosinus ist vor allem noch der Tangens

$$\tan x = \frac{\sin x}{\cos x}$$

wichtig; im rechtwinkligen Dreieck gibt er das Verhältnis zwischen Gegenkathete und Ankathete eines Winkels an und damit das, was man üblicherweise als *Steigung* bezeichnet: das Verhältnis der Höhendifferenz zur horizontalen Entfernung.

Da $\cos x$ für $x = \pi/2 + k\pi$ mit $k \in \mathbb{Z}$ verschwindet, ist $\tan x$ an diesen Stellen nicht definiert; in ihrer Umgebung wird er links, wo Sinus und Cosinus dasselbe Vorzeichen haben (+ für gerades k, $-$ für ungerades) beliebig groß, rechts beliebig negativ; siehe dazu auch Abbildung 18.

Aus den speziellen Werten von Sinus und Cosinus berechnet man leicht auch die entsprechenden Werte für Tangens:

x	0	$\frac{\pi}{6}$	$\frac{\pi}{4}$	$\frac{\pi}{3}$	$\frac{\pi}{2}$

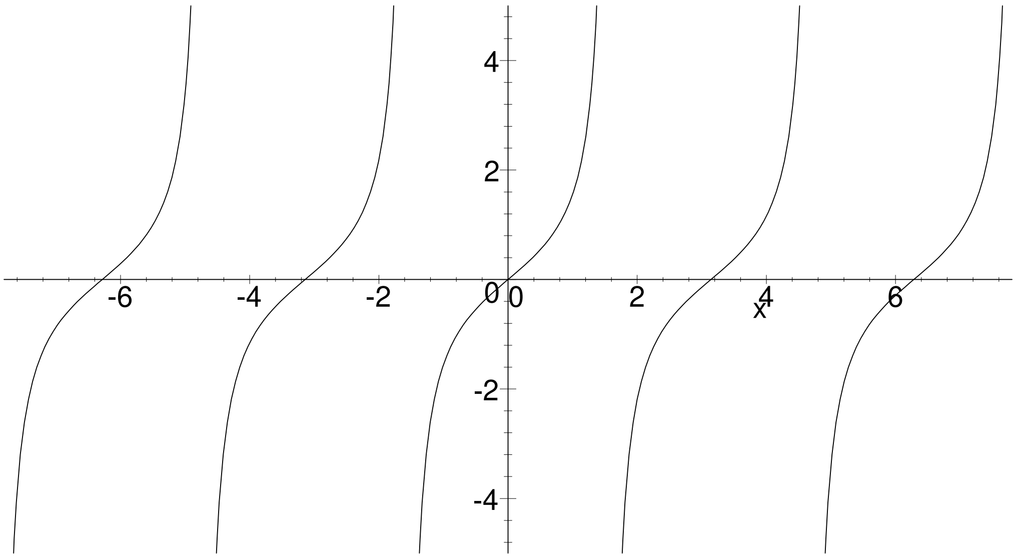


Abb. 18: Der Tangens

$\tan x$	0	$\frac{\sqrt{3}}{3}$	1	$\sqrt{3}$	$-$

Damit haben drei der Seitenverhältnisse im rechtwinkligen Dreieck Namen erhalten; die restlichen drei haben natürlich auch welche, sind aber weniger wichtig, da sie gerade die Kehrwerte der drei bereits definierten sind:

$$\cot x = \frac{\text{Ankathete}}{\text{Gegenkathete}} = \frac{1}{\tan x}\,,$$

$$\sec x = \frac{\text{Hypothenuse}}{\text{Ankathete}} = \frac{1}{\cos x} \qquad \text{und}$$

$$\operatorname{cosec} x = \frac{\text{Hypothenuse}}{\text{Gegenkathete}} = \frac{1}{\sin x}\,.$$

Sekans und Cosekans spielten bei der klassischen Navigation nach Sonnenstand und Sternen eine Rolle; für uns sind sie überflüssig.

Offensichtlich haben Sekans und Cosekans genau wie Sinus und Cosinus die Periode 2π, Tangens und Cotangens allerdings haben Periode π, denn

$$\tan(x+\pi) = \frac{\sin(x+\pi)}{\cos(x+\pi)} = \frac{-\sin x}{-\cos x} = \tan x\,.$$

Damit genug über die trigonometrische Funktionen selbst; kommen wir zum eigentlichen Thema dieses Abschnitts, ihren Umkehrfunktionen.

Da alle trigonometrischen Funktionen periodisch sind, können sie keine global definierten Umkehrfunktionen haben; wir müssen sie also jeweils auf ein Intervall einschränken, auf dem sie injektiv oder – besser noch – monoton sind.

Im Fall des Sinus bietet sich das Intervall von $-\pi/2$ bis $\pi/2$ an, in dem er monoton von -1 auf 1 ansteigt; wir definieren daher den *Hauptwert* der Umkehrfunktion *Arkussinus* für dieses Intervall:

$$\arcsin\colon [-1,1] \to \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$$

ist die Umkehrfunktion von

$$\sin\colon \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \to [-1,1]\,.$$

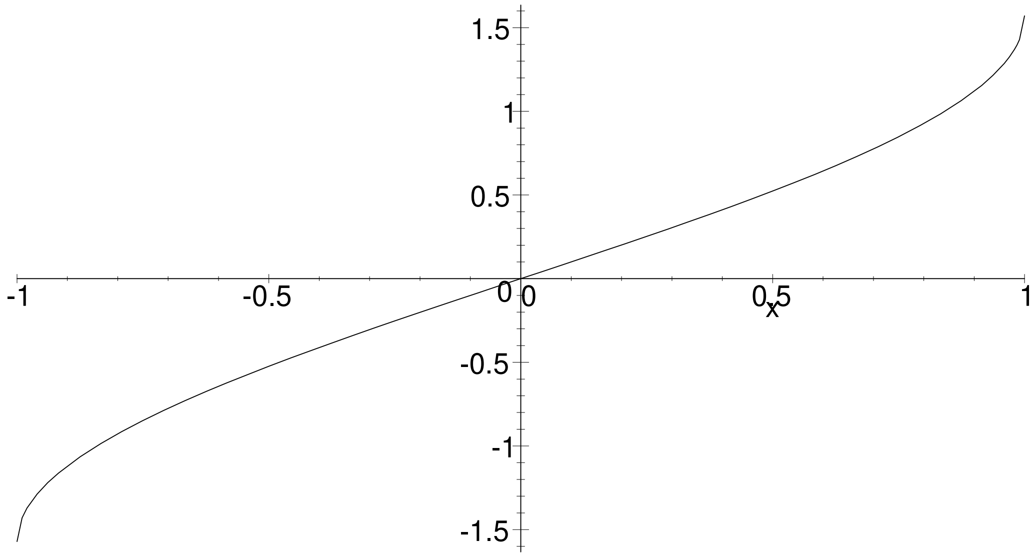


Abb. 19: Der Arkussinus

Der Cosinus fällt von seinem Maximalwert 1 an der Stelle Null monoton ab, bis er bei π den Wert -1 erreicht hat. Wir definieren den *Hauptwert* der Umkehrfunktion *Arkuscosinus* daher für dieses Intervall:

$$\arccos\colon [-1,1] \to [0,\pi]$$

ist die Umkehrfunktion von

$$\cos\colon [0,\pi] \to [-1,1]\,.$$

Aufgrund der Beziehung $\cos\varphi = \sin\left(\frac{\pi}{2} - \varphi\right)$ und der Wahl der Wertebereiche ist

$$\arccos x = \frac{\pi}{2} - \arcsin x\,.$$

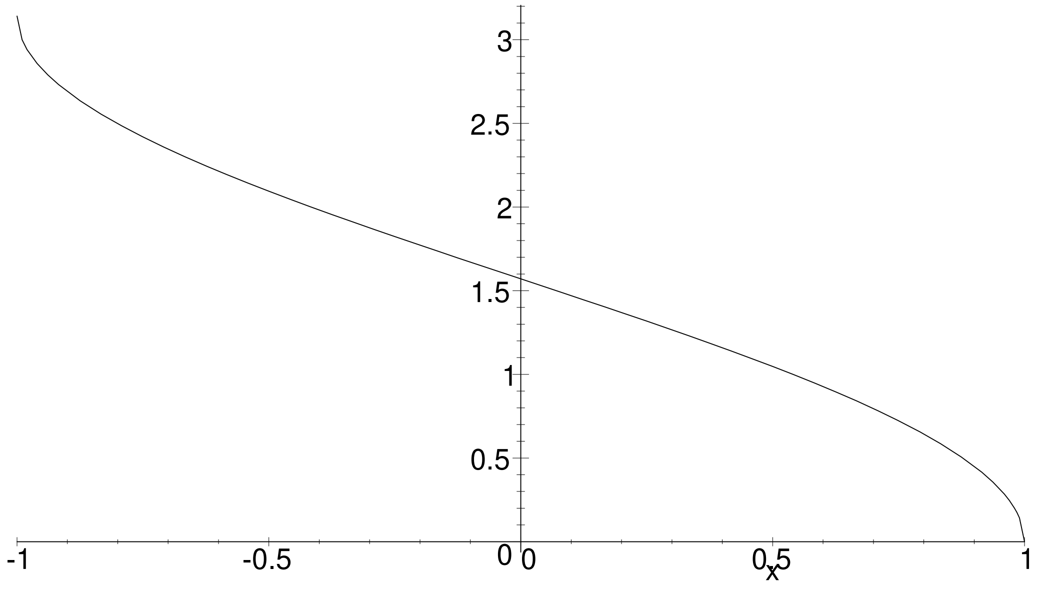


Abb. 20: Der Arkuscosinus

Tangens schließlich ist monoton ansteigend im offenen Intervall $(-\pi/2, \pi/2)$ und nimmt dort jeden reellen Wert an; wir definieren den *Hauptwert* der Umkehrfunktion *Arkustangens* daher für dieses Intervall:

$$\arctan\colon \mathbb{R} \to \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$$

ist die Umkehrfunktion von

$$\tan\colon \left(-\frac{\pi}{2}, \frac{\pi}{2}\right) \to \mathbb{R}\,.$$

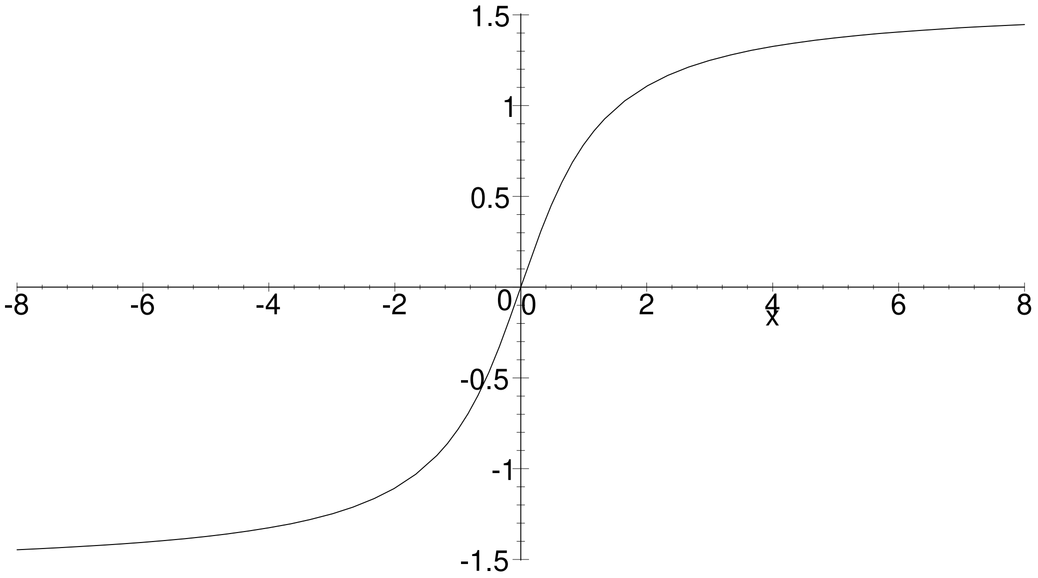


Abb. 21: Der Arkustangens

Gelegentlich ganz nützlich ist der Wert an der Stelle 1: Da nach obiger Tabelle $\tan\frac{\pi}{4}$ den Wert 1 hat, ist $\arctan(1) = \frac{\pi}{4}$ oder

$$\pi = 4\ \arctan(1)\,.$$

Falls man in einem Programm den Wert π benötigt und ihn über diese Formel als Programmkonstante definiert, kann man – eine saubere Implementierung der Arkusfunktionen vorausgesetzt – sicher sein, daß π mit der maximal möglichen Stellenzahl der jeweiligen Gleitkommaarithmetik dargestellt wird – ohne daß man diese Stellenzahl oder gar π mit der entsprechenden Genauigkeit zu kennen braucht.

Abgesehen vom gerade definierten Hauptwert lassen sich natürlich noch in offensichtlicher Weise weitere Zweige definieren; wir wollen hier auf Einzelheiten verzichten. Der Leser sollte sich allerdings klarmachen, daß es beispielsweise für den Arkussinus sowohl monoton wachsende als auch monoton fallende Zweige gibt.

Mit Hilfe der Arkusfunktionen lassen sich die Polarkoordinatendarstellung und die kartesische Darstellung der komplexen Zahlen ineinander umrechnen: Da $z = re^{i\varphi}$ Realteil $x = r\cos\varphi$ und Imaginärteil

$y = r \sin\varphi$ hat, sollte außer der bereits bekannten Formel

$$r = \sqrt{x^2 + y^2}$$

für $z \neq 0$ auch so etwas wie

$$\varphi = \arctan\frac{y}{x} = \arccos\frac{x}{r} = \arcsin\frac{y}{r}$$

gelten. Dies im Prinzip auch richtig, allerdings nicht immer für den Hauptwert: Da alle drei Umkehrfunktionen nur Werte in einem Intervall der Länge π annehmen, während das Argument (genauer: der Hauptwert des Arguments) einer komplexen Zahl im doppelt so langen Intervall $(-\pi, \pi]$ liegt, kann obige Gleichung bei keiner der drei Funktionen für alle z gelten. Eine korrekte Formel mit dem Hauptwert des Arkuscosinus ist beispielsweise

$$\arg z = \begin{cases} \arccos\frac{x}{r} & \text{falls } \mathfrak{Im}\, z \geq 0 \\ -\arccos\frac{x}{r} & \text{falls } \mathfrak{Im}\, z < 0 \end{cases}.$$

Der für uns interessanteste Aspekt der Arkusfunktionen sind allerdings nicht solche Rechenregeln, sondern die *Ableitungen* dieser Funktionen, die uns neue Integrationsregeln und Stammfunktionen liefern werden.

Wie bei der Ableitung des Logarithmus gehen wir aus von der Formel

$$g'(x) = \frac{1}{f'\big(g(x)\big)}$$

für die Umkehrfunktion g einer Funktion f. Da die Ableitung des Sinus der Cosinus ist, erhalten wir also

$$\arcsin'(x) = \frac{1}{\cos(\arcsin x)}.$$

Wegen $\sin^2 y + \cos^2 y = 1$ ist

$$\cos^2(\arcsin x) = 1 - \sin^2(\arcsin x) = 1 - x^2\,;$$

da der Cosinus im Intervall $[-\pi/2, \pi/2]$, in dem der Arkussinus seine Werte annimmt, größer oder gleich Null ist, folgt

$$\cos(\arcsin x) = \sqrt{1 - x^2} \quad \text{und} \quad \arcsin'(x) = \frac{1}{\sqrt{1 - x^2}}.$$

Genauso ist

$$\arccos'(x) = \frac{-1}{\sin(\arccos x)} = \frac{-1}{\sqrt{1-x^2}}\,,$$

da der Sinus zwischen 0 und π keine negativen Werte annimmt.

Für den Arkustangens schließlich müssen wir zunächst Tangens selbst ableiten; nach der Quotientenregel ist

$$\tan'(x) = \frac{d}{dx}\frac{\sin x}{\cos x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = 1 + \tan^2 x$$

und dementsprechend

$$\arctan'(x) = \frac{1}{1+\tan^2(\arctan x)} = \frac{1}{1+x^2}\,.$$

i) Die Hyperbelfunktionen und ihre Umkehrfunktionen

Als letzte Klasse spezieller Funktionen wollen uns noch die Hyperbelfunktionen anschauen; sie sind in Analogie zu den EULERschen Formeln für die trigonometrischen Funktionen definiert durch

$$\sinh x = \frac{e^x - e^{-x}}{2}, \quad \cosh x = \frac{e^x + e^{-x}}{2}$$

und

$$\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}\,;$$

sie heißen jeweils *Sinus hyperbolicus, Cosinus hyperbolicus* und *Tangens hyperbolicus.* Offensichtlich ist

$$\cosh^2 x - \sinh^2 x = \frac{1}{4}\left((e^x + e^{-x})^2 - (e^x - e^{-x})^2\right) = \frac{4}{4}e^x e^{-x} = 1\,,$$

d.h. die Punkte $(\cosh x, \sinh x)$ liegen auf der Hyperbel $u^2 - v^2 = 1$, was den Namen Hyperbelfunktionen erklärt.

Unmittelbar aus der Definition einsichtig sind die Ableitungsregeln

$$\sinh'(x) = \cosh x \quad \text{und} \quad \cosh'(x) = \sinh x\,;$$

$$\tanh'(x) = \frac{\cosh^2 x - \sinh^2 x}{\cosh^2 x} = \frac{1}{\cosh^2 x} = 1 - \tanh^2 x\,.$$

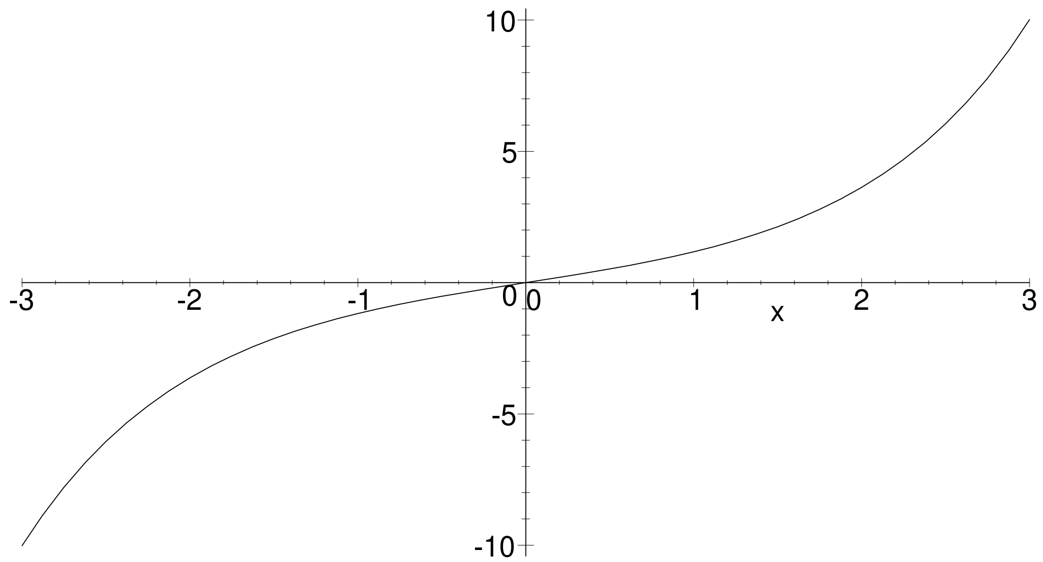


Abb. 22: Der Sinus hyperbolicus

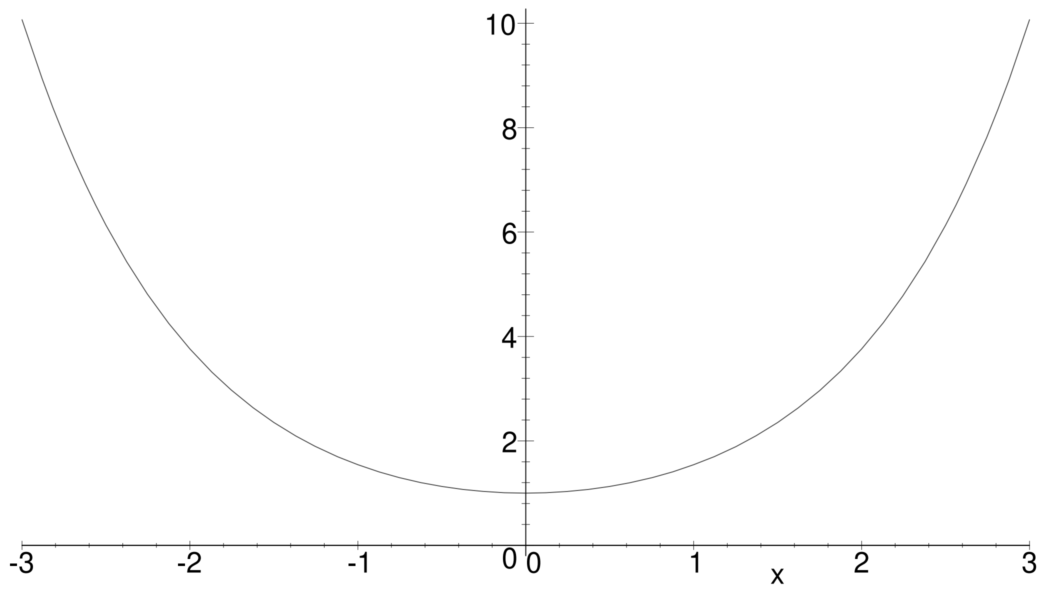


Abb. 23: Der Cosinus hyperbolicus

Die Umkehrfunktionen der Hyperbelfunktionen werden als *Areafunk-*

tionen bezeichnet und heißen *Areasinus hyperbolicus, Areacosinus hyperbolicus* und *Areatangens hyperbolicus.* Um sie exakt definieren zu können, müssen wir zunächst die Hyperbelfunktionen etwas genauer betrachten.

Für große Werte von x wird e^{-x} beliebig klein, sowohl $\sinh x$ als auch $\cosh x$ unterscheiden sich für solche Werte also beliebig wenig von $\frac{1}{2}e^x$ und gehen insbesondere gegen unendlich. Für stark negative Werte wird e^x beliebig klein und das Verhalten beider Funktionen wird dominiert durch den Term $\pm e^{-x}$; daher geht $\sinh x$ für $x \to -\infty$ gegen $-\infty$ und $\cosh x$ gegen $+\infty$. (Dies sieht man auch daran, daß $\sinh x$ eine ungerade, $\cosh x$ aber eine gerade Funktion ist.)

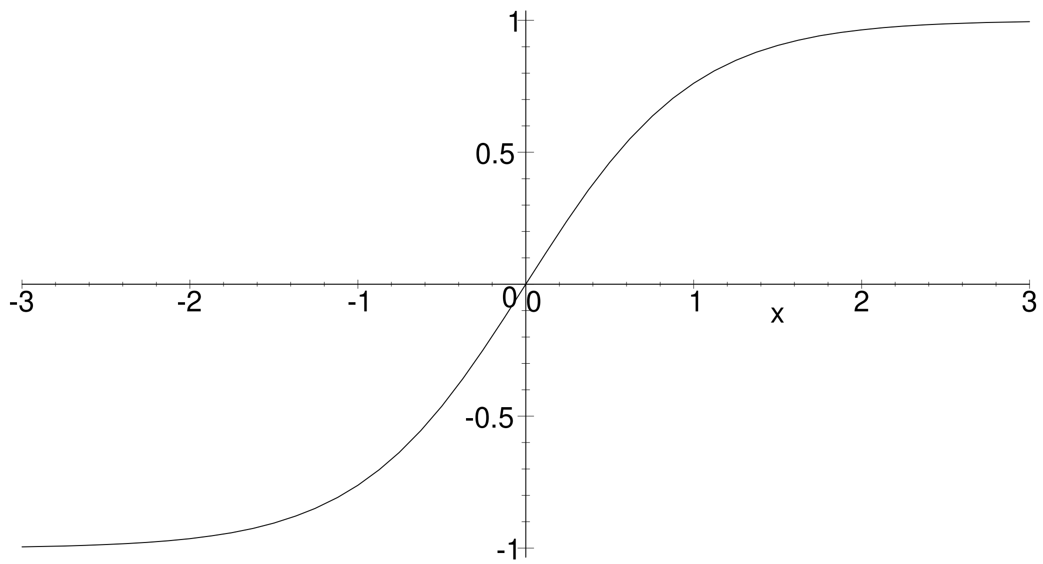


Abb. 24: Der Tangens hyperbolicus

$\cosh x$ ist für alle reellen Zahlen x positiv; daher ist $\sinh x$ überall strikt monoton ansteigend und hat somit eine auf ganz $\mathbb{R}$ definierte Umkehrfunktion

$$\operatorname{arsinh}: \mathbb{R} \to \mathbb{R}\,.$$

$\sinh x$ ist nur für positive x positiv, $\cosh x$ also nur für $x \geq 0$ ansteigend; die Funktion wächst von $\cosh 0 = 1$ aus über alle Grenzen und hat somit eine Umkehrfunktion

$$\operatorname{arcosh}\colon \mathbb{R}_{\geq 1} \to \mathbb{R}_{\geq 0}\,.$$

Der Tangens hyperbolicus schließlich ist monoton wachsend auf $\mathbb{R}$, da seine Ableitung überall positiv ist; die Grenzwerte für $x \to \pm\infty$ sind offensichtlich ± 1, so daß es eine Umkehrfunktion

$$\operatorname{artanh}\colon (-1, 1) \to \mathbb{R}$$

gibt.

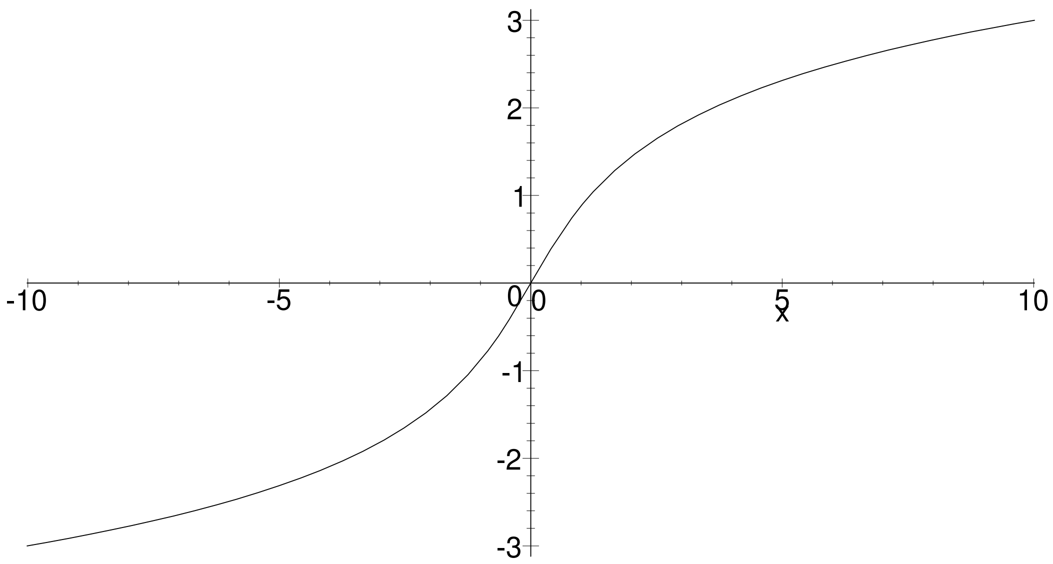


Abb. 25: Der Areasinus hyperbolicus

Auch diese Umkehrfunktionen sind wieder vor allem wegen ihrer Ableitungen interessant; eine weitgehend zum Fall der Arkusfunktionen analoge Rechnung zeigt, daß

$$\operatorname{arsinh}'(x) = \frac{1}{\cosh(\operatorname{arsinh} x)} = \frac{1}{\sqrt{1+x^2}}$$

und

$$\operatorname{arcosh}'(x) = \frac{1}{\sinh(\operatorname{arcosh} x)} = \frac{1}{\sqrt{x^2-1}}$$

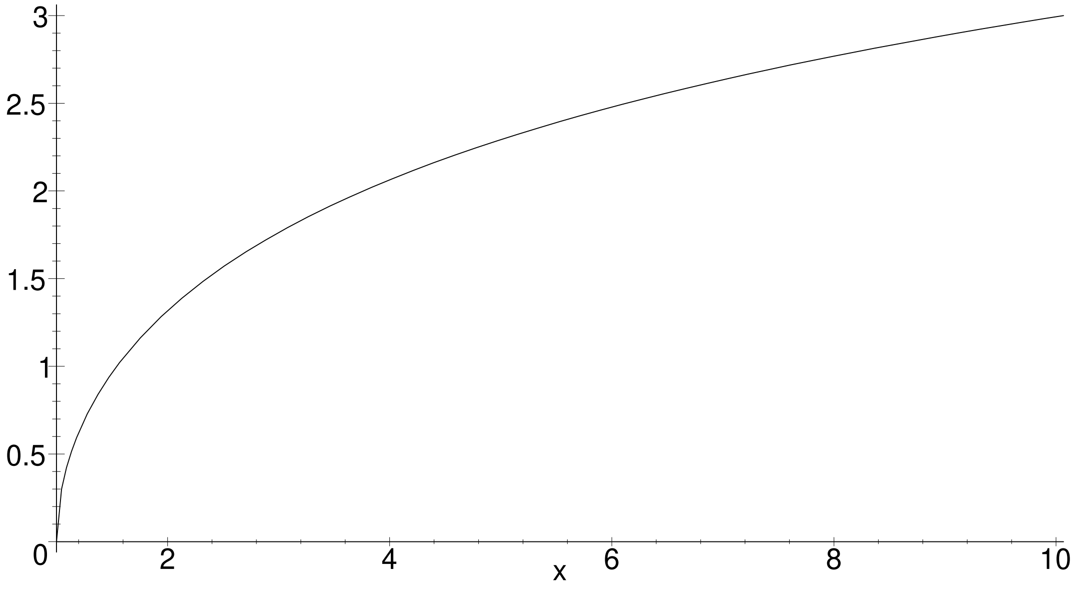


Abb. 26: Der Areacosinus hyperbolicus

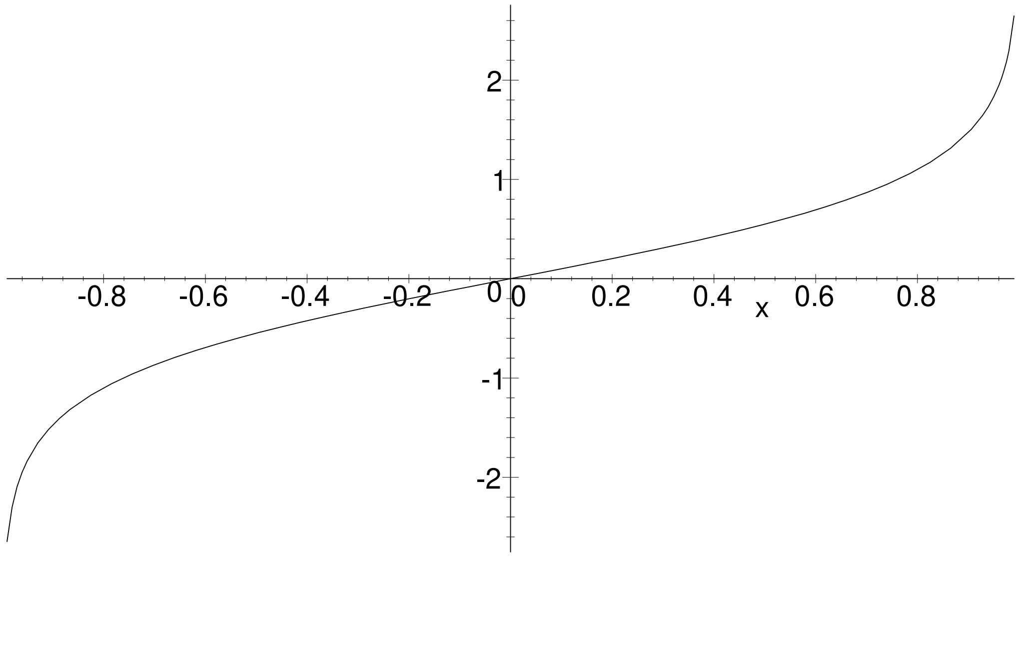


Abb. 27: Der Areatangens hyperbolicus

ist. Für den Areatangens hyperbolicus schließlich ist

$$\operatorname{artanh}'(x) = \frac{1}{1 + \tanh^2(\operatorname{artanh} x)} = \frac{1}{1 - x^2} .$$

Genau wie die Hyperbelfunktionen durch Exponentialfunktionen ausgedrückt werden können, lassen sich auch die Areafunktionen durch Logarithmen ausdrücken. Ist beispielsweise $y = \operatorname{arsinh}(x)$, so ist

$$x = \sinh y = \frac{e^y - e^{-y}}{2} = \frac{a - 1/a}{2} \quad \text{mit} \quad a = e^y .$$

Somit ist

$$2ax = a^2 - 1 \quad \text{oder} \quad a^2 - 2ax - 1 = 0 .$$

Fassen wir dies als eine quadratische Gleichung für a auf, so folgt

$$(a - x)^2 = x^2 + 1 \quad \text{oder} \quad a = x \pm \sqrt{x^2 + 1} .$$

Da die Quadratwurzel größer als x ist, gibt es eine positive und eine negative Lösung; hier ist $a = e^y > 0$, also gleich der positiven Lösung. Somit ist

$$y = \operatorname{arsinh}(x) = \ln(x + \sqrt{x^2 + 1}) .$$

Genauso findet man, daß

$$\operatorname{arcosh}(x) = \ln(x + \sqrt{x^2 - 1})$$

und

$$\operatorname{artanh}(x) = \frac{1}{2} \ln\left(\frac{1 + x}{1 - x}\right)$$

ist.

§5: Ausbau der Integralrechnung

§4 hat uns eine ganze Reihe neuer Stammfunktionen beschert; insbesondere kennen wir nun

$$\int \frac{dx}{x} = \ln x + C ,$$

$$\int \sin x \, dx = -\cos x + C \quad \text{und} \quad \int \cos x \, dx = \sin x + C ,$$

$$\int \sinh x\,dx = \cosh x + C \quad \text{und} \quad \int \cosh x\,dx = \sinh x + C\,,$$

$$\int \frac{dx}{1+x^2} = \arctan x + C \quad \text{und} \quad \int \frac{dx}{1-x^2} = \operatorname{artanh} x + C\,,$$

$$\int \frac{dx}{\sqrt{1+x^2}} = \operatorname{arsinh} x + C \quad \text{und} \quad \int \frac{dx}{\sqrt{1-x^2}} = \arcsin x + C\,.$$

Am interessantesten aber sind die im vorigen Paragraphen eingeführten Funktionen für die Integralrechnung im Zusammenhang mit der Substitutionsregel, die erst jetzt richtig anwendbar wird.

a) Der Spezialfall logarithmischer Ableitungen

Für $f(x) = 1/x$ führt die Substitutionsregel

$$\int f\big(g(x)\big)g'(x)\,dx = F\big(g(x)\big) + C \quad \text{mit} \quad F'(x) = f(x)$$

zur Integrationsregel

$$\int \frac{g'(x)}{g(x)}\,dx = \ln|g(x)| + C\,.$$

Als erste Anwendung betrachten wir

$$\tan x = \frac{\sin x}{\cos x} = -\frac{g'(x)}{g(x)}$$

mit $g(x) = \cos x$. Nach der gerade bewiesenen Regel ist daher

$$\int \tan x\,dx = -\ln|\cos(x)| + C\,,$$

eine Funktion, die genau wie der Tangens selbst an den Nullstellen der Cosinusfunktion nicht definiert ist.

Beim Tangens hyperbolicus gibt es keine Probleme mit dem Vorzeichen; hier ist ganz entsprechend

$$\int \tanh x\,dx = \ln|\cosh(x)| + C = \ln\cosh(x) + C\,,$$

da der Cosinus hyperbolicus nur positive Werte annimmt.

Auch Integrale wie

$$\int \frac{x}{1+x^2}\,dx$$

lassen sich nach dieser Regel ausrechnen: Da die Ableitung des Nenners $2x$ ist, folgt

$$\int \frac{x}{1+x^2}\,dx = \tfrac{1}{2}\ln\left|1+x^2\right| + C = \tfrac{1}{2}\ln(1+x^2) + C\,.$$

Damit können wir dann beispielsweise den Arkustangens integrieren: Partielle Integration zeigt, daß

$$\begin{aligned}\int 1\cdot\arctan x\,dx &= x\arctan x - \int \frac{x}{1+x^2}\,dx + C\\ &= x\arctan x - \tfrac{1}{2}\ln(1+x^2) + C\end{aligned}$$

ist.

b) Substitutionen mit linearen Funktionen

Eine der elementarsten Anwendungen der Substitutionsregel, auf die man üblicherweise auch ohne diese Regel kommt, ist die Substitution mit linearen Funktionen $g(x) = ax+b$; hier besagt die Substitutionsregel, daß

$$\int f(ax+b)\cdot a\,dx = F(ax+b) \quad \text{mit} \quad F'(x) = f(x)$$

ist, oder besser

$$\int f(ax+b)\,dx = \frac{F(ax+b)}{a} \quad \text{mit} \quad F'(x) = f(x)\,.$$

Somit ist beispielsweise

$$\int \sin(\omega t+\varphi)\,dt = \frac{-\cos(\omega t+\varphi)}{\omega} + C$$

oder

$$\int \frac{dx}{x^2+a^2} = \int \frac{\frac{dx}{a^2}}{\left(\frac{x}{a}\right)^2+1} = \frac{1}{a}\arctan\left(\frac{x}{a}\right) + C\,.$$

Entsprechend berechnet man auch

$$\int \frac{dx}{x^2 - a^2} = -\frac{1}{a} \operatorname{artanh}\left(\frac{x}{a}\right) + C$$

und sogar ganz allgemein $\displaystyle\int \frac{dx}{ax^2 + bx + c}$:

Da der Fall $a = 0$ uninteressant ist (*bzw.* eine unbedingt notwendige Übungsaufgabe für alle, die nicht sofort einsehen warum), schließen wir dies aus und schreiben

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 + c - \frac{b^2}{4a}\,.$$

Mit

$$y = x + \frac{b}{2a} \quad \text{und} \quad d = \frac{c}{a} - \frac{b^2}{4a^2}$$

ist dann

$$\begin{aligned}
\int \frac{dx}{ax^2 + bx + c} &= \frac{1}{a} \int \frac{dy}{y^2 + d} \\
&= \frac{1}{a} \begin{cases} \frac{1}{\sqrt{d}} \arctan\left(\frac{y}{\sqrt{d}}\right) + C & \text{falls } d > 0 \\ \frac{-1}{y} & \text{falls } d = 0 \\ \frac{1}{\sqrt{-d}} \operatorname{artanh}\left(\frac{y}{\sqrt{-d}}\right) + C & \text{falls } d < 0\,, \end{cases}
\end{aligned}$$

und je nach Vorzeichen von d kann dies entweder über eine der obigen Formeln oder (für $d = 0$) elementar ausgerechnet werden. Insgesamt erhalten wir mit $\Delta = b^2 - 4ac$ (und damit $d = -\Delta/4a^2$)

$$\int \frac{dx}{ax^2 + bx + c} = \begin{cases} \frac{2}{\sqrt{-\Delta}} \arctan\left(\frac{2ax+b}{\sqrt{-\Delta}}\right) + C & \text{falls } \Delta < 0 \\ \frac{-2}{2ax+b} + C & \text{falls } \Delta = 0 \\ \frac{-2}{\sqrt{\Delta}} \operatorname{artanh}\left(\frac{2ax+b}{\sqrt{\Delta}}\right) + C & \text{falls } \Delta > 0\,. \end{cases}$$

c) Substitutionen mit trigonometrischen und Hyperbelfunktionen

Wegen der Beziehungen

$$\sin^2 x + \cos^2 x = 1 \quad \text{und} \quad \cosh^2 - \sinh^2 x = 1$$

bieten sich Substitutionen mit trigonometrischen Funktionen und Hyperbelfunktionen an bei Integralen, in denen Ausdrücke der Form $\sqrt{1 \pm x^2}$ und ähnliche vorkommen.

Betrachten wir als einfachstes Beispiel die Stammfunktion von $\sqrt{1-x^2}$ selbst. Mit der Substitution $x = \sin t$ mit $-\frac{\pi}{2} \leq t \leq \frac{\pi}{2}$ erhalten wir (durch Rückwärtslesen der Substitutionsregel)

$$\int \sqrt{1-x^2}\,dx = \int \sqrt{1-\sin^2 t}\cos t\,dt\,,$$

und dies ist $\int \cos^2 t\,dt$, da Cosinus zwischen $-\frac{\pi}{2}$ und $\frac{\pi}{2}$ nur nichtnegative Werte annimmt.

Dieses Integral kennen wir von Aufgabe *1b)* des dritten Übungsblatts:

$$\int \cos^2 t\,dt = \tfrac{1}{2}(\sin t\cos t + t) + C\,.$$

Mit der Rücksubstitution $t = \arcsin x$ erhalten wir schließlich (da wir schon aus §4*h)* wissen, daß $\cos(\arcsin x) = \sqrt{1-x^2}$ ist)

$$\int \sqrt{1-x^2}\,dx = \tfrac{1}{2}\left(x\sqrt{1-x^2} + \arcsin x\right) + C\,.$$

Aus ähnliche Weise läßt sich auch die Stammfunktion von $\sqrt{1+x^2}$ bestimmen: Hier bietet sich die Substitution $x = \sinh t$ an und wir erhalten

$$\int \sqrt{1+x^2}\,dx = \int \sqrt{1-\sinh^2 t}\cosh t\,dt = \int \cosh^2 t\,dt\,,$$

da $\cosh t$ nur positive Werte annimmt.

Leider kennen wir das rechtsstehende Integral noch nicht; angesichts der großen Ähnlichkeiten zwischen trigonometrischen Funktionen und Hyperbelfunktionen lohnt es sich aber sicherlich, als ersten Ansatz vor einer partiellen Integration zu schauen, ob vielleicht etwas analoges gilt wie oben. In der Tat ist nach der Produktregel

$$\frac{d}{dt}\tfrac{1}{2}(\sinh t\cosh t + t) = \cosh^2 t + C\,,$$

also

$$\int \cosh^2 t \, dt = \tfrac{1}{2}(\sinh t \cosh t + t) + C \, .$$

Somit ist

$$\int \sqrt{1+x^2} \, dx = \tfrac{1}{2}(x\sqrt{1+x^2} + \operatorname{arsinh} x) + C \, .$$

Als letztes Beispiel, in dem x einmal nicht quadratisch vorkommt, betrachten wir noch

$$\int \sqrt{\frac{1-x}{1+x}} \, dx \, .$$

Da es in diesem Abschnitt um Substitutionen mit trigonometrischen Funktionen geht, versuchen wir es wieder mit dem Ansatz $x = \sin t$ für $-\frac{\pi}{2} \le t \le \frac{\pi}{2}$ und erhalten

$$\int \sqrt{\frac{1-x}{1+x}} \, dx = \int \sqrt{\frac{1-\sin t}{1+\sin t}} \cos t \, dt \, .$$

Hier hilft nun ein Trick weiter: Erweitern wir den Bruch unter der Quadratwurzel mit $1 - \sin t$, so wird das Integral zu

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

Damit ist

$$\begin{aligned} \int \sqrt{\frac{1-x}{1+x}} \, dx &= \arcsin x + \cos(\arcsin x) + C \\ &= \arcsin x + \sqrt{1-x^2} + C \, . \end{aligned}$$

Bei einem so komplizierten Integral empfiehlt es sich, das Ergebnis durch Differentiation zu überprüfen; wir erhalten

$$\frac{d}{dx}\left(\arcsin x + \sqrt{1-x^2}\right) = \frac{1}{\sqrt{1-x^2}} - \frac{x}{\sqrt{1-x^2}} \, ,$$

was zunächst so aussieht, als sei irgend etwas falsch gelaufen. Beachtet man aber, daß

$$\sqrt{1+x} \cdot \sqrt{1-x} = \sqrt{1-x^2}$$

ist, so ist

$$\frac{1}{\sqrt{1-x^2}} - \frac{x}{\sqrt{1-x^2}} = \frac{1-x}{\sqrt{1-x^2}} = \frac{\sqrt{1-x}}{\sqrt{1+x}}\,,$$

was erheblich vertrauenserweckender aussieht.

d) Integrale rationaler Funktionen

Bekanntlich ist ein Polynom in x ein Ausdruck der Form

$$f(x) = \sum_{\nu=0}^{n} a_\nu x^\nu = a_n x^n + \cdots + a_1 x + a_0\,;$$

ist $a_n \neq 0$, so bezeichnet man n als den *Grad* des Polynoms, in Zeichen

$$n = \deg f\,.$$

Polynome vom Grad eins, zwei, drei *bzw.* vier heißen linear, quadratisch, kubisch *bzw.* biquadratisch.

Unter einer *rationalen Funktion* versteht man eine Funktion der Form

$$x \mapsto \frac{f(x)}{g(x)}\,,$$

wobei $f(x)$ und $g(x)$ Polynome sind. Es ist klar, daß diese Funktion nur in jenen Punkten $x \in \mathbb{R}$ definiert ist, in denen $g(x) \neq 0$ ist.

Die Integrationstheorie rationaler Funktionen ist seit langem bekannt; im Prinzip kann man zu jeder rationalen Funktion eine Stammfunktion angeben, indem man wie folgt vorgeht: (Auf Einzelheiten und Beweise muß hier weitgehend verzichtet werden, da wir nicht über das notwendige mathematische Instrumentarium verfügen.)

1. Schritt: Durch den vielleicht aus der Schule oder sonstwie bekannten Algorithmus zur Polynomdivision mit Rest kann man Polynome $q(x)$, $r(x)$ finden, so daß

$$f(x) = q(x) \cdot g(x) + r(x) \quad \text{mit} \quad \deg r(x) < \deg g(x)\,;$$

damit ist

$$\frac{f(x)}{g(x)} = q(x) + \frac{r(x)}{g(x)}\,.$$

Da es nicht schwierig ist, eine Stammfunktion des Polynoms $q(x)$ zu finden, können wir uns im folgenden auf die rationale Funktion $r(x)/g(x)$ beschränken, deren Zähler kleineren Grad hat als der Nenner.

2. Schritt: Man zerlege den Nenner in ein Produkt aus linearen und quadratischen Polynomen. Die ist zumindest grundsätzlich möglich, denn nach dem *Fundamentalsatz der Algebra* läßt sich ein Polynom vom Grad n stets in der Form

$$g(x) = a_n(x - z_1)(x - z_2) \cdot \cdots \cdot (x - z_n)$$

schreiben, wobei die *komplexen* Zahlen z_i die Nullstellen von g sind. Wie man sich leicht überlegt, folgt aus $g(z) = 0$ sofort, daß auch $g(\overline{z}) = 0$ ist, d.h. die nichtreellen Nullstellen treten als Paare konjugiert komplexer Zahlen auf. Das Produkt der zugehörigen Linearfaktoren ist

$$(x - z)(x - \overline{z}) = x^2 - (z + \overline{z})x + z\overline{z} = x^2 - 2\,\mathfrak{Re}\, z \cdot x + |z|^2 \ ,$$

ein quadratisches Polynom mit reellen Koeffizienten. Also läßt sich $g(x)$ als Produkt linearer und quadratischer Polynome mit reellen Koeffizienten schreiben. Faßt man gleiche Faktoren zusammen, so erhält man demnach eine Darstellung

$$g(x) = a_n \ell_1(x)^{e_1} \cdot \cdots \cdot \ell_r(x)^{e_r} \cdot q_1(x)^{f_1} \cdot \cdots \cdot q_s(x)^{f_s}$$

mit linearen Polynomen ℓ_i, quadratischen Polynomen q_j und natürlichen Zahlen e_i, f_j. Dabei sind sowohl die ℓ_i als auch die q_j paarweise voneinander verschieden.

Diese Zerlegung kann algorithmisch sehr aufwendig sein; daher kommt der obige Zusatz „im Prinzip“ für die Durchführbarkeit der Integration rationaler Funktionen.

3. Schritt: Man schreibe die Funktion $r(x)/g(x)$ als Summe von *Partialbrüchen* der Form

$$\frac{\alpha_{ik}}{\ell_i(x)^k} \quad \text{und} \quad \frac{\beta_{jk}x + \gamma_{jk}}{q_j(x)^k} \ ,$$

wobei k von eins bis e_i *bzw.* f_j läuft und α_{ik}, β_{jk} und γ_{jk} reelle Zahlen sind, d.h.

$$\frac{r(x)}{g(x)} = \sum_{i=1}^{r} \sum_{k=1}^{e_i} \frac{\alpha_{ik}}{\ell_i(x)^k} + \sum_{j=1}^{s} \sum_{k=1}^{f_j} \frac{\beta_{jk}x + \gamma_{jk}}{q_j(x)^k} .$$

Die Bestimmung der Zähler der Partialbrüche sowie auch der Nachweis, daß eine Zerlegung in eine Summe obiger Form stets möglich ist, erfordern Methoden aus der linearen Algebra, die wir erst im nächsten Kapitel kennenlernen werden; auf Einzelheiten muß hier deshalb verzichtet werden. (Die Existenz der Partialbruchzerlegung folgt tatsächlich am einfachsten mit Methoden der Funktionentheorie, die wir vielleicht im nächsten Semester behandeln werden.)

4. Schritt: Integration der Partialbrüche: Die Integrale der Funktionen $\dfrac{\alpha_{ik}}{\ell_i(x)^k}$ lassen sich über die Substitution $u = \ell_i(x)$ sofort auf $\int \frac{du}{u^k}$ zurückführen und sind somit problemlos. Weiter ist

$$\int \frac{\beta x + \gamma}{ax^2 + bx + c}\, dx = \frac{\beta}{2a} \int \frac{2ax + b}{ax^2 + bx + c}\, dx + \int \frac{\gamma - \frac{\beta b}{2a}}{ax^2 + bx + c}\, dx ,$$

wobei das erste Integral nach der Regel über die logarithmische Ableitung aus Abschnitt *a)* berechnet werden kann, während wir das zweite in Abschnitt *b)* behandelt haben.

Bleiben schließlich noch die Integrale mit Potenzen eines quadratischen Polynoms im Nenner; hier wird k durch partielle Integration rekursiv erniedrigt.

Als ganz einfaches Beispiel, in dem alles explizit durchführbar ist, betrachten wir

$$\int \frac{dx}{x^2 - 5x + 6} .$$

Da hier sechs das Produkt und fünf die Summe der Nullstellen des Nenners ist, errät man leicht, daß $x^2 - 5x + 6 = (x-2)(x-3)$ ist. Die Partialbruchzerlegung bestimmen wir nun durch probieren: Da

$$\frac{1}{x-2} - \frac{1}{x-3} = \frac{(x-3)-(x-2)}{(x-2)(x-3)} = \frac{-1}{(x-2)(x-3)}$$

ist, folgt

$$\int \frac{dx}{(x-2)(x-3)} = -\int \frac{dx}{x-2} + \int \frac{dx}{x-3} = \ln\left|\frac{x-3}{x-2}\right| + C\,.$$

e) Integrale der Form $\int h(e^{ax})\,dx$

Bei solchen Integralen führt oft die Substitution $u = e^{ax}$ oder, anders ausgedrückt, $x = g(u) = \frac{1}{a}\ln u$ zu einer Vereinfachung. Die Substitutionsregel

$$\int f\big(g(u)\big)g'(u)\,du = \int f(x)\,dx$$

wird wegen $g'(u) = \dfrac{dx}{du} = \dfrac{1}{au}$ zu

$$\int h(u)\,\frac{du}{au} = \int h(e^{ax})\,dx\,;$$

insbesondere wird das Integral also bei rationalem h auf ein Integral einer rationalen Funktion zurückgeführt, und dieses kann nach dem im vorigen Abschnitt skizzierten Verfahren der Partialbruchzerlegung berechnet werden.

Im allgemeinen schreibt man eine Substitution wie die obige kurz in der Form

$$u = e^{ax}, \quad dx = \frac{du}{au}\,,$$

wobei man letztere Beziehung auch aus

$$\frac{du}{dx} = ae^{ax} = au \quad \text{oder} \quad du = au\,dx$$

herleiten kann.

Betrachten wir als Beispiel das Integral $\displaystyle\int \frac{e^{3x}}{e^{2x}-1}\,dx$: Da sowohl e^{2x} als auch e^{3x} im Integranden vorkommen, empfiehlt sich die Substitution $u = e^x$ mit $du = u\,dx$; damit wird

$$\int \frac{e^{3x}}{e^{2x}-1}\,dx = \int \frac{u^3}{u^2-1}\,\frac{du}{u} = \int \frac{u^2}{u^2-1}\,du\,,$$

und wir sind beim Integral einer rationalen Funktion angelangt. Die im vorigen Abschnitt skizzierte Methode der Partialbruchzerlegung führt auf

$$\frac{u^2}{u^2-1} = 1 + \frac{1}{u^2-1} = 1 + \frac{\frac{1}{2}}{u-1} - \frac{\frac{1}{2}}{u+1}\,,$$

d.h.

$$\frac{u^2}{u^2-1}\,du = \int du + \tfrac{1}{2}\left(\int \frac{du}{u-1} - \int \frac{du}{u+1}\right) = u + \ln\left|\frac{u-1}{u+1}\right| + C\,.$$

Nun muß nur noch die Substitution $u = e^x$ rückgängig gemacht werden, und wir erhalten als Ergebnis

$$\int \frac{e^{3x}}{e^{2x}-1}\,dx = e^x + \tfrac{1}{2}\ln\left|\frac{e^x-1}{e^x+1}\right| + C\,.$$

Dieselbe Technik funktioniert natürlich auch bei Integralen über Hyperbelfunktionen, da man diese genausogut als Ausdrücke in Exponentialfunktionen schreiben kann. Beispielsweise ist

$$\begin{aligned}\int \frac{dx}{\sinh x} &= 2\int \frac{dx}{e^x - e^{-x}} = 2\int \frac{du}{u\left(u-\frac{1}{u}\right)} = 2\int \frac{du}{u^2-1}\\ &= \ln\left|\frac{u-1}{u+1}\right| + C = \ln\left|\frac{e^x-1}{e^x+1}\right| + C\,.\end{aligned}$$

f) Integrale der Form $\int h(\sin x, \cos x)\,dx$

Als letzte Anwendung der Substitutionsregel in diesem Paragraphen wollen wir noch Integrale betrachten, die von trigonometrischen Funktionen abhängen. Hier führt oft die Substitution

$$x = 2\arctan t, \qquad dx = \frac{2\,dt}{t^2+1}$$

zum Erfolg; sie macht $\sin x$ zu $\sin(2\arctan t)$ und $\cos x$ zu $\cos(2\arctan t)$, was wir wie folgt ausrechnen können: Aus

$$\tan^2 x = \frac{sin^2 x}{\cos^2 x} = \frac{1-\cos^2 x}{\cos^2 x} = \frac{1}{\cos^2 x} - 1$$

folgt, daß

$$\cos(2x) = 2cos^2(x) - 1 = \frac{2}{1 + tan^2 x} - 1 = \frac{1 - tan^2 x}{1 + \tan^2 x}$$

ist und somit

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

Da der Arkustangens nur Werte zwischen $-\frac{\pi}{2}$ und $\frac{\pi}{2}$ annimmt, liegt $2 \arctan t$ zwischen $-\pi$ und π, ein Intervall, in dem der Sinus dasselbe Vorzeichen hat wie sein Argument. Daher ist

$$\sin(2 \arctan t) = \sqrt{1 - \cos^2(2 \arctan t)} = \sqrt{1 - \frac{(1 - t^2)^2}{(1 + t^2)^2}} = \frac{2t}{1 + t^2} \,.$$

Als Beispiel betrachten wir

$$\int \frac{dx}{\cos x} = \int \frac{1 + t^2}{1 - t^2} \, \frac{2\,dt}{1 + t^2} = -2 \int \frac{dt}{t^2 - 1} \,.$$

Das letzte Integral kennen wir aus dem vorigen Abschnitt:

$$\int \frac{dt}{t^2 - 1} = \frac{1}{2} \ln \left| \frac{1 - t}{1 + t} \right| + C \,.$$

Also ist

$$\int \frac{dx}{\cos x} = -\frac{2}{2} \ln \left| \frac{1 - t}{1 + t} \right| + C = \ln \left| \frac{1 + t}{1 - t} \right| + C = \ln \left| \frac{1 - \tan \frac{x}{2}}{1 + \tan \frac{x}{2}} \right| + C \,.$$

g) Symmetrie

Bei bestimmten Integralen kann man sich gelegentlich einen Großteil der Mühe des Ausrechnens ersparen, indem man Symmetrien berücksichtigt: Ist etwa die Funktion f spiegelsymmetrisch zur Achse $x = c$, d.h. $f(c + x) = f(c - x)$ für alle x, für die beide Seiten definiert sind, so ist auch

$$\int_c^{c+a} f(x)\,dx = \int_{c-a}^{c} f(x)\,dx \,,$$

d.h.

$$\int\limits_{c-a}^{c+a} f(x)\,dx = 2\int\limits_{c}^{c+a} f(x)\,dx\,.$$

Viel interessanter wird es, wenn f *punktsymmetrisch* zu einem Punkt $(c, 0)$ auf der x-Achse ist, d.h. $f(c+x) = -f(c-x)$. Dann ist

$$\int\limits_{c}^{c+a} f(x)\,dx = -\int\limits_{c-a}^{c} f(x)\,dx \quad \text{und} \quad \int\limits_{c-a}^{c+a} f(x)\,dx = 0\,,$$

ohne daß man irgendetwas über die Stammfunktion von f wissen müßte. So ist beispielsweise

$$\int\limits_{-1}^{1} \frac{\sin(x)}{|x|}\,dx = 0\,,$$

aber ein Leser, der mit den uns bislang bekannten Methoden nach einer Stammfunktion suchen möchte, dürfte eine sehr harte Zeit und wenig Erfolg haben.

h) Einige nicht elementar integrierbare Funktionen

Tatsächlich ist die Stammfunktion von $\dfrac{\sin x}{x}$ nicht durch die in §4 eingeführten elementaren transzendenten Funktionen und/oder Wurzeln und Grundrechenarten darstellbar, und sie teilt diese Eigenschaft mit einer ganzen Reihe weiterer Stammfunktionen.

Da einige dieser Stammfunktionen trotzdem für Anwendungen wichtig sind, behilft man sich damit, daß man einigen speziellen dieser Funktionen Namen gibt und versucht, alles darauf zurückzuführen. Die so definierten Funktionen können dann, genau wie der Sinus und die Exponentialfunktion, in Tabellenwerke und Unterprogrammbibliotheken aufgenommen werden, so daß man mit den dadurch abgedeckten Integralen genauso rechnen kann wie mit den elementaren transzendenten Funktionen.

Einige wichtige dieser neuen Funktionen sind

1) Der Integralsinus: Da $\lim\limits_{t\to 0} \dfrac{\sin t}{t} = 1$ ist, ist das bestimmte Integral

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

für alle $x \in \mathbb{R}$ wohldefiniert und eine stetige Funktion von x, der *Integralsinus.* Er ist beispielsweise wichtig für Tiefpaßfilterungen, wo Faltungsintegrale mit $\frac{\sin x}{x}$ berechnet werden müssen.

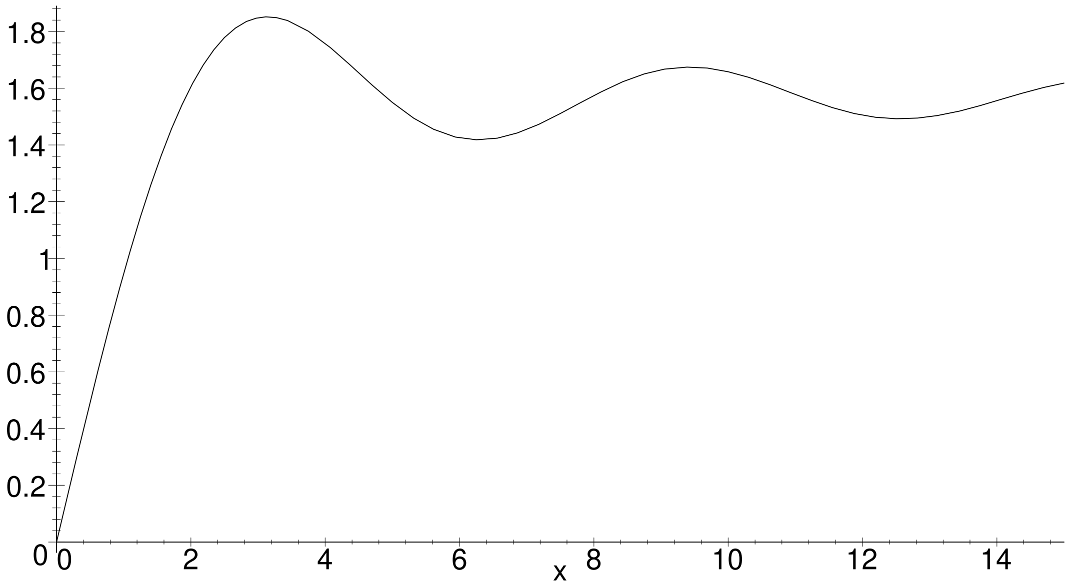


Abb. 28: Der Integralsinus

2) Die Fehlerfunktion: Auch die Funktion e^{-x^2} hat keine elementar angebbare Stammfunktion; hier definiert man als eine Stammfunktion die *Fehlerfunktion* oder *error function* durch

$$\mathrm{Erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt\,.$$

(Der Vorfaktor sorgt dafür, daß $\lim_{x\to\infty} \mathrm{Erf}(x) = 1$ ist, was wir im Augenblick allerdings noch nicht beweisen können.)

Sie ist wichtig für die Statistik, denn für eine normalverteilte Zufallsvariable (dazu gehören die meisten Meßgrößen) mit Mittelwert $\overline{x}$ und Standardabweichung σ ist die Wahrscheinlichkeit dafür, daß ein Meßwert zwischen a und b liegt, gleich

$$\frac{1}{2\pi}\int_a^b e^{\frac{1}{2}\left(\frac{x-\overline{x}}{\sigma}\right)^2}\,dx\,,$$

ein Integral, von dem man sich leicht überlegt, daß es mit Hilfe von Erf berechnet werden kann als

$$\frac{1}{2}\operatorname{Erf}\left(\frac{\sqrt{2}}{2}\,\frac{\overline{x}-a}{\sigma}\right)-\frac{1}{2}\operatorname{Erf}\left(\frac{\sqrt{2}}{2}\,\frac{\overline{x}-b}{\sigma}\right)\,.$$

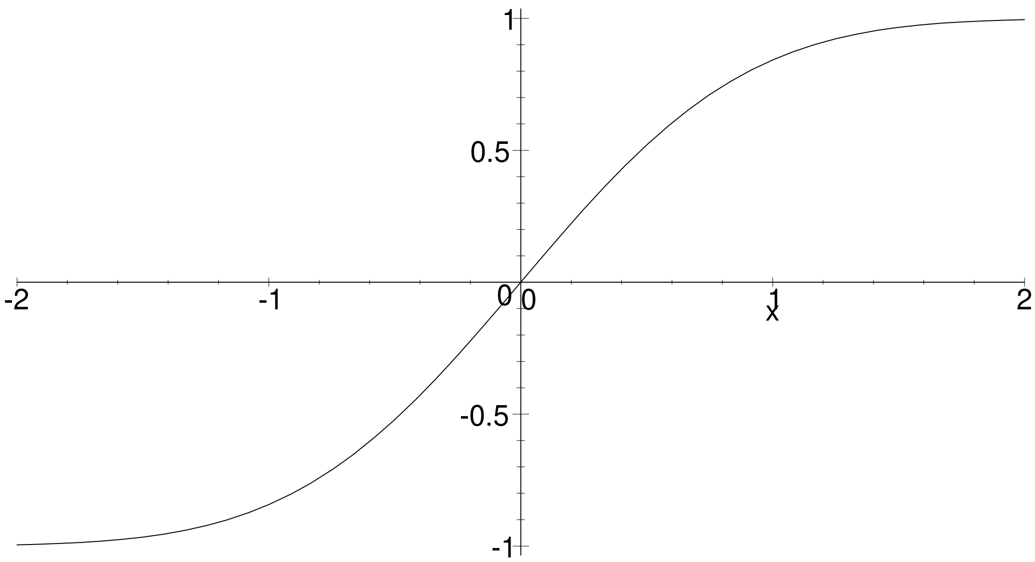


Abb. 29: Die Fehlerfunktion

3) Elliptische Integrale: Bei der Berechnung der Bogenlänge eines Ellipsensegments kommt man, je nach Art des Ansatzes und der Substitution, auf Integrale der Form

$$\int\frac{dx}{\sqrt{x^3+ax^2+bx+c}}\,,$$

wobei das Polynom im Nenner keine mehrfachen Nullstellen hat, oder auf Integrale der Form

$$\int \frac{dx}{\sqrt{1-k^2\sin^2 x}} \quad \text{oder} \quad \int \sqrt{1-k^2\sin^2 x}\,dx$$

mit $0 < k < 1$. Alle diese Integrale heißen *elliptische Integrale* und sind nicht elementar ausdrückbar. Eine hauptsächlich auf ADRIEN MARIE LEGENDRE (1752–1833) und KARL THEODOR WILHELM WEIERSTRASS (1815–1897) zurückgehende Theorie der elliptischen Funktionen stellt das Instrumentarium bereit, mit dem man diese vor allem in der Geodäsie, Kartographie und im Maschinenbau wichtigen Integrale berechnen kann.

4) Algebraische Integrale: Die erste Form der elliptischen Integrale ist ein Spezialfall eines sogenannen algebraischen Integrals; das sind Integrale, deren Integranden durch Wurzeln und Grundrechenarten (oder allgemeiner implizit durch Lösungen von Polynomgleichungen) gegeben sind. Die komplizierteren dieser Integrale sind fast alle nicht elementar ausdrückbar; es gibt inzwischen algorithmische Verfahren, die entscheiden, wann ein solches Integral elementar ausdrückbar ist, und die dann auch eine Stammfunktion finden können. Natürlich gibt es auch hier spezielle Funktionen, mit denen sich weitere dieser Integrale ausdrücken lassen.

§6: Erste komplexe Integrale und ihre Anwendungen

Wir wollen uns in diesem Paragraphen damit begnügen, *bestimmte* Integrale einer komplexwertigen Funktion über ein reelles Intervall zu definieren; alles andere erfordert neue Techniken, die wir erst später kennenlernen werden.

Hauptanwendung sind Orthogonalitätsrelationen für Exponentialfunktionen imaginärer Argumente und – was fast dasselbe ist – reelle trigonometrische Funktionen; wie wir zumindest andeutungsweise sehen werden, sind dies die Grundbausteine der harmonischen Analyse.

a) Bestimmte Integrale komplexer Funktionen

Sei also $f\colon [a,b] \to \mathbb{C}$ eine stetige Funktion; das für uns wichtigste Beispiel ist $f(t) = e^{i\omega t}$ mit $\omega \in \mathbb{R}$.

Definition:

$$\int_a^b f(t)\,dt = \int_a^b \mathfrak{Re}\, f(t)\,dt + i \int_a^b \mathfrak{Im}\, f(t)\,dt\,.$$

Insbesondere ist also für $\omega \neq 0$

$$\begin{aligned}\int_a^b e^{i\omega t}\,dt &= \int_a^b \cos\omega t\,dt + i\int_a^b \sin\omega t\,dt = \frac{1}{\omega}\sin\omega t\,\Big|_a^b - \frac{i}{\omega}\cos\omega t\,\Big|_a^b \\ &= \frac{1}{i\omega}\left(\cos\omega t\,\Big|_a^b + i\sin\omega t\,\Big|_a^b\right) = \frac{1}{i\omega}e^{i\omega t}\,\Big|_a^b \\ &= \frac{1}{i\omega}(e^{i\omega b} - e^{i\omega a})\,,\end{aligned}$$

genau wie man es vom Reellen her erwartet hätte. Trotzdem wollen wir in diesem Zusammenhang noch nicht von $\frac{1}{i\omega}e^{i\omega t}$ als einer *Stammfunktion* von $e^{i\omega t}$ reden, denn von einer komplexen Stammfunktion werden wir später mehr verlangen als nur die Anwendbarkeit auf Integrale entlang eines reellen Intervalls.

Genau wie bei den trigonometrischen Funktionen verschwindet also das Integral über eine Periode der Funktion: Für $e^{i\omega t}$, $\cos\omega t$ und $\sin\omega t$ ist die Periode jeweils

$$T = \frac{2\pi}{\omega}\,,$$

und

$$\int_0^T e^{i\omega t}\,dt = 0\,.$$

Dasselbe gilt natürlich auch für die Oberschwingungen, d.h. die Funktionen $e^{ki\omega t}$ mit $k \in \mathbb{Z} \setminus \{0\}$, und insbesondere ist auch für zwei solche Funktionen

$$\int_0^T e^{ki\omega t} e^{\ell i\omega t}\,dt = \int_0^T e^{(k+\ell)i\omega t}\,dt = \begin{cases} 0 & \text{falls } k \neq -\ell \\ T & \text{falls } k = -\ell \end{cases} .$$

Bei allen hier betrachteten Integralen ist offenbar nur wichtig, daß wir über ein Intervall der Länge T integrieren; die speziellen Grenzen Null und T könnten genausogut etwa durch $\pm T/2$ ersetzt werden.

b) Orthogonalitätsrelationen für trigonometrische Funktionen

Aufgrund der EULERschen Formeln ist

$$\begin{aligned} e^{ki\omega t} e^{\ell i\omega t} &= (\cos k\omega t \cos \ell\omega t - \sin k\omega t \sin \ell\omega t) \\ &\qquad + i(\cos k\omega t \sin \ell\omega t + \sin k\omega t \cos \ell\omega t) . \end{aligned}$$

Reell geschrieben wird die letzte Formel des vorigen Abschnitts daher zu

$$\int_0^T (\cos k\omega t \cos \ell\omega t - \sin k\omega t \sin \ell\omega t)\,dt = \begin{cases} 0 & \text{falls } k \neq -\ell \\ T & \text{falls } k = -\ell \end{cases}$$

und

$$\int_0^T (\cos k\omega t \sin \ell\omega t + \sin k\omega t \cos \ell\omega t) = 0 .$$

Ersetzt man hierin ℓ durch $-\ell$, ändert sich nichts bei den Cosinustermen, aber $\sin \ell\omega t$ wird zu $-\sin \ell\omega t$, d.h. wir erhalten die Formeln

$$\int_0^T (\cos k\omega t \cos \ell\omega t + \sin k\omega t \sin \ell\omega t)\,dt = \begin{cases} 0 & \text{falls } k \neq \ell \\ T & \text{falls } k = \ell \end{cases}$$

und

$$\int_0^T (-\cos k\omega t \sin \ell\omega t + \sin k\omega t \cos \ell\omega t) = 0 .$$

Addiert *bzw.* subtrahiert man hier die einander entsprechenden Formeln voneinander, ergeben sich als Ergebnis dieses Abschnitts für $k, \ell \geq 0$ die sogenannten *Orthogonalitätsrelationen* (was dieser Begriff bedeutet, werden wir im nächsten Kapitel über lineare Algebra genauer lernen) für trigonometrische Funktionen:

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

und

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

sowie

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

Auch hier bleiben alle Formeln gültig, wenn wir statt von 0 bis T über irgendein anderes Intervall der Länge T integrieren.

c) Harmonische Analyse im Komplexen

Aufgabe der harmonischen Analyse ist es, eine aus einer Grund- und verschiedenen Oberschwingungen zusammengesetzte Funktion wieder in ihre Bestandteile zu zerlegen. Wir beginnen mit den komplexen Fall, wo alles – wie üblich – etwas einfacher ist. Wir nehmen also an, wir hätten eine Funktion der Form

$$f(t) = \sum_{k=-N}^{N} c_k e^{ki\omega t}$$

in irgendeiner Weise gegeben (zum Beispiel als Meßkurve), und wir möchten die Koeffizienten c_k bestimmen. Nach der Orthogonalitätsrelation am Ende von Abschnitt *a)* ist (mit $T = \frac{2\pi}{\omega}$)

$$\int_0^T f(t) e^{\ell i\omega t} \, dt = \sum_{k=-N}^{N} c_k \int_0^T e^{ki\omega t} e^{\ell i\omega t} \, dt = c_{-\ell} T \, ,$$

d.h.

$$c_k = \frac{1}{T} \int_0^T f(t) e^{-ki\omega t}\, dt\ .$$

Die Bedingung, daß $f(t)$ als Summe endlich vieler Schwingungen gegeben sein soll, schränkt die Brauchbarkeit dieses Resultats leider ganz erheblich ein; zum Glück ist sie jedoch in dieser Strenge nicht notwendig. Zumindest formal kann man für *jede* periodische Funktion $f(t)$ mit Periode T Zahlen

$$c_k = \frac{1}{T} \int_0^T f(t) e^{-ki\omega t}\, dt$$

definieren und die Reihe

$$\sum_{k=-\infty}^{\infty} c_k e^{ki\omega t}$$

betrachten; sie heißt FOURIERreihe zur Funktion $f(t)$ nach dem französischen Mathematiker JEAN-BAPTISTE-JOSEPH FOURIER (1768–1830).

Es gibt dabei allerdings zwei Probleme:

1.) Wir wissen nichts über die Konvergenz dieser Reihe.

2.) Selbst wenn sie konvergiert, muß sie nicht gegen $f(t)$ konvergieren.

Für praktische Anwendungen ist beides glücklicherweise nur selten ein Problem: Für alle halbwegs vernünftigen Funktionen $f(t)$ konvergiert die Reihe, und abgesehen von den Unstetigkeitsstellen von f konvergiert sie im allgemeinen auch gegen $f(t)$. Für exaktere Aussagen und Beweise siehe **[D]**, §18, oder **[MV]**, *Band 2, Kapitel 11, §3,* besonders *Satz 3.2+3,* oder – vielleicht – die Höhere Mathematik II im nächsten Semester.

d) Harmonische Analyse im Reellen

Im Falle einer reellen Funktion $f(t)$ ist es manchmal – wenn auch beileibe nicht immer – nützlicher, eine Zerlegung in reelle Bestandteile, d.h. in Sinus- und Cosinusschwingungen zu betrachten. Eine solche

Zerlegung läßt sich mit Hilfe der EULERschen Formeln leicht aus dem Ergebnis des letzten Abschnitts herleiten.

Beginnen wir wieder mit dem Fall einer Funktion, die sich als endliche Summe

$$f(t) = \sum_{k=-N}^{N} c_k e^{ki\omega t}$$

darstellen läßt – schließlich sind das die einzigen Funktionen, für die wir bislang etwas beweisen können.

Zunächst sieht man fast sofort, daß für eine reelle Funktion $f(t)$

$$c_{-k} = \frac{1}{T}\int_0^T f(t)e^{ki\omega t}\,dt = \frac{1}{T}\int_0^T f(t)\overline{e^{-ki\omega t}}\,dt = \overline{\frac{1}{T}\int_0^T f(t)e^{-ki\omega t}\,dt} = \overline{c_k}$$

ist; es bietet sich daher an, in der Zerlegung

$$f(t) = \sum_{k=-N}^{N} c_k e^{ki\omega t}$$

die Terme mit den Indizes k und $-k$ zusammenzufassen. Mit $c_k = \alpha_k + i\beta_k$ ist dann

$$\begin{aligned} f(t) &= c_0 + \sum_{k=1}^{N}(c_k e^{ki\omega t} + c_{-k}e^{-ki\omega t}) \\ &= c_0 + \sum_{k=1}^{N}(2\alpha_k \cos k\omega t - 2\beta_k \sin k\omega t)\,, \end{aligned}$$

wobei

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

eine reelle Zahl ist. Somit ist

$$f(t) = c_0 + \sum_{k=1}^{N} a_k \cos k\omega t + \sum_{k=1}^{N} b_k \sin k\omega t$$

mit

$$a_k = 2\alpha_k = 2\,\mathfrak{Re}\,c_k = \frac{2}{T}\,\mathfrak{Re}\int_0^T f(t)e^{-ki\omega t}\,dt = \frac{2}{T}\int_0^T f(t)\cos k\omega t\,dt$$

und

$$b_k = -2\beta_k = -2\,\mathfrak{Im}\,c_k = -\frac{2}{T}\,\mathfrak{Im}\int_0^T f(t)e^{-ki\omega t}\,dt = \frac{2}{T}\int_0^T f(t)\sin k\omega t\,dt\,.$$

Als Beispiel betrachten wir den Rechteckimpuls

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

periodisch fortgesetzt mit Periode T. Da $f(t)$ eine ungerade Funktion ist, sind (mit $\omega = 2\pi/T$) die

$$a_k = \frac{2}{T}\int_0^T f(t)\cos k\omega t\,dt = 0$$

für alle k, und auch $c_0 = 0$. Bleiben noch die

$$\begin{aligned} b_k &= \frac{2}{T}\int_0^T f(t)\sin k\omega t\,dt \\ &= \frac{2}{T}\left(\int_0^{T/2} h\sin k\omega t\,dt + \int_{T/2}^0 (-h)\sin k\omega t\,dt\right) \\ &= \frac{4h}{T}\int_0^{T/2}\sin k\omega t\,dt = \frac{4h}{T}\frac{-1}{k\omega}(\cos(k\omega T/2) - \cos 0) \\ &= \frac{-4h}{2k\pi}(\cos(k\pi) - 1) = \begin{cases} \frac{4h}{k\pi} & \text{für ungerades } k \\ 0 & \text{für gerades } k \end{cases}\,. \end{aligned}$$

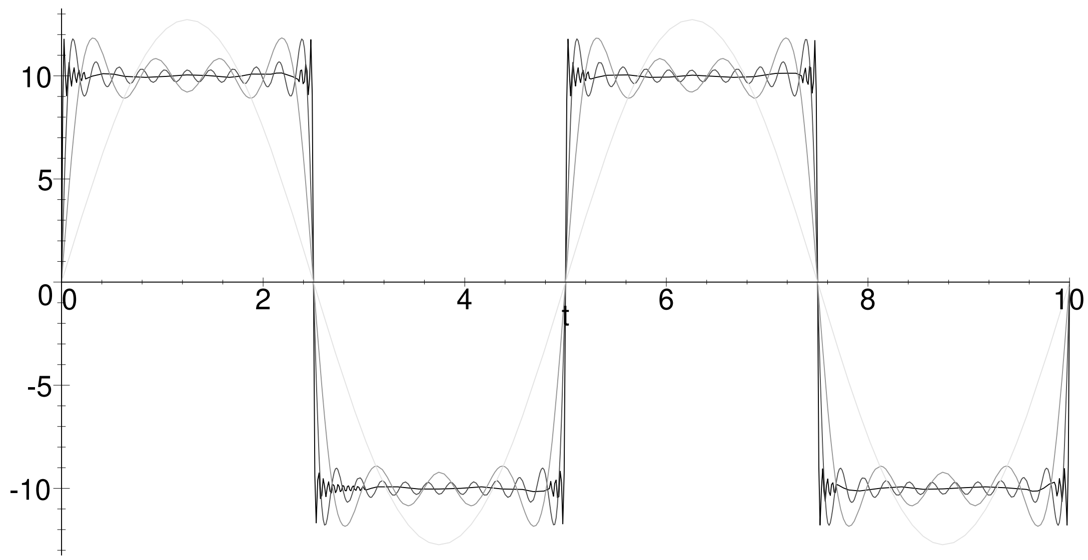


Abb. 30: FOURIERpolynome für Rechteckimpulse

Die FOURIERreihe für $f(t)$ ist also

$$\frac{4h}{\pi}\sum_{k=0}^{\infty}\frac{\sin(2k+1)\omega t}{2k+1}\,.$$

Abbildung 30 zeigt die Teilsummen mit oberer Grenze $1,\ 3,\ 10$ und 50, die die Funktion f in der Tat 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 und wohlbekannte Eigenschaft von FOURIERreihen stückweise stetiger Funktionen; es ist benannt nach dem amerikanischen mathematischen Physikers JOSIAH WILLARD GIBBS (1839–1903).

e) Ausblick: Die Fourier-Transformation

Die Zerlegung einer Funktion in Sinus- und Cosinusschwingungen 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. In diesem Abschnitt soll angedeutet werden, wie dies zumindest für relativ viele Funktionen möglich ist.

Für eine beliebige reelle Funktion $f\colon \mathbb{R} \to \mathbb{R}$ wählen wir dazu zunächst eine (große) Periode T und betrachten die Funktion $\widetilde{f}\colon \mathbb{R} \to \mathbb{R}$, die auf dem Intervall $(-T/2, T/2]$ mit f übereinstimmt und dann periodisch mit Periode T auf ganz $\mathbb{R}$ fortgesetzt wird. Für diese können wir die FOURIERreihe

$$\sum_{k=-\infty}^{\infty} c_k e^{ki\omega t} \quad \text{mit} \quad c_k = \frac{1}{T} \int_{-T/2}^{T/2} f(t) e^{-ki\omega t}\, dt \quad \text{und} \quad \omega = \frac{2\pi}{T}$$

betrachten, die – wenn alles gut geht – fast überall gegen $\widetilde{f}(t)$ konvergieren sollte.

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 eine Funktion $C(\nu)$ durch

$$C(\nu) \underset{\text{def}}{=} \frac{1}{2\pi} \int_{-T/2}^{T/2} f(t) e^{-i\nu t}\, dt\,;$$

dann ist

$$c_k = \frac{2\pi}{T} C(k\omega) = \omega C(k\omega)\,,$$

und die FOURIERreihe läßt sich schreiben als

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

Wäre dies eine endliche Summe, etwa

$$\sum_{k=-N}^{N} C(k\omega) e^{i(k\omega)t} \omega\,,$$

so könnten wir sie auffassen als RIEMANN-Summe für

$$\int_{-N\omega}^{(N+1)\omega} C(\nu) e^{i\nu t}\, d\nu$$

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, sollte die FOURIERreihe konvergieren gegen das (bislang noch nicht definierte) Integral

$$\int_{-\infty}^{\infty} C(\nu)e^{i\nu t}\, d\nu ,$$

das für „vernünftige" Funktionen f fast überall mit $f(t)$ übereinstimmen sollte. Im Idealfall ist also

$$f(t) = \int_{-\infty}^{\infty} C(\nu)e^{i\nu t}\, d\nu \quad \text{mit} \quad C(\nu) \underset{\text{def}}{=} \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)e^{-i\nu t}\, dt .$$

Die Funktion $C(\nu)$, so dieses Integral existiert, wird als FOURIER-Transformierte von $f(t)$ bezeichnet; sie gibt an, wie stark die Frequenz ν beim Aufbau der Funktion $f(t)$ aus Sinus- und Cosinusschwingungen berücksichtigt werden muß.

Einzelheiten über die Existenz der FOURIER-Transformierten und über Bedingungen, in denen $f(t)$ in obiger Form darstellbar ist, müssen wieder auf später verschoben werden; mit unseren bisherigen Mitteln können wir gerade im nächsten Paragraphen definieren, was Integrale von $-\infty$ bis ∞ bedeuten sollen.

§7: Uneigentliche Integrale

Sei $a > 0$ und $b > a$. Dann ist für eine reelle Zahl $r \neq 1$

$$\int_a^b \frac{dx}{x^r} = \frac{-1}{(r-1)x^{r-1}} \bigg|_a^b = \frac{1}{r-1} \left(\frac{1}{a^{r-1}} - \frac{1}{b^{r-1}} \right) .$$

Falls $r > 1$ ist, können wir hiervon den Grenzwert für b gegen unendlich betrachten und es liegt nahe, diesen als Wert des Integrals von a bis unendlich zu bezeichnen:

$$\int_a^{\infty} \frac{dx}{x^r} = \frac{1}{r-1} \frac{1}{a^{r-1}} \quad \text{für} \quad r > 1 .$$

Auch wenn wir nicht bis unendlich integrieren wollen, gibt es Beispiele von Integralen, denen wir via Grenzwertbetrachtung einen sinnvollen Wert zuordnen können, ohne daß das Integral im Sinne unserer bisherigen Definitionen existieren würde: Beispielsweise ist für $a \le c < b$ und eine relle Zahl $0 < r < 1$

$$\int_a^c \frac{dx}{(b-x)^r} = \frac{(b-x)^{1-r}}{r-1}\Bigg|_a^c = \frac{b-a^{1-r}}{r-1} - \frac{b-c^{1-r}}{r-1}\,,$$

und auch hier liegt es nahe, den Grenzwert für $c \to b$ als Wert des Integrals von a bis b zu bezeichnen. Wir müssen hier allerdings vorsichtig sein mit dem Grenzübergang, denn die obige Formel gilt natürlich nur für $c < b$; für $c > b$ ist das Integral undefiniert. In einer solchen Situation können wir daher nicht den gewöhnlichen Limes betrachten, sondern brauchen einen eingeschränkten Grenzwertbegriff, der nur Folgen von einer der beiden Seiten berücksichtigt: Allgemein schreiben wir

$$y = \lim_{x \to b-} f(x)\,,$$

wenn es für jedes $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß $|f(x) - y| < \varepsilon$ für alle x mit $b - \delta < x < b$. (Zur Erinnerung: Beim gewöhnlichen Grenzwert fordert man dies für alle x mit $|x - b| < \delta$, d.h. $b - \delta < x < b + \delta$.)

Völlig analog läßt sich natürlich auch ein rechtsseitiger Grenzwert

$$y = \lim_{x \to a+} f(x)$$

definieren durch die Bedingung, daß es für jedes $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß $|f(x) - y| < \varepsilon$ für alle x mit $b < x < b + \delta$.

Schließlich sollten wir, wenn wir schon beim Definieren sind, zur Bezeichnungsökonomie noch vereinbaren, daß halboffene Intervalle wie

$$[a, b) = \left\{ x \in \mathbb{R} \;\middle|\; a \le x < b \right\}$$

auch für unendliche b definiert sein sollen durch

$$[a, \infty) \underset{\text{def}}{=} \left\{ x \in \mathbb{R} \;\middle|\; a \le x \right\};$$

entsprechend auch

$$(-\infty, b] \underset{\text{def}}{=} \left\{ x \in \mathbb{R} \;\middle|\; x \le b \right\} \quad \text{und} \quad (-\infty, \infty) = \mathbb{R}\,.$$

Für unendliche Intervallgrenzen sind die Definitionen für links- und rechtsseitige Grenzwerte nicht sinnvoll anwendbar; wir vereinbaren daher, daß für $b = \infty$ der Ausdruck $\lim\limits_{c\to b-} \int_a^c f(x)\,dx$ für den gewöhnlichen Grenzwert $\lim\limits_{c\to\infty}$ stehen soll; entsprechend bei $\lim\limits_{x\to a+} f(x)$ für $a = -\infty$.

Mit all diesen Definitionen können wir dann eine Funktion f betrachten, die in einem halboffenen Intervall $[a, b)$ mit $b \in \mathbb{R} \cup \{\infty\}$ definiert und stückweise stetig ist; für diese definieren wir das rechtsseitig uneigentliche Integral

$$\int_a^b f(x)\,dx = \lim_{c\to b-} \int_a^c f(x)\,dx\,,$$

falls dieser Grenzwert existiert; andernfalls sagen wir, das Integral sei *divergent.*

Völlig analog definieren wir linksseitige uneigentliche Integrale: f sei stückweise stetig auf dem halboffenen Intervall $(a, b]$ mit $a \in \mathbb{R} \cup \{-\infty\}$; dann ist

$$\int_a^b f(x)\,dx = \lim_{c\to a+} \int_c^b f(x)\,dx\,,$$

falls dieser Grenzwert existiert; andernfalls sagen wir, das Integral sei *divergent.*

Diese Definitionen sind immer noch nicht allgemein genug: Eine Funktion könnte auch an *beiden* Enden eines Intervalls (a, b) undefiniert sein, wobei wir auch die Sonderfälle $a = -\infty$ und/oder $b = \infty$ zulassen wollen, und zusätzlich könnte sie auch noch Undefiniertheitsstellen $c_1 < \cdots < c_r$ im Intervallinnern haben.

In diesem Fall läßt sich das Intervall so in Teilintervalle zerlegen, daß f in jedem Teilintervall höchstens an *einem* der beiden Intervallenden uneigentlich ist: Falls es keine Undefiniertheitsstellen im Intervallinnern gibt, wählen wir willkürlich ein c_0 zwischen a und b und betrachten die beiden Intervalle $(a, c]$ und $[c, b)$. Im anderen Fall können zwei

Zusatzpunkte notwendig sein: ein Punkt c_0 zwischen a und c_1 sowie ein Punkt c_{r+1} zwischen c_r und b.

Wir sagen dann, das uneigentliche Integral $\int_a^b f(x)\,dx$ konvergiere, wenn *jedes* der Integrale

$$\int_a^{c_0} f(x)\,dx, \quad \int_{c_0}^{c_1} f(x)\,dx, \quad \dots, \quad \int_{c_{r+1}}^{b} f(x)\,dx$$

konvergiert; die Summe ihrer Werte bezeichnen wir als den Wert des Integrals von a bis b.

Als Beispiel betrachten wir das an beiden Grenzen uneigentliche Integral

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^2} .$$

Da der Integrand eine gerade Funktion ist, empfiehlt es sich, das Integrationsintervall, das hier aus ganz $\mathbb{R}$ besteht, bei Null zu unterteilen, und wir erhalten

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{dx}{1+x^2} &= \int_{-\infty}^{0} \frac{dx}{1+x^2} + \int_{0}^{\infty} \frac{dx}{1+x^2} = 2\int_{0}^{\infty} \frac{dx}{1+x^2} \\ &= 2 \lim_{c\to\infty} \int_0^c \frac{dx}{1+x^2} = 2(\lim_{c\to\infty} \arctan c - \arctan 0) = \pi .\end{aligned}$$

Obige Forderung, daß *jedes* der Teilintegrale einzeln konvergieren soll, ist gelegentlich etwas sehr restriktiv. Für das bei Null uneigentliche Integral

$$\int_{-2}^{4} \frac{dx}{x^3}$$

könnte man etwa argumentieren, daß der Integrand eine ungerade Funktion ist, so daß aus Symmetriegründen das Integral von -2 bis 2 verschwinden sollte und

$$\int_{-2}^{4} \frac{dx}{x^3} = \int_{-2}^{2} \frac{dx}{x^3} + \int_{2}^{4} \frac{dx}{x^3} = 0 + \frac{3}{32} = \frac{3}{32}$$

sein sollte. Diese Art der Argumentation ist durch das, was wir bislang gelernt haben, nicht gedeckt, und es gibt auch gute Gründe, sie zu verbieten: Schließlich geht die Stammfunktion $-\frac{1}{2}x^{-2}$ für $x \to 0$ gegen minus unendlich, und wenn eine Größe mit Relevanz in der realen Welt unendlich groß wird, hat dies im Allgemeinen zu gravierende Konsequenzen, als daß man einfach durch diesen Punkt hindurch weitergehen könnte.

Andererseits sind aber viele der mathematischen Formeln, die in den Naturwissenschaften und der Technik angewandt werden, nur näherungsweise gültig: Mathematische Modelle sind praktisch immer *vereinfachende* Modelle der Wirklichkeit, beispielsweite gilt das OHMsche Gesetz sicher nicht mehr, wenn man einen 5 Ω-Widerstand aus einer auf 5 V Spannung ausgelegten Schaltung im Hochspannungslabor mit 100 kV belastet, und es gilt auch nicht mehr ohne Korrekturterme, wenn man einen Wechselspannung mit 500 MHz anlegt.

Entsprechend gibt es durchaus Situationen, in denen das mathematische Modell einen unendlich großen Wert vorhersagt, wohingegen in der Realität limitierende Faktoren, die für Werte im „üblichen“ Größenbereich noch keine nennenswerte Rolle spielen, für eine Begrenzung sorgen. Falls man in einer solchen Situation sicher sein kann, daß auch in der realen Situation noch die Symmetrie zum Nullpunkt erhalten bleibt, kann man so wie oben argumentieren; falls allerdings die Symmetrie *nicht* erhalten bleibt, können durch die Begrenzung der Funktion beliebig große Abweichungen erzeugt werden, über die man mit dem vereinfachten mathematischen Modell nichts aussagen kann.

Da somit alles von der Anwendung abhängt, kann die Mathematik hier nicht mehr bieten als eine *Definition:* Falls für die Funktion f, die auf

$[a, c) \cup (c, b]$ definiert ist, der Grenzwert

$$\lim_{h \to 0} \int_a^{c-h} f(x)\,dx + \int_{c-h}^{b} f(x)\,dx$$

existiert, bezeichnen wir ihn als CAUCHYschen Hauptwert von $\int_a^b f(x)\,dx$ nach dem aus der Analysis I bekannten französischen Mathematiker Baron AUGUSTIN LOUIS CAUCHY (1789–1857). Entsprechend reden wir auch in komplizierteren Situationen mit mehreren Unstetigkeitsstellen vom CAUCHYschen Hauptwert, falls sich eine Aufteilung in Teilintervalle finden läßt, so daß für jedes Teilintervall der CAUCHYsche Hauptwert existiert. Im obigen Beispiel wäre also $\frac{3}{32}$ der CAUCHYsche Hauptwert des Integrals, wohingegen das Integral selbst undefiniert ist.

Die Frage, wann der CAUCHYsche Hauptwert für ein eigentlich divergentes Integral verwendet werden sollte, ist keine mathematische Frage: Unter rein mathematischen Gesichtspunkten gibt es **nie** eine Rechtfertigung für die Verwendung des CAUCHYschen Hauptwerts. Der CAUCHYsche Hauptwert ist nur dann sinnvoll anwendbar, wenn man davon ausgeht, daß ein mathematisches Modell eine Situation nur für nicht zu große Funktionswerte (ungefähr) korrekt beschreibt, und wenn man gleichzeitig sicher ist, daß die Unendlichkeitsstelle des mathematischen Modells für die Anwendung unproblematisch ist und gleichzeitig die Symmetrie, die der Berechnung des CAUCHYschen Hauptwerts zugrundeliegt, auch in der realen Anwendung gilt.

Der CAUCHYsche Hauptwert darf auch *nie* als eine Rechtfertigung dafür verstanden werden, daß man unbesonnen

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

setzt, wobei F eine zwar in den Punkten a und b, nicht aber auch für jeden Zwischenwert $a \le x \le b$ definierte Stammfunktion von f ist: So etwas kann zu Ergebnissen wie

$$\int_{-2}^{2} \frac{dx}{x^2} = -\frac{1}{x}\Big|_{-2}^{2} = \frac{-1}{2} - \frac{-1}{-2} = -1$$

führen, und natürlich ist keine Anwendung denkbar, in der eine negative Zahl in sinnvoller Weise als Integral über eine überall positive Funktion angesehen werden kann. In der Tat existiert im obigen Beispiel weder das Integral noch dessen CAUCHYscher Hauptwert, da sich die Unendlichkeiten links und rechts der Null hier nicht wegheben, sondern verstärken.

Zu einer etwas systematischeren Untersuchung uneigentlicher Integrale empfiehlt es sich, zunächst die Potenzen zu betrachten. Für positive Exponenten r ist x^r überall definiert, so daß Integrale über einen *endlichen* Bereich unproblematisch sind; für Integrale über einen *unendlichen* Bereich rechnet man leicht nach, daß sie immer divergieren. Für $r = 0$ ändert sich nichts an dieser Situation; interessant ist also nur der Fall $r < 0$, wo sowohl der Wert $x = 0$ als auch unendliche obere und/oder untere Grenzen zu Problemen führen können. Für negatives r ist $x^r = 1/x^{-r}$, wir interessieren uns also für

$$\int_a^b \frac{dx}{x^r} = \frac{1}{(1-r)x^{r-1}}\Bigg|_a^b \quad \text{für} \quad r > 0, \quad r \neq 1\,.$$

Für $a > 0$ und $b \to \infty$ existiert ein Grenzwert genau dann, wenn $r - 1 > 0$, also $r > 1$ ist; alsdann ist

$$\int_a^\infty \frac{dx}{x^r} = \lim_{b\to\infty} \int_a^b \frac{dx}{x^r} = \frac{1}{(r-1)a^{r-1}}\,.$$

Für $b > 0$ und $a \to \infty$ existiert ein Grenzwert genau dann, wenn $r - 1 < 0$, also $r < 1$ ist; alsdann ist

$$\int_0^b \frac{dx}{x^r} = \lim_{a\to 0} \int_a^b \frac{dx}{x^r} = \frac{1}{1-r} b^{1-r}\,.$$

Zusammen mit der Monotonieeigenschaft des RIEMANN-Integrals aus §3*a)* ergeben sich hieraus zwei allgemeine Kriterien für die Konvergenz uneigentlicher Integrale:

Satz: Die Funktion f sei stetig für $x \geq a$ und g sei stetig für $0 < x \leq b$. Dann gilt:

1.) Falls es eine reelle Zahl K und eine reelle Zahl $r > 1$ gibt, so daß $|f(x)| \leq \frac{K}{x^r}$ ist, konvergiert $\int_a^\infty f(x)\,dx$.

2.) Falls es eine reelle Zahl K und eine reelle Zahl $0 < r < 1$ gibt, so daß $|f(x)| \leq \frac{K}{x^r}$ ist, konvergiert $\int_0^b f(x)\,dx$.

Beweis: 1.) Nach der Monotonieregel ist für jedes $b \geq a$

$$\int_a^b |f(x)|\;dx \leq K \int_a^b \frac{dx}{x^r}\,,$$

und letzteres Integral konvergiert, wie wir gerade nachgerechnet haben, unter den angenommenen Voraussetzungen. Das linksstehende Integral ist somit beschränkt durch eine von b unabhängige Konstante; da es wegen der Nichtnegativität des Betrags zusätzlich eine monoton wachsende Funktion von b ist, existiert daher der Grenzwert nach dem bekannten, schon für die Existenz des RIEMANN-Integrals verwendeten Satz aus der Analysis I, wonach jede monotone und beschränkte Folge reeller Zahlen konvergent ist (**[Bö]**, *Folgerung 15.4;* **[D]**, *Satz 2.15;* **[MV]**, *Kap. 2, Satz 5.4*).

Genau dasselbe gilt für die ebenfalls nichtnegative Funktion $f(x)+|f(x)|$, die durch $\frac{2K}{x^r}$ beschränkt ist; also existieren die uneigentlichen Integrale

$$\int_a^\infty \big(f(x) + |f(x)|\big)\,dx \quad \text{und} \quad \int_a^\infty |f(x)|\;dx\,,$$

und damit auch das gesuchte Integral als ihre Differenz.

2.) geht völlig analog. ■

Als Beispiel hierfür betrachten wir die Strahlung eines schwarzen Körpers: Nach dem RAYLEIGHschen Strahlungsgesetz, wonach alle mit der Geometrie des Körpers verträglichen Wellenzahlvektoren gleichwahrscheinlich sind, wäre die Energiedichte proportional zum Quadrat der Frequenz, die Gesamtenergie also proportional zu

$$\int_0^\infty \nu^2\,d\nu\,,$$

einem offensichtlich divergenten Integral: Das ist die sogenannte „UV-Katastrophe“, die dieses Modell zum Widerspruch führt.

Die Abhilfe besteht bekanntlich darin, daß man die Gleichverteilung der Frequenzen durch eine BOSE-EINSTEIN-Statistik ersetzt und als neue Energiedichte nun nach dem PLANCKschen Strahlungsgesetz

$$\frac{8\pi h}{c^2}\frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}$$

erhält, wobei h das PLANCKsche Wirkungsquantum, k die BOLTZMANN-Konstante, c die Lichtgeschwindigkeit und T die absolute Temperatur bezeichnet. Hier ist nun die Gesamtenergie

$$\frac{8\pi h}{c^2}\int_0^\infty \frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}\,d\nu\,,$$

und die UV-Katastrophe wird genau dann vermieden, wenn dieses Integral konvergiert.

Mit der Substitution $x = \dfrac{h\nu}{kT}$ erhalten wir

$$\frac{8\pi h}{c^2}\int_0^\infty \frac{\nu^3}{e^{\frac{h\nu}{kT}}-1}\,d\nu = \frac{8\pi h}{c^2}\left(\frac{kT}{h}\right)^4\int_0^\infty \frac{x^3}{e^x-1}\,dx\,.$$

Dieses Integral ist an beiden Grenzen uneigentlich; betrachten wir also eine positive Konstante a und getrennt die Integrale von 0 bis a und von a bis ∞.

Letzteres Integral existiert nach dem gerade bewiesenen Satz, wenn wir K, r finden können, so daß

$$\frac{x^3}{e^x-1} \le \frac{K}{x^r} \quad \text{für alle } x \ge a\,.$$

Diese Ungleichung ist äquivalent zu

$$\frac{1}{e^x-1} \le \frac{K}{x^{r+3}} \quad \text{oder} \quad e^x \ge 1 + \frac{x^{r+3}}{K}\,,$$

und dies gilt mit geeignetem a, K und (beispielsweise) $r = 0$, da die Exponentialfunktion schneller wächst als jedes Polynom.

Entsprechend gilt für $0 \le x \le 1$ die Ungleichung etwa mit $r = \frac{1}{2}$, da dort $e^x - 1$ stärker wächst als jede x-Potenz. Also konvergiert das Integral sowohl an seiner unteren als auch seiner oberen Grenze, und dazwischen ist ohnehin alles unproblematisch, da der Integrand stetig ist.

■

Als (vorerst) letztes Beispiel eines uneigentlichen Integrals sei noch

$$\Gamma(x) \underset{\text{def}}{=} \int_0^\infty e^{-t} t^{x-1}\, dt \quad \text{für } x > 0$$

betrachtet. Dieses Integral ist natürlich immer uneigentlich an der oberen Grenze; für $x < 1$ zusätzlich auch noch an der unteren.

Diese untere Grenze ist völlig harmlos, denn für $0 < x < 1$ ist

$$e^{-t} t^{x-1} = \frac{e^{-t}}{t^{1-x}} \le \frac{1}{t^{1-x}} \quad \text{mit} \quad 0 < 1 - x < 1\,,$$

so daß obiger Satz sofort die Konvergenz zeigt.

Auch die obere Grenze ist unproblematisch, denn die dazu notwendige Abschätzung

$$e^{-t} t^{x-1} \le \frac{K}{t^r} \iff e^t \ge t^{r+x-1}$$

folgt wieder, da die Exponentialfunktion stärker wächst als jede Potenz.

Die somit für alle $x > 0$ definierte Funktion $x \mapsto \Gamma(x)$ heißt EULERsche *Gamma-Funktion* nach dem Schweizer Mathematiker LEONHARD EULER (1707–1783).

Ihre wesentliche Eigenschaft ergibt sich durch partielle Integration:

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1}\, dt = e^{-t} \frac{t^x}{x} \Big|_0^\infty + \frac{1}{x} \int_0^\infty e^{-t} t^x\, dt = \frac{\Gamma(x+1)}{x}$$

oder

$$\Gamma(x+1) = x\Gamma(x)\,.$$

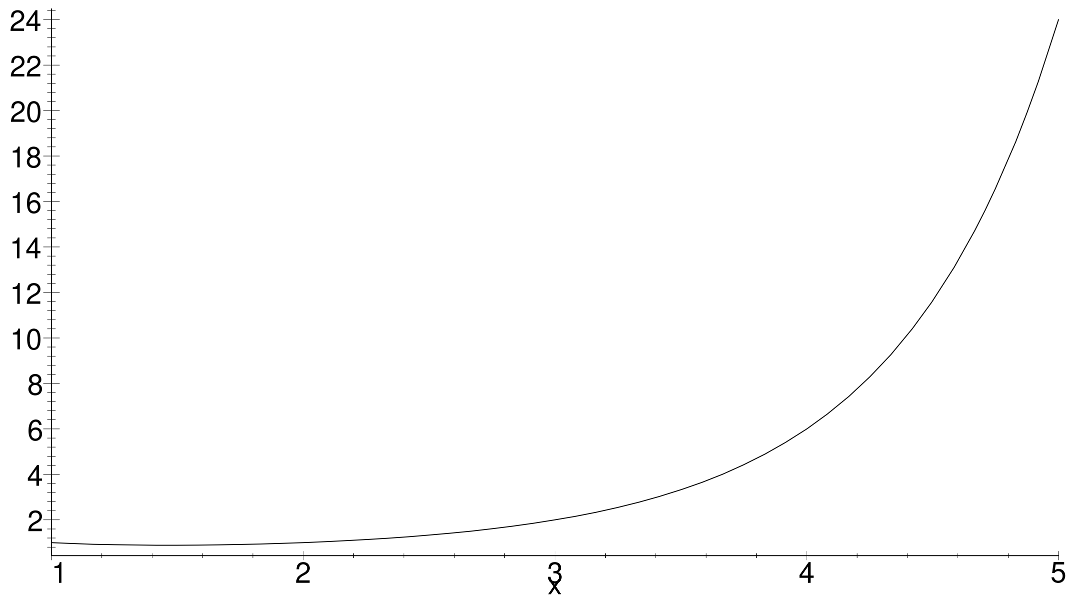


Abb. 31: Die Γ-Funktion

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 eine Art stetig gemachter Fakultätsfunktion. CARL FRIEDRICH GAUSS (1777–1855) 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.

Kapitel 2
Lineare Algebra

In der linearen Algebra geht es um Vektoren, Matrizen, Systeme linearer Gleichungen und den darauf aufbauenden Abstraktionen wie Vektorräumen und linearen Abbildungen. Genau wie Vektoren und Matrizen sind sie Methoden der linearen Algebra nicht wegzudenken aus der mehrdimensionalen Analysis, der Theorie der Differentialgleichungen sowie zahlreichen weiteren Gebieten innerhalb und außerhalb der Mathematik. Wir beginnen mit den wohl allgemein vertrauten

§1: Vektoren im $\mathbb{R}^3$

Unter einem *Pfeil* im $\mathbb{R}^3$ wollen wir die gerichtete Verbindungsstrecke zwischen zwei Punkten verstehen; er ist bestimmt nicht nur durch die Angabe von Anfangs- und Endpunkt, sondern beispielsweise auch durch seinen Anfangspunkt, seine Länge und seine Richtung, wobei diese Richtung allerdings für Pfeile der Länge Null nicht definiert ist – aber dieses Phänomen kennen wir ja schon vom Argument einer komplexen Zahl z, das auch nur für $z \neq 0$ definiert ist.

Zwei Pfeile lassen sich addieren, falls der Endpunkt des ersten gleich dem Anfangspunkt des zweiten ist; die Summe ist dann derjenige Pfeil, der den Anfangspunkt des ersten mit dem Endpunkt des zweiten Pfeils verbindet. Auch läßt sich ein Pfeil mit einer reellen Zahl multiplizieren, wenn wir vereinbaren, daß das Ergebnis jeder Pfeil sein soll, der denselben Anfangspunkt und diesselbe Richtung hat wie der ursprüngliche Pfeil, dessen Länge aber mit der reellen Zahl multipliziert wurde. (Eine Multiplikation mit einer negativen Zahl soll dabei bedeuten, daß der

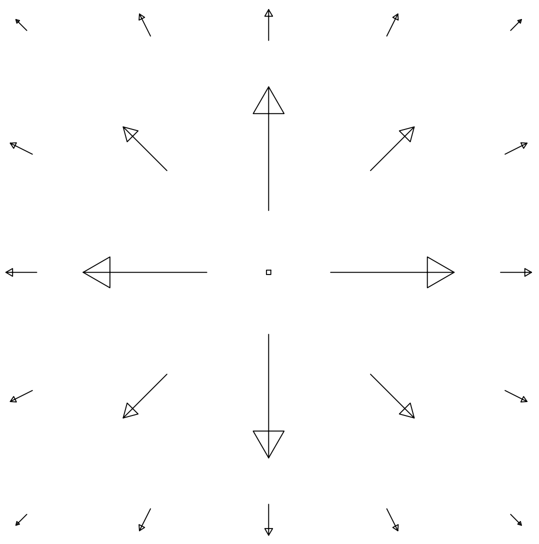


Abb. 32: Abstoßendes Feld einer Punktladung

Pfeil an seinem Anfangspunkt gespiegelt und dann mit dem Betrag der Zahl multipliziert wird.)

Pfeile dieser Art können beispielsweise verwendet werden, um wie in Abbildung 32 das Feld einer Punktladung zu veranschaulichen oder ein Kraftfeld *usw.* Trotzdem eignen sie sich nicht besonders gut als Grundbegriff einer mathematischen Theorie, da sich zwei Pfeile nur addieren lassen, wenn der Endpunkt des ersten gleich dem Anfangspunkt des zweiten ist, was etwa im Fall eines Kraft- oder elektromagnetischen Felds nicht auf den richtigen Additionsbegriff führt: Bei der Überlagerung der Felder zweier Punktladungen entsteht das Gesamtfeld schließlich daraus, daß man in jedem Punkt die beiden dort beginnenden Pfeile betrachtet und diese gemäß dem „Parallelogramm der Kräfte" kombiniert, d.h. der eine Pfeil wird so lange verschoben, bis sein Anfangspunkt gleich dem Endpunkt des anderen Pfeils ist.

Wir definieren daher als neuen Begriff den des *Vektors,* wobei ein Vektor zwar wie ein Pfeil eine Länge und eine Richtung haben soll, aber keinen Anfangspunkt. Mathematisch exakt ausgedrückt ist also ein Vektor eine Äquivalenzklassen von Pfeilen, wobei zwei Pfeile genau dann äquivalent heißen, wenn sie dieselbe Länge und (so die Länge von Null verschieden ist) dieselbe Richtung haben.

Vektoren werden meist durch Fraktur- oder Fettbuchstaben bezeichnet; da sich Fettdruck schlecht auf einer Tafel realisieren läßt und die Hörer dieser Vorlesung keine Frakturbuchstaben lernen wollen, verwenden wir hier stattdessen lateinische Buchstaben, die mit einem Pfeil überstrichen sind, also $\vec{u}, \vec{v}, \vec{w}, \dots$. Die Länge eines Vektors $\vec{v}$ bezeichnen wir mit $|\vec{v}|$; man sollte sich kurz überlegen, daß das für „Vektoren“ in $\mathbb{R}^1$ genau mit dem gewohnten Betrag übereinstimmt.

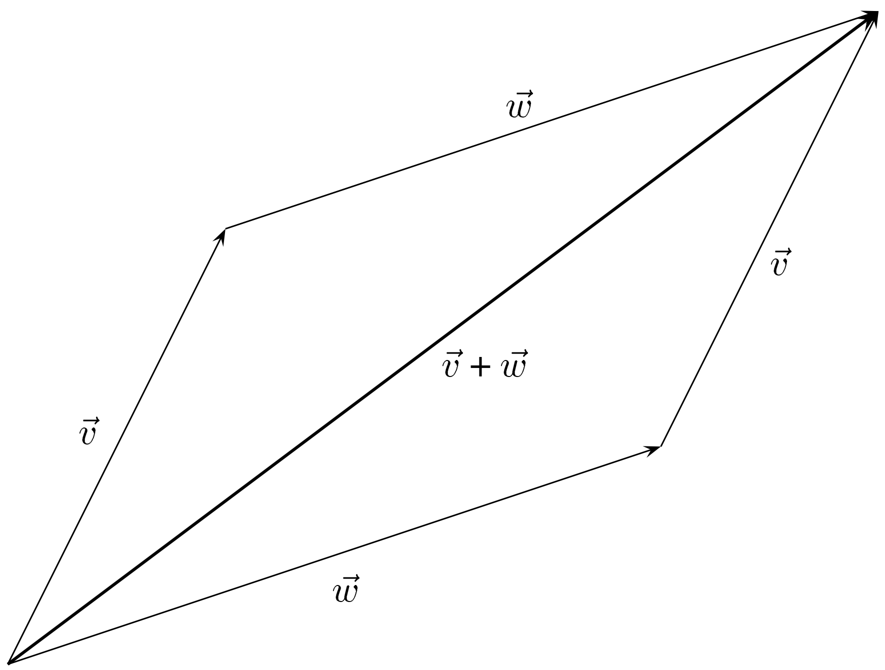


Abb. 33: Das „Parallelogramm der Kräfte“

Zwei Vektoren werden addiert, indem man sich zu jedem der beiden einen Pfeil wählt, wobei der Anfangspunkt des zweiten Pfeils gleich dem Endpunkt des ersten ist; die Summe dieser beiden Pfeile soll dann ein Pfeil für den Summenvektor sein. (Man veranschaulicht sich leicht, daß der resultierende Vektor nicht von der speziellen Wahl der Pfeile abhängt.) Aus dem „Parallelogramm der Kräfte“ in Abbildung 33, die diese Addition veranschaulicht, liest man sofort ab, daß

$$\vec{v}+\vec{w}=\vec{w}+\vec{v}$$

für alle Vektoren $\vec{v}, \vec{w} \in \mathbb{R}^3$.

Genauso definiert man die Multiplikation eines Vektors mit einer reellen Zahl dadurch, daß man einen der Pfeile zum gegebenen Vektor mit der Zahl multipliziert und den Vektor zum resultierenden Pfeil nimmt; die Rechenregel

$$\lambda(\vec{v} + \vec{w}) = \lambda\vec{v} + \lambda\vec{w}$$

folgt, indem man das linke Dreieck in Abbildung 33 um den Faktor λ streckt.

Schließlich gibt es noch zwei Produkte von Vektoren im $\mathbb{R}^3$, das *Skalarprodukt* und das *Vektorprodukt.*

Beginnen wir mit dem Skalarprodukt. Es hat seinen Namen daher, daß es zwei Vektoren zu einem *Skalar,* d.h. einer reellen Zahl, verknüpft; diese wird mit $\vec{v} \cdot \vec{w}$ oder auch kurz $\vec{v}\vec{w}$ bezeichnet. Nach Definition ist

$$\vec{v} \cdot \vec{w} \underset{\text{def}}{=} |\vec{v}|\,|\vec{w}| \measuredangle(\vec{v}, \vec{w})\,,$$

wobei $\measuredangle(\vec{v}, \vec{w})$ für den Winkel zwischen $\vec{v}$ und $\vec{w}$ steht. Da

$$\cos(2\pi - \varphi) = \cos(-\varphi) = \cos\varphi$$

ist, folgt sofort das *Kommutativgesetz*

$$\vec{v} \cdot \vec{w} = \vec{w} \cdot \vec{v}$$

für das Skalarprodukt; außerdem folgt wegen $\cos 0 = 1$, daß

$$\vec{v} \cdot \vec{v} = |\vec{v}|^2 \quad \text{oder} \quad |\vec{v}| = \sqrt{\vec{v} \cdot \vec{v}}$$

ist. Weiter ist $\cos(\pi/2) = \cos(3\pi/2) = 0$ und damit $\vec{v} \cdot \vec{w} = 0$, falls $\vec{v}$ und $\vec{w}$ aufeinander senkrecht stehen; falls keiner der beiden Vektoren der Nullvektor ist, stehen $\vec{v}$ und $\vec{w}$ *genau dann* senkrecht aufeinander, wenn $\vec{v} \cdot \vec{w} = 0$ ist.

Das Skalarprodukt kann auch geometrisch interpretiert werden: Da der Cosinus eines Winkels gleich Ankathete durch Hypothenuse ist, zeigt Abbildung 34, daß $|\vec{w}| \cos \measuredangle(\vec{v}, \vec{w})$ gerade die Länge des senkrecht auf die von $\vec{v}$ erzeugte Gerade projizierten Vektors $\vec{w}$ ist; in der Zeichnung ist dies die fett eingezeichnete Strecke.

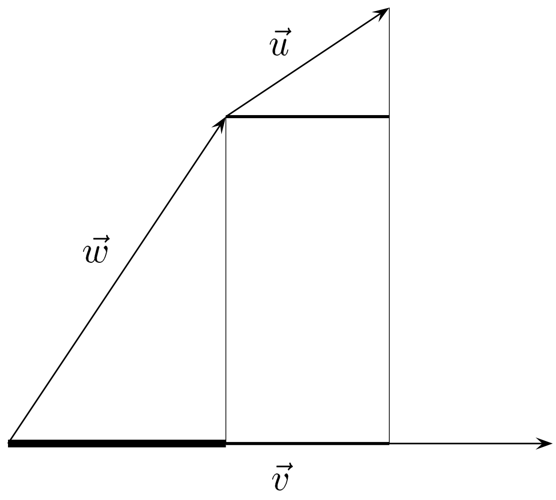


Abb. 34: Geometrische Interpretation des Skalarprodukts

Addieren wir zu $\vec{w}$ einen weiteren Vektor $\vec{u}$, so ist auch hier wieder $|\vec{u}| \cos \measuredangle(\vec{v}, \vec{u})$ die (in Abbildung 34 halbfett eingezeichnete) Länge des auf die von $\vec{v}$ erzeugte Gerade projizierten Vektors $\vec{u}$ und entsprechendes gilt für $\vec{w} + \vec{u}$; wir erhalten somit die Regel

$$\vec{v} \cdot (\vec{w} + \vec{u}) = \vec{v} \cdot \vec{w} + \vec{v} \cdot \vec{u}\,,$$

und damit wegen des Kommutativgesetzes auch

$$(\vec{v} + \vec{w}) \cdot \vec{u} = \vec{v} \cdot \vec{u} + \vec{w} \cdot \vec{u}\,.$$

Zum Berechnen des Skalarprodukts ist die bisherige Definition für viele Fälle recht unhandlich, da Vektoren oft in einer solchen Weise gegeben sind, daß man den Winkel zwischen ihnen *nicht* ohne weiteres kennt. Typischerweise sind Vektoren bezüglich eines *Koordinatensystems* gegeben, etwa eines kartesischen Koordinatensystems aus drei paarweise aufeinander senkrecht stehenden Vektoren $\vec{e}_1, \vec{e}_2, \vec{e}_3$ der Länge eins. Für diese Vektoren läßt sich das Skalarprodukt leicht ausrechnen: Da zwei verschiedene stets aufeinander senkrecht stehen und jeder einzelne die Länge eins hat, ist

$$\vec{e}_i \cdot \vec{e}_j = \begin{cases} 0 & \text{falls } i \neq j \\ 1 & \text{falls } i = j \end{cases}\,.$$

Die rechte Seite wird oft abgekürzt durch δ_{ij}, was nach dem deutschen Mathematiker LEOPOLD KRONECKER (1823–1891) als KRONECKER-δ bezeichnet wird.

Bezüglich eines solchen Koordinatensystems läßt sich ein beliebiger Vektor $\vec{v}$ in der Form

$$\vec{v} = v_1\vec{e}_1 + v_2\vec{e}_2 + v_3\vec{e}_3 \quad \text{mit} \quad e_i \in \mathbb{R}$$

schreiben; falls das Koordinatensystem feststeht (was meistens der Fall ist, da es sonst nicht viel nützt), schreiben wir dies auch kurz als

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ b_3 \end{pmatrix} .$$

Für zwei Vektoren

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \quad \text{und} \quad \vec{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}$$

ist dann

$$\vec{v} \cdot \vec{w} = (v_1\vec{e}_1 + v_2\vec{e}_2 + v_3\vec{e}_3) \cdot (w_1\vec{e}_1 + w_2\vec{e}_2 + w_3\vec{e}_3) ,$$

und nach obigen Rechenregeln läßt sich dies berechnen als

$$\vec{v} \cdot \vec{w} = \sum_{i=1}^{3}\sum_{j=1}^{3} v_i w_j \; \vec{e}_i\vec{e}_j = \sum_{i=1}^{3}\sum_{j=1}^{3} v_i w_j \delta_{ij} = \sum_{i=1}^{3} v_i w_i .$$

In dieser Form werden Skalarprodukte meistens ausgerechnet.

Das zweite Produkt von Vektoren im $\mathbb{R}^3$ ist das *Vektorprodukt* oder *Kreuzprodukt.* Wie die beiden Namen sagen, liefert es als Ergebnis einen Vektor und wird mit $\vec{v} \times \vec{w}$ bezeichnet. Dieser Vektor soll die Länge

$$|\vec{v} \times \vec{w}| = |\vec{v}|\,|\vec{w}|\,|\sin \measuredangle(\vec{v}, \vec{w})|$$

haben und senkrecht stehen sowohl auf $\vec{v}$ als auch auf $\vec{w}$. Insbesondere ist damit $\vec{v} \times \vec{w} = \vec{0}$ falls $\vec{v}$ und $\vec{w}$ dieselbe (oder entgegengesetzte) Richtung haben oder (mindestens) einer der beiden Vektoren der Nullvektor ist; in allen anderen Fällen spannen $\vec{v}$ und $\vec{w}$ eine Ebene auf, und $\vec{v} \times \vec{w}$ muß auf dieser Ebenen senkrecht stehen. Dazu gibt es zwei Möglichkeiten: Entweder er steht ab nach oben oder nach unten (wie immer man auch oben und unten festlegen möchte; für eine beliebige Ebene ist das natürlich völlig willkürlich).

Um diese beiden Fälle voneinander zu unterscheiden, benötigen wir einen neuen Begriff: Wir sagen, daß drei Vektoren $\vec{v}_1, \vec{v}_2$ und $\vec{v}_3$, die nicht in einer Ebenen liegen, ein *Rechtssystem* bilden, wenn sich die Finger der *rechten* Hand so ausrichten lassen, daß der Daumen in Richtung von $\vec{v}_1$ zeigt, der Zeigefinger in Richtung von $\vec{v}_2$ und der Mittelfinger in Richtung von $\vec{v}_3$. Alternativ kann man es, wenn der dritte Vektor so wie hier senkrecht auf den beiden ersten steht, auch so definieren, daß sich ein von $\vec{v}_1$ nach $\vec{v}_2$ gedrehter Korkenzieher in Richtung $\vec{v}_3$ in den Kork bohrt. Ähnlich kann man das Rechtssystem auch mit Schrauben definieren; da es allerdings neben den (üblichen) Rechtsschrauben auch die (seltenen) Linksschrauben gibt, ist diese Definition eventuell zirkulär: Alles hängt davon ab, wie man Rechtsschrauben definiert.

Aus jeder dieser Regeln folgt sofort die *Antikommutativität* des Vektorprodukts:

$$\vec{v} \times \vec{w} = -\vec{w} \times \vec{v}\,.$$

Auch für das Vektorprodukt lassen sich leicht weitere Rechenregeln geometrisch ableiten: Da der Sinus eines Winkels gleich Gegenkathete durch Hypothenuse ist, ist in der von $\vec{v}$ und $\vec{w}$ aufgespannten Ebenen $|\vec{w}|\,|\sin \measuredangle(\vec{v}, \vec{w})|$ gleich der Länge des auf die senkrecht auf $\vec{v}$ stehenden Geraden projizierten Vektors $\vec{w}$, das heißt also gleich der Höhe des in Abbildung 35 eingezeichneten Rechtecks. Die Länge des Vektors $\vec{v} \times \vec{w}$ ist daher gleich dem Flächeninhalt dieses Rechtecks und damit – wie eine Scherung zeigt – gleich der Fläche des von $\vec{v}$ und $\vec{w}$ aufgespannten Parallelogramms.

Wie oben beim Skalarprodukt folgt daher, daß

$$\vec{v} \times (\vec{w} + \vec{u}) = \vec{v} \times \vec{w} + \vec{v} \times \vec{u}$$

und wegen der Antikommutativität entsprechend auch

$$(\vec{u} + \vec{v}) \times \vec{w} = \vec{u} \times \vec{w} + \vec{v} \times \vec{w}\,.$$

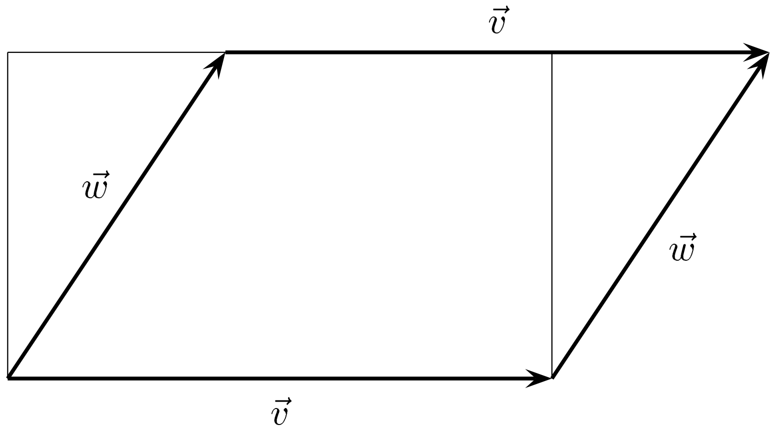


Abb. 35: Geometrische Interpretation des Vektorprodukts

Um das Vektorprodukt in Koordinaten ausrechnen zu können, müssen wir zunächst die Produkte der $\vec{e}_i$ kennen; diese können wir aber nur ausrechnen, wenn wir wissen, ob $\vec{e}_1, \vec{e}_2$ und $\vec{e}_3$ ein Rechtssystem oder ein (analog zu definierendes) Linkssystem bilden. Wir nehmen an, daß sie ein Rechtssystem bilden; dann folgt sofort, daß $\vec{e}_1 \times \vec{e}_2 = \vec{e}_3$ ist, und nach einigen Fingerübungen auch, daß

$$\vec{e}_2 \times \vec{e}_3 = \vec{e}_1 \quad \text{und} \quad \vec{e}_1 \times \vec{e}_3 = -\vec{e}_2$$

ist. Die Produkte mit vertauschten Faktoren sind natürlich gerade das negative davon, und $\vec{e}_i \times \vec{e}_i = 0$. Für

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \quad \text{und} \quad \vec{w} = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}$$

ist also

$$\vec{v} \times \vec{w} = (v_1\vec{e}_1 + v_2\vec{e}_2 + v_3\vec{e}_3) \times (w_1\vec{e}_1 + w_2\vec{e}_2 + w_3\vec{e}_3)\,,$$

und nach obigen Rechenregeln ist dies gleich

$$\sum_{i=1}^{3}\sum_{j=1}^{3} v_i w_j\, \vec{e}_i \times \vec{e}_j$$

$$= (v_2w_3 - v_3w_2)\vec{e}_1 + (v_3w_1 - v_1w_3)\vec{e}_2 + (v_1w_2 - v_2w_1)\vec{e}_3\,.$$

Dies läßt sich dadurch merken, daß man im Schema

$$\begin{matrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 & \vec{e}_1 & \vec{e}_2 \\ v_1 & v_2 & v_3 & v_1 & v_2 \\ w_1 & w_2 & w_3 & w_1 & w_2 \end{matrix}$$

von $\vec{e}_i$ ausgeht und als dessen Koeffizient das Produkt schräg nach rechts unten *positiv* und das Produkt schräg nach links unten *negativ* nimmt.

§2: Vektoren und Matrizen über beliebigen Körpern

Vektoren gibt es natürlich nicht nur im $\mathbb{R}^3$; in diesem Paragraphen wollen wir sie so verallgemeinern, daß die reellen Zahlen durch einen beliebigen Körper und 3 durch eine beliebige Dimension ersetzt werden können.

Sei also k ein beliebiger Körper; die für uns im Augenblick wichtigsten Beispiele sind $k = \mathbb{R}$ und $k = \mathbb{C}$; da aber ein Computer in keinem dieser beiden Körper exakt rechnen kann, sind auch $k = \mathbb{Q}$ und einige seiner Erweiterungskörper wichtig, besonders für Anwendungen in der Computergraphik. Für die bit-orientierte Informationsverarbeitung schließlich ist auch der Körper $k = \mathbb{F}_2 = \{0, 1\}$ mit zwei Elementen wichtig; die Rechenregeln hier sind

$$0 + 0 = 1 + 1 = 0 \quad \text{und} \quad 0 + 1 = 1 + 0 = 1$$

sowie

$$0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0 \quad \text{und} \quad 1 \cdot 1 = 1 \,.$$

Auch dieser Körper hat interessante Erweiterungskörper; der REED-SOLOMON-Decoder in einem CD-Player oder Laufwerk etwa verwendet einen Körper, der bezüglich der Addition gleich $\mathbb{F}_2^8$ ist und somit 64 Elemente hat.

Am anschaulichsten ist alles, wenn man sich weiterhin unter k den Körper der reellen Zahlen vorstellt, allerdings müssen wir darauf achten, daß ab jetzt nur noch Operationen und Begriffe erlaubt sind, die für beliebige Körper definiert sind: Winkel und Winkelfunktionen beispielsweise sind nun genausowenig definiert wie Quadratwurzeln (die schon für $k = \mathbb{Q}$ im allgemeinen nicht existieren) und damit Längen von Vektoren.

Ein Vektor $\vec{v} \in k^n$ ist daher zunächst ganz formal ein n-Tupel, das wir in Spaltenform als

$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$$

schreiben. Vektoren werden komponentententenweise addiert und mit Skalaren, d.h. Elementen des Körpers k, multipliziert: Für $\lambda \in k$ ist

$$\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} + \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} v_1 + w_1 \\ \vdots \\ v_n + w_n \end{pmatrix} \quad \text{und} \quad \lambda \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} \lambda v_1 \\ \vdots \\ \lambda v_n \end{pmatrix}.$$

Auch ein Skalarprodukt kann über die naheliegende Verallgemeinerung der für den $\mathbb{R}^3$ berechneten Formel als

$$\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = v_1 w_1 + \cdots + v_n w_n$$

definiert werden; das Vektorprodukt ist allerdings eine Ausnahmeerscheinung der Dimension drei, die vor allem im Reellen nützlich ist; es läßt sich natürlich über seine Koordinatendarstellung für jeden k^3 definieren, nicht aber in anderen Dimensionen.

Unmittelbar aus den Definitionen folgt, daß für alle Vektoren $\vec{v}, \vec{w} \in k^n$ und alle $\lambda \in k$ gilt:

$$\begin{aligned} \vec{v} + \vec{w} &= \vec{w} + \vec{v} \\ \lambda(\vec{v} + \vec{w}) &= \lambda\vec{v} + \lambda\vec{w} \\ \vec{v} \cdot \vec{w} &= \vec{w} \cdot \vec{v} \qquad \text{und} \\ \lambda(\vec{v} \cdot \vec{w}) &= (\lambda\vec{v}) \cdot \vec{w}\,. \end{aligned}$$

Vektoren aus k^n sind zunächst einmal nichts anderes als n-Tupel oder, anders ausgedrückt, eindimensionale Anordnungen von Körperelementen. Genauso können wir auch zweidimensionale Anordnungen definieren, die sogenannten *Matrizen:*

Definition: Eine $n \times m$-Matrix A über dem Körper k ist eine zweidimensionale Anordnung von Körperelementen in n Zeilen und m Spalten, d.h.

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{pmatrix}.$$

Die Menge aller $n \times m$-Matrizen A über k bezeichnen wir mit $k^{n\times m}$.

Genau wie Vektoren lassen sich auch $n \times m$-Matrizen addieren und mit Körperelementen $\lambda \in k$ multiplizieren gemäß

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \dots & a_{1m}+b_{1m} \\ a_{21}+b_{21} & a_{22}+b_{22} & \dots & a_{2m}+b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}+b_{n1} & a_{n2}+b_{n2} & \dots & a_{nm}+b_{nm} \end{pmatrix}$$

und

$$\lambda \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} & \dots & \lambda a_{1m} \\ \lambda a_{21} & \lambda a_{22} & \dots & \lambda a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda a_{n1} & \lambda a_{n2} & \dots & \lambda a_{nm} \end{pmatrix} .$$

Auch hier folgen wieder unmittelbar die Rechenregeln

$$A + B = B + A \qquad \text{und}$$
$$\lambda(A + B) = \lambda A + \lambda B .$$

Schließlich lassen sich Matrizen gelegentlich auch miteinander multiplizieren; dies läßt sich am besten verstehen, wenn wir zunächst Matrizen mit *linearen Abbildungen* in Verbindung bringen:

Definition: Eine lineare Abbildung $\varphi\colon k^m \to k^n$ ist eine Abbildung, die für alle $\lambda, \mu \in k$ und alle Vektoren $\vec{v}, \vec{w} \in k^m$ die Gleichung

$$\varphi(\lambda\vec{v} + \mu\vec{w}) = \lambda\varphi(\vec{v}) + \mu\varphi(\vec{w})$$

erfüllt.

Diese Definition läßt sich natürlich sofort durch vollständige Induktion ausdehnen zu der Aussage, daß φ genau dann linear ist, wenn für beliebige Körperelemente λ_i und Vektoren $\vec{v}_i$ auch

$$\varphi(\lambda_1\vec{v}_1 + \cdots \lambda_r\vec{v}_r) = \lambda_1\varphi(\vec{v}_1) + \cdots \lambda_r\varphi(\vec{v}_r)$$

ist. Insbesondere gilt dies für die *Koordinatendarstellung* eines Vektors: Definieren wir in k^m *Einheitsvektoren* $\vec{e}_i$ durch

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots, \quad \vec{e}_m = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

so läßt sich ein beliebiger Vektor $\vec{v} \in k^m$ schreiben als

$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = v_1\vec{e}_1 + \cdots + v_n\vec{e}_n .$$

Für eine lineare Abbildung $\varphi\colon k^m \to k^n$ ist daher

$$\varphi(\vec{v}) = v_1\varphi(\vec{e}_1) + \cdots + v_m\varphi(\vec{e}_m) .$$

Falls wir die Vektoren $\varphi(e_i)$ kennen, können wir also $\varphi(\vec{v})$ für jeden Vektor $\vec{v} \in k^m$ ausrechnen: Ist

$$\varphi(\vec{e}_i) = \begin{pmatrix} a_{1i} \\ \vdots \\ a_{ni} \end{pmatrix},$$

so ist

$$\varphi(\vec{v}) = v_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix} + \cdots + v_m \begin{pmatrix} a_{1m} \\ \vdots \\ a_{nm} \end{pmatrix} = \begin{pmatrix} a_{11}v_1 + \cdots + a_{1m}v_m \\ \vdots \\ a_{n1}v_1 + \cdots + a_{nm}v_m \end{pmatrix} .$$

Die Elemente a_{ij} können wir natürlich zu einer Matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{pmatrix} \in k^{n\times m}$$

zusammenfassen; diese Matrix bezeichnen wir als die *Abbildungsmatrix* von φ. Ihre Spalten sind, wenn man sie als Vektoren auffaßt, gerade die Bilder der Einheitsvektoren.

Wie obige Rechnung zeigt, ist eine lineare Abbildung φ durch ihre Abbildungsmatrix eindeutig bestimmt. Um dies in möglichst kompakter Weise darzustellen, definieren wir das Produkt einer $n \times m$-Matrix $A \in k^{n\times m}$ mit einem Vektor $\vec{v} \in k^m$ als

$$A\vec{v} \underset{\text{def}}{=} \begin{pmatrix} a_{11}v_1 + \cdots + a_{1m}v_m \\ \vdots \\ a_{n1}v_1 + \cdots + a_{nm}v_m \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^m a_{1j}v_j \\ \vdots \\ \sum_{j=1}^m a_{nj}v_j \end{pmatrix} ;$$

damit läßt sich also jede lineare Abbildung $\varphi\colon k^m \to k^n$ in der Form

$$\varphi(\vec{v}) = A\vec{v}$$

schreiben, wobei $A \in k^{n\times m}$ die Abbildungsmatrix von φ ist.

Ist umgekehrt $A \in k^{n\times m}$ eine beliebige Matrix, so überzeugt man sich leicht, daß die Abbildung

$$\varphi\colon \begin{cases} k^n \to k^m \\ \vec{v} \mapsto A\vec{v} \end{cases}$$

linear ist; die linearen Abbildungen $k^m \to k^n$ entsprechen also eineindeutig den Matrizen aus $k^{n\times m}$.

Zwei lineare Abbildungen $\varphi\colon k^m \to k^n$ und $\psi\colon k^p \to k^m$ lassen sich verknüpfen zu einer Abbildung

$$\varphi \circ \psi\colon \begin{cases} k^p \to k^n \\ \vec{v} \mapsto \varphi\bigl(\psi(\vec{v})\bigr) \end{cases} .$$

Die zu definierende Matrixmultiplikation soll dieser Hintereinanderausführung von Abbildungen entsprechen; ist also $A \in k^{n\times m}$ die Abbildungsmatrix zu φ und $B \in k^{m\times p}$ die zu ψ, so soll $AB \in k^{n\times p}$ die Abbildungsmatrix zu $\varphi \circ \psi$ sein.

Um AB explizit berechnen zu können, müssen wir diese Abbildungsmatrix ausrechnen: Für den Vektor

$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix} \in k^p$$

ist

$$\vec{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} \underset{\text{def}}{=} \psi(\vec{v}) = B\vec{v} = \begin{pmatrix} \sum_{j=1}^p b_{1j} v_j \\ \vdots \\ \sum_{j=1}^p b_{mj} v_j \end{pmatrix}$$

und damit

$$\varphi \circ \psi(\vec{v}) = \varphi\big(\psi(\vec{v})\big) = \varphi(\vec{w}) = A\vec{w} = \begin{pmatrix} \sum_{\ell=1}^m a_{1\ell} w_\ell \\ \vdots \\ \sum_{\ell=1}^m a_{n\ell} w_\ell \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{\ell=1}^m a_{1\ell} \sum_{j=1}^p b_{\ell j} v_j \\ \vdots \\ \sum_{\ell=1}^m a_{n\ell} \sum_{j=1}^p b_{\ell j} v_j \end{pmatrix} = \begin{pmatrix} \sum_{\ell=1}^m \sum_{j=1}^p a_{1\ell} b_{\ell j} v_j \\ \vdots \\ \sum_{\ell=1}^m \sum_{j=1}^p a_{n\ell} b_{\ell j} v_j \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{j=1}^p \left(\sum_{\ell=1}^m a_{1\ell} b_{\ell j}\right) v_j \\ \vdots \\ \sum_{j=1}^p \left(\sum_{\ell=1}^m a_{n\ell} b_{\ell j}\right) v_j \end{pmatrix} = P\vec{v}\,,$$

wobei die Matrix $P = (p_{ij}) \in k^{n\times p}$ die Einträge

$$p_{i\ell} = \sum_{\ell=1}^m a_{i\ell} b_{\ell j}$$

hat. Nach der gerade beendeten Rechnung ist dies die Abbildungsmatrix von $\varphi\circ\psi$, wir *definieren* sie daher als das Produkt $P = AB$ von A und B:

Definition: $A = (a_{i\ell}) \in k^{n\times m}$ und $B = (b_{\ell j}) \in k^{m\times p}$ seien zwei Matrizen derart, daß die Spaltenzahl m von A mit der Zeilenzahl von B übereinstimmt. Dann bezeichnen wir die Matrix $P = (p_{ij}) \in k^{n\times p}$ mit

$$p_{ij} = \sum_{\ell=1}^m a_{i\ell} b_{\ell j}$$

als *Produktmatrix* $P = AB$ von A und B.

Da alle $b_{\ell j}$ linear in den Ausdrücken für p_{ij} vorkommen, folgt sofort, daß für eine weitere Matrix $C \in k^{m\times p}$ gilt

$$A(B+C) = AB + AC\,;$$

ganz entsprechend ist für eine weitere Matrix $D \in k^{n\times m}$

$$(A+D)B = AB + DB\,.$$

(Hier kann man natürlich *nicht* mit irgendeiner Art von Kommutativität argumentieren: Falls nicht zufälligerweise $n = p$ sein sollte, ist das Matrixprodukt BA nicht einmal definiert, geschweige denn gleich AB. Allgemein ist Kommutativität bei der Hintereinanderausführung von Abbildungen eine sehr seltene Ausnahmeerscheinung; schließlich ist auch $\sin(x^2) \neq \sin^2 x$ für fast jedes x, und so ist auch bei Matrizen, selbst wenn beide Produkte definiert sind, im allgemeinen $AB \neq BA$.)

Problemlos sind Regeln wie

$$(\lambda A)B = \lambda(AB) = A(\lambda B) \quad \text{für alle } \lambda \in k\,;$$

nicht so klar und rechnerisch erheblich aufwendiger ist das *Assoziativitätsgesetz*

$$A(BC) = (AB)C \quad \text{für alle } A \in k^{n\times m},\ B \in k^{m\times p},\ C \in k^{p\times q}\,.$$

Zum Glück müssen wir das aber nicht nachrechnen, denn wir wissen ja, daß Matrizen genau den linearen Abbildungen entsprechen und das Matrizenprodukt der Hintereinanderausführung von Abbildungen. Sind aber

$$\varphi\colon k^m \to k^n, \quad \psi\colon k^n \to k^p \quad \text{und} \quad \omega\colon k^q \to k^p$$

die linearen Abbildungen mit Abbildungsmatrizen A, B, C, so ist für jeden Vektor $\vec{v} \in k^q$ sowohl

$$\big(\varphi\circ(\psi\circ\omega)\big)(\vec{v}) = \varphi\big(\psi\circ\omega)(\vec{v})\big) = \varphi\Big(\psi\big(\omega(\vec{v})\big)\Big)$$

als auch

$$\big((\varphi\circ\psi)\circ\omega\big)\,(\vec{v}) = (\varphi\circ\psi)\big(\omega(\vec{v})\big) = \varphi\Big(\psi\big(\omega(\vec{v})\big)\Big)\,,$$

d.h. für die Hintereinanderausführung von Abbildungen (egal ob linear oder nicht) ist das Assoziativgesetz

$$\varphi \circ (\psi \circ \omega) = (\varphi \circ \psi) \circ \omega$$

automatisch erfüllt.

Da nun $A(BC)$ die Abbildungsmatrix von $\varphi \circ (\psi \circ \omega)$ ist und $(AB)C$ die von $(\varphi \circ \psi) \circ \omega$, und da diese beiden Abbildungen übereinstimmen, müssen auch die Abbildungsmatrizen gleich sein, wir haben also gezeigt, daß

$$A(BC) = (AB)C \quad \text{für alle } A \in k^{n\times m},\ B \in k^{m\times p},\ C \in k^{p\times q}\ ,$$

ohne daß wir ein einziges Matrixprodukt explizit ausrechnen mußten.

In dieser schönen Situation ist man natürlich nur selten; als nächstes müssen wir uns unbedingt überlegen, wie man Matrixprodukte konkret am geschicktesten ausrechnet.

Die in der Definition des Matrixprodukts gegebene Formel ist natürlich explizit genug für jede Rechnung; zur praktischen Durchführung empfiehlt es sich, den zweiten Faktor des Produkts etwas höher zu stellen nach dem Schema

$$\begin{array}{cc} & \begin{pmatrix} \cdots & b_{1j} & \cdots \\ & \vdots & \\ \cdots & b_{mj} & \cdots \end{pmatrix} \\ \begin{pmatrix} \vdots & & \vdots \\ a_{i1} & \cdots & a_{im} \\ \vdots & & \vdots \end{pmatrix} & \begin{pmatrix} & \vdots & \\ \cdots & \sum\limits_{\ell=1}^{m} a_{i\ell} b_{\ell j} & \cdots \\ & \vdots & \end{pmatrix} \end{array} \quad ;$$

dadurch behält man einerseits den Überblick, welcher Rechenschritt als nächster auszuführen ist; außerdem sehen wir anhand dieser Darstellung, daß der Eintrag an der Stelle (i, j) der Produktmatrix gerade das Skalarprodukt des i-ten Zeilenvektors von A und mit dem j-ten Spaltenvektors von B ist:

$$p_{ij} = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{im} \end{pmatrix} \cdot \begin{pmatrix} b_{1j} \\ \vdots \\ b_{mj} \end{pmatrix} .$$

Entsprechend steht in der j-ten Spalte der Produktmatrix AB gerade das Produkt von A mit der als Vektor aufgefaßten j-ten Spalte von B.

Somit kann man das oben definierte Produkt einer Matrix mit einem Vektor als Spezialfall des Matrixprodukts auffassen, weshalb Vektoren vor allem in der physikalischen und technischen Literatur gerne mit $n \times 1$-Matrizen identifiziert werden. Auch das Skalarprodukt zweier Vektoren läßt sich als Spezialfall des Matrixprodukts interpretieren, wenn man den ersten Vektor als eine $1 \times n$- und den zweiten als eine $n \times 1$-Matrix betrachtet. Dazu muß man allerdings zunächst aus einer $n \times 1$-Matrix eine $1 \times n$-Matrix machen, und dafür definieren wir ganz allgemein eine Operation, die die Zeilen und Spalten einer Matrix vertauscht:

Definition: Die transponierte Matrix zur Matrix $A = (a_{ij}) \in k^{n\times m}$ ist jene Matrix ${}^tA = (b_{ij}) \in k^{m\times n}$ mit

$$b_{ij} = a_{ji} \,.$$

Für zwei $n \times 1$-Matrizen $A, B \in k^{n\times 1}$ ist daher ${}^tA \cdot B$ gleich dem Skalarprodukt der entsprechenden Vektoren.

Für Matrizen $A \in k^{n\times m}$ und $B \in k^{m\times p}$ rechnet man leicht nach, daß ${}^t(AB) = {}^tB \cdot {}^tA$ ist.

Für eine quadratische Matrix $A \in k^{n\times n}$ kann es vorkommen, daß ${}^tA = A$ ist; in diesem Fall nennen wir die Matrix A *symmetrisch.* Entsprechend könnte ${}^tA = -A$ sein; dann bezeichnen wir A als *schiefsymmetrisch.*

Symmetrische und schiefsymmetrische Matrizen treten in Physik und Technik oft in natürlicher Weise auf; beispielsweise ist der *Trägheitstensor* eines starren Körpers durch eine symmetrische Matrix gegeben, ebenso der Dielektrizitätstensor eines optisch inaktiven Kristalls. Ein Beispiel einer schiefsymmetrischen Matrix ist etwa der F-Tensor

$$F = \begin{pmatrix} 0 & H_x & -H_y & E_x \\ -H_z & 0 & H_x & E_y \\ H_y & -H_x & 0 & E_z \\ -E_x & -E_y & -E_y & 0 \end{pmatrix} ,$$

durch den man in der relativistischen Elektrodynamik ein elektromagnetisches Feld kovariant beschreiben kann.

Es ist kein Zufall, daß in der Hauptdiagonalen dieser Matrix lauter Nullen stehen: Wegen der Schiefsymmetrie muß ein Diagonalelement a_{ii} die Gleichung $a_{ii} = -a_{ii}$ erfüllen, und Null ist die einzige reelle Zahl, die gleich ihrem Negativen ist.

Für beliebige Körper muß man etwas vorsichtiger sein: Im zu Beginn dieses Paragraphen eingeführten Körper $k = \mathbb{F}_2$ mit zwei Elementen ist $1 + 1 = 0$, also $1 = -1$, so daß $a_{ii} = -a_{ii}$ keine Bedingung an a_{ii} stellt. In einem Körper, in dem man durch $1 + 1$ dividieren kann, in dem also $1 + 1 \neq 0$ ist, gibt es dagegen keine Probleme: Dort gilt

$$a_{ii} = -a_{ii} \Rightarrow (1 + 1)a_{ii} = 0 \Rightarrow a_{ii} = 0\,.$$

Unter den Körpern, die wir bislang kennen, ist $\mathbb{F}_2$ der einzige, in dem $1 + 1 = 0$ ist; es gibt allerdings weitere solche Körper, die etwa in der Kodierungstheorie durchaus auch eine praktische Rolle spielen.

§3: Lineare Gleichungssysteme

Die häufigste und wohl auch wichtige Anwendung der linearen Algebra ist das Lösen linearer Gleichungssysteme. Unter einem linearen Gleichungssystem versteht man ein Gleichungssystem der Form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m &= b_2 \\ \vdots \qquad &= \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m &= b_n\,, \end{aligned}$$

wobei die a_{ij} und die b_i bekannte Elemente des Grundkörpers k sind, während die x_i für unbekannte Elemente von k stehen, die so bestimmt werden sollen, daß alle m Gleichungen zu wahren Aussagen werden. Ein m-Tupel $(x_1, \ldots, x_m) \in k^m$ mit dieser Eigenschaft bezeichnen wir als eine *Lösung* des linearen Gleichungssystems; die Menge $\mathcal{L} \subseteq k^m$ aller Lösungen heißt *Lösungsmenge.*

Falls alle $b_i = 0$ sind, sprechen wir von einem *homogenen* Gleichungssystem, ansonsten von einem *inhomogenen.* Für ein homogenes lineares

Gleichungssystem ist $(0, \dots, 0)$ offenbar immer eine Lösung; ein inhomogenes lineares Gleichungssystem kann dagegen auch unlösbar sein, d.h. $\mathcal{L} = \emptyset$, etwa im Beispiel

$$\begin{aligned} x_1 + x_2 &= 0 \\ x_1 + x_2 &= 1 \,. \end{aligned}$$

a) Das Gaußsche Eliminationsverfahren

Aus der Schule sind vielleicht spezielle Verfahren bekannt, wie man für $n, m \leq 3$ die Lösungsmenge eines linearen Gleichungssystems bestimmen kann. Hier wollen wir ein Verfahren betrachten, das für beliebige n, m zum Ziel führt und die Lösung auch, falls das Gleichungssystem nicht gerade eine spezielle Gestalt hat, im allgemeinen auf die effizienteste Weise liefert. Gemeint ist der GAUSS-*Algorithmus,* benannt wieder nach CARL FRIEDRICH GAUSS (1777–1855), den wir schon von der GAUSS*sche Zahlenebenen* und der Π-Funktion her kennen, und der uns im Laufe der Vorlesung *Höhere Mathematik* noch mehrfach in den verschiedensten Zusammenhängen begegnen wird.

Die Grundidee ist sehr einfach: Im Falle einer einzigen Gleichung mit einer einzigen Unbekannten x können wir das „Gleichungssystem“

$$ax = b$$

sofort lösen: Für $a \neq 0$ ist $x = -b/a$; ansonsten gibt es für $b \neq 0$ keine Lösung, d.h. $\mathcal{L} = \emptyset$, und für $a = b = 0$ ist $\mathcal{L} = k$.

Das GAUSSsche Eliminationsverfahren führt ein allgemeines lineares Gleichungssystem sukzessive zurück auf solche lineare Gleichungen in einer Unbekannten, ausgehend von zwei trivialen Beobachtungen:

1. Die Lösungsmenge eines linearen Gleichungssystems ändert sich nicht, falls wir zwei Gleichungen miteinander vertauschen.
2. Die Lösungsmenge ändert sich auch dann nicht, wenn wir ein Vielfaches einer Gleichung zu einer anderen addieren, d.h. wenn wir die Gleichung

$$\ell_j(x_1, \dots, x_m) \underset{\text{def}}{=} a_{j1}x_1 + \dots + a_{jm}x_m = b_j$$

ersetzen durch

$$\ell_j(x_1,\dots,x_m) + \lambda\ell_i(x_1,\dots,x_m) = b_j + \lambda b_i\,, \qquad (*)$$

denn unter der Nebenbedingung

$$\ell_i(x_1,\dots,x_m) = b_i$$

ist $(*)$ äquivalent zu $\ell_j(x_1,\dots,x_m) = 0$.

Mit Hilfe dieser beiden Beobachtungen läßt sich nun die Variablenanzahl wie folgt sukzessive reduzieren: Beginnen wir mit der Elimination von x_1. Falls x_1 im Gleichungssystem überhaupt nicht vorkommt, falls also alle $a_{i1} = 0$ sind, gibt es nichts zu tun: Wir haben ein Gleichungssystem in $x_2,\dots,x_m$, und für jede Lösung $(x_2,\dots,x_m)$ dieses Systems sowie jedes beliebige $x_1 \in k$ ist $(x_1,x_2,\dots,x_m)$ eine Lösung des ursprünglichen Systems.

Ansonsten können wir, indem wir nötigenfalls zwei Gleichungen miteinander vertauschen, annehmen, daß $a_{11} \neq 0$ ist. Dann lassen wir die erste Gleichung so stehen, wie sie ist, und ersetzen jede weitere Gleichung $\ell_j(x_1,\dots x_m) = b_j$ durch

$$\ell_j(x_1,\dots x_m) - \frac{a_{j1}}{a_{11}}\ell_1(x_1,\dots x_m) = b_j - \frac{a_{j1}}{a_{11}}b_1\,;$$

in diesen Gleichungen kommt x_1 offenbar nicht mehr vor. Wir haben somit ein Gleichungssystem in einer Variablen weniger, plus der Gleichung

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m = b_1\,.$$

Sobald wir das Gleichungssytem für $x_2,\dots,x_m$ gelöst haben, wird diese Gleichung nach Einsetzen einer Lösung $(x_2,\dots,x_m)$ zu einer linearen Gleichung für x_1, die wir lösen können.

Das Gleichungssystem für $x_2,\dots,x_m$ wird nun, falls $m > 2$ ist, nach genau derselben Methode weiterreduziert: Falls x_2 in keiner der Gleichungen vorkommt, haben wir tatsächlich ein Gleichungssystem in $x_3,\dots,x_m$, andernfalls können wir durch Vertauschen zweier Gleichungen annehmen, daß x_2 in der ersten Gleichung mit einem von Null verschiedenen Vorfaktor auftritt, und wir können wie oben x_2 aus allen weiteren Gleichungen eliminieren *usw.*

Da in jedem Eliminationsschritt ein Nenner auftritt, können die Nenner vor allem bei großem m gelegentlich schnell unübersichtlich groß werden. Obwohl es *im Prinzip* nicht notwendig ist, kann man um dies zu vermeiden noch als dritte Operation die Multiplikation einer Gleichung mit einem Körperelement (z.B. einem Hauptnenner der Koeffizienten) zulassen. Dies ist gleichbedeutend damit, daß man anstelle der Operation $(*)$ allgemeiner die Ersetzung von $\ell_j(x_1, \dots x_m)$ durch

$$\mu\ell_j(x_1, \dots x_m) + \lambda\ell_i(x_1, \dots x_m)$$

mit beliebigem $\mu \neq 0$ aus k zuläßt.

c) Erste Beispiele

Betrachten wir dazu einige Beispiele; der Grundkörper k sei dabei jeweils der Körper der reellen Zahlen oder einer seiner Teilkörper, etwa $k = \mathbb{Q}$.

Sei zunächst

$$\begin{array}{rcrcrcr} 3x_1 & + & 2x_2 & + & x_3 & = & 5 \\ x_1 & + & 2x_2 & + & 4x_3 & = & 3 \\ 5x_1 & - & 3x_2 & + & 7x_2 & = & 19 \end{array}$$

das zu lösende Gleichungssystem. Da x_1 in der ersten Gleichung tatsächlich vorkommt, müssen wir nichts vertauschen (obwohl sich gleich herausstellen wird, daß wir vielleicht doch besser die ersten beiden Gleichungen miteinander vertauscht hätten) und können ein Drittel der ersten Gleichung von der zweiten und fünf Drittel der ersten Gleichung von der dritten subtrahieren, um x_2 aus diesen beiden Gleichungen zu eliminieren. Dies führt auf das neue Gleichungssystem

$$\begin{array}{rcrcrcr} 3x_1 & + & 2x_2 & + & x_3 & = & 5 \\ & & \frac{4}{3}x_2 & + & \frac{11}{3}x_3 & = & \frac{4}{3} \\ & & -\frac{19}{3}x_2 & + & \frac{16}{3}x_3 & = & \frac{32}{3}\,, \end{array}$$

von dem uns zunächst nur die letzten beiden Zeilen interessieren. Diese bilden ein lineares Gleichungssystem in x_2 und x_3, aus dem wir x_2 in einer der beiden Gleichungen eliminieren möchten.

Da x_2 in der zweiten Gleichung wirklich vorkommt, addieren wir $19/4$ mal diese Gleichung zur dritten und erhalten als neues System

$$\begin{array}{rcrcrcl} 3x_1 & + & 2x_2 & + & x_3 & = & 5 \\ & & \frac{4}{3}x_2 & + & \frac{11}{3}x_3 & = & \frac{4}{3} \\ & & & & \frac{91}{4}x_3 & = & 17\,. \end{array}$$

Hier ist die letzte Gleichung eine gewöhnliche lineare Gleichung für x_3, aus der wir sofort ablesen, daß

$$x_3 = \frac{68}{91}$$

ist. Dies setzen wir in die vorletzte Gleichung ein:

$$\frac{4}{3}x_2 + \frac{748}{273} = \frac{4}{3}$$

ist eine lineare Gleichung für x_2 mit Lösung

$$x_2 = -\frac{96}{91}\,.$$

Dies sowie x_3 setzen wir schließlich in die erste Gleichung ein:

$$3x_1 - \frac{124}{91} = 5$$

hat die Lösung

$$x_1 = \frac{193}{91}\,,$$

so daß ingesamt

$$\left(\frac{193}{91}, -\frac{96}{91}, \frac{68}{91}\right)$$

die (einzige) Lösung des Gleichungssystems ist.

Als nächstes Beispiel betrachten wir

$$\begin{array}{rcrcrcrcrcrcrcr} & & & & 5x_3 & & & & & - & 2x_6 & & & = & 3 \\ & & 4x_2 & & & & & & & + & 3x_6 & - & 7x_7 & = & -2 \\ x_1 & - & x_2 & + & 2x_3 & - & 3x_4 & + & x_5 & & & & & = & 0 \end{array}$$

Hier kommt x_1 in der ersten Gleichung nicht vor, wohl aber in der dritten. Wir vertauschen daher die erste Gleichung mit der dritten und erhalten

$$\begin{array}{rcrcrcrcrcrcrcr} x_1 & - & x_2 & + & 2x_3 & - & 3x_4 & + & x_5 & & & & & = & 0 \\ & & 4x_2 & & & & & & & + & 3x_6 & - & 7x_7 & = & -2 \\ & & & & 5x_3 & & & & & - & 2x_6 & & & = & 3\,. \end{array}$$

In diesem Gleichungssystem kommt x_1 nur in der ersten Gleichung vor, x_2 nicht mehr hinter der zweiten und x_3 (natürlich) nicht mehr hinter der dritten.

Die dritte Gleichung $5x_3 - 2x_6 = 3$ hat offensichtlich keine eindeutig bestimmte Lösung: Wir können entweder nach x_3 oder nach x_6 auflösen, so daß eine der beiden Unbekannten beliebig bewählt werden kann. Wählen wir etwa für x_6 irgendein Element $\lambda_6 \in k$, so wird

$$x_6 = \lambda_6, \quad x_3 = \frac{2}{5}\lambda_6 + \frac{3}{5}\,.$$

Die zweite Gleichung wird nach Einsetzen von $x_6 = \lambda_6$ zu

$$4x_2 + 3\lambda_6 - 7x_7 = -2\,,$$

also zu einer Beziehung zwischen x_2 und x_7, denn λ_6 ist ein bereits fest gewähltes und somit bekanntes Element des Grundkörpers k. Wieder können wir eine der beiden Variablen willkürlich festlegen, etwa x_7 auf ein $\lambda_7 \in k$ setzen, und erhalten dann

$$x_7 = \lambda_7, \quad x_2 = -\frac{3}{4}\lambda_6 + \frac{7}{4}\lambda_7 - \frac{1}{2}\,.$$

In der ersten Gleichung schließlich müssen wir alle bereits berechneten oder festgelegten Unbekannten einsetzen und erhalten dann die Gleichung

$$x_1 - 3x_4 + x_5 + \frac{31}{20}\lambda_6 - \frac{7}{4}\lambda_7 = -\frac{17}{10}\,.$$

Hier können wir noch $x_4 = \lambda_4$ und $x_5 = \lambda_5$ beliebig festlegen und erhalten dann für die letzte noch verbleibende Variable

$$x_1 = -\frac{17}{10} + 3\lambda_4 - \lambda_5 - \frac{31}{20}\lambda_6 + \frac{7}{4}\lambda_7\,.$$

Insgesamt hängt die Lösung dieses Gleichungssystems also ab von vier willkürlich wählbaren Körperelementen $\lambda_4, \lambda_5, \lambda_6, \lambda_7 \in k$ oder, wie wir auch sagen werden, von vier *Parametern.*

Im ersten Beispiel brauchten wir von den beiden Operationen des GAUSS-Algorithmus nur den Eliminationsschritt, im zweiten nur den

Vertauschungsschritt. Als nächstes betrachten wir ein Beispiel, in dem beide notwendig sind, das lineare Gleichungssystem

$$\begin{array}{rcrcrcrcrcrcr}
 & & x_2 & & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10\\
2x_1 & + & 2x_2 & & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12\\
6x_1 & - & 3x_2 & + & x_3 & + & 5x_4 & & & - & 3x_6 & = & 15\\
4x_1 & - & 7x_2 & & & - & 3x_4 & & & + & x_6 & = & 0\\
2x_1 & - & 7x_2 & + & x_3 & + & 13x_4 & + & 12x_5 & - & 7x_6 & = & -8
\end{array}$$

Hier kommt x_1 ausgerechnet in der ersten Gleichung nicht vor, dafür aber in allen folgenden; wir müssen also die erste Gleichung mit einer der anderen vertauschen, etwa der zweiten:

$$\begin{array}{rcrcrcrcrcrcr}
2x_1 & + & 2x_2 & & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12\\
 & & x_2 & & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10\\
6x_1 & - & 3x_2 & + & x_3 & + & 5x_4 & & & - & 3x_6 & = & 15\\
4x_1 & - & 7x_2 & & & - & 3x_4 & & & + & x_6 & = & 0\\
2x_1 & - & 7x_2 & + & x_3 & + & 13x_4 & + & 12x_5 & - & 7x_6 & = & -8
\end{array}$$

Nun kommt x_1 in der zweiten Gleichung nicht mehr vor; aus der dritten, vierten *bzw.* fünften Gleichung kann x_1 eliminiert werden, indem man dreimal, zweimal *bzw.* einmal die erste Gleichung subtrahiert:

$$\begin{array}{rcrcrcrcrcrcr}
2x_1 & + & 2x_2 & & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12\\
 & & x_2 & & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10\\
 & & -9x_2 & + & x_3 & + & 17x_4 & + & 18x_5 & - & 9x_6 & = & -21\\
 & & -3x_2 & & & + & 5x_4 & + & 12x_5 & - & 3x_6 & = & -24\\
 & & -9x_2 & + & x_3 & + & 17x_4 & + & 18x_5 & - & 9x_6 & = & -20
\end{array}$$

Als nächstes muß x_2 aus der dritten bis fünften Gleichung eliminiert werden; da x_2 in der zweiten Gleichung mit Koeffizient eins vorkommt, müssen wir einfach jenes Vielfache der zweiten Gleichung subtrahieren, das dem jeweiligen x_2-Koeffizienten entspricht. Da hier alle diese Koeffizienten negativ sind, bedeutet dies, daß wir dasjenige Vielfache der zweiten Gleichung *addieren,* das dem Betrag des Koeffizienten entspricht:

$$\begin{array}{rcrcrcrcrcrcr}
2x_1 & + & 2x_2 & & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12\\
 & & x_2 & & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10\\
 & & & & x_3 & + & 62x_4 & & & & & = & 69\\
 & & & & & & 20x_4 & + & 6x_5 & & & = & 6\\
 & & & & x_3 & + & 62x_4 & & & & & = & 70\,.
\end{array}$$

Eigentlich sollte man hier schon sehen, was los ist, aber wir rechnen zur Veranschaulichung des GAUSS-Algorithmus trotzdem stur weiter nach Schema F: Als nächstes muß x_3 aus der vierten und fünften Gleichung eliminiert werden; in der vierten kommt es ohnehin nicht vor, bei der fünften muß die dritte subtrahiert werden:

$$\begin{array}{rcrcrcrcrcrcrcr}
2x_1 & + & 2x_2 & & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12 \\
& & x_2 & & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10 \\
& & & & x_3 & + & 62x_4 & & & & & = & 69 \\
& & & & & & 20x_4 & + & 6x_5 & & & = & 6 \\
& & & & & & & & & & 0 & = & 1\,.
\end{array}$$

Damit wird nun endgültig klar, daß jede Lösung $(x_1, \dots, x_6)$ des gegebenen Gleichungssystems insbesondere auch die Gleichung $0 = 1$ erfüllen muß, d.h. die Lösungsmenge ist leer.

Nachdem wir soviel Arbeit in dieses Beispiel investiert haben, sollten wir zumindest einen Teil der Rechnungen recyclen zu einem Beispiel, in dem es Lösungen gibt. Dazu muß nur die letzte der fünf ursprünglichen Gleichungen auf der rechten Seite leicht abgeändert werden: Wir betrachten nun das System

$$\begin{array}{rcrcrcrcrcrcr}
& & x_2 & & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10 \\
2x_1 & + & 2x_2 & & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12 \\
6x_1 & - & 3x_2 & + & x_3 & + & 5x_4 & & & - & 3x_6 & = & 15 \\
4x_1 & - & 7x_2 & & & - & 3x_4 & & & + & x_6 & = & 0 \\
2x_1 & - & 7x_2 & + & x_3 & + & 13x_4 & + & 12x_5 & - & 7x_6 & = & -9\,.
\end{array}$$

Hierauf lassen sich genau dieselben Umformungen anwenden wie oben, anstelle des Systems mit fünfter Gleichung $0 = 1$ führen diese nun aber auf

$$\begin{array}{rcrcrcrcrcrcr}
2x_1 & + & 2x_2 & & & - & 4x_4 & - & 6x_5 & + & 2x_6 & = & 12 \\
& & x_2 & & & + & 5x_4 & - & 2x_5 & + & x_6 & = & 10 \\
& & & & x_3 & + & 62x_4 & & & & & = & 69 \\
& & & & & & 20x_4 & + & 6x_5 & & & = & 6 \\
& & & & & & & & & & 0 & = & 0\,.
\end{array}$$

Diese letzte Gleichung ist natürlich für jedes Tupel $(x_1, \dots, x_6)$ erfüllt. Die vorletzte gibt eine Beziehung zwischen x_4 und x_5, wir können also

$$x_5 = \lambda_5 \in k$$

beliebig wählen und erhalten dann

$$x_4 = -\frac{3}{10}\lambda_5 + \frac{3}{10}\,.$$

Wenn wir dies in die dritte Gleichung einsetzen, bleibt dort nur noch x_3 als Variable stehen, und aus

$$x_3 - \frac{93}{5}\lambda_5 + \frac{93}{5} = 69$$

lesen wir sofort ab, daß

$$x_3 = \frac{93}{5}\lambda_5 + \frac{252}{5}$$

ist. Damit gehen wir in die zweite Gleichung:

$$x_2 - \frac{7}{5}\lambda_5 + x_6 + \frac{3}{2} = 10\,.$$

Hier können wir wieder eine der beiden noch verbliebenen Variablen auf einen beliebigen Wert setzen, etwa

$$x_6 = \lambda_6 \in k\,.$$

Dann wird

$$x_2 = \frac{7}{2}\lambda_5 - \lambda_6 + \frac{17}{2}\,,$$

was wir schließlich zusammen mit all den anderen bereits berechneten x_i in die erste Gleichung einsetzen können:

$$2x_1 + \frac{11}{5}\lambda_5 + \frac{79}{5} = 12$$

hat die Lösung

$$x_1 = -\frac{11}{10}\lambda_5 - \frac{19}{10}\,.$$

Damit hängt also die allgemeine Lösung dieses linearen Gleichungssystems von zwei frei wählbaren Parametern λ_5 und λ_6 ab.

Als letztes Beispiel schließlich wollen wir ein System betrachten, daß von zwar festen, aber nicht numerisch gegebenen Parametern abhängt:

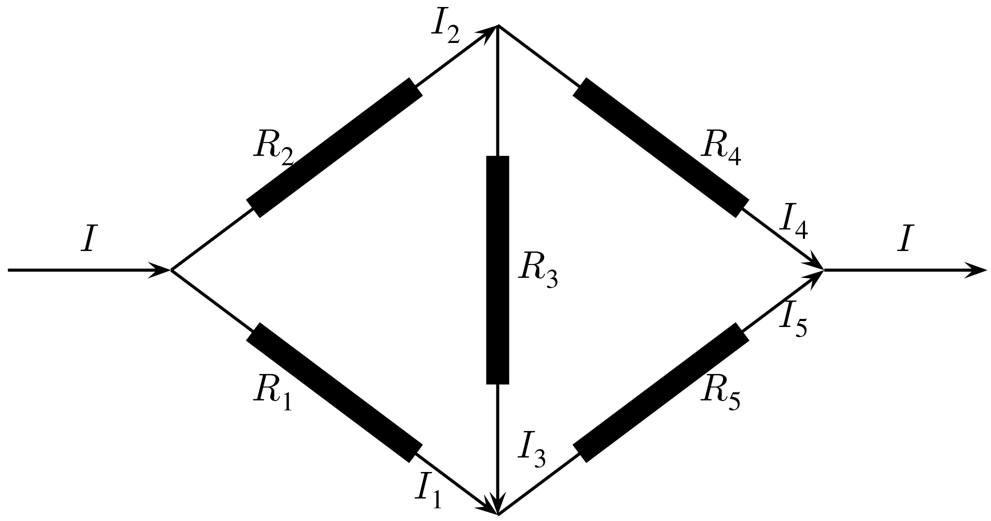


Abb. 36: Ein Gleichstromnetzwerk

Wir betrachten das Gleichstromnetzwerk aus Abbildung 36 mit bekannten Widerständen $R_1, \ldots, R_5$ und bekanntem Eingangsstrom I; gesucht sind die Ströme $I_1, \ldots, I_5$.

Nach den KIRCHHOFFschen Gesetzen müssen sich zunächst an allen Knoten die eingehenden und die ausgehenden Ströme zu Null ergänzen, d.h. wir erhalten die Gleichungen

$$I_1 + I_2 = I = I_4 + I_5$$

für Anfang und Ende sowie

$$I_2 = I_3 + I_4 \quad \text{und} \quad I_1 + I_3 = I_5$$

für oben und unten. Außerdem müssen sich die Spannungen in den beiden Dreiecken zu Null summieren; nach dem OHMschen Gesetz führt dies auf die beiden Gleichungen

$$R_2I_2 + R_3I_3 - R_1I_1 = 0 \quad \text{und} \quad R_3I_3 + R_5I_5 - R_4I_4 = 0\,.$$

Ordnen wir dies nach den Strömen I_ν, erhalten wir also das lineare Gleichungssystem

$$\begin{array}{rcrcrcrcrcl}
I_1 & + & I_2 & & & & & & & = & I \\
 & & & & & & I_4 & + & I_5 & = & I \\
 & & I_2 & - & I_3 & - & I_4 & & & = & 0 \\
I_1 & & & + & I_3 & & & - & I_5 & = & 0 \\
-R_1I_1 & + & R_2I_2 & + & R_3I_3 & & & & & = & 0 \\
 & & & & -R_3I_3 & + & R_4I_4 & - & R_5I_5 & = & 0\,.
\end{array}$$

Die Elimination von I_1 aus allen Gleichungen ab der zweiten ist einfach: Wir müssen die erste Gleichung von der vierten subtrahieren und ihr R_1-faches zur fünften addieren; das neue System ist

$$\begin{array}{rcrcrcrcrcr}
I_1 & + & I_2 & & & & & & & = & I \\
 & & & & & & I_4 & + & I_5 & = & I \\
 & & I_2 & - & I_3 & - & I_4 & & & = & 0 \\
 & & -I_2 & + & I_3 & & & - & I_5 & = & -I \\
 & & (R_1+R_2)I_2 & + & R_3I_3 & & & & & = & R_1 I \\
 & & & & -R_3I_3 & + & R_4I_4 & - & R_5I_5 & = & 0\,,
\end{array}$$

Da I_2 in der zweiten Gleichung nicht vorkommt, vertauschen wir diese mit der dritten; danach können wir letztere zur vierten addieren und ihr $(R_1 + R_2)$-faches von der fünften subtrahieren, um I_2 aus allen weiteren Gleichungen zu eliminieren.

Damit weiterhin jede Gleichung in eine Zeile paßt, setzen wir zur Abkürzung

$$R_{12} \underset{\text{def}}{=} R_1 + R_2 \quad \text{und} \quad R_{123} \underset{\text{def}}{=} R_1 + R_2 + R_3$$

und erhalten

$$\begin{array}{rcrcrcrcrcr}
I_1 & + & I_2 & & & & & & & = & I \\
 & & I_2 & - & I_3 & - & I_4 & & & = & 0 \\
 & & & & & & I_4 & + & I_5 & = & I \\
 & & & & & & -I_4 & + & -I_5 & = & -I \\
 & & & & R_{123}I_3 & + & R_{12}I_4 & & & = & R_1 I \\
 & & & & -R_3I_3 & + & R_4I_4 & - & R_5I_5 & = & 0\,.
\end{array}$$

Zur Elimination von I_3 bietet sich an, die dritte Gleichung, in der I_3 nicht vorkommt, mit der sechsten zu vertauschen (diese hat einen einfacheren I_3-Koeffizienten als die fünfte), und dann das R_{123}/R_3-fache dieser Gleichung zur fünften zu addieren. Dazu müssen wir uns allerdings zunächst überlegen, ob das überhaupt zulässig ist: Falls $R_3 = 0$ ist, dürfen wir natürlich nicht durch R_3 dividieren; wenn dieser Fall nicht ausgeschlossen werden kann, muß er also ab hier getrennt behandelt werden.

Im vorliegenden Beispiel wollen aber davon ausgehen, daß alle fünf Widerstände tatsächlich vorhanden und somit durch positive reelle Zahlen gebeben sind. Dann können wir dividieren und das System wird zu

$$\begin{array}{ccccccccccc}
I_1 & + & I_2 & & & & & & & = & I \\
& & I_2 & - & I_3 & - & I_4 & & & = & 0 \\
& & & & -R_3I_3 & + & R_4I_4 & - & R_5I_5 & = & 0 \\
& & & & & & -I_4 & + & -I_5 & = & -I \\
& & & & & & \alpha I_4 & + & \beta I_5 & = & R_1I \\
& & & & & & I_4 & + & I_5 & = & I
\end{array}$$

mit

$$\alpha = \frac{R_4R_{123}}{R_3} \quad \text{und} \quad \beta = \frac{-R_5R_{123}}{R_3}\,.$$

Schließlich muß noch I_4 aus den letzten beiden Gleichungen eliminiert werden; wir addieren also die vierte Gleichung unverändert zur letzten und ihr α-faches zur vorletzten:

$$\begin{array}{ccccccccccc}
I_1 & + & I_2 & & & & & & & = & I \\
& & I_2 & - & I_3 & - & I_4 & & & = & 0 \\
& & & & -R_3I_3 & + & R_4I_4 & - & R_5I_5 & = & 0 \\
& & & & & & -I_4 & + & -I_5 & = & -I \\
& & & & & & & & (\beta-\alpha)I_5 & = & (R_1-\alpha)I \\
& & & & & & & & 0 & = & 0
\end{array}$$

Damit können wir nacheinander

$$\begin{aligned}
I_5 &= \frac{R_1-\alpha}{\beta-\alpha}\cdot I \\
I_4 &= I - I_5 \\
I_3 &= \frac{R_4I_4 - R_5I_5}{R_3} \\
I_2 &= I_3 + I_4 \qquad \text{und} \\
I_1 &= I - I_2
\end{aligned}$$

bestimmen, wobei sich der Leser noch überlegen sollte, warum die Division durch $\beta - \alpha$ unproblematisch ist.

Zum Berechnen der Ströme in konkreten Beispielen reicht diese Lösungsformel aus; ist man allerdings an einem symbolischen Ausdruck interessiert, muß man die Definitionen von $\alpha, \beta, R_{12}, R_{123}$ einsetzen und nacheinander alle rechte Seiten auf Ausdrücke nur in I und den R_i reduzieren. Dies ist eine *im Prinzip* einfache Übungsaufgabe in Bruchrechnung, die man allerdings im vorliegenden Fall besser einem Computeralgebrasystem überläßt. Als Ergebnis erhält man die expliziten Formeln

$$I_1 = \frac{\big(R_2R_5 + R_2R_3 + R_2R_4 + R_3R_4\big) \cdot I}{R_1R_5 + R_2R_5 + R_3R_5 + R_1R_3 + R_2R_3 + R_1R_4 + R_2R_4 + R_3R_4}$$
$$I_2 = \frac{\big(R_1R_4 + R_1R_5 + R_3R_5 + R_1R_3\big) \cdot I}{R_1R_5 + R_2R_5 + R_3R_5 + R_1R_3 + R_2R_3 + R_1R_4 + R_2R_4 + R_3R_4}$$
$$I_3 = \frac{\big(R_1R_4 - R_2R_5\big) \cdot I}{R_1R_5 + R_2R_5 + R_3R_5 + R_1R_3 + R_2R_3 + R_1R_4 + R_2R_4 + R_3R_4}$$
$$I_4 = \frac{\big(R_2R_3 + R_1R_4 + R_2R_4 + R_3R_4\big) \cdot I}{R_1R_5 + R_2R_5 + R_3R_5 + R_1R_3 + R_2R_3 + R_1R_4 + R_2R_4 + R_3R_4}$$
$$I_5 = \frac{\big(R_2R_3 + R_1R_4 + R_2R_4 + R_3R_4\big) \cdot I}{R_1R_5 + R_2R_5 + R_3R_5 + R_1R_3 + R_2R_3 + R_1R_4 + R_2R_4 + R_3R_4},$$

die für die meisten *konkreten* Anwendungen erheblich weniger nützlich sind als die obige Form des Ergebnisses.

c) Lösbarkeitskriterien für lineare Gleichungssysteme

Nach diesen Beispielen ist es an der Zeit, wieder zu den theoretischen Grundlagen zurückzukehren. Sei also wieder

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m &= b_2 \\ \vdots \qquad\qquad &= \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m &= b_n\,, \end{aligned}$$

ein allgemeines lineares Gleichungssystem. Wenn wir die a_{ij} zu einer Matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix}$$

zusammenfassen und die Unbekannten und rechten Seiten zu Vektoren

$$\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \quad \text{und} \quad \vec{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} ,$$

läßt sich das Gleichungssystem auch kurz als $A\vec{x} = \vec{b}$ schreiben. Wir bezeichen A als die *Matrix des Gleichungssystems* und

$$\left(A \mid \vec{b}\,\right) \underset{\text{def}}{=} \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} & b_1 \\ a_{21} & a_{22} & \dots & a_{2m} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} & b_m \end{pmatrix}$$

als die *erweiterte Matrix.*

Die verschiedenen Schritte des GAUSS-Algorithmus wirken sich auf die Matrix wie auch auf die erweiterte Matrix so aus, daß entweder zwei Zeilen miteinander vertauscht werden, oder aber ein Vielfaches einer Zeile zu einer anderen Zeile addiert wird – Addition bedeutet dabei die gewöhnliche komponentenweise Addition, d.h. die Addition von Zeilenvektoren.

Nachdem wir diese Operationen solange wie möglich ausgeführt haben, bekommt die erweiterte Matrix die Form

$$\left(\begin{array}{cccccccccccc}
[0] & \bullet & [\star] & \star & [\star] & \star & \dots & [\star] & \star & [\star] & \star \\
[0] & 0 & [0] & \bullet & [\star] & \star & \dots & [\star] & \star & [\star] & \star \\
[0] & 0 & [0] & 0 & [0] & \bullet & \dots & [\star] & \star & [\star] & \star \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
[0] & 0 & [0] & 0 & [0] & 0 & \dots & [0] & \bullet & [\star] & \star \\
\hline
[0] & 0 & [0] & 0 & [0] & 0 & \dots & [0] & 0 & [0] & \bullet \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
[0] & 0 & [0] & 0 & [0] & 0 & \dots & [0] & 0 & [0] & \bullet \\
\hline
[0] & 0 & [0] & 0 & [0] & 0 & \dots & [0] & 0 & [0] & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
[0] & 0 & [0] & 0 & [0] & 0 & \dots & [0] & 0 & [0] & 0
\end{array}\right) ,$$

wobei $\bullet$ für ein von Null verschiedenes und $\star$ für ein beliebiges Element des Körpers k steht; [0] und $[\star]$ stehen für keine, eine oder mehrere Nullen *bzw.* beliebige Elemente, deren Anzahl in untereinanderstehenden Termen jeweils die gleiche sein soll.

Die Zeilen unterhalb der ersten eingezeichneten Linie, in denen alle Einträge der Matrix des Gleichungssystems verschwinden, müssen natürlich nicht wirklich auftreten, genauso wenig die Zeilen unterhalb der zweiten Linie, wo sogar alle Einträge der erweiterten Matrix verschwinden.

Falls zwischen den beiden eingezeichneten Linien wirklich eine oder mehrere Zeilen stehen, ist das Gleichungssystem *unlösbar,* denn dann gibt es ja in der Endform des Gleichungssystems Gleichungen, in denen links alle Koeffizienten Null sind, während rechts eine von Null verschiedene Zahl steht. Indem man gegebenenfalls ein Vielfaches der ersten solchen Gleichung von den etwa vorhandenen weiteren Gleichungen subtrahiert, kann man erreichen, daß alle weiteren Gleichungen zu $0 = 0$ werden, d.h. wir können erreichen, daß zwischen den beiden Linien *höchstens eine* Zeile steht, und das lineare Gleichungssystem ist genau dann lösbar, wenn keine dort steht.

Um diese Situation durch ein numerisches Kriterium zu beschreiben, definieren wir die Anzahl der Zeilen oberhalb der ersten Linie als *Rang der Matrix A* und die Anzahl der Zeilen oberhalb der zweiten Linie (nachdem man dafür Sorge getragen hat, daß höchstens eine Zeile zwischen den beiden Linien steht) als *Rang der erweiterten Matrix.* Der Rang ist also jeweils die Anzahl von Zeilen, die auch nach Anwendung des GAUSS-Algorithmus einschließlich der gerade diskutierten weiteren Operationen noch von Null verschiedene Elemente enthalten, und der Rang des erweiterten Matrix ist entweder gleich dem Rang der Matrix oder um eins größer.

Genauer gilt:

Satz: *a)* Das lineare Gleichungssystem $A\vec{x} = \vec{b}$ hat genau dann (mindestens) eine Lösung, wenn der Rang der Matrix A gleich dem Rang der erweiterten Matrix $\left(A \mid \vec{b}\,\right)$ ist.

b) Falls $A \in k^{n\times m}$ und $n = m = \text{Rang}\, A$, hat das Gleichungssystem genau eine Lösung $(x_1, \ldots, x_m)$.

Beweis: Der Rang der erweiterten Matrix kann natürlich nicht kleiner sein als der der Matrix selbst, und wenn er größer wird, bekommen wir Gleichungen der Form $0 = a$ mit $a \neq 0$, aus denen die Unlösbarkeit des Gleichungssystems folgt; falls es also Lösungen gibt, müssen die beiden Ränge gleich sein.

Falls sie gleich sind, gibt es keine Zeilen zwischen den beiden Strichen, und unter dem zweiten Strich stehen nur Nullen, die trivialerweise erfüllten Gleichungen $0 = 0$ entsprechen; wir können uns also auf die Zeilen oberhalb des ersten Strichs beschränken. Jede dieser Zeilen beginnt nach einer Reihe von Nullen mit einem $\bullet$, entspricht also einer Gleichung der Form

$$c_i x_i + c_{i+1} x_{i+1} + \ldots + c_m x_m = d$$

mit $c_i \neq 0$. Wenn wir von unten anfangen und uns dann nach oben hocharbeiten, sind die x_j mit $j > i$ entweder bereits aus weiter unten stehenden Gleichungen bekannt, oder aber sie treten zum ersten Mal auf und können beliebig gewählt werden; nach dieser Wahl ist

$$x_i = \frac{d - c_{i+1} x_{i+1} - \ldots - c_m x_m}{c_i}$$

eindeutig bestimmt. Insbesondere gibt es also Lösungen, und in der Situation von Teil *b),* wo die Matrix der Endgestalt die Form

$$\begin{pmatrix} \bullet & \star & \ldots & \star & \star \\ 0 & \bullet & \ldots & \star & \star \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & \bullet & \star \end{pmatrix}$$

hat, ist jedes x_i eindeutig bestimmt.

■

d) Ausblick: Numerische Lösung linearer Gleichungssysteme

Mancher Leser mag sich in die Mittelstufe zurückversetzt gefühlt haben, als bei den Beispielen dieses Paragraphen plötzlich Brüche auftauchten anstelle der Dezimalzahlen, mit denen der „echte Praktiker“ seinen Taschenrechner füttert.

Tatsächlich sollte der GAUSS-Algorithmus, so wie er hier vorgestellt wurde, nur für Körper benutzt werden, in denen man *exakt* rechnen kann; beim Rechnen mit reellen Zahlen, vor allem wenn es per Computer oder Taschenrechner geschieht, muß man sich meist mit Näherungslösungen begnügen, und auch kleine Rundungsfehler können schon große Auswirkungen auf das Ergebnis haben.

Hier sei nur ein noch ziemliches harmloses Beispiel eines linearen Gleichungssystems mit zwei Gleichungen in zwei Unbekannten betrachtet, nämlich

$$\begin{array}{rcrcl} 0.001x & + & 0.00001y & = & 1 \\ 0.00002x & - & 0.0005y & = & 1\,. \end{array}$$

Für seine Lösung ist

$$x \approx 1019{,}6 \quad \text{und} \quad y \approx -1959{,}2\,.$$

Stört man aber die rechte Seite nur im 0,01 in der zweiten Gleichung, ersetzt man dort also die 1 durch 1,01, wird die Lösung zu

$$x \approx 1019{,}8 \quad \text{und} \quad y \approx -1979{,}2\,,$$

und ersetzt man sie gar durch 1,1, erhält man

$$x \approx 1021{,}6 \quad \text{und} \quad y \approx -2159{,}1\,.$$

Schon kleine Störungen der Konstanten, wie sie durch allfällige Rundungsfehler immer wieder auftreten, können also das Ergebnis stark verfälschen.

Beim Lösen von linearen Gleichungssystemen muß man daher unbedingt darauf achten, daß Rundungsfehler auf das absolut notwendige Minimum beschränkt werden, und selbst dann muß man noch vorsichtig sein.

Die notwendigen Strategien zur Minimierung der Effekte von Rundungsfehler werden in der Vorlesung *Numerik I* behandelt; da aber bereits in der *Praktischen Informatik II* etwas auf rundungsfehlerbewußtes Rechnen eingegangen wurde, seien hier im Vorgriff auf die Numerik wenigstens zwei Faustregeln erwähnt:

1.) Die verschiedenen Gleichungen des Systems sollten Koeffizienten in ähnlichen Größenordnungen enthalten. Da sich nichts an der Lösungsmenge ändert, wenn man eine der Gleichungen mit einer von Null verschiedenen Konstanten multipliziert, kann man dies etwa dadurch erreichen, daß man alle Zeilenvektoren der Matrix des Gleichungssystems auf dieselbe Länge (z.B. Eins) normiert.
2.) Falls man ein Vielfaches einer Gleichung zu einer anderen addiert, sollte der Koeffizient, mit dem die erste Gleichung multipliziert wird, möglichst klein sein, damit die zweite Gleichung nicht zu stark gestört wird. Dies läßt sich etwa dadurch erreichen, daß bei der Elimination einer Variablen genau die Gleichung unverändert läßt, in der diese Variable mit dem betragsgrößten Koeffizienten vorkommt.

Für alles weitere sei auf die *Numerik I* verwiesen.

§4: Die inverse Matrix

Nachdem wir aus §2 wissen, wann und wie sich Matrizen miteinander multiplizieren lassen, liegt es nahe, auch danach zu fragen, wann und wie sie sich durcheinander *dividieren* lassen. Da das Matrizenprodukt nicht kommutativ ist, hat die Frage in dieser Form natürlich nur wenig Sinn. Wir können sie aber sinnvoll machen, indem wir etwa fragen, ob es zu einer gegebenen Matrix A eine Matrix X gibt, so daß $AX = E$ die Einheitsmatrix ist.

Die Techniken, die wir im folgenden kennenlernen werden, gestatten eine Antwort auf diese Frage für beliebige $n \times m$-Matrizen A (mit $m \times n$-Matrizen X und der $n \times n$-Einheitsmatrix), aber wir wollen uns hier mit dem für die Anwendungen wichtigsten Fall $n = m$ begnügen.

a) Matrixgleichungen und lineare Gleichungssysteme

Wir gehen also aus von einer gegebenen $n \times n$-Matrix $A = (a_{ij}) \in k^{n\times n}$ und einer unbekannten Matrix $X = (x_{ij}) \in k^{n\times n}$; wir versuchen die Matrixgleichung $AX = E$ zu lösen.

Sind

$$\vec{x}^{\,(j)} = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \quad \text{und} \quad \vec{e}^{\,(j)} = \begin{pmatrix} \delta_{1j} \\ \vdots \\ \delta_{nj} \end{pmatrix}$$

mit

$$\delta_{ij} = \begin{cases} 1 & \text{für } i = j \\ 0 & \text{für } i \neq j \end{cases}$$

die Spaltenvektoren von X und E, so ist die Matrixgleichung $AX = E$ äquivalent zu den n linearen Gleichungssystemen

$$A\vec{x}^{\,(j)} = \vec{e}^{\,(j)} \quad \text{für} \quad j = 1, \ldots, n\,,$$

die allesamt dieselbe Matrix A haben. Falls diese Gleichungen alle lösbar sind, gibt es also eine Matrix X mit $AX = E$.

Entsprechend lassen sich auch allgemeinere Matrizengleichungen der Form $AX = B$ behandeln und gegebenenfalls lösen.

b) Beispiel einer Matrixinversion

Betrachten wir als Beispiel die Matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 7 \\ 8 & 9 & 12 \end{pmatrix}.$$

Wir müssen das lineare Gleichungssystem $A\vec{x} = \vec{b}$ lösen für die drei rechten Seiten

$$\vec{b} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Dazu wenden wir den GAUSS-Algorithmus simultan auf alle drei Gleichungssysteme an, indem wir die drei rechten Seiten nebeneinander schreiben, d.h. wir gehen aus von einer um alle drei rechten Seiten erweiterten Matrix

$$\begin{pmatrix} 1 & 2 & 3 & & 1 & 0 & 0 \\ 4 & 5 & 7 & & 0 & 1 & 0 \\ 8 & 9 & 12 & & 0 & 0 & 1 \end{pmatrix}$$

und wenden darauf die üblichen Eliminationsschritte an:

Als erstes müssen wir in der ersten Spalte alle Koeffizienten bis auf einen zu Null machen; dazu subtrahieren wir das vier- *bzw.* achtfache der ersten Zeile von der zweiten *bzw.* dritten und erhalten

$$\begin{pmatrix} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -5 & -4 & 1 & 0 \\ 0 & -7 & -12 & -8 & 0 & 1 \end{pmatrix}.$$

Als nächstes sollte der zweite Eintrag der letzten Spalte eliminiert werden; dazu können wir $7/3$ der mittleren Zeile von der letzten subtrahieren oder aber, um Nenner zu vermeiden, sieben mal die zweite Gleichung von dreimal der dritten subtrahieren; dies führt auf

$$\begin{pmatrix} 1 & 2 & 3 & 1 & 0 & 0 \\ 0 & -3 & -5 & -4 & 1 & 0 \\ 0 & 0 & -1 & 4 & -7 & 3 \end{pmatrix}.$$

In dieser Gestalt lassen sich die drei Gleichungssysteme nun leicht durch Rücksubstitution lösen; wir wollen aber stattdessen lieber weiter mit Zeilenoperationen arbeiten: Subtrahieren wir fünfmal die dritte Zeile von der zweiten und addieren sie dreimal zur ersten, erhalten wir

$$\begin{pmatrix} 1 & 2 & 0 & 13 & -21 & 9 \\ 0 & -3 & 0 & -24 & 36 & -15 \\ 0 & 0 & -1 & 4 & -7 & 3 \end{pmatrix}.$$

Diese Matrix wird deutlich übersichtlicher, wenn wir die zweite Zeile durch -3 dividieren und die dritte durch -1 und schließlich noch zweimal die neue zweite Zeile von der ersten subtrahieren; sie bekommt dann die Form

$$\begin{pmatrix} 1 & 0 & 0 & -3 & 3 & -1 \\ 0 & 1 & 0 & 8 & -12 & 5 \\ 0 & 0 & 1 & -4 & 7 & -3 \end{pmatrix},$$

aus der man die Lösungsvektoren sofort ablesen kann: Da links die Einheitsmatrix steht, sind die Lösungen der Gleichungssysteme gerade die jeweiligen rechten Seiten, d.h. die rechte Matrix

$$X = \begin{pmatrix} -3 & 3 & -1 \\ 8 & -12 & 5 \\ -4 & 7 & -3 \end{pmatrix}$$

ist die Lösung von $AX = E$, wovon man sich zur Vorsicht vielleicht noch einmal durch Nachrechnen überzeugen sollte.

c) Bedingungen für die Invertierbarkeit einer Matrix

Das Verfahren aus dem obigen Beispiel läßt sich natürlich für beliebige Matrizen anwenden; nicht klar ist allerdings, ob gegebenenfalls wie viele Lösungen es gibt. Diese Frage wollen wir in diesem Abschnitt untersuchen.

Aus dem vorigen Paragraphen wissen wir, daß ein lineares Gleichungssystem genau dann lösbar ist, wenn der Rang der erweiterten Matrix gleich dem der Matrix ist. Da weder der Rang einer $n \times n$-Matrix noch der einer $n \times (n + 1)$-Matrix größer als n sein können, ist diese Bedingung insbesondere dann erfüllt, wenn der Rang von $A \in k^{n\times n}$ gleich n ist; alsdann ist X nach dem obigen Satz sogar eindeutig bestimmt.

Bevor wir uns überlegen, wie man unter dieser Voraussetzung die Matrix X am besten berechnet, wollen wir uns zunächst davon überzeugen, daß die gerade aufgestellte *hinreichende* Bedingung für die Existenz von X tatsächlich sogar *notwendig* ist. Dazu verwenden wir wieder den schon bei der Matrizenmultiplikation als nützlich erkannten Zusammenhang zwischen Matrizen und linearen Abbildungen: Die Matrix A entspricht einer linearen Abbildung $\varphi\colon k^n \to k^n$; falls es dazu eine lineare Abbildung $\psi\colon k^n \to k^n$ gibt mit

$$\varphi \circ \psi = \mathrm{id}\colon \begin{cases} k^n & \to & k^n \\ \vec{x} & \mapsto & \vec{x} \end{cases},$$

ist $AX = E$ für die Abbildungsmatrix X von ψ.

Ist aber $\varphi \circ \psi = \mathrm{id}$, so ist ϕ surjektiv, da $\psi(\vec{w})$ ein Urbild von $\vec{w}$ ist. Somit hat das lineare Gleichungssystem $A\vec{x} = \vec{b}$ für jede rechte Seite mindestens eine Lösung, d.h. unabhängig von der rechten Seite ist der Rang der erweiterten Matrix gleich dem Rang der Matrix A. Damit muß der Rang von A aber gleich n sein, denn hätte A nach Anwendung des GAUSS-Algorithmus Nullzeilen, könnte man dort auf der rechten Seite einen Wert ungleich Null hinschreiben und alle Transformationen rückgängig machen: Dies würde auf ein unlösbares Gleichungssystem der Form $A\vec{x} = \vec{b}$ führen.

Satz: *a)* Zu einer $n \times n$-Matrix A gibt es genau dann eine $n \times n$-Matrix X mit $AX = E$, wenn Rang $A = n$ ist.
b) Unter diesen Voraussetzungen ist auch $XA = E$.

Beweis: a) haben wir gerade gezeigt, bleibt also noch *b)*.

Aus $AX = E$ folgt, daß $(AX)A = EA = A$ ist; da die Matrizenmultiplikation assoziativ ist, folgt $A(XA) = A$. Außerdem ist natürlich $AE = A$; falls wir zeigen können, daß die Matrizengleichung $AY = A$ nur eine einzige Lösung $Y \in k^{n\times n}$ hat, muß also $XA = E$ sein, wie behauptet.

Die Eindeutigkeit von Y folgt aber aus dem Satz am Ende von §3*c)*, denn wir können $AY = A$ wieder als n lineare Gleichungssysteme für die Spalten von Y betrachten, und jedes dieser Gleichungssysteme hat wegen Rang $A = n$ eine eindeutig bestimmte Lösung. ∎

Definition: Falls es zu $A \in k^{n\times n}$ eine Matrix $X \in k^{n\times n}$ gibt mit $AX = E$, nennen wir A *invertierbar* und bezeichnen X als die *inverse Matrix* A^{-1}.

Lemma: Sind $A, B \in k^{n\times n}$ invertierbar, so auch ihr Produkt, und $(AB)^{-1} = B^{-1}A^{-1}$.

Beweis: $(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AEA^{-1} = AA^{-1} = E$; also ist AB invertierbar und $B^{-1}A^{-1}$ ist die inverse Matrix. ∎

Man beachte, daß im allgemeinen $A^{-1}B^{-1}$ *nicht* invers zu AB ist: Für

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \quad \text{und} \quad AB = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix}$$

sind

$$A^{-1} = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}, \quad B^{-1} = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}$$

und damit

$$(AB)^{-1} = \begin{pmatrix} 1 & -2 \\ -2 & 5 \end{pmatrix} \neq \begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix} = A^{-1}B^{-1}\,.$$

Insbesondere unterscheidet sich auch

$$ABA^{-1}B^{-1} = \begin{pmatrix} 21 & -8 \\ 8 & -3 \end{pmatrix}$$

deutlich von der Einheitsmatrix.

d) Die LR-Zerlegung einer Matrix

Das oben angegebene Verfahren zur Berechnung der inversen Matrix durch simultanes Lösen mehrerer linearer Gleichungssysteme, die sich nur durch ihre rechten Seiten unterscheiden, läßt sich etwas modifizieren und führt dann auf eine Zerlegung, die sich nicht nur bei der Berechnung der inversen Matrix als nützlich erweist, sondern allgemein bei der Lösung von Gleichungssystemen; auch später, bei der Berechnung von Determinanten wird sie uns ein sehr effizientes Verfahren liefern.

Dazu betrachten wir den wesentlichen Schritt des GAUSS-Algorithmus, das Addieren von Vielfachen einer Gleichung zu einer anderen, als Matrizenmultiplikation: Ist

$$L_1 = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & c & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix}$$

eine Matrix, die in der Hauptdiagonalen lauter Einsen stehen hat und ansonsten als einzigen von Null verschiedenen Eintrag ein Element $c \in k$ an der Stelle (j, i) mit $i < j$, so ist für eine beliebige $n \times n$-Matrix $A = (a_{\mu\nu})$

$$L_1 A = A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{(j-1)1} & a_{(j-1)2} & \cdots & a_{(j-1)m} \\ ca_{i1} + a_{j1} & ca_{i2} + a_{j2} & \cdots & ca_{im} + a_{jm} \\ a_{(j+1)1} & a_{(j+1)2} & \cdots & a_{(j+1)m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}$$

gerade jene Matrix, die aus A entsteht, wenn man das c-fache der i-ten Zeile zur j-ten Zeile addiert.

Falls also bei der Anwendung des GAUSS-Algorithmus nie eine Vertauschung zweier Zeilen notwendig ist, verändert er die Matrix des Gleichungssystems, indem er sie sukzessive mit Matrizen obiger Form multipliziert, insgesamt gesehen also mit einem Produkt L solcher Matrizen.

Um etwas über Gestalt dieses Produkts zu erfahren, beachten wir, daß L_1 oberhalb der Diagonalen nur Nullen enthält; L_1 ist daher eine untere Dreiecksmatrix im Sinne der folgenden

Definition: Eine Matrix $A = (a_{ij}) \in k^{n\times n}$ heißt *untere Dreiecksmatrix,* falls $a_{ij} = 0$ für $i < j$; sie heißt *obere Dreiecksmatrix,* falls $a_{ij} = 0$ wann immer $i > j$.

Lemma: *a)* Das Produkt zweier unterer Dreiecksmatrizen ist wieder eine untere Dreiecksmatrix.
b) Das Produkt zweier oberer Dreiecksmatrizen ist wieder eine obere Dreiecksmatrix.

Beweis: a) $A = (a_{i\ell})$ und $B = (b_{\ell j})$ seien untere Dreiecksmatrizen, und $C = (c_{ij})$ sei ihr Produkt, d.h.

$$c_{ij} = \sum_{\ell=1}^{n} a_{i\ell} b_{\ell j} \ .$$

Für $i > j$ ist für jedes $\ell < i$ der Eintrag $a_{i\ell} = 0$, da A eine untere Dreiecksmatrix ist, und für $\ell \geq i > j$ ist $b_{\ell j} = 0$, da B eine untere Dreiecksmatrix ist. Also ist auch c_{ij} als Summe von lauter Nullen gleich Null.

b) geht genauso.

■

Damit wissen wir also, daß das Produkt L der Matrizen zu den Zeilenoperationen des GAUSS-Algorithmus eine untere Dreiecksmatrix ist; falls wir ohne Zeilenvertauschungen auskommen, ist die Anwendung des GAUSS-Algorithmus auf ein lineares Gleichungssystem für dessen

Matrix also gleichbedeutend mit der Multiplikation mit einer unteren Dreiecksmatrix L.

Da der GAUSS-Algorithmus aus der Matrix eines Gleichungssystems eine obere Dreiecksmatrix R macht, erhalten wir somit eine untere Dreiecksmatrix L und eine obere Dreiecksmatrix R derart, daß

$$LA = R$$

ist. Dabei ist die Matrix L invertierbar, da alle ihre Faktoren invertierbar sind: Die Addition des c-fachen der i-ten Zeile zur j-ten Zeile wird rückgängig gemacht durch die Subtraktion des c-fachen der i-ten Zeile von der j-ten Zeile, d.h.

$$\begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & c & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & -c & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix}.$$

Daher können wir auch schreiben

$$A = L^{-1}R\,.$$

Diese Zerlegung einer Matrix in ein Produkt zweier Dreiecksmatrizen heißt *LR-Zerlegung* der Matrix A. Da man beim GAUSS-Algorithmus im allgemeinen auch Zeilenvertauschungen benötigt, muß sie nicht immer existieren, sie existiert aber, nachdem man in A gegebenenfalls die Zeilen permutiert hat, und das ist für praktische Anwendungen gut genug.

Falls man die LR-Zerlegung einer Matrix kennt, folgt beispielsweise, daß A genau dann invertierbar ist, wenn R invertierbar ist; alsdann ist

$$A^{-1} = R^{-1}L\,.$$

Da die Inversion einer oberen Dreiecksmatrix sehr viel einfacher ist als die einer beliebigen Matrix, läßt sich auf diese Weise die inverse Matrix effizient berechnen.

Auch wenn man mehrere Gleichungssysteme

$$A\vec{x} = \vec{b}_i$$

mit verschiedenen rechten Seiten $\vec{b}_i$ lösen muß, ist die Kenntnis der LR-Zerlegung nützlich: Durch Multiplikation mit L werden die Gleichungssysteme zu

$$LA\vec{x} = R\vec{x} = L\vec{b}_i ,$$

und diese Gleichungssysteme lassen sich problemlos lösen.

§5: Vektorräume und lineare Abbildungen

Wir begannen dieses Kapitel mit Vektoren im $\mathbb{R}^3$, haben dann gesehen, daß sich zumindest ein Teil des dort Gelernten auf Vektoren beliebiger Dimension $n \in \mathbb{N}$ über einem beliebigen Körper k übertragen läßt und wollen jetzt schließlich als letzte Stufe der Verallgemeinerung sehen, daß sich auch viele andere mathematische Objekte wie reelle oder komplexe Funktionen, Matrizen *usw.* gelegentlich mit Erfolg als „Vektoren" interpretieren lassen. Das formale Hilfsmittel dazu ist der Begriff des abstrakten *Vektorraums.*

a) Definition des Vektorraums

Definition: k sei ein Körper. Eine Menge V heißt *Vektorraum* über k oder k-Vektorraum, wenn gilt:

I. V ist eine abelsche Gruppe, d.h.

1.) Für alle $\vec{u}, \vec{v} \in V$ gilt

$$\vec{u} + \vec{v} = \vec{v} + \vec{u} .$$

2.) Für alle $\vec{u}, \vec{v}.\vec{w} \in V$ gilt

$$(\vec{u} + \vec{v}) + \vec{w} = \vec{u} + (\vec{v} + \vec{w}) .$$

3.) Es gibt einen Vektor $\vec{0} \in V$, so daß für jeden Vektor $\vec{v} \in V$ gilt

$$\vec{v} + \vec{0} = \vec{0} + \vec{v} .$$

4.) Zu jedem Vektor $\vec{v} \in V$ gibt es einen Vektor $-\vec{v} \in V$, so daß

$$\vec{v} + -\vec{v} = -\vec{v} + \vec{v} = \vec{0} .$$

II. V hat eine k-Modulstruktur, d.h. es gibt eine Verknüpfung

$$\begin{cases} k \times V \to V \\ (\lambda, \vec{v}\,) \mapsto \lambda\vec{v} \end{cases},$$

so daß gilt:

1.) Für alle $\lambda, \mu \in k$ und alle $\vec{v} \in V$ gilt

$$(\lambda + \mu)\vec{v} = \lambda\vec{v} + \mu\vec{v}\,.$$

2. Für alle $\lambda \in k$ und alle $\vec{v}, \vec{w} \in V$ gilt

$$\lambda(\vec{v} + \vec{w}) = \lambda\vec{v} + \lambda\vec{w}\,.$$

3. Für alle $\vec{v} \in V$ ist

$$1\vec{v} = \vec{v}\,.$$

4. Für alle $\vec{v} \in V$ und alle $\lambda \in k$ ist

$$0\vec{v} = \vec{0} \quad \text{und} \quad \lambda\vec{0} = \vec{0}\,.$$

b) Erste Beispiele

Da Vektorräume Verallgemeinerungen von $\mathbb{R}^3$ und k^n sein sollen, sollte man sich als allererstes davon überzeugen, daß beides in der Tat Vektorräume im Sinne obiger Definition sind; aus den in §§1–2 zusammengestellten Rechenregeln folgt dies sofort.

Ebenfalls aus den in §2 zusammengestellten Rechenregeln folgt auch, daß die Menge $k^{n\times m}$ aller $n \times m$-Matrizen mit der Matrizenaddition und der Multiplikation einer Matrix mit einem Skalar ein k-Vektorraum ist.

Weiter sind die komplexen Zahlen ein $\mathbb{R}$-Vektorraum, wenn wir als Addition die gewöhnliche Addition komplexer Zahlen betrachten und als Multiplikation mit einem Skalar die Multiplikation einer reellen mit einer komplexen Zahl.

Genauso ist auch $\mathbb{R}$ ein $\mathbb{Q}$-Vektorraum: Vektoraddition ist die Addition reeller Zahlen und Multiplikation mit Skalaren die Multiplikation von rationalen und reellen Zahlen.

Schließlich ist auch für jede offene Teilmenge $U \subseteq \mathbb{R}$ und jede natürliche Zahl $n \in \mathbb{N}$ die Menge

$$\mathcal{C}^n(U,\mathbb{R}) = \bigl\{f\colon U \to \mathbb{R} \bigm| f \text{ ist } n\text{-mal stetig differenzierbar}\bigr\}$$

ein $\mathbb{R}$-Vektorraum mit der Addition

$$f+g\colon \begin{cases} U \to \mathbb{R} \\ x \mapsto f(x)+g(x) \end{cases}$$

und der Skalarmultiplikation

$$\lambda f\colon \begin{cases} U \to \mathbb{R} \\ x \mapsto \lambda f(x)\,. \end{cases}$$

Ganz entsprechend lassen sich auch die Mengen

$$\mathcal{C}^0(U,\mathbb{R}) = \bigl\{f\colon U \to \mathbb{R} \bigm| f \text{ ist stetig}\bigr\}$$

und

$$\mathcal{C}^\infty(U,\mathbb{R}) = \bigl\{f\colon U \to \mathbb{R} \bigm| f \text{ ist beliebig oft differenzierbar}\bigr\}$$

sowie

$$\mathcal{C}^\omega(U,\mathbb{R}) = \left\{f\colon U \to \mathbb{R} \;\middle|\; \begin{array}{l} f \text{ ist um jeden Punkt } x \in U \text{ durch} \\ \text{eine TAYLOR-Reihe darstellbar} \end{array}\right\}$$

zu $\mathbb{R}$-Vektorräumen machen.

Als trivialstes Beispiel überhaupt haben wir schließlich noch über jedem Körper k den Nullvektorraum, der nur aus dem Nullvektor $\vec{0}$ besteht.

c) Lineare Abhängigkeit

Im $\mathbb{R}^3$ definieren zwei Vektoren eine Ebene – es sei denn, sie liegen, wenn man sie am gleichen Anfangspunkt beginnen läßt, auf einer Geraden, d.h einer der beiden Vektoren ist ein Vielfaches des anderen.

Entsprechend spannen drei Vektoren im allgemeinen den ganzen $\mathbb{R}^3$ auf – es sei denn, sie liegen, wenn man sie am gleichen Anfangspunkt beginnen läßt, in einer Ebenen, d.h. einer der drei ist als Summe von Vielfachen der anderen beiden darstellbar.

Der Begriff der *linearen Abhängigkeit* verallgemeinert diese Ausnahmebedingungen so, daß sie auf beliebige Vektorräume angewandet werden können:

Definition: M sei eine Teilmenge des k-Vektorraums V.

a) Eine *Linearkombination* von Vektoren aus M ist eine Summe der Form

$$\lambda_1\vec{v}_1 + \cdots + \lambda_n\vec{v}_n$$

mit $n \in \mathbb{N}$, $\lambda_i \in k$ und paarweise verschiedenen Vektoren $\vec{v}_i \in M$; ist diese Summe gleich dem Vektor $\vec{v} \in V$, so sagen wir, $\vec{v}$ sei als Linearkombination von Vektoren aus M darstellbar.

b) Die Menge aller Vektoren, die sich als Linearkombination von Vektoren aus M darstellen lassen, bezeichnen wir mit $[M]$; wir nennen sie das *Erzeugnis* von M.

c) Eine Linearkombination wie in *a)* heißt *nichttrivial,* falls mindestens ein λ_i von Null verschieden ist; ansonsten heißt sie *trivial.*

d) Die Menge M heißt *linear unabhängig,* wenn der Nullvektor nicht als eine nichttriviale Linearkombination von Vektoren aus M darstellbar ist, d.h. wann immer

$$\lambda_1\vec{v}_1 + \cdots + \lambda\vec{v}_n = \vec{0}$$

für Vektoren $\vec{v}_i \in M$ und Skalare $\lambda_i \in k$, müssen alle λ_i verschwinden.

e) Ist M *nicht* linear unabhängig, nennen wir die Menge M *linear abhängig.*

Im (häufigsten) Fall einer endlichen Menge $M = \{\vec{v}_1, \ldots, \vec{v}_r\}$ sagen wir kurz, die *Vektoren* $\vec{v}_1, \ldots, \vec{v}_r$ seien linear (un)abhängig, falls M diese Eigenschaft hat.

Beispielsweise sind also die Vektoren

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix} \in \mathbb{R}^3$$

linear abhängig, da der zweite das Zweifache des ersten ist, wohingegen

$$\begin{pmatrix}1\\0\\0\end{pmatrix}, \quad \begin{pmatrix}0\\1\\0\end{pmatrix}, \quad \begin{pmatrix}0\\0\\1\end{pmatrix} \in \mathbb{R}^3$$

linear unabhängig sind, denn

$$\lambda_1 \begin{pmatrix}1\\0\\0\end{pmatrix} + \lambda_2 \begin{pmatrix}0\\1\\0\end{pmatrix} + \lambda_3 \begin{pmatrix}0\\0\\1\end{pmatrix} = \begin{pmatrix}\lambda_1\\\lambda_2\\\lambda_3\end{pmatrix}$$

ist genau dann gleich dem Nullvektor, wenn alle λ_i verschwinden.

Auch in komplizierteren Vektorräumen können wir leicht Beispiele für lineare Abhängigkeit und Unabhängigkeit finden. In $\mathcal{C}^0(\mathbb{R}, \mathbb{R})$ sind etwa Sinus und Cosinus linear unabhängig, denn gäbe es $\lambda_{1/2} \in \mathbb{R}$ mit

$$\lambda_1 \sin x + \lambda_2 \cos x = 0 \quad \text{für alle } x \in \mathbb{R}$$

mit $\lambda_1 \neq 0$, wäre

$$\tan x = \frac{\sin x}{\cos x} = -\frac{\lambda_2}{\lambda_1}$$

eine konstante Funktion; wäre $\lambda_2 \neq 0$ könnten wir entsprechend auf die Konstanz des Cotangens schließen.

Entsprechend sind auch die Funktionen $\sin^2 x$ und $\cos^2 x$ linear unabhängig, denn auch die Quadrate von Tangens und Cotangens sind nichtkonstante Funktionen. Dagegen sind die drei Funktionen $\sin^2 x,\ \cos^2 x$ und die konstante Funktion 1 linear abhängig, denn

$$\sin^2 x + \cos^2 x - 1 = 0 \quad \text{für alle } x \in \mathbb{R}\,.$$

Lemma: Falls die Menge $M \subseteq V$ linear abhängig ist, gibt es ein Element $\vec{v} \in M$, das sich als Linearkombination der übrigen, d.h. von Vektoren aus $M \setminus \{\vec{v}\}$ schreiben läßt.

Beweis: Wenn M linear abhängig ist, gibt es eine nichttriviale Linearkombination von Vektoren $\vec{v}_i \in M$, so daß

$$\lambda_1\vec{v}_1 + \cdots + \lambda_n\vec{v}_n = \vec{0}$$

ist mit Körperelementen λ_i, die nicht alle gleich Null sind. Sei zum Beispiel $\vec{\lambda}_j \neq 0$. Dann kann obige Gleichung nach $\vec{v}_j$ aufgelöst werden; für $1 < j < n$ etwa ist

$$\vec{v}_j = -\frac{\lambda_1}{\lambda_j}\vec{v}_1 - \cdots - \frac{\lambda_{j-1}}{\lambda_j}\vec{v}_{j-1} - \frac{\lambda_{j+1}}{\lambda_j}\vec{v}_{j+1} - \cdots - \frac{\lambda_n}{\lambda_j}\vec{v}_n\,,$$

und entsprechend läßt sich $\vec{v}_j$ auch im Falle $j = 1$ oder $j = n$ als Linearkombination der übrigen $\vec{v}_i$ schreiben.

■

d) Die Dimension eines Vektorraums

Die Dimension eines Vektorraums soll natürlich so definiert werden, daß k^n die Dimension n hat; wir müssen die Zahl n also irgendwie als Eigenschaft von (Mengen von) Vektoren aus k^n rekonstruieren.

Oben wurde gezeigt, daß die drei Vektoren

$$\begin{pmatrix}1\\0\\0\end{pmatrix},\quad \begin{pmatrix}0\\1\\0\end{pmatrix},\quad \begin{pmatrix}0\\0\\1\end{pmatrix},\ \in\ \mathbb{R}^3$$

linear unabhängig sind; genauso folgt natürlich die lineare Unabhängigkeit von

$$\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix},\quad \begin{pmatrix}0\\1\\\vdots\\0\end{pmatrix},\ \ldots,\ \begin{pmatrix}0\\0\\\vdots\\1\end{pmatrix}\ \in\ k^n\,.$$

Es fällt allerdings schwer, sich $n+1$ linear unabhängige Vektoren in k^n vorzustellen, und in der Tat wird für $n+1$ Vektoren $\vec{v}_1, \ldots, \vec{v}_{n+1} \in k^n$ die Gleichung

$$\lambda_1\vec{v}_1 + \cdots + \lambda_{n+1}\vec{v}_{n+1} = \vec{0}$$

zu einem homogenen linearen Gleichungssystem für die λ_i, das aus n Gleichungen für $n+1$ Variablen besteht. Wenden wir den GAUSS-Algorithmus auf ein solches Gleichungssystem an, kann die Endform explizite Gleichungen für höchstens n der Variablen liefern, mindestens eine ist also frei wählbar und kann damit insbesondere auf ein von Null verschiedenes Element von k gesetzt werden. Damit gibt es also nichttriviale Linearkombinationen der Null aus $\vec{v}_1, \ldots, \vec{v}_{n+1}$.

Diese Eigenschaft machen wir uns nun zunutze, um allgemein Dimensionen zu definieren:

Definition: Wir sagen, der k-Vektorraum V habe die Dimension n, in Zeichen

$$n = \dim_k V \quad \text{oder kurz} \quad n = \dim V ,$$

wenn es in V einerseits n linear unabhängige Vektoren $\vec{v}_1, \ldots, \vec{v}_n$ gibt, andererseits aber jede Teilmenge aus mehr als n Vektoren linear abhängig ist. Dem Nullvektorraum ordnen wir formal die Dimension Null zu.

Falls ein Vektorraum die Dimension Null oder eine Dimension $n \in \mathbb{N}$ hat, sagen wir, er sei *endlichdimensional;* andernfalls nennen wir ihn *unendlichdimensional.*

In diesem Sinne ist also k^n ein n-dimensionaler Vektorraum. Unendlichdimensionale Vektorräume sind beispielsweise $\mathcal{C}^0(\mathbb{R}, \mathbb{R})$, denn offensichtlich ist die unendliche Menge der Potenzfunktionen $x \mapsto x^r$ linear unabhängig, da es ansonsten ein von Null verschiedenes Polynom gäbe, das nur den Wert Null annimmt.

Weniger offensichtlich ist, daß auch $\mathbb{R}$ als $\mathbb{Q}$-Vektorraum unendlichdimensional ist: Dazu betrachten wir die unendliche Menge M aller Logarithmen $\ln p$ der Primzahlen. Wäre diese Menge linear abhängig, gäbe es eine nichttriviale Linearkombination

$$\lambda_1 \ln p_1 + \cdots + \lambda_r \ln p_r = 0$$

mit $\lambda_i \in \mathbb{Q}$. Multipliziert man diese Gleichung mit dem Hauptnenner der λ_i, so erhält man eine entsprechende Gleichung mit $\lambda_i \in \mathbb{Z}$. Dann ist

$$\lambda_1 \ln p_1 + \cdots + \lambda_r \ln p_r = \ln(p_1^{\lambda_1} \cdot \cdots \cdot p_r^{\lambda_r}) = 0$$

gleichbedeutend mit

$$p_1^{\lambda_1} \cdot \cdots \cdot p_r^{\lambda_r} = 1 ,$$

was wegen der Eindeutigkeit der Primzerlegung in $\mathbb{Z}$ nur gelten kann, wenn alle λ_i verschwinden.

Also ist $\mathbb{R}$ als $\mathbb{Q}$-Vektorraum unendlichdimensional, und dies erklärt, warum Computer so große Schwierigkeiten mit reellen Zahlen haben: Exakt rechnen kann ein Computer nur in Teilkörpern von $\mathbb{R}$, die endlichdimensionale $\mathbb{Q}$-Vektorräume sind. Vor allem in der Computergraphik, wo man die logische und geometrische Konsistenz der Daten nur erhalten kann, wenn man *sicher* entscheiden kann, ob etwa zwei als Schnittpunkte von Kurven gegebene Punkte gleich sind oder nicht, ist dies ein großes, wenn auch kein unlösbares, Problem.

e) Basen

Im $\mathbb{R}^3$ läßt sich jeder Vektor

$$\vec{v} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} = a \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + b \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + c \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

auf genau eine Weise als Linearkombination der drei Einheitsvektoren

$$\vec{e}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{und} \quad \vec{e}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

schreiben; die entspricht der Tatsache, daß wir im $\mathbb{R}^3$ drei Koordinaten haben.

Mit dem Begriff der *Basis* soll dieser Sachverhalt (soweit möglich) auf beliebige Vektorräume verallgemeinert werden.

Definition: Eine Teilmenge $\mathcal{B} \subset V$ eines k-Vektorraums V heißt *Basis* von V, wenn gilt:

1. $V = [\mathcal{B}]$
2. $\mathcal{B}$ ist linear unabhängig.

In diesem Sinne ist $\{\vec{e}_1, \vec{e}_2, \vec{e}_3\}$ offenbar eine Basis des $\mathbb{R}^3$. Ihre wesentliche Eigenschaft der eindeutigen Darstellbarkeit eines jeden Vektors als Linearkombination teilt sie mit jeder beliebigen Basis:

Lemma: Ist $\mathcal{B}$ eine Basis eines Vektorraums V, so läßt sich jeder Vektor $\vec{v} \in V$ auf genau eine Weise als Linearkombination

$$\vec{v} = \lambda_1 \vec{b}_1 + \cdots + \lambda_r \vec{b}_r$$

von Basisvektoren $\vec{b}_i \in \mathcal{B}$ darstellen.

Beweis: Nach der ersten Eigenschaft in der Definition einer Basis ist $V = [\mathcal{B}]$, also läßt sich jeder Vektor $\vec{v} \in V$ als eine Linearkombination von endlich vielen Elementen aus $\mathcal{B}$ darstellen. Auch wenn wir von zwei solchen Darstellungen ausgehen, ist die Menge der daran beteiligten Vektoren aus $\mathcal{B}$ noch endlich; wir können also annehmen, daß es r Vektoren $\vec{b}_1, \dots, \vec{b}_r$ gibt, so daß

$$\begin{aligned}\vec{v} &= \lambda_1\vec{b}_1 + \cdots + \lambda_r\vec{b}_r \\ &= \mu_1\vec{b}_1 + \cdots + \mu_r\vec{b}_r\end{aligned}$$

ist, wobei wir einfach λ_i oder μ_i auf Null setzen, wenn $\vec{b}_i$ in der entsprechenden Darstellung nicht vorkommt. Subtrahieren wir die beiden Darstellungen voneinander, erhalten wir eine Darstellung des Nullvektors als Linearkombination

$$\vec{0} = (\lambda_1 - \mu_1)\vec{b}_1 + \cdots + (\lambda_r - \mu_r)\vec{b}_r$$

von Vektoren aus der Menge $\mathcal{B}$. Da diese nach der zweiten definierenden Eigenschaft einer Basis linear unabhängig ist, müssen alle Koeffizienten $\lambda_i - \mu_i$ verschwinden. Damit sind die beiden betrachteten Darstellungen von $\vec{v}$ als Linearkombination der Vektoren aus $\mathcal{B}$ gleich, mit anderen Worten: Es gibt genau eine solche Darstellung.

■

Als nächstes stellt sich die Frage, wann es Basen gibt. Glücklicherweise hat *jeder* Vektorraum eine Basis, der Beweis ist allerdings für unendlichdimensionale Vektorräume logisch nicht ganz einfach. Für diese Vorlesung wollen wir uns daher begnügen mit dem Beweis für endlichdimensionale Vektorräume. Wir beweisen dazu den etwas allgemeineren, tatsächlich auch für beliebige Vektorräume gültigen

Basisergänzungssatz: $\emptyset \neq M \subset V$ sei eine linear unabhängige Teilmenge des endlichdimensionalen Vektorraums V. Dann gibt es eine Basis $\mathcal{B}$ von V, die M enthält.

Beweis: Wir überlegen uns zunächst, daß es überhaupt eine endliche Menge $E \subset V$ gibt, so daß $V = [E]$ von E erzeugt wird.

Als endlichdimensionaler Vektorraum hat V eine Dimension $n \in \mathbb{N}$, d.h. es gibt eine Menge $E = \{\vec{v}_1, \ldots, \vec{v}_n\}$ von n linear unabhängigen Vektoren aus V, wohingegen jede Menge aus mehr als n Vektoren linear abhängig ist. Insbesondere ist also für jeden Vektor $\vec{v} \notin E$ die Menge $E \cup \{\vec{v}\}$ linear abhängig, d.h. es gibt Elemente $\lambda_0, \ldots, \lambda_n \in k$, die nicht alle verschwinden, so daß

$$\lambda_0\vec{v} + \lambda_1\vec{v}_1 + \cdots + \lambda_n\vec{v}_n = \vec{0}$$

ist. Wäre in dieser Gleichung $\lambda_0 = 0$, müßte mindestens ein anderes λ_i von Null verschieden sein, und wir hätten eine lineare Abhängigkeit zwischen den Vektoren aus E, im Widerspruch zur Wahl von E. Somit ist $\lambda_0 \neq 0$ und

$$\vec{v} = -\frac{\lambda_1}{\lambda_0}\vec{v}_1 - \cdots - \frac{\lambda_n}{\lambda_0}\vec{v}_n$$

ist als Linearkombination von Elementen aus E darstellbar. Also ist $V = [E]$ von E erzeugt und damit erst recht auch von $M \cup E$.

Sei $M = \{\vec{b}_1, \ldots, \vec{b}_r\}$; wenn wir die $\vec{v}_i$ so anordnen, daß die eventuell vorhandenen $\vec{v}_i \in M$ an den letzten Stellen stehen, gibt es dann ein $s \leq n$, so daß

$$M \cup E = \{\vec{b}_1, \ldots, \vec{b}_r, \vec{v}_1, \ldots, \vec{v}_s\}$$

ist. Falls diese Menge linear unabhängig ist, sind wir fertig, denn dann ist sie eine Basis von V, die M enthält.

Andernfalls gibt es Elemente $\lambda_1, \ldots, \lambda_r$ und $\mu_1, \ldots, \mu_s$, die nicht alle gleichzeitig verschwinden, so daß

$$\lambda_1\vec{b}_1 + \cdots + \lambda_r\vec{b}_r + \mu_1\vec{v}_1 + \cdots + \mu_s\vec{v}_s = \vec{0}$$

ist. In dieser Gleichung können nicht alle μ_i Null sein, denn sonst wären die $\vec{b}_i$ linear abhängig, im Widerspruch zur Voraussetzung. Also gibt es (mindestens) ein $\mu_i \neq 0$, und der zugehörige Vektor $\vec{v}_i$ läßt sich als Linearkombination der restlichen $\vec{v}_j$ und der $\vec{b}_\ell$ ausdrücken.

Damit wird $V = [M \cup E \setminus \{\vec{v}_i\}]$ auch von der um $\vec{v}_i$ verminderten Menge erzeugt; falls diese linear unabhängig ist, sind wir fertig. Andernfalls können wir auf genau die gleiche Weise weitere Vektoren $\vec{v}_h$ eliminieren, bis die Menge – spätestens nach Elimination aller Vektoren aus E – linear unabhängig wird. Da sich durch diese Eliminationen nichts an der Erzeugung des gesamten Vektorraums V ändert, haben wir damit eine Basis gefunden, die M enthält.

■

Korollar: Jeder endlichdimensionale Vektorraum $V \neq \{\vec{0}\}$ hat eine Basis.

Beweis: Man wende den obigen Satz an auf eine Menge M, die aus einem einzigen Vektor $\vec{v} \neq \vec{0}$ besteht.

■

Um die Sonderrolle des Nullvektorraums zu eliminieren, vereinbaren wir, daß er die leere Menge als Basis haben soll; dies ist kompatibel mit der üblichen Interpretation von leeren Summen.

Um wenigstens anhand eines Beispiels zu sehen, daß auch unendlichdimensionale Vektorräume Basen haben, betrachten wir den Vektorraum V aller Polynome mit reellen Koeffizienten. Da sich ein Polynom P vom Grad d als

$$P = a_0 + a_1 X + \cdots + a_d X^d$$

schreiben läßt, erzeugt die Menge

$$\mathcal{B} = \{1, X, x^2, X^3, \dots\}$$

diesen Vektorraum, und jede Linearkombination des Nullvektors, des Polynoms $P = 0$ also, aus Elementen von $\mathcal{B}$ wäre ein Polynom

$$\lambda_0 + \lambda_1 X + \cdots + \lambda_n X^n ,$$

dessen Koeffizienten zumindest teilweise von Null verschieden sind, während es selbst identisch Null wäre. Da es kein solches Polynom gibt, ist die Menge $\mathcal{B}$ linear unabhängig, also eine Basis von V.

Die Schwierigkeiten, die bei unendlichdimensionalen Vektorräumen auftreten können, sieht man, wenn man in diesem Beispiel die Polynome durch Potenzreihen (egal ob formal oder konvergent) ersetzt: Da Potenzreihen *unendliche* Summen sind, während bei Linearkombinationen nur *endliche* Summen erlaubt sind, reichen nun die X-Potenzen nicht mehr aus zur Erzeugung. Nach dem Basisergänzungssatz, der, auch wenn wir das nicht bewiesen haben, auch für unendlichdimensionale Vektorräume gilt, gibt es eine Menge von Potenzreihen, die zusammen mit der obigen Menge $\mathcal{B}$ eine Basis bilden; explizit angeben konnte diese Menge noch niemand.

Kehren wir also zurück zum überschaubareren endlichdimensionalen Fall, und beweisen wir dort zunächst einen anschaulich fast selbstverständlichen Satz:

Satz: Jede Basis $\mathcal{B}$ eines n-dimensionalen Vektorraums besteht aus n Vektoren.

Beweis: Es ist klar, daß keine Basis *mehr* als n Vektoren enthalten kann; denn in einem n-dimensionalen Vektorraum sind alle Mengen aus mehr als n Vektoren linear abhängig. Wenn $\mathcal{B} = \{\vec{b}_1, \ldots, \vec{b}_m\}$ *weniger* als n Vektoren enthält, erwarten wir, daß es Probleme mit der Erzeugung des Vektorraums gibt: Nach Definition gibt es n linear unabhängige Vektoren $\vec{v}_1, \ldots, \vec{v}_n$ in V, und wenn $\mathcal{B}$ eine Basis ist, müssen auch alle $\vec{v}_i$ als Linearkombinationen der $\vec{b}_j$ darstellbar sein, also etwa

$$\vec{v}_i = a_{i1}\vec{b}_1 + \cdots + a_{im}\vec{b}_m \,.$$

Eine Linearkombination

$$x_1 \cdot \vec{v}_1 + \cdots + x_n \cdot \vec{v}_n$$

läßt sich daher in den $\vec{b}_j$ ausdrücken als

$$\sum_{i=1}^{n} x_i\vec{v}_i = \sum_{i=1}^{n} x_i \sum_{j=1}^{m} a_{ij}\vec{b}_j \,.$$

Eine solche Summe ist wegen der linearen Unabhängigkeit der $\vec{b}_j$ genau dann gleich dem Nullvektor, wenn die Koeffizienten aller $\vec{b}_j$ verschwinden, wenn also gilt

$$a_{1j}x_1 + \cdots + a_{nj}x_n = 0 \quad \text{für } j = 1, \ldots, m.$$

Dies ist ein homogenes lineares Gleichungssystem von $m < n$ Gleichungen für n Unbekannte; nachdem die GAUSSschen Eliminationsschritte ausgeführt sind, können also höchstens m der x_i festgelegt sein, so daß noch mindestens $n - m$ frei und damit insbesondere ungleich Null gewählt werden können.

Damit gibt es also eine nichttriviale Möglichkeit, den Nullvektor aus den linear unabhängigen Vektoren $\vec{v}_i$ linear zu kombinieren, ein Widerspruch. Somit kann $\mathcal{B}$ keine Basis sein.

Insgesamt haben wir also gesehen, daß weder Basen mit mehr noch Basen mit weniger als n Elementen möglich sind, d.h. jede Basis besteht aus n Vektoren.

■

Nach diesem Satz läßt sich die Dimension eines Vektorraums einfach dadurch bestimmen, daß man eine Basis findet und deren Elemente zählt. Beispielsweise hat $\mathbb{C}$ als $\mathbb{R}$-Vektorraum die Dimension zwei, da $\{1, i\}$ offensichtlich eine Basis ist.

f) Untervektorräume

Definition: Eine Teilmenge $U \subseteq V$ eines Vektorraums V heißt *Untervektorraum,* in Zeichen $U \leq V$, wenn U bezüglich der Einschränkungen der Addition und der Skalarmultiplikation von V ein Vektorraum ist. Ist zusätzlich $U \neq V$, sprechen wir von einem *echten* Untervektorraum und schreiben $U < V$.

Beim Nachweis der Vektorraumaxiome für U kann man natürlich Eigenschaften wie das Assoziativgesetz oder die Distributivgesetze weglassen, denn da diese für alle Vektoren von V gelten, gelten sie erst recht auch für alle aus U. Wirklich gezeigt werden muß nur, daß weder die Addition noch die Skalarmultiplikation aus U herausführen, was man auch zusammenfassen kann zum

Untervektorraumkriterium: Eine nichtleere Teilmenge $U \subseteq V$ eines Vektorraums V ist genau dann ein Untervektorraum, wenn mit zwei Vektoren $\vec{u}, \vec{v} \in U$ und zwei Skalaren $\lambda, \mu \in k$ auch $\lambda\vec{u} + \mu\vec{v}$ in U liegt. ■

Anschaulich klar ist

Satz: Für einen echten Untervektorraum $U < V$ eines endlichdimensionalen Vektorraums V ist $\dim U < \dim V$.

Beweis: Da linear unabhängige Vektoren aus U natürlich auch in V linear unabhängig sind, ist auf jeden Fall $\dim U \leq \dim V$.

Falls beide Vektorräume die gleiche Dimension n haben, gibt es n linear unabhängige Vektoren in U. Wie oben beim Beweis des Basisergänzungssatzes folgt dann, daß jeder Vektor aus V als Linearkombination dieser Vektoren darstellbar ist, d.h. $U = V$.

■

g) Lineare Abbildungen

Lineare Abbildungen zwischen beliebigen Vektorräumen werden genauso definiert wie die zwischen den speziellen Vektorräumen k^n:

Definition: *a)* Eine Abbildung $\varphi: V \to W$ heißt *linear,* wenn für alle Vektoren $\vec{u}, \vec{v} \in V$ und alle $\lambda, \mu \in k$ gilt:

$$\varphi(\lambda\vec{u} + \mu\vec{v}) = \lambda\varphi(\vec{u}) + \mu\varphi(\vec{v})\,.$$

b) Unter dem *Kern* von φ verstehen wir die Menge

$$\operatorname{Kern}\varphi \underset{\text{def}}{=} \left\{\vec{v} \in V \;\middle|\; \varphi(\vec{v}) = \vec{0}\right\}.$$

c) Das *Bild* von φ ist die Menge

$$\operatorname{Bild}\varphi \underset{\text{def}}{=} \left\{\vec{w} \in W \;\middle|\; \text{Es gibt } \vec{v} \in V \text{ mit } \varphi(\vec{v}) = \vec{w}\right\}.$$

Auch wenn Kern und Bild zunächst nur als Mengen definiert sind, überzeugt man sich sofort, daß tatsächlich $\operatorname{Kern}\varphi$ ein Untervektorraum von V und $\operatorname{Bild}\varphi$ ein Untervektorraum von W ist.

Lemma: $\varphi: V \to W$ ist genau dann injektiv, wenn $\operatorname{Kern}\varphi$ der Nullvektorraum ist, und genau dann surjektiv, wenn $\operatorname{Bild}\varphi = W$ ist.

Beweis: Die zweite Aussage ist zu trivial, als daß man etwas dazu sagen müßte, betrachten wir also die erste. Falls φ injektiv ist, hat insbesondere der Nullvektor nur ein einziges Urbild, d.h. der Kern besteht nur aus dem Nullvektor, der natürlich immer im Kern liegt. Ist umgekehrt $\operatorname{Kern}\varphi$ der Nullraum und haben zwei Vektoren $\vec{u}, \vec{v} \in V$ dasselbe Bild, so ist

$$\varphi(\vec{u} - \vec{v}) = \varphi(\vec{v}) - \varphi(\vec{u}) = \vec{0}\,,$$

d.h. $\vec{u} - \vec{v}$ liegt im Kern und muß daher gleich dem Nullvektor sein, so daß $\vec{u} = \vec{v}$ ist. Dies zeigt die Injektivität von φ.

■

Satz: Für einen endlichdimensionalen Vektorraum V und eine lineare Abbildung $\varphi: V \to W$ ist

$$\dim \operatorname{Bild} \varphi = \dim V - \dim \operatorname{Kern} \varphi \,.$$

Beweis: $\vec{b}_1, \dots, \vec{b}_r$ sei eine Basis von $\operatorname{Kern} \varphi$; falls φ injektiv ist, setzen wir $r = 0$. Nach dem Basisergänzungssatz oder (falls $r = 0$) wegen der Existenz von Basen lassen sich dann $n - r$ Vektoren $\vec{b}_{r+1}, \dots, \vec{b}_n$ finden mit $n = \dim V$, so daß $\vec{b}_1, \dots, \vec{b}_n$ eine Basis von V ist.

Das Bild eines beliebigen Vektors $\vec{v} = \lambda_1 \vec{b}_1 + \cdots + \lambda_n \vec{b}_n$ ist dann

$$\varphi(\vec{v}) = \lambda_1 \varphi(\vec{b}_1) + \cdots + \lambda_n \varphi(\vec{b}_n) = \lambda_{r+1} \varphi(\vec{b}_{r+1}) + \cdots + \lambda_n \varphi(\vec{b}_n) \,,$$

da $\vec{b}_1 \dots, \vec{b}_r$ ja auf den Nullvektor abgebildet werden. Also wird $\operatorname{Bild} \varphi$ von den Vektoren $\varphi(\vec{b}_{r+1}), \dots, \varphi(\vec{b}_n)$ erzeugt.

Diese Vektoren sind auch linear unabhängig in W, denn ist

$$\lambda_{r+1} \varphi(\vec{b}_{r+1}) + \cdots + \lambda_n \varphi(\vec{b}_n) = \varphi(\lambda_{r+1} \vec{b}_{r+1} + \cdots + \lambda_n \vec{b}_n) = \vec{0} \,,$$

liegt $\lambda_{r+1} \vec{b}_{r+1} + \cdots + \lambda_n \vec{b}_n$ im Kern von φ.

Im Fall einer injektiven Abbildung ist $\lambda_{r+1} \vec{b}_{r+1} + \cdots + \lambda_n \vec{b}_n$ daher gleich dem Nullvektor, und damit müssen alle $\lambda_i = 0$ sein, denn die $\vec{b}_i$ sind Basisvektoren.

Falls φ nicht injektiv ist, können wir nur sagen, daß der Vektor $\lambda_{r+1} \vec{b}_{r+1} + \cdots + \lambda_n \vec{b}_n$ im Kern von φ liegt; er ist also darstellbar als Linearkombination der Basisvektoren $\vec{b}_1, \dots, \vec{b}_r$ des Kerns:

$$\lambda_{r+1} \vec{b}_{r+1} + \cdots + \lambda_n \vec{b}_n = \lambda_1 \vec{b}_1 + \cdots + \lambda_r \vec{b}_r \,.$$

Auch daraus folgt wegen der linearen Unabhängigkeit der $\vec{b}_i$, daß alle λ_i Null sein müssen.

Damit ist $\{\vec{b}_{r+1}, \dots, \vec{b}_n\}$ eine Basis von $\operatorname{Bild} \varphi$, d.h.

$$\dim \operatorname{Bild} \varphi = n - r = \dim V - \dim \operatorname{Kern} \varphi \,,$$

wie behauptet.

■

Wir werden diese Aussage gelegentlich als den *Homomorphiesatz* bezeichnen. Der „echte“ Homomorphiesatz ist zwar eine schärfere Aussage über den Bildraum, die auch für unendlichdimensionale Vektorräume gilt, die wir mit dem uns bislang zur Verfügung stehenden Begriffsapparat aber noch nicht formulieren können. Im endlichdimensionalen Fall ist obiger Satz eine unmittelbare Folgerung aus dem Homomorphiesatz.

g) Isomorphie von Vektorräumen

Da lineare Abbildungen die Vektorraumstrukturen ihres Urbild- und Zielraums respektieren, liegt es nahe, Vektorräume, zwischen denen es eine *bijektive* lineare Abbildung gibt, als nicht wesentlich verschieden zu betrachten: Schließlich überträgt sich jede Rechnung, die wir in einem der beiden Vektorräume machen, sofort über die Abbildung oder ihre Umkehrabbildung auf den anderen. Der mathematische Begriff für „nicht wesentlich verschieden“ ist *isomorph,* was ins Deutsche etwa mit „von gleicher Gestalt“ übersetzt werden kann:

Definition: *a)* Ein *Isomorphismus* $\varphi: V \to W$ zwischen zwei k-Vektorräumen V und W ist eine bijektive lineare Abbildung.
b) Zwei Vektorräume V, W heißen *isomorph,* in Zeichen $V \cong W$, wenn es einen Isomorphismus $\varphi: V \to W$ gibt.

Aus der Existenz von Basen für endlichdimensionale Vektorräume folgt sofort

Satz: Jeder n-dimensionale Vektorraum V über dem Körper k ist isomorph zu k^n.

Beweis: Als n-dimensionaler Vektorraum hat V eine n-elementige Basis $\mathcal{B} = \{\vec{b}_1, \dots, \vec{b}_n\}$, bezüglich derer sich jeder Vektor $\vec{v} \in V$ in eindeutiger Weise schreiben läßt als

$$\vec{v} = \lambda_1(\vec{v})\vec{b}_1 + \cdots + \lambda_n(\vec{v})\vec{b}_n$$

mit von $\vec{v}$ abhängigen Koeffizienten $\lambda_i(\vec{v}) \in k$. Offensichtlich ist die Abbildung

$$\varphi: V \to k^n; \quad \vec{v} \mapsto \begin{pmatrix} \lambda_1(\vec{v}) \\ \vdots \\ \lambda_n(\vec{v}) \end{pmatrix}$$

ein Isomorphismus.

■

h) Basiswechsel und Abbildungsmatrizen

Im Fall der Vektorräume k^n können wir lineare Abbildungen durch Matrizen rechnerisch gut in den Griff bekommen. Nachdem wir gerade gesehen haben, daß jeder endlichdimensionale Vektorraum isomorph ist zu einem Vektorraum der Form k^n, sollte es auch für beliebige lineare Abbildungen zwischen endlichdimensionalen Vektorräumen Abbildungsmatrizen geben.

Es gibt allerdings einen wesentlichen Unterschied: Zu einer linearen Abbildung $\varphi: k^n \to k^m$ gehört auf ganz natürliche Weise genau eine Abbildungsmatrix; einer lineare Abbildung $\varphi: V \to W$ zwischen zwei beliebigen endlichdimensionalen Vektorräumen dagegen können wir erst dann eine Abbildungsmatrix zuordnen, wenn wir vorher Isomorphismen $V \to k^n$ und $W \to k^m$ fixiert haben, oder, was nach dem vorigen Abschnitt auf dasselbe hinausläuft, wenn wir sowohl in V also auch in W eine Basis festgelegt haben.

Dies kompliziert die Lage zwar etwas, hat aber andererseits den Vorteil, daß man durch geschickte Basiswahl auf rechnerisch einfache Matrizen kommen kann; entsprechende Strategien werden wir im nächsten Paragraphen kennenlernen.

Hier wollen wir uns zunächst genau überlegen, wie Abbildungsmatrizen erklärt werden können und wie sie sich unter Basiswechsel verhalten. Dabei beschränken wir uns zunächst auf den besonders wichtigen Fall einer linearen Selbstabbildung $\varphi: V \to V$ eines Vektorraums; die Verallgemeinerung auf den allgemeinen Fall wird nachher eine reine Formalität sein.

Beginnen wir mit einem Beispiel: Der Körper $\mathbb{C}$ der komplexen Zahlen kann als Vektorraum über $\mathbb{R}$ aufgefaßt werden; praktisch nach Definition von $\mathbb{C}$ ist $\{1, i\}$ eine Basis, so daß $\mathbb{C}$ als $\mathbb{R}$-Vektorraum isomorph zu $\mathbb{R}^2$ ist.

Für jede komplexe Zahl $c = a + ib$ ist die Abbildung

$$\varphi\colon \mathbb{C} \to \mathbb{C}; \qquad z \mapsto cz$$

linear; was ist ihre Abbildungsmatrix bezüglich der Basis $\mathcal{B} = \{1, i\}$?

Für eine Antwort auf diese Frage müssen wir ausrechnen, wie φ auf die Basisdarstellung von Vektoren wirkt: Mit $z = x + iy$ ist

$$\varphi(z) = cz = (a + ib)(x + iy) = (ax - by) + i(bx + ay)\,;$$

der Vektor $\binom{x}{y} \in \mathbb{R}^2$, der der komplexen Zahl z entspricht, wird also abgebildet auf

$$\begin{pmatrix} ax - by \\ ay + bx \end{pmatrix} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},$$

und die Abbildungsmatrix von φ bezüglich der Basis $\mathcal{B}$ ist $\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$.

Formal gesehen haben wir also auf einen Vektor $\binom{x}{y} \in \mathbb{R}^2$ zunächst die Umkehrabbildung des Isomorphismus $\mathbb{C} \to \mathbb{R}^2$ angewandt, um ihn zur komplexen Zahl $z = x + iy$ zu machen; als nächstes wurde diese auf die komplexe Zahl $\varphi(z) = cz$ abgebildet, und darauf schließlich wurde der Isomorphismus $\mathbb{C} \to \mathbb{R}^2$ angewandt.

Entsprechend können wir auch für eine beliebige lineare Abbildung $\varphi\colon V \to V$ vorgehen: Wir wählen eine Basis $\mathcal{B}$ von V und damit, wie oben ausgeführt, einen Isomorphismus $\psi_{\mathcal{B}}\colon V \to k^n$. Über die Abbildungskette

$$k^n \xrightarrow{\psi_{\mathcal{B}}^{-1}} V \xrightarrow{\varphi} V \xrightarrow{\psi_{\mathcal{B}}} k^n$$

definieren wir dann eine lineare Abbildung

$$\varphi_{\mathcal{B}} \underset{\text{def}}{=} \psi_{\mathcal{B}} \circ \varphi \circ \psi_{\mathcal{B}}^{-1}\colon k^n \to k^n$$

und bezeichnen ihre Abbildungsmatrix $A_{\mathcal{B}} \in k^{n \times n}$ als die *Abbildungsmatrix von φ bezüglich der Basis $\mathcal{B}$.*

Für eine andere Basis $\mathcal{C}$ von V haben wir entsprechend eine Abbildung

$$\varphi_{\mathcal{C}} = \psi_{\mathcal{C}} \circ \varphi \circ \psi_{\mathcal{C}}^{-1} : k^n \to k^n ,$$

und da $\psi_{\mathcal{B}}, \psi_{\mathcal{C}}$ Isomorphismen sind, folgt

$$\begin{aligned}\varphi_{\mathcal{C}} &= \psi_{\mathcal{C}} \circ \varphi \circ \psi_{\mathcal{C}}^{-1} = \psi_{\mathcal{C}} \circ (\psi_{\mathcal{B}}^{-1} \circ \psi_{\mathcal{B}}) \circ \varphi \circ (\psi_{\mathcal{B}}^{-1} \circ \psi_{\mathcal{B}}) \circ \psi_{\mathcal{C}}^{-1} \\ &= (\psi_{\mathcal{C}} \circ \psi_{\mathcal{B}}^{-1}) \circ \varphi_{\mathcal{B}} \circ (\psi_{\mathcal{C}} \circ \psi_{\mathcal{B}}^{-1})^{-1} .\end{aligned}$$

Die eingeklammerte lineare Abbildung

$$\psi_{\mathcal{C}} \circ \psi_{\mathcal{B}}^{-1} : k^n \to k^n$$

bildet einen Vektor $\begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} \in k^n$ zunächst via $\psi_{\mathcal{B}}^{-1}$ ab auf den Vektor

$$\vec{v} = \lambda_1 \vec{b}_1 + \cdots + \lambda_n \vec{b}_n \in V ,$$

wobei $\mathcal{B} = \{\vec{b}_1, \ldots, \vec{b}_n\}$. Dieser Vektor wird dann via $\psi_{\mathcal{C}}$ wieder nach k^n abgebildet, d.h. ist $\mathcal{C} = \{\vec{c}_1, \ldots, \vec{c}_n\}$ und

$$\vec{v} = \mu_1 \vec{c}_1 + \cdots + \mu_n \vec{c}_n ,$$

so läßt sich die Abbildung explizit schreiben als

$$\psi_{\mathcal{C}} \circ \psi_{\mathcal{B}}^{-1} : k^n \to k^n ; \qquad \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} \mapsto \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} .$$

Wie jede lineare Abbildung $k^n \to k^n$ hat sie eine Abbildungsmatrix M, und wie üblich stehen in den Spalten dieser Matrix die Bilder der Basisvektoren $\vec{b}_i$. Ist also

$$\vec{b}_i = \mu_{1i} \vec{c}_1 + \cdots + \mu_{1n} \vec{c}_n ,$$

so ist

$$M = \begin{pmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2n} \\ \vdots & \cdots & \ddots & \vdots \\ \mu_{n1} & \mu_{n2} & \cdots & \mu_{nn} \end{pmatrix} .$$

Mithin ist

$$A_{\mathcal{C}} = M A_{\mathcal{B}} M^{-1} \quad \text{oder} \quad A_{\mathcal{B}} = M^{-1} A_{\mathcal{C}} M \,.$$

Meist sind nicht die „alten“ Basisvektoren $\vec{b}_i$ als Linearkombinationen der „neuen“ Vektoren $\vec{c}_i$ bekannt, sondern umgekehrt die $\vec{c}_i$ definiert als gewisse Linearkombinationen

$$\vec{c}_i = \lambda_{1i}\vec{b}_1 + \cdots + \lambda_{1n}\vec{b}_n \,.$$

Vertauscht man in obiger Formel die Rolle von $\mathcal{B}$ und $\mathcal{C}$ folgt sofort, daß für die Matrix

$$L = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2n} \\ \vdots & \dots & \ddots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \dots & \lambda_{nn} \end{pmatrix}$$

gilt

$$A_{\mathcal{C}} = L^{-1} A_{\mathcal{B}} L \quad \text{und} \quad A_{\mathcal{B}} = L A_{\mathcal{B}} L^{-1} \,;$$

insbesondere ist also $L = M^{-1}$.

Betrachten wir als einfaches Beispiel hierzu die Basis $\mathcal{C} = \{1+i, 1-i\}$ von $\mathbb{C}$ über $\mathbb{R}$ und wieder die Abbildung

$$\varphi\colon \mathbb{C} \to \mathbb{C}; \quad z \mapsto cz \quad \text{mit} \quad c = a + ib \in \mathbb{C} \,.$$

Bezüglich $\mathcal{B} = \{1, i\}$ kennen wir die Abbildungsmatrix

$$A_{\mathcal{B}} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

bereits; die Matrix des Basiswechsels ist offenbar

$$L = M^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \text{also ist} \quad M = L^{-1} = \begin{pmatrix} \frac12 & \frac12 \\ \frac12 & -\frac12 \end{pmatrix}$$

und damit

$$A_{\mathcal{C}} = MAL = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \,.$$

Zum Abschluß dieses Abschnitts sei noch, wie angekündigt, kurz auf die rein formalen Verallgemeinerungen eingegangen, die für Abbildungen $\varphi: V \to W$ zwischen zwei nicht notwendigerweise gleichen Vektorräumen notwendig sind. Natürlich können wir weiterhin nur endlichdimensionale Vektorräume behandeln; sei also $\dim V = n$ und $\dim W = m$.

Als nächstes wählen wir eine Basis $\mathcal{B}$ von V sowie eine Basis $\mathcal{B}'$ von W und haben damit Isomorphismen $\psi_{\mathcal{B}}: V \to k^n$ und $\psi_{\mathcal{B}'}: W \to k^m$. Nun geht alles fast wörtlich genauso weiter wie oben im Fall $V = W$: Über die Abbildungskette

$$k^n \xrightarrow{\psi_{\mathcal{B}}^{-1}} V \xrightarrow{\varphi} W \xrightarrow{\psi_{\mathcal{B}'}} k^m$$

definieren wir eine lineare Abbildung

$$\varphi_{\mathcal{B},\mathcal{B}'} \underset{\text{def}}{=} \psi_{\mathcal{B}'} \circ \varphi \circ \psi_{\mathcal{B}}^{-1}: k^n \to k^m$$

und bezeichnen ihre Abbildungsmatrix $A_{\mathcal{B},\mathcal{B}'} \in k^{n\times n}$ als die *Abbildungsmatrix von φ bezüglich der Basen $\mathcal{B}$ und $\mathcal{B}'$*.

Auch den Effekt von Basiswechseln auf die Abbildungsmatrix können wir genau wie oben berechnen: Für weitere Basen $\mathcal{C}$ von V und $\mathcal{C}'$ von W ist

$$\begin{aligned}\varphi_{\mathcal{C},\mathcal{C}'} &= \psi_{\mathcal{C}'} \circ \varphi \circ \psi_{\mathcal{C}}^{-1} = \psi_{\mathcal{C}'} \circ (\psi_{\mathcal{B}'}^{-1} \circ \psi_{\mathcal{B}'}) \circ \varphi \circ (\psi_{\mathcal{B}}^{-1} \circ \psi_{\mathcal{B}}) \circ \psi_{\mathcal{C}}^{-1}\\ &= (\psi_{\mathcal{C}'} \circ \psi_{\mathcal{B}'}^{-1}) \circ \varphi_{\mathcal{B},\mathcal{B}'} \circ (\psi_{\mathcal{B}} \circ \psi_{\mathcal{C}}^{-1})\,;\end{aligned}$$

sind also M und $L = M^{-1}$ die oben definierten Matrizen, durch die sich die Basisdarstellungen bezüglich $\mathcal{B}$ und $\mathcal{C}$ ineinander umrechnen lassen, und sind M', $L' = M'^{-1}$ die entsprechend definierten $m \times m$-Matrizen für $\mathcal{B}'$ und $\mathcal{C}'$, so ist

$$A_{\mathcal{C},\mathcal{C}'} = M' A_{\mathcal{B},\mathcal{B}'} L \quad \text{und} \quad A_{\mathcal{B},\mathcal{B}'} = L' A_{\mathcal{C},\mathcal{C}'} M\,.$$

i) Direkte Summen von Vektorräumen

Wir schreiben schon die ganze Zeit den Vektorraum k^n als eine Potenz; in diesem Abschnitt soll erklärt werden, daß er tatsächlich als eine Art Potenz des eindimensionalen Vektorraums k aufgefaßt werden kann.

Dazu beginnen wir mit der direkten Summe zweier Vektorräume V, W über demselben Körper k.

Definition: Die direkte Summe $V \oplus W$ zweier k-Vektorräume V, W ist die Menge $V \times W$ mit der Addition

$$(v, w) + (v', w') \underset{\text{def}}{=} (v + v', w + w')$$

und der Skalarmultiplikation

$$\lambda(v, w) \underset{\text{def}}{=} (\lambda v, \lambda w)$$

für alle $v, v' \in V$ und $w, w' \in W$ sowie alle $\lambda \in k$.

Da alles komponentenweise definiert ist, lassen sich alle Vektorraumaxiome problemlos abhaken; $V \oplus W$ ist also in der Tat ein Vektorraum.

Ist $\mathcal{B}$ eine Basis von V und $\mathcal{C}$ eine Basis von W, so bilden offensichtlich die Vektoren $(\vec{b}, \vec{0})$ mit $\vec{b} \in \mathcal{B}$ zusammen mit den Vektoren $(\vec{0}, \vec{c})$ mit $\vec{c} \in \mathcal{C}$ eine Basis von $V \oplus W$; insbesondere ist also für endlichdimensionale Vektorräume V, W

$$\dim V \oplus W = \dim V + \dim W\,.$$

Insbesondere ist also

$$\underbrace{k \oplus \cdots \oplus k}_{n \text{ mal}}$$

ein n-dimensionaler Vektorraum und somit isomorph zu k^n.

Ein weiteres Beispiel liefern elektrische Netzwerke: Der Strom, der ein einzelnes Element durchfließt, wird beschrieben durch ein Element eines eindimensionalen Vektorraums über $\mathbb{R}$ oder $\mathbb{C}$, je nachdem ob man Gleichströme oder Wechselströme (einer konstanten Frequenz) betrachtet. Setzt man die einzelnen Elemente zu einem Netzwerk zusammen, wird dessen Zustand beschrieben durch ein Element der direkten Summe der Vektorräume zu den einzelnen Elementen, wobei allerdings die KIRCHHOFFschen Gesetze dafür sorgen, daß tatsächlich nur Zustände aus einem Untervektorraum dieser direkten Summe möglich sind. (Für die genaue Bestimmung dieses Untervektorraums und einer Basis davon sei auf die topologische Netzwerktheorie verwiesen, die in der

Vorlesung *Elektrotechnik* behandelt wird.) Entsprechend ist, wenn zwei oder mehrere elektrische Netzwerke miteinander in Verbindung gebracht werden, der Zustandsraum ein Untervektorraum der direkten Summe der Zustandsräume der einzelnen Netzwerke.

§7: Determinanten, Eigenwerte und Eigenvektoren

In diesem Paragraphen soll genauer erklärt werden, wie man für eine gegebene lineare Abbildung eine Basis findet, bezüglich derer ihre Abbildungsmatrix eine besonders einfache Form hat; solche Basen bestehen im Idealfall aus den sogenannten *Eigenvektoren* der linearen Abbildung. Als zwar nicht unbedingt notwendiges aber nützliches Hilfsmittel zu ihrer Berechnung betrachten wir zunächst *Determinanten,* die uns insbesondere ein kompaktes Kriterium dafür geben, wann n Vektoren aus k^n linear unabhängig sind; außerdem sind sie nützlich für Volumenberechnungen, für die Lösung nichtlinearer Gleichungssysteme und viele weitere Anwendungen.

a) Determinanten in $\mathbb{R}^2$ und $\mathbb{R}^3$

Beginnen wir mit den Vektorräumen $\mathbb{R}^2$ und $\mathbb{R}^3$, wo wir bereits Kriterien für die lineare Abhängigkeit kennen:

Zwei Vektoren im $\mathbb{R}^3$ sind genau dann linear abhängig, wenn sie dieselbe Richtung haben oder (mindestens) einer der beiden der Nullvektor ist, d.h. genau dann, wenn ihr Vektorprodukt verschwindet. Wenden wir dies insbesondere auf zwei Vektoren mit dritter Komponente Null an, so sind diese linear abhängig genau dann, wenn

$$\begin{pmatrix} u_1 \\ u_2 \\ 0 \end{pmatrix} \times \begin{pmatrix} v_1 \\ v_2 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ u_1v_2 - u_2v_1 \end{pmatrix} = \vec{0}$$

ist, also genau dann, wenn $u_1v_2 - u_2v_1$ verschwindet.

Da wir Vektoren im $\mathbb{R}^2$ mit Vektoren aus einer festen Ebenen des $\mathbb{R}^3$ identifizieren können, gilt also:

$\vec{u} = \binom{u_1}{u_2}$ und $\vec{v} = \binom{v_1}{v_2} \in \mathbb{R}^2$ sind genau dann linear abhängig, wenn $u_1v_2 - u_2v_1 = 0$ ist.

Diesen Ausdruck bezeichnen wir als *Determinante*

$$\det(\vec{u},\vec{v}) \underset{\text{def}}{=} u_1v_2 - u_2v_1$$

der beiden Vektoren.

Sowohl aus der expliziten Form wie auch aus den Eigenschaften des Vektorprodukts liest man sofort die folgenden Eigenschaften ab:

(D1) $\det(\vec{u},\vec{v}) = -\det(\vec{v},\vec{u})$

(D2) $\det(\vec{u},\vec{v})$ ist linear in jedem seiner beiden Argumente, d.h.

$$\det(\lambda\vec{u}_1 + \mu\vec{u}_2, \vec{v}) = \lambda\det(\vec{u}_1,\vec{v}) + \mu\det(\vec{u}_2,\vec{v})$$

und

$$\det(\lambda(\vec{u}, \lambda\vec{v}_1 + \mu\vec{v}_2) = \lambda\det(\vec{u},\vec{v}_1) + \mu\det(\vec{u},\vec{v}_2)\,.$$

Außerdem gilt natürlich

(D3) $\det(\vec{u},\vec{v}) = 0$ genau dann, wenn $\vec{u}$ und $\vec{v}$ linear abhängig sind.

Für den $\mathbb{R}^3$ haben wir in Aufgabe eins des achten Übungsblatts gesehen, daß drei Vektoren $\vec{u},\vec{v},\vec{w}$ genau dann linear abhängig sind, wenn ihr *Spatprodukt* $(\vec{u}\times\vec{v})\cdot\vec{w}$ verschwindet. Wir konnten dieses Produkt geometrisch interpretieren in dem Sinne, daß sein Betrag gleich dem Volumen des von $\vec{u},\vec{v},\vec{w}$ aufgespannten Parallelepipeds ist, und wir hatten es auch aus den Koordinaten der Vektoren berechnet.

Da das Volumen eines Parallelepipeds unabhängig davon ist, in welcher Reihenfolge man die aufspannenden Vektoren benennt, muß offensichtlich gelten

$$\begin{aligned} &\left|(\vec{u}\times\vec{v})\cdot\vec{w}\right| = \left|(\vec{v}\times\vec{w})\cdot\vec{u}\right| = \left|(\vec{w}\times\vec{u})\cdot\vec{v}\right| \\ =&\left|(\vec{v}\times\vec{u})\cdot\vec{w}\right| = \left|(\vec{w}\times\vec{v})\cdot\vec{u}\right| = \left|(\vec{u}\times\vec{w})\cdot\vec{v}\right|\ . \end{aligned}$$

Definieren wir also

$$\det(\vec{u},\vec{v},\vec{w}) \underset{\text{def}}{=} (\vec{u}\times\vec{v})\cdot\vec{w}\,,$$

so hängt der Betrag nicht von der Reihenfolge der Vektoren ab.

Das Vorzeichen hängt natürlich von der Reihenfolge ab, denn wegen der Schiefsymmetrie des Vektorprodukt ist beispielsweise

$$\det(\vec{v},\vec{u},\vec{w}) = -\det(\vec{u},\vec{v},\vec{w})\,;$$

entsprechend ändert sich das Vorzeichen, wie man sich leicht überzeugt, bei jeder Vertauschung der Reihenfolge zweier Vektoren. Da sowohl das Vektorprodukt als auch das Skalarprodukt linear in ihren Argumenten sind, ist auch det linear in jedem seiner drei Argumente, wir haben also wieder wie im $\mathbb{R}^2$ die drei Eigenschaften

(D1) $\det(\vec{u}, \vec{v}, \vec{w})$ ändert sein Vorzeichen, nicht aber seinen Betrag, wenn irgendwelche zwei der Argumente miteinander vertauscht werden.

(D2) $\det(\vec{u}, \vec{v}, \vec{w})$ ist linear in jedem seiner drei Argumente.

(D3) $\det(\vec{u}, \vec{v}, \vec{w}) = 0$ genau dann, wenn $\vec{u}, \vec{v}$ und $\vec{w}$ linear abhängig sind.

b) Forderungen an eine allgemeinere Determinante

Wir möchten gerne für beliebige endlichdimensionale Vektorräume V eine Determinantenfunktion nach Art des vorigen Abschnitts definieren, d.h. wir suchen für einen n-dimensionalen Vektorraum V eine Abbildung

$$\det\colon \underbrace{V \times \cdots \times V}_{n \text{ mal}} \to k; \quad (\vec{v}_1, \ldots, \vec{v}_n) \mapsto \det(\vec{v}_1, \ldots, \vec{v}_n)$$

mit den Eigenschaften

(D1) $\det(\vec{v}_1, \ldots, \vec{v}_n)$ ändert sein Vorzeichen, wenn irgendwelche zwei der Argumente miteinander vertauscht werden.

(D2) $\det(\vec{v}_1, \ldots, \vec{v}_n)$ ist linear in jedem seiner n Argumente.

(D3) $\det(\vec{v}_1, \ldots, \vec{v}_n) = 0$ genau dann, wenn $\vec{v}_1, \ldots, \vec{v}_n$ linear abhängig sind.

Um zu zeigen, daß es eine solche Abbildung auch tatsächlich gibt, folgen wir einer auch sonst bei Existenzbeiweisen oftmals nützlichen Strategie: Wir nehmen an, wir *hätten* eine Funktion det mit den Eigenschaften (D1) bis (D3) und versuchen dann, die Funktion aufgrund dieser Eigenschaften möglichst explizit auszurechnen. Dies wird auf eine Formel für det führen, die wir dann als *Definition* nehmen können, nachdem wir nachgewiesen haben, daß sie (D1) bis (D3) erfüllt.

Wir nehmen also an, wir hätten eine Funktion det mit den Eigenschaften (D1) bis (D3). Zu ihrer Berechnung wählen wir eine Basis $\vec{e}_1, \ldots, \vec{e}_n$

von V; nach Eigenschaft (D3) ist dann $\det(\vec{e}_1, \dots, \vec{e}_n) \neq 0$. Über den genauen Wert können uns die drei Bedinungen nichts sagen, denn mit det erfüllt offensichtlich auch jedes skalare Vielfache $\lambda \cdot \det$ diese Forderungen.

Für beliebige Vektoren $\vec{v}_1, \dots, \vec{v}_n$ schreiben wir

$$\vec{v}_i = \sum_{j=1}^{n} a_{ij}\vec{e}_j \ ;$$

dann ist

$$\begin{aligned}
\det(\vec{v}_1, \dots, \vec{v}_n) &= \det\Big(\sum_{j_1=1}^{n} a_{1j_1}\vec{e}_{j_1}, \vec{v}_2, \dots, \vec{v}_n\Big) \\
&\underset{(D2)}{=} \sum_{j_1=1}^{n} a_{1j_1} \det(\vec{e}_{j_1}, \vec{v}_2, \dots, \vec{v}_n)\,. \\
&= \sum_{j_1=1}^{n} a_{1j_1} \det\Big(\vec{e}_{j_1}, \sum_{j_2=1}^{n} a_{2j_2}\vec{e}_{j_2}, \vec{v}_3, \dots, \vec{v}_n\Big) \\
&\underset{(D2)}{=} \sum_{j_1=1}^{n}\sum_{j_2=1}^{n} a_{1j_1}a_{2j_2} \det(\vec{e}_{j_1}, \vec{e}_{j_2}, \vec{v}_3, \dots, \vec{v}_n) \\
&\ \ \vdots \\
&= \sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\cdots\sum_{j_n=1}^{n} a_{1j_1}a_{2j_2}\cdot\dots\cdot a_{nj_n} \det(\vec{e}_{j_1}, \cdots, \vec{e}_{j_n})\,.
\end{aligned}$$

Um dies noch weiter ausrechnen zu können, müssen wir (D1) anwenden, um die Vektoren $\vec{e}_{j_\nu}$ in jedem der Summanden in die natürliche Reihenfolge zu bringen. Nach (D1) ist

$$\det(\vec{e}_{j_1}, \cdots, \vec{e}_{j_n}) = 0\,,$$

wenn (mindestens) zwei der $\vec{e}_{j_\nu}$ übereinstimmen; ein von Null verschiedener Wert ist also nur möglich, wenn $\{j_1, \dots, j_n\} = \{1, \dots, n\}$ ist, wenn also die Abbildung

$$\pi\colon \{1, \dots, n\} \to \{1, \dots, n\};\quad \nu \mapsto j_\nu$$

bijektiv ist.

Unsere nächste Aufgabe wird daher sein, solche Abbildungen zu studieren und sie auf Vertauschungen zweier Elemente zurückzuführen.

c) Permutationen

Definition: *a)* Eine Permutation ist eine bijektive Abbildung

$$\pi: \{1, \dots, n\} \to \{1, \dots, n\}\,.$$

b) Eine Transposition ist eine Permutation τ, die zwei Elemente von $\{1, \dots, n\}$ vertauscht und den Rest festläßt. Sind i, j die beiden Elemente, die von τ vertauscht werden, so schreiben wir kurz $\tau = (i\ j)$.

c) Die Menge aller Permutationen $\pi: \{1, \dots, n\} \to \{1, \dots, n\}$ wird als *symmetrische Gruppe* $\mathfrak{S}_n$ bezeichnet.

Das Wort *Gruppe* kennen wir bereits aus den Definitionen von Körpern und Vektorräumen als Name für eine Struktur, deren Elemente man so miteinander verknüpfen kann, daß die „üblichen“ Rechenregeln gelten; wenn wir die Menge aller Permutationen als *symmetrische Gruppe* bezeichnen, sollten wir uns also zunächst davon überzeugen, daß sie tatsächlich eine Gruppe ist.

Die Gruppenoperation ist hier natürlich die Hintereinanderausführung von Abbildungen, das Produkt zweier Permutationen π und ω ist also die Abbildung $\pi \circ \omega$, die eine Zahl i abbildet auf $(\pi \circ \omega)(i) = \pi\big(\omega(i)\big)$.

Die Assoziativität der Verknüpfung ist, wie stets bei der Hintereinanderausführung von Abbildungen, klar; Einselement ist die identische Permutation, und Inverse gibt es, da eine Permutation nach Definition bijektiv ist und somit eine Umkehrabbildung hat.

Speziell für Transpositionen ist die Bestimmung dieser Umkehrabbildung besonders einfach: Da eine Transposition nicht anderes tut, als zwei Elemente miteinander zu vertauschen, macht sie sich selbst rückgängig und ist somit ihr eigenes Inverses.

Satz: Jede Permutation π kann als Produkt

$$\pi = \tau_1 \circ \cdots \circ \tau_r$$

von Transpositionen geschrieben werden.

Beweis: Jeder gängige Sortieralgorithmus liefert ein konstruktives Verfahren, um Permutationen als Produkte von Transpositionen zu schreiben: Soll etwa π zerlegt werden, so sortiere man durch fortgesetztes Vertauschungen zweier Zahlen die Folge

$$\pi^{-1}(1),\ \pi^{-1}(2),\ \pi^{-1}(3),\ \ldots, \pi^{-1}(n-1),\ \pi^{-1}(n)$$

der Größe nach. Für das Produkt der angewandeten Vertauschungen τ_1 bis τ_r ist dann $\pi^{-1} \circ \tau_1 \circ \cdots \circ \tau_r$ = Identität und damit

$$\pi = \pi \circ \text{Identität} = \tau_1 \circ \cdots \circ \tau_r \pi\ ,$$

wie gewünscht. ∎

Die Anzahl n der benötigten Vertauschungen hängt natürlich von der Effiziens des gewählten Sortierverfahrens ab; die Darstellung von π als Produkt von Transpositionen ist also alles andere als eindeutig. Wir wollen einsehen, daß wenigstens die Parität von r eindeutig bestimmt ist, d.h. r ist für gegebenes π entweder immer gerade oder immer ungerade.

Dazu definieren wir für jede Permutation $\pi \in \mathfrak{S}_n$ ein Vorzeichen:

Definition: Das *Vorzeichen* der Permutation $\pi \in \mathfrak{S}_n$ ist die Zahl

$$\varepsilon(\pi) \underset{\text{def}}{=} \prod_{\{i,j\}\subset\{1,\ldots,n\}} \frac{\pi(j)-\pi(i)}{j-i}\ ,$$

wobei das Produkt über die sämtlichen zweielementigen Teilmengen von $\{1,\ldots,n\}$ genommen wird.

$\varepsilon(\pi)$ ist wirklich ein Vorzeichen, das heißt gleich ± 1, denn

$$\left|\varepsilon(\pi)\right| = \frac{\prod\limits_{\{i,j\}\subset\{1,\ldots,n\}} |\pi(j)-\pi(i)|}{\prod\limits_{\{i,j\}\subset\{1,\ldots,n\}} |j-i|} = 1\ ,$$

weil mit $\{i,j\}$ wegen der Bijektivität von π auch $\{\pi(i),\pi(j)\}$ die sämtlichen zweielementigen Teilmengen von $\{1,\ldots,n\}$ durchläuft, so daß im Zähler und Nenner bis auf Reihenfolge genau dieselben Faktoren stehen.

Für eine Transposition $\tau = (k\ \ell)$ ist

$$\frac{\tau(j)-\tau(i)}{j-i} = \frac{j-i}{j-i} = 1 \quad \text{falls} \quad \{i,j\} \cap \{k,l\} = \emptyset\,,$$

da $(k\ \ell)$ alle Zahlen außer k und ℓ festläst. Die Zahlen k und ℓ werden vertauscht, daher ist

$$\frac{\tau(k)-\tau(\ell)}{k-\ell} = \frac{\ell-k}{k-\ell} = -1\,.$$

Bleibt noch der Fall, daß eine der beiden Zahlen i oder j gleich k oder ℓ ist; da wir zweielementige *Mengen* betrachten und $\{i,j\} = \{j,i\}$ ist, können wir vereinbaren, daß diese Zahl i sein soll.

Ist $i = k$, so ist das Produkt der Terme für $\{k,j\}$ und $\{\ell,j\}$ gleich

$$\frac{\tau(k)-\tau(j)}{k-j} \cdot \frac{\tau(\ell)-\tau(j)}{\ell-j} = \frac{\ell-j}{k-j} \cdot \frac{k-j}{\ell-j} = +1\,;$$

und genau entsprechend kann man natürlich auch für $i = \ell$ argumentieren.

Bildet man daher das Produkt über alle zweielementigen Teilmengen von $\{1,\ldots,n\}$, so erhält man

$$\varepsilon(\tau) = -1\,,$$

wir haben also gezeigt:

Lemma: Das Vorzeichen einer Transposition ist -1. ■

Der Zusammenhang zwischen dem Vorzeichen und der Transpositionsdarstellung einer Permutation ergibt sich aus

Lemma: Für zwei Permutationen $\pi, \pi' \in \mathfrak{S}_n$ ist

$$\varepsilon(\pi \circ \pi') = \varepsilon(\pi) \cdot \varepsilon(\pi')\,.$$

Beweis: Nach Definition ist

$$\varepsilon(\pi \circ \pi') = \prod_{\{i,j\}\subset\{1,\ldots,n\}} \frac{(\pi \circ \pi')(j) - (\pi \circ \pi')(i)}{j-i}$$
$$= \prod_{\{i,j\}\subset\{1,\ldots,n\}} \frac{\pi\big(\pi'(j)\big) - \pi\big(\pi'(i)\big)}{j-i}\,.$$

Nach Erweiterung mit

$$\prod_{\{i,j\}\subset\{1,\ldots,n\}} \frac{\pi'(j) - \pi'(i)}{\pi'(j) - \pi'(i)}$$

und Umordnung wird dies zu

$$\varepsilon(\pi \circ \pi') = \prod_{\{i,j\}} \frac{\pi\big(\pi'(j)\big) - \pi\big(\pi'(i)\big)}{\pi'(j) - \pi'(i)} \cdot \prod_{\{i,j\}} \frac{\pi'(j) - \pi'(i)}{j-i}\,.$$

Das zweite dieser Produkte ist natürlich $\varepsilon(\pi')$, und das erste ist $\varepsilon(\pi)$, da mit $\{i,j\}$ auch $\{\pi'(i), \pi'(j)\}$ die sämtlichen zweielementigen Teilmengen von $\{1,\ldots,n\}$ durchläuft.

■

Sind nun

$$\pi = \tau_1 \circ \cdots \circ \tau_r = \sigma_1 \circ \cdots \circ \sigma_s$$

zwei Darstellungen einer Permutation π als Produkt von Transpositionen, so ist nach den gerade bewiesenen Lemmata einerseits

$$\varepsilon(\pi) = \prod_{i=1}^{r} \varepsilon(\tau_i) = (-1)^r,$$

andererseits aber auch

$$\varepsilon(\pi) = \prod_{i=1}^{s} \varepsilon(\sigma_i) = (-1)^s,$$

also muß $(-1)^r = (-1)^s$ sein und damit r, s entweder beide gerade oder beide ungerade.

Definition: Eine Permutation $\pi \in \mathfrak{S}_n$ heißt *gerade,* wenn sie als Produkt einer geraden Anzahl von Transpositionen geschrieben werden kann, ansonsten heißt sie *ungerade.*

Das Vorzeichen $\varepsilon(\pi)$ von π ist also +1 für gerades π und -1 für ungerades.

d) Existenz von Determinanten

Nach diesem Einschub über Permutationen und Transpositionen haben wir das nötige Rüstzeug, um die Rechnung aus Abschnitt *b)* zu Ende zu führen:

Da die Determinante bei jeder Vertauschung zweier Elemente ihr Vorzeichen wechselt, folgt sofort aus den obigen Betrachtungen, daß für jede Permutation $\pi \in \mathfrak{S}_n$ gilt

$$\det\bigl(\vec{e}_{\pi(1)}, \dots, \vec{e}_{\pi(n)}\bigr) = \begin{cases} +1 & \text{falls } \pi \text{ gerade} \\ -1 & \text{falls } \pi \text{ ungerade} \end{cases},$$

d.h. $\det\bigl(\vec{e}_{\pi(1)}, \dots, \vec{e}_{\pi(n)}\bigr) = \varepsilon(\pi)$. Wenn wir die Rechnung aus Abschnitt *b)* fortführen, erhalten wir daher

$$\begin{aligned} \det\bigl(\vec{v}_1, \dots, \vec{v}_n\bigr) &= \sum_{\pi \in \mathfrak{S}_n} a_{1\pi(1)} \cdots a_{n\pi(n)} \cdot \det\bigl(\vec{e}_{\pi(1)}, \dots, \vec{e}_{\pi(n)}\bigr) \\ &= \left(\sum_{\pi \in \mathfrak{S}_n} \varepsilon(\pi) a_{1\pi(1)} \cdots a_{n\pi(n)} \right) \det\bigl(\vec{e}_1, \dots, \vec{e}_n\bigr) . \end{aligned}$$

Somit ist $\det\bigl(\vec{v}_1, \dots, \vec{v}_n\bigr)$ eindeutig bestimmt bis auf den Wert von $\det\bigl(\vec{e}_1, \dots, \vec{e}_n\bigr)$; falls es überhaupt Determinanten gibt, müssen sie also so aussehen. Wir definieren daher

Definition: Die Determinante von n Vektoren

$$\vec{v}_1 = a_{11}\vec{e}_1 + \cdots + a_{1n}\vec{e}_n, \quad \dots, \quad \vec{v}_n = a_{n1}\vec{e}_1 + \cdots + a_{nn}\vec{e}_n$$

des n-dimensionalen Vektorraums V bezüglich der Basis $\vec{e}_1, \dots, \vec{e}_n$ ist

$$\det\bigl(\vec{v}_1, \dots, \vec{v}_n\bigr) = \sum_{\pi \in \mathfrak{S}_n} \varepsilon(\pi) a_{1\pi(1)} \cdots a_{n\pi(n)} .$$

Insbesondere ist also $\det(\vec{e}_1, \dots, \vec{e}_n) = 1$.

Satz: Die so definierte Determinante hat die Eigenschaften (D1) bis (D3).

Beweis: Beginnen wir mit (D1), d.h. die Determinante wechselt ihr Vorzeichen, wenn zwei der Argumente vertauscht werden. Vertauscht man etwa die Argumente $\vec{v}_i$ und $\vec{v}_j$, so werden anstelle der Summanden

$$\varepsilon(\pi)a_{1\pi(1)}\cdots a_{i\pi(i)}\cdots a_{j\pi(j)}\cdots a_{n\pi(n)}$$

die Summanden

$$\varepsilon(\pi')a_{1\pi(1)}\cdots a_{j\pi(j)}\cdots a_{i\pi(i)}\cdots a_{n\pi(n)}$$

aufaddiert, wobei

$$\pi'(k) = \begin{cases} \pi(k) & \text{falls } k \neq i, j \\ \pi(j) & \text{falls } k = i \\ \pi(i) & \text{falls } k = j \end{cases}$$

ist, d.h. $\pi' = \pi \circ (i\ j)$. Als Produkt von Transpositionen geschrieben hat π' also einen Faktor mehr als π; falls π ungerade war, ist daher π' gerade, und umgekehrt. Somit ist

$$\varepsilon(\pi') = -\varepsilon(\pi)$$

und damit auch

$$\begin{aligned} &\varepsilon(\pi')a_{1\pi(1)}\cdots a_{j\pi(i)}\cdots a_{i\pi(j)}\cdots a_{n\pi(n)} \\ &= -\varepsilon(\pi)a_{1\pi(1)}\cdots a_{i\pi(i)}\cdots a_{j\pi(j)}\cdots a_{n\pi(n)}\,. \end{aligned}$$

Durch Aufsummieren folgt

$$\det\big(\vec{v}_1,\ldots,\vec{v}_j,\ldots,\vec{v}_i,\ldots,\vec{v}_n\big) = -\det\big(\vec{v}_1,\ldots,\vec{v}_i,\ldots,\vec{v}_j,\ldots,\vec{v}_n\big)\,.$$

Ist in dieser Formel $\vec{v}_i = \vec{v}_j$, so ändert sich durch die Vertauschung von $\vec{v}_i$ und $\vec{v}_j$ garnichts, wir haben also auch die Formel

$$\det\big(\vec{v}_1,\ldots,\vec{v}_j,\ldots,\vec{v}_i,\ldots,\vec{v}_n\big) = \det\big(\vec{v}_1,\ldots,\vec{v}_i,\ldots,\vec{v}_j,\ldots,\vec{v}_n\big)\,,$$

und daraus folgt zumindest für $k = \mathbb{R}$, daß

$$\det\big(\vec{v}_1,\ldots,\vec{v}_i,\ldots,\vec{v}_i,\ldots,\vec{v}_n\big) = 0$$

sein muß.

Für den Beweis von Eigenschaft (D3) werden wir diese Formel benötigen, allerdings haben wir sie bislang nur bewiesen für $k = \mathbb{R}$. Tatsächlich funktioniert obiger Beweis für jeden Körper, in dem man aus $x = -x$ folgern kann, daß $x = 0$ sein muß. Leider spielen aber in der Informationsverarbeitung auch Körper eine Rolle, in denen $x + x = 0$ ist für *jedes* Körperelement x, und für solche Körper folgt aus $x = -x$ überhaupt nichts.

Im Hinblick auf solche Körper muß die Formel daher noch einmal bewiesen werden, und zwar auf eine Art und Weise, die für jeden Körper funktioniert. Dazu beachten wir, daß im Falle $\vec{v}_i = \vec{v}_j$ auch für alle ℓ die Koeffizienten $a_{i\ell}$ und $a_{j\ell}$ übereinstimmen, d.h. in der Summe

$$\det A = \sum_{\pi \in \mathfrak{S}_n} \varepsilon(\pi) a_{1\pi(1)} \cdots a_{i\pi(i)} \cdots a_{j\pi(j)} \cdots a_{n\pi(n)}$$

ist stets

$$a_{1\pi(1)} \cdots a_{i\pi(i)} \cdots a_{j\pi(j)} \cdots a_{n\pi(n)} = a_{1\pi(1)} \cdots a_{i\pi(j)} \cdots a_{j\pi(i)} \cdots a_{n\pi(n)} .$$

Hier ist die rechte Seite gerade der Summand zur Permutation $\pi \circ (i\ j)$, die genau dann gerade ist, wenn π ungerade ist, und umgekehrt. Also haben beide Terme verschiedenes Vorzeichen und heben sich gegenseitig weg; die Summe über alle Permutationen ist also Null.

Eigenschaft (D2), die Linearität in jedem der Argumente, ist klar, denn jedes der Monome $a_{1\pi(1)} \cdots a_{n\pi(n)}$ enthält *genau eine* Komponente des Vektors $\vec{v}_i$ und ist damit linear in $\vec{v}_i$.

Bleibt noch (D3), d.h. $\det\bigl(\vec{v}_1, \ldots, \vec{v}_n\bigr)$ soll genau dann verschwinden, wenn die $\vec{v}_i$ linear abhängig sind.

Sind zunächst die Vektoren $\vec{v}_1, \ldots, \vec{v}_n$ linear abhängig, so läßt sich einer von ihnen als Linearkombination der anderen schreiben; durch Umordnen (wobei die Determinante höchstens ihr Vorzeichen ändert) können wir annehmen, daß dies $\vec{v}_1$ sei und etwa

$$\vec{v}_1 = \sum_{i=2}^{n} \lambda_i \vec{v}_i .$$

Dann ist

$$\det\bigl(\vec{v}_1,\ldots,\vec{v}_n\bigr) = \det\left(\sum_{i=2}^{n}\lambda_i\vec{v}_i,\vec{v}_2,\ldots,\vec{v}_n\right)$$

$$\underset{(D2)}{=} \sum_{i=2}^{n}\lambda_i \det\bigl(\vec{v}_i,\vec{v}_2,\ldots,\vec{v}_n\bigr) \underset{(D1)}{=} 0\,,$$

denn im i-ten Summanden der letzten Summe tritt der Vektor $\vec{v}_i$ zweimal als Argument von det auf. Also verschwindet die Determinante, falls die Vektoren linear abhängig sind.

Nun seien $\vec{v}_1,\ldots,\vec{v}_n$ linear unabhängig; für dieses Fall müssen wir zeigen, daß die Determinante *nicht* verschwindet.

Wir wissen, daß für die Basisvektoren $\vec{e}_1,\ldots,\vec{e}_n$

$$\det\bigl(\vec{e}_1,\ldots,\vec{e}_n\bigr) = 1 \neq 0$$

ist; außerdem wissen wir, daß n linear unabhängige Vektoren eines n-dimensionalen Vektorraums eine Basis bilden, so daß wir die $\vec{e}_i$ als Linearkombinationen der $\vec{v}_j$ schreiben können, etwa als

$$\vec{e}_i = \sum_{j=1}^{n} b_{ij}\vec{v}_j\ .$$

Nach der Rechnung des letzten Paragraphen ist dann, da det den Forderungen (D1) und (D2) genügt und bei zwei gleichen Agrumenten verschwindet,

$$\det\bigl(\vec{e}_1,\ldots,\vec{e}_n\bigr) = \left(\sum_{\pi\in\mathfrak{S}_n}\varepsilon(\pi)b_{1\pi(1)}\cdots b_{n\pi(n)}\right)\det\bigl(\vec{v}_1,\ldots,\vec{v}_n\bigr)\ .$$

In dieser Gleichung steht links die Zahl eins, also kann keiner der beiden Faktoren auf der rechten Seite verschwinden; insbesondere ist also $\det\bigl(\vec{v}_1,\ldots,\vec{v}_n\bigr) \neq 0$.

Damit ist der Satz vollständig bewiesen. ■

Fassen wir zusammen, was wir bislang in diesem Paragraphen gemacht haben:

- Wir haben anhand des Beispiels von Determinanten im $\mathbb{R}^2$ und $\mathbb{R}^3$ drei zentrale Forderungen an eine allgemeine Determinante aufgestellt.
- Unter der Annahme, daß es Funktionen gibt, die diese drei Forderungen erfüllen, haben wir ausgerechnet, wie solche Funktionen aussehen müssen; inbsbesondere haben wir gesehen, daß sie durch ihren Wert auf den Vektoren einer Basis eindeutig bestimmt sind.
- Dann haben wir die so erhaltene explizite Formel als *Definition* einer allgemeinen Determinanten genommen und gezeigt, daß die so definierte Funktion tatsächlich die drei Forderungen erfüllt.

Insgesamt haben wir also gesehen, daß es Funktionen gibt, die den drei Forderungen (D1) bis (D3) genügen und daß solche Funktionen bis auf eine multiplikative Konstante eindeutig bestimmt sind.

e) Die Determinante einer Matrix

Definition: Unter der Determinante einer Matrix $A \in k^{n\times n}$ verstehen wir die Determinante ihrer Spaltenvektoren; ist also

$$A = (a_{ij}) \quad \text{und} \quad \vec{v}_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix},$$

so ist

$$\det A \underset{\text{def}}{=} \det\bigl(\vec{v}_1, \dots, \vec{v}_n\bigr) .$$

Wir schreiben auch kurz

$$\det A = |A| = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix} .$$

Um etwas Übung im Umgang mit Determinanten zu bekommen, wollen wir dies für kleine n explizit ausrechnen:

Der Fall $n = 1$ einer 1×1-Matrix ist natürlich völlig uninteressant: Die einzige Permutation einer einelementigen Menge ist die Identität, die für jedes n gerade ist, d.h. für $A = (a) \in k^{1\times 1}$ ist $\det A = a$. Die Schreibweise $\det A = |a|$ ist hier irreführend, da es zumindest für $k = \mathbb{R}$ oder $\mathbb{C}$ zu Verwechslungen mit der Betragsfunktion kommen kann. Wegen der geringen Bedeutung von 1×1-Matrizen und ihren Determinanten ist dies jedoch nicht weiter schlimm.

Für $n = 2$ gibt es zwei Permutationen: Die (gerade) Identität und die (ungerade) Transposition (1 2). Nach Definition der Determinanten ist

$$\det\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}\,,$$

wie beim bereits in Abschnitt *a)* betrachtete Kreuzprodukt der Vektoren
$\begin{pmatrix} a_{11} \\ a_{21} \\ 0 \end{pmatrix}$ und $\begin{pmatrix} a_{12} \\ a_{22} \\ 0 \end{pmatrix}$.

Für $n = 3$ gibt es bereits $3! = 6$ Permutationen, darunter natürlich die (gerade) Identität.

Jede weitere Permutation kann höchstens ein Element festlassen, denn würde sie zwei festlassen, müßte wegen der Bijektivität der Abbildung auch das dritte auf sich selbst abgebildet werden, wir hätten also die Identität.

Falls ein Element festgehalten wird, bleibt einer nichtidentischen Permutation daher nichts anderes übrig, als die beiden anderen zu vertauschen; dies führt auf die drei (ungeraden) Transpositionen (2 3), (1 3) und (1 2).

Falls kein Element auf sich selbst abgebildet wird, muß die Eins entweder auf 2 oder auf 3 abgebildet werden. Im ersten Fall geht dann 2 auf 3, denn sonst müßte zwei entweder festbleiben oder wir hätten die Transposition (1 2), die drei auf sich selbst abbildet. Wegen der Bijektivität der Abbildung gibt es dann keine andere Möglichkeit als daß drei auf eins abgebildet wird; wir haben also die zyklische Vertauschung

$$1 \mapsto 2 \mapsto 3 \mapsto 1\,.$$

Diese Permutation ist gerade, denn sie läßt sich beispielsweise als Produkt (1 3)(1 2) zweier Transpositionen schreiben: Wenden wir dieses

Produkt an auf 1, so wird die Eins zunächst von (1 2) auf zwei abgebildet, und zwei bleibt fest unter (1 3), so daß insgesamt eins auf zwei abgebildet wird, wie verlangt.

Zwei wird von (1 2) auf die Eins abgebildet, und die wiederum von (1 3) auf drei, also haben wir auch hier insgesamt das richtige Ergebnis. Drei schließlich kann dann wegen der Bijektivität von (1 3)(1 2) nur auf eins abgebildet werden, was man auch schnell direkt sieht, denn (1 2) läßt die Drei unverändert, während (1 3) sie auf eins abbildet.

Die noch verbleibende Permutation bildet eins auf drei und daher drei auf zwei ab, sie ist also die zyklische Vertauschung

$$1 \mapsto 3 \mapsto 2 \mapsto 1 \quad \text{oder} \quad 3 \mapsto 2 \mapsto 1 \mapsto 3\,,$$

d.h. die Umkehrabbildung zur gerade betrachteten Permutation. Daher ist sie als Produkt von Transpositionen gleich

$$\big((1\ 2)(1\ 3)\big)^{-1} = (1\ 3)^{-1}(1\ 2)^{-1} = (1\ 3)(1\ 2)$$

und somit auch gerade.

Zur Berechnung der Determinanten einer 3×3-Matrix (a_{ij}) beginnen wir mit den geraden Permutationen: Die Identität liefert den Term $a_{11}a_{22}a_{33}$, d.h. das Produkt der drei Diagonalelemente, die zyklische Vertauschung $1 \to 2 \to 3 \to 1$ liefert das Produkt $a_{12}a_{23}a_{31}$, und ihr Inverses $a_{13}a_{21}a_{32}$.

Die drei Transpositionen führen auf die negativ zu nehmenden Produkte $a_{11}a_{23}a_{32}$, $a_{13}a_{22}a_{31}$ und $a_{12}a_{21}a_{33}$, insgesamt ist also

$$\begin{aligned}\det A = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} &= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ &\qquad - a_{11}a_{23}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33}\,.\end{aligned}$$

Diese Formel läßt sich leicht merken nach der Regel von PIERRE SARRUS

(1798–1861): Schreibt man die Einträge der Matrix in der Form

$$\begin{array}{ccccccccc}
a_{11} & & a_{12} & & a_{13} & & a_{11} & & a_{12} \\
& \searrow & & \swarrow\!\!\!\!\searrow & & \swarrow\!\!\!\!\searrow & & \swarrow & \\
a_{21} & & a_{22} & & a_{23} & & a_{21} & & a_{22} \\
& \swarrow & & \swarrow\!\!\!\!\searrow & & \swarrow\!\!\!\!\searrow & & \searrow & \\
a_{31} & & a_{32} & & a_{33} & & a_{31} & & a_{32}
\end{array},$$

so sind die Produkte entlang der drei schrägen Linien von links oben nach rechts unten, der Hauptdiagonalen und ihrer Parallelen also, positiv zu rechnen und die entlang der drei schrägen Linien von rechts oben nach links unten negativ.

Für die Determinanten von 4×4-Matrizen gibt es keine entsprechende Regel mehr; da dann bereits $4! = 24$ Summanden berücksichtigt werden müssen, eignet sich die Formel aus der Definition der Determinanten nicht mehr gut zur Berechnung, von noch größeren Matrizen ganz zu schweigen: Wie wir aus Kapitel I, §4*g2* wissen, ist $\log n! \approx n \log n$, d.h. die Anzahl $n!$ der Summanden wächst mit größer werdendem n noch schneller als die Exponentialfunktion.

Wir werden im folgenden daher auf Verfahren hinarbeiten, die es erlauben, auch größere Determinanten mit vertretbarem Aufwand zu berechnen. Davor sollten wir aber zunächst einige grundlegende Rechenregeln für Determinanten von Matrizen kennenlernen.

Fast unmittelbar aus der Definition folgt

Lemma: Für jede $n \times n$-Matrix A ist $\det A = \det {}^tA$.

Beweis: Nach Definition ist für $A = (a_{ij})$

$$\det A = \sum_{\pi \in \mathfrak{S}_n} \varepsilon(\pi) a_{1\pi(1)} \cdots a_{n\pi(n)}$$

und

$$\det {}^tA = \sum_{\pi \in \mathfrak{S}_n} \varepsilon(\pi) a_{\pi(1)1} \cdots a_{\pi(n)n} .$$

Ordnet man die Faktoren des Produkts $a_{\pi(1)1} \cdots a_{\pi(n)n}$ nach dem ersten Index, so erhält man

$$a_{\pi(1)1} \cdots a_{\pi(n)n} = a_{1\pi^{-1}(1)} \cdots a_{1\pi^{-1}(n)}$$

und damit

$$\det {}^tA = \sum_{\pi\in\mathfrak{S}_n} \varepsilon(\pi) a_{1\pi^{-1}(1)} \cdots a_{1\pi^{-1}(n)} .$$

Da mit π auch π^{-1} die gesamte Gruppe $\mathfrak{S}_n$ durchläuft und $\varepsilon(\pi) = \varepsilon(\pi^{-1})$ ist, stimmt die Summe dieser Terme mit der aus der Formel für $\det A$ überein, die beiden Determinanten sind also gleich.

■

Für Leser, denen nicht klar ist, warum $\varepsilon(\pi)$ und $\varepsilon(\pi^{-1})$ gleich sind, seien zur Übung des Umgangs mit Permutationen hier kurz zwei Beweise angegeben: Am einfachsten sieht man diese Formel aufgrund des obigen Lemmas, wonach für zwei beliebige Permutationen π und π' gilt: $\varepsilon(\pi \circ \pi') = \varepsilon(\pi) \cdot \varepsilon(\pi')$. Setzt man hier $\pi' = \pi^{-1}$, so folgt

$$1 = \varepsilon(\text{Identität}) = \varepsilon(\pi) \cdot \varepsilon(\pi^{-1}) ,$$

so daß die beiden Vorzeichen gleich sein müssen.

Alternativ können wir uns auch explizit überlegen, daß sich π^{-1} als Produkt von r Transpositionen schreiben läßt, falls dies für π der Fall ist; genauer gilt: Ist $\pi = \tau_1 \circ \cdots \circ \tau_r$, so ist $\pi^{-1} = \tau_r \circ \cdots \circ \tau_1$. Wie wir bereits wissen, ist nämlich jede Transposition zu sich selbst invers, d.h.

$$\begin{aligned}
\pi \circ (\tau_r \circ \cdots \circ \tau_1) &= (\tau_1 \circ \cdots \circ \tau_r) \circ (\tau_r \circ \cdots \circ \tau_1) \\
&= \tau_1 \circ \cdots \circ \tau_{r-1} \circ (\tau_r \circ \tau_r) \circ \tau_{r-1} \circ \cdots \circ \tau_1 \\
&= \tau_1 \circ \cdots \circ \tau_{r-2} \circ (\tau_{r-1} \circ \tau_{r-1}) \circ \tau_{r-2} \circ \cdots \circ \tau_1 \\
&= \tau_1 \circ \cdots \circ \tau_{r-3} \circ (\tau_{r-2} \circ \tau_{r-2}) \circ \tau_{r-3} \cdots \circ \tau_1 \\
&= \cdots = \tau_1 \circ (\tau_2 \circ \tau_2) \circ \tau_1 = \tau_1 \circ \tau_1 = \text{Identität} ,
\end{aligned}$$

so daß in der Tat $\pi^{-1} = \tau_r \circ \cdots \circ \tau_1$ ist.

Aus obigem Lemma folgt insbesondere, daß man die Determinante einer Matrix auch als Determinante der Zeilenvektoren definieren kann und daß die Spaltenvektoren einer Matrix genau dann linear unabhängig sind, wenn es die Zeilenvektoren sind. Mit nur wenig mehr Aufwand könnte man sogar zeigen, daß der im Zusammenhang mit der Lösung linearer Gleichungssysteme definierte *Rang* einer Matrix genausogut über Zeilen- wie über Spaltenoperationen berechnet werden kann.

Wichtiger ist der *Multiplikationssatz:*

Satz: *a)* Für zwei Matrizen $A, B \in k^{n\times n}$ ist

$$\det(AB) = \det A \cdot \det B\ .$$

b) Für eine invertierbare Matrix $A \in k^{n\times n}$ ist

$$\det(A^{-1}) = (\det A)^{-1}\ .$$

Beweis: a) Die Determinante von AB ist die Determinante der Spaltenvektoren $\vec{v}_1, \ldots, \vec{v}_n$ von AB, die von B ist gleich der Determinante der Spaltenvektoren $\vec{b}_1, \ldots, \vec{b}_n$ von B, und nach Definition der Matrizenmultiplikation ist mit $A = (a_{ij})$

$$\vec{v}_i = \sum_{j=1}^{n} a_{ij}\vec{b}_j\ .$$

Damit sind wir in der Situation der Abschnitte *b)* und *d)* (wobei hier die $\vec{b}_j$ die Rolle der dortigen $\vec{e}_j$ übernehmen) und können aus den Eigenschaften (D1) bis (D3) folgern, daß

$$\det\left(\vec{v}_1, \ldots, \vec{v}_n\right) = \left(\sum_{\pi\in\mathfrak{S}_n} \varepsilon(\pi) a_{1\pi(1)} \cdots a_{n\pi(n)}\right) \cdot \det\left(\vec{b}_1, \ldots, \vec{b}_n\right)$$

ist. Das ist aber genau die Formel

$$\det(AB) = (\det A) \cdot (\det B)\ .$$

b) Für eine invertierbare Matrix A ist $A \cdot A^{-1} = E$ die Einheitsmatrix, und deren Determinante ist gleich der Determinante der Basisvektoren in ihrer natürlichen Reihenfolge, also 1. Daher ist nach Teil *a)*

$$\det A \cdot \det\left(A^{-1}\right) = \det E = 1\ ,$$

woraus die Behauptung folgt.

■

Als erste Anwendung dieses Satzes können wir Determinanten von linearen Abbildungen $\varphi\colon V \to V$ definieren: Wir setzen einfach

$$\det \varphi \underset{\text{def}}{=} \det A_{\mathcal{B}}\ ,$$

wobei A_φ die Abbildungsmatrix von φ bezüglich irgendeiner Basis $\mathcal{B}$ von V ist. Diese Abbildungsmatrix hängt natürlich stark von der Basis $\mathcal{B}$ ab: Ist $\mathcal{C}$ eine andere Basis, so wissen wir aus §5*h,* daß es eine invertierbare Matrix M gibt, so daß $A_\mathcal{C} = MA_\mathcal{B}M^{-1}$; aus dem Multiplikationssatz folgt daher

$$\det(A_\mathcal{C}) = \det(MA_\mathcal{B}M^{-1}) = \det(M)\det(A_\mathcal{B})\det(M)^{-1} = \det(A_\mathcal{B})\,,$$

d.h. die beiden Determinanten sind gleich.

Der Multiplikationssatz läßt sich auch in ein Berechnungsverfahren für Determinanten übersetzen: Falls wir die LR-Zerlegung $LA = R$ der Matrix A kennen, sagt uns der Satz, daß

$$\det A = \frac{\det R}{\det L}$$

ist. Die Determinanten von L und R lassen sich leicht ausrechnen nach dem folgenden

Lemma: Die Determinante einer (unteren oder oberen) Dreiecksmatrix ist gleich dem Produkt ihrer Diagonalelemente.

Beweis: $A = (a_{ij})$ sei eine Dreiecksmatrix. Wie üblich ist

$$\det A = \sum_{\pi\in\mathfrak{S}_n} \varepsilon(\pi) a_{1\pi(1)} \cdots a_{n\pi(n)}\,.$$

Nun ist aber für eine untere *bzw.* obere Dreiecksmatrix $a_{ij} = 0$ falls $i < j$ *bzw.* $i > j$ ist. Da Permutationen bijektive Abbildungen sind, muß es für jede Permutation außer der Identität mindestens ein i geben mit $\pi(i) < i$ und mindestens ein j mit $\pi(j) > j$, d.h. das Produkt $a_{1\pi(1)} \cdots a_{n\pi(n)}$ enthält für jede nichtidentische Permutation mindestens einen Faktor Null. Damit kann höchstens die Identität einen von Null verschiedenen Beitrag zur Summe liefern; da sie Vorzeichen +1 hat ist also

$$\det A = a_{11} \cdots a_{nn}$$

das Produkt der Diagonalelemente von A. ∎

So wie wir die LR-Zerlegung definiert haben, ist L eine untere Dreiecksmatrix mit lauter Einsen in der Hauptdiagonale und R eine beliebige obere Dreiecksmatrix. Also ist $\det L = 1$, die Determinante von R ist gleich dem Produkt der Diagonalelemente von R, und damit ist auch

$$\det A = \text{Produkt der Diagonalelemente von } R\,.$$

Zumindest für große n ermöglicht dies eine erheblich effizientere Berechnung von Determinanten als die definierende Formel: Der Aufwand für eine LR-Zerlegung ist ungefähr proportional zu n^3, was für große n weitaus günstiger ist als der überexponentiell steigende Aufwand proportional $n!$ für die Summe aus der definierenden Formel.

Für mittlere n, für die man Determinanten noch von Hand ausrechnen kann, ist allerdings oft ein anderes $n!$-Verfahren günstiger als die LR-Zerlegung; damit wollen wir uns im nächsten Abschnitt beschäftigen.

f) Der Entwicklungssatz von Laplace

Hier geht es um ein Berechnungsverfahren, das die definierenden Eigenschaften der Determinante ausnutzt, die „Entwicklung nach einer Zeile oder Spalte“.

Dazu schreiben wir, wenn wir beispielsweise nach der i-ten Spalte entwickeln wollen, den i-ten Spaltenvektor

$$\vec{v}_i = \begin{pmatrix} a_{1i} \\ \vdots \\ a_{ni} \end{pmatrix}$$

als Linearkombination der Einheitsvektoren

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \dots \quad \vec{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix},$$

d.h. als

$$\vec{v}_i = a_{1i}\vec{e}_1 + \cdots + a_{ni}\vec{e}_n\,.$$

Dann ist wegen der Linearität der Determinante in ihrem i-ten Argument

$$\det A = \det\bigl(\vec{v}_1, \ldots, \vec{v}_i, \ldots, \vec{v}_n\bigr) = \det\Bigl(\vec{v}_1, \ldots, \sum_{j=1}^{n} a_{ji}\vec{e}_j, \ldots, \vec{v}_n\Bigr)$$

$$= \sum_{j=1}^{n} a_{ji} \det\bigl(\vec{v}_1, \ldots, \vec{e}_j, \ldots, \vec{v}_n\bigr) .$$

Um daraus die Determinante von A zu berechnen, müssen wir die Determinanten

$$D_{ij} \underset{\text{def}}{=} \det\bigl(\vec{v}_1, \ldots, \vec{e}_j, \ldots, \vec{v}_n\bigr)$$

kennen. Dazu wenden wir die definierende Formel an: In den Produkten $a_{1\pi(1)} \cdots a_{n\pi(n)}$ ist für die gewählte Stelle i, wo wir den Vektor $\vec{e}_j$ eingefügt haben, $a_{i\ell} = 0$ für $\ell \neq j$ und $a_{ij} = 1$. Also verschwindet $a_{1\pi(1)} \cdots a_{n\pi(n)}$ für jede Permutation π mit $\pi(i) \neq j$.

Am einfachsten läßt sich dies im Fall $i = j = n$ weiter ausrechnen: Dann müssen wir nur Permutationen mit $\pi(n) = n$ betrachten, und die entsprechen eineindeutig den Permutationen auf der Menge aller Zahlen von 1 bis $n - 1$; was übrigbleibt ist also (wegen $a_{ij} = 1$) gerade die Determinante jener $(n-1) \times (n-1)$-Matrix, die durch Streichung der letzten Spalte und der letzten Zeile aus A entsteht.

Der allgemeine Fall läßt sich durch Vertauschungen darauf zurückführen: Wir können zwar nicht einfach die i-te Spalte mit der n-ten vertauschen, denn dann steht ja die n-te Spalte an i-ter Stelle, aber wir können die i-te Spalte durch sukzessive Vertauschung mit ihrem rechten Nachbarn immer weiter nach außen bringen: Wir vertauschen zunächst die i-te Spalte mit der $(i+1)$-ten, dann die neue $(i+1)$-te (=alte i-te) Spalte mit der $(i+2)$-ten *usw.,* bis schließlich die alte i-te Spalte an n-ter Stelle steht, ohne daß sich vor oder nach der Stelle i irgendetwas an der relativen Reihenfolge der Spalten geändert hätte. Die Anzahl der dazu notwendigen Vertauschungen ist $(n - i)$, die Determinante wurde bei der gesamten Prozedur also mit $(-1)^{n-i}$ multipliziert.

Als nächstes müssen wir auch noch die j-te Zeile nach unten bringen; dazu verwenden wir die bereits gezeigte Formel $\det A = \det {}^tA$. Die j-te Zeile von A ist gleich der j-ten Spalte von ${}^tA$; wie wir oben gesehen haben, läßt sie sich durch $(n-j)$ Spaltenvertauschungen zur letzen Spalte machen, wobei die Determinante mit $(-1)^{n-j}$ multipliziert wird, und durch nochmaliges Transponieren erhalten wir schließlich jene Matrix, in der die j-te Zeile zur n-ten geworden ist, ohne daß sich vor oder nach der Stelle j irgendetwas an der relativen Reihenfolge der Spalten geändert hätte.

Die Determinante der so entstehenden Matrix ist, wie wir oben beim Spezialfall $i = j = n$ gesehen haben, gleich der Determinante jener $(n-1) \times (n-1)$-Matrix A_{ij}, die durch Streichen der letzten Zeile und Spalte entsteht; bezogen auf die ursprüngliche Matrix A entsteht sie durch Streichen der i-ten Spalte und der j-ten Zeile. Also ist, wenn wir noch die Vertauschungen berücksichtigen,

$$D_{ij} = (-1)^{n-i} \cdot (-1)^{n-j} \cdot \det A_{ij} = (-1)^{i+j} \det A_{ij} ,$$

denn $(n-i) + (n-j) = 2n - (i+j)$ hat dieselbe Parität wie $(i+j)$.

Fassen wir alles zusammen und beachten wir, daß wir wegen $\det A = \det {}^tA$ auch völlig analog nach einer Zeile entwickeln können, so erhalten wir den nach dem französischen Mathematiker PIERRE-SIMON DE LAPLACE (1749–1827) benannten *Entwicklungssatz,* der allerdings auch schon GOTTFRIED WILHELM VON LEIBNIZ (1646–1716) bekannt war:

Entwicklungssatz von Laplace: A sei eine $n \times n$-Matrix und A_{ij} sei jene $(n-1) \times (n-1)$-Matrix, die durch Streichen der i-ten Spalte und der j-ten Zeile von A entsteht. Dann ist für jedes feste i

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij}$$

und für jedes feste j

$$\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij} ,$$

wobei A_{ij} jene $(n-1) \times (n-1)$-Matrix ist, die aus A durch Streichen der i-ten Zeile und der j-ten Spalte entsteht.

■

Die erste Formel aus diesem Satz bezeichnet man als die *Entwicklung der Determinante nach ihrer i-ten Zeile,* die zweite (die durch Anwendung der ersten auf die transponierte Matrix entsteht) entsprechend als die *Entwicklung nach der j-ten Spalte.*

Rein theoretisch ist es gleichgültig, nach welcher Zeile oder Spalte man eine Determinante entwickelt, aber man wird natürlich versuchen, eine zu finden, in der möglichst viele Nullen stehen, so daß in obiger Summe möglichst wenige Summanden übrigbleiben.

Falls es keine Zeile oder Spalte mit vielen Nullen gibt, kann man gelegentlich etwas nachhelfen, indem man ein Vielfaches einer Spalte zu einer anderen addiert (oder entprechend für eine Zeile):

$$\begin{aligned} &\det\bigl(\vec{v}_1, \ldots, \vec{v}_i + \lambda\vec{v}_j, \ldots, \vec{v}_n\bigr) \\ =&\det\bigl(\vec{v}_1, \ldots, \vec{v}_i, \ldots, \vec{v}_n\bigr) + \lambda \det\bigl(\vec{v}_1, \ldots, \vec{v}_j, \ldots, \vec{v}_n\bigr) \\ =&\det\bigl(\vec{v}_1, \ldots, \vec{v}_i, \ldots, \vec{v}_n\bigr)\ , \end{aligned}$$

da der zweite Summand das Argument $\vec{v}_j$ doppelt enthält und somit verschwindet.

Betrachten wir als Beispiel die Determinante der Matrix

$$A = \begin{pmatrix} 1 & 2 & -2 & 1 \\ -1 & -3 & 2 & -2 \\ -4 & -2 & -1 & 0 \\ 3 & 2 & -1 & 1 \end{pmatrix} .$$

Sie hat zwar einen Eintrag Null, aber da in der dritten und vierten Spalte gleich in zwei Zeilen die Einträge entgegengesetz gleich sind, ist es besser, zunächst die dritte Spalte durch die Summe von dritter und vierter Spalte zu ersetzen:

$$\begin{vmatrix} 1 & 2 & -2 & 1 \\ -1 & -3 & 2 & -2 \\ -4 & -2 & -1 & 0 \\ 3 & 2 & -1 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 2 & -1 & 1 \\ -1 & -3 & 0 & -2 \\ -4 & -2 & -1 & 0 \\ 3 & 2 & 0 & 1 \end{vmatrix} .$$

Wenn wir jetzt nach der dritten Spalte entwickeln, kommen nur zwei der Summanden wirklich vor: Der erste ($i = 1$ und $j = 3$) und der dritte ($i = j = 3$). In beiden Fällen ist $i + j$ gerade, der Vorfaktor $(-1)^{i+j}$

ist also jeweils +1. Die 3×3-Determinanten, die durch Streichung der dritten Spalte und der ersten *bzw.* dritten Zeile entstehen, können nach der Regel von SARRUS berechnet werden:

$$\begin{vmatrix} -1 & -3 & -2 \\ -4 & -2 & 0 \\ 3 & 2 & 1 \end{vmatrix} = (2+0+16) - (12+0+12) = -6$$

und

$$\begin{vmatrix} 1 & 2 & 1 \\ -1 & -3 & -2 \\ 3 & 2 & 1 \end{vmatrix} = (-3-12-2) - (-9-4-2) = -2\,.$$

Also ist

$$\det A = (-1)(-6) + (-1)(-2) = 8\,.$$

g) Die Cramersche Regel

Determinanten können auch angewandt werden, um die Lösungen eines linearen Gleichungssystems vom Rang n aus n Gleichungen in n Unbekannten in geschlossener Form als Funktion der Koeffizienten darzustellen. Dazu schreiben wir das Gleichungssystem

$$A\vec{x} = \vec{b}$$

in der Form

$$x_1\vec{a}_1 + \cdots + x_n\vec{a}_n = \vec{b}\,,$$

wobei die $\vec{a}_i$ die Spaltenvektoren der Matrix A seien und $(x_1, \ldots, x_n)$ eine Lösung des linearen Gleichungssystems.

Nun ersetzen wir in

$$\det A = \det\bigl(\vec{a}_1, \ldots, \vec{a}_n\bigr)$$

rechts den Vektor $\vec{a}_i$ durch die rechte Seite $\vec{b}$ des Gleichungssystems. (Es gibt eigentlich keinen vernünftigen Grund, warum wir das tun sollten; auch dieser Trick wird, wie so viele, erst nachträglich durch das Ergebnis gerechtfertigt.) Die so entstehende Determinante ist

$$\begin{aligned}
&\det\big(\vec{a}_1, \dots, \vec{b}, \dots, \vec{a}_n\big)\\
&= \det\Big(\vec{a}_1, \dots, \sum_{j=1}^{n} x_j \vec{a}_j, \dots, \vec{a}_n\Big)\\
&= \sum_{j=1}^{n} x_j \det\big(\vec{a}_1, \dots, \vec{a}_j, \dots, \vec{a}_n\big)\\
&= x_i \det\big(\vec{a}_1, \dots, \vec{a}_i, \dots, \vec{a}_n\big) = x_i \det A\,,
\end{aligned}$$

da für jeden Index $j \neq i$ der Vektor $\vec{a}_j$ zweimal als Argument der Determinante auftritt, so daß alle Summanden bis auf den i-ten verschwinden.

Falls $\det A = 0$ ist, nützt uns diese Formel überhaupt nichts; ist allerdings $\det A \neq 0$, wissen wir bereits, daß das Gleichungssystem eindeutig lösbar ist, und wir erhalten die Formel

$$x_i = \frac{\det\big(\vec{a}_1, \dots, \vec{b}, \dots, \vec{a}_n\big)}{\det A}$$

für die Lösung.

Diese Formel heißt CRAMERsche Regel nach dem Schweizer Mathematiker GABRIEL CRAMER (1704–1752); für den Fall $n = 2$ ist sie bereits 1545 in der *Ars magna* des italienischen Mathematikers GIROLAMO CARDANO (1501–1576) zu finden.

Die CRAMERsche Regel ist sicherlich kein Verfahren, das man oft anwendet zur Lösung eines einzelnen Gleichungssystems: Der Aufwand für die Berechnung von $n + 1$ Determinanten ist weit größer als eine Lösung nach dem GAUSS-Algorithmus. Falls man allerdings eine ganze Familie ähnlicher Gleichungssysteme hat, in der sich nur wenige Parameter ändern, kann es sich lohnen, mit Hilfe der CRAMERschen Regel eine (von den Parametern abhängige) Lösungsformel zu berechnen und dann diese anzuwenden.

h) Eigenwerte und Eigenvektoren

Ein endlichdimensionaler k-Vektorraum V ist, wie wir gesehen haben, immer isomorph zu einem Vektorraum der Form k^n mit $n = \dim V$;

die möglichen Isomorphismen entsprechen genau den verschiedenen Basen von V. Allerdings empfiehlt sich im allgemeinen nicht, den Vektorraum V von vornherein mit k^n zu identifizieren, denn die Freiheit bei der Wahl einer Basis erlaubt bei geschickter Anwendung oftmals beträchtliche Vereinfachungen bei expliziten Rechnungen.

Eigenvektoren einer linearen Abbildung sind Beispiele von Kandidaten für Basisvektoren, bezüglich derer eine lineare Abbildung eine besonders einfache Abbildungsmatrix hat:

Definition: *a)* 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.

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 Eigenwert 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

$$\lambda_s \alpha_1 \vec{v}_1 + \cdots + \lambda_s \alpha_s \vec{v}_s = \vec{0}\,.$$

Bei Subtraktion der letzten beiden Gleichungen voneinander fällt der letzte Term auf der linken Seite weg, und wir erhalten 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 hier alle Koeffizienten verschwinden. Da die Eigenwerte $\lambda_1, \cdots, \lambda_s$ allesamt verschieden sind, ist dies nur möglich, wenn $\alpha_1 = \cdots = \alpha_{s-1} = 0$ ist. 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.

■

Bevor wir zu nichttrivialen Beispielen kommen können, benötigen wir ein Verfahren, um Eigenwerte und Eigenvektoren effektiv zu berechnen. Nach Definition ist $\lambda \in k$ genau dann Eigenwert der linearen Abbildung φ, wenn es dazu einen Eigenvektor gibt, d.h. einen vom Nullvektor verschiedenen Vektor $\vec{v}$, so daß $\varphi(\vec{v}) = \lambda\vec{v}$ ist. Diese Bedingung wird, wenn man sie bezüglich einer Basis hinschreibt, zu einem linearen Gleichungssystem für die Komponenten von $\vec{v}$: Ist A die Abbildungsmatrix von φ bezüglich der gewählten Basis, haben wir die Gleichung $A\vec{v} = \lambda\vec{v}$ oder $(A - \lambda E)\vec{v} = \vec{0}$, wobei E die $n \times n$-Einheitsmatrix bezeichnet.

In letzterer Form ist dies ein homogenes lineares Gleichungssystem für die Komponenten von $\vec{v}$; wie jedes homogene lineare Gleichungssystem hat es den Nullvektor als Lösung. Weitere Lösungen gibt es genau dann, wenn die Matrix $A - \lambda E$ des Gleichungssystems singulär ist, d.h. also, wenn $\det(A - \lambda E)$ verschwindet.

Damit ist klar, wie man Eigenwerte und Eigenvektoren berechnen kann: Man löse die Gleichung $\det(A - \lambda E) = 0$ und dann für jede Wurzel λ_i dieser Gleichung das lineare Gleichungssystem $(A - \lambda E)\vec{x} = \vec{0}$

Als Beispiel betrachten wir die Matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 8 & 7 & 6 & 5 \\ 4 & 3 & 2 & 1 \end{pmatrix} .$$

Hier ist

$$A - \lambda E = \begin{pmatrix} 1-\lambda & 2 & 3 & 4 \\ 5 & 6-\lambda & 7 & 8 \\ 8 & 7 & 6-\lambda & 5 \\ 4 & 3 & 2 & 1-\lambda \end{pmatrix}$$

und nach dem Entwicklungssatz ist (bei Entwicklung nach der ersten Zeile)

$$\begin{aligned} \det(A - \lambda E) = (1-\lambda) &\begin{vmatrix} 6-\lambda & 7 & 8 \\ 7 & 6-\lambda & 5 \\ 3 & 2 & 1-\lambda \end{vmatrix} \\ & - 2 \begin{vmatrix} 5 & 7 & 8 \\ 8 & 6-\lambda & 5 \\ 4 & 2 & 1-\lambda \end{vmatrix} \\ & + 3 \begin{vmatrix} 5 & 6-\lambda & 8 \\ 8 & 7 & 5 \\ 4 & 3 & 1-\lambda \end{vmatrix} - 4 \begin{vmatrix} 5 & 6-\lambda & 7 \\ 8 & 7 & 6-\lambda \\ 4 & 3 & 2 \end{vmatrix} \\ &= (1-\lambda)(-\lambda^3 + 13\lambda^2 + 35\lambda) - 2(5\lambda^2 + 53\lambda) \\ &\quad + 3(-8\lambda^2 + \lambda) - 4(4\lambda^2 - 17\lambda) \\ &= \lambda^4 - 14\lambda^3 - 72\lambda^2 = \lambda^2(\lambda^2 - 14\lambda - 72) . \end{aligned}$$

Letzterer Ausdruck verschwindet genau dann, wenn entweder λ verschwindet oder der Klammerausdruck ganz hinten. Letzterer ist eine quadratische Gleichung für λ, die man entweder nach der üblichen Formel lösen kann, oder aber man beachtet den Wurzelsatz von VIÈTE, wonach die Summe der beiden Lösungen 14 und ihr Produkt -72 sein muß; da $72 = 4 \times 18$ ist, können die beiden Lösungen daher nur -4 und 18 sein.

Die Eigenwerte von A sind also $\lambda_1 = 0$, $\lambda_2 = -4$ und $\lambda_3 = 18$; für jeden dieser Eigenwerte müssen wir nun über ein lineares Gleichungssystem die Eigenvektoren bestimmen.

Beginnen wir mit $\lambda_1 = 0$. Hier ist das Gleichungssystem einfach $A\vec{x} = \vec{0}$, d.h.

$$\begin{aligned} x_1 + 2x_2 + 3x_3 + 4x_4 &= 0 \\ 5x_1 + 6x_2 + 7x_3 + 8x_4 &= 0 \\ 8x_1 + 7x_2 + 6x_3 + 5x_4 &= 0 \\ 4x_1 + 3x_2 + 2x_3 + x_4 &= 0\,. \end{aligned}$$

Aufgrund der speziellen Struktur der Matrix A ergänzen sich hier die erste und die vierte Gleichung zu

$$5(x_1 + x_2 + x_3 + x_4) = 0\,;$$

entsprechend addieren sich die beiden mittleren Gleichungen zu

$$13(x_1 + x_2 + x_3 + x_4) = 0\,.$$

Da außerdem noch die Differenz zwischen den ersten beiden Gleichungen gleich

$$4(x_1 + x_2 + x_3 + x_4) = 0$$

ist, genügt es, außer der ersten Gleichung noch als einzige weitere Gleichung

$$x_1 + x_2 + x_3 + x_4 = 0 \quad \text{oder} \quad x_1 = -(x_2 + x_3 + x_4)$$

zu betrachten. Subtrahiert man diese Gleichung von der ersten oder, was dasselbe ist, setzt man sie ein, so bleibt

$$x_2 + 2x_3 + 3x_4 = 0$$

als einzige Relation zwischen x_2, x_3 und x_4. Das gegebene Gleichungssystem ist also äquivalent zu

$$\begin{aligned} x_1 + 2x_2 + 3x_3 + 4x_4 &= 0 \\ x_2 + 2x_3 + 3x_4 &= 0\,; \end{aligned}$$

insbesondere hat es also den Rang zwei und damit einen Lösungsraum der Dimension $4 - 2 = 2$.

Im noch verbliebenen Gleichungssystem kann beispielsweise x_4 beliebig gewählt werden. Ist $x_4 = 0$, so bleibt noch die Relation $x_2 + 2x_3 = 0$ übrig, in der etwa x_3 beliebig gewählt werden kann; danach sind $x_2 = -2x_3$ und $x_1 = -x_3$ eindeutig festgelegt; die entsprechenden Lösungen sind also die Vielfachen der Lösung $(-1, 2, 1, 0)$.

Genausogut könnte etwa $x_3 = 0$ gesetzt werden; dann wäre $x_2 = -3x_4$ und $x_1 = -2x_4$; dies führt auf die Vielfachen der Lösung $(-2, -3, 0, 1)$.

Damit haben wir zwei linear unabhängige Lösungen gefunden; da der Lösungsraum zweidimensional ist, sind die sämtlichen Lösungen des Gleichungssystems daher die Linearkombinationen

$$x_1 = -\lambda - 2\mu, \quad x_2 = 2\lambda - 3\mu, \quad x_3 = \lambda \quad \text{und} \quad x_4 = \mu$$

dieser beiden Lösungen. Insbesondere besteht also der Eigenraum von A zum Eigenwert Null genau aus den sämtlichen Linearkombinationen

$$\lambda \begin{pmatrix} -1 \\ 2 \\ 1 \\ 0 \end{pmatrix} + \mu \begin{pmatrix} -2 \\ 3 \\ 0 \\ 1 \end{pmatrix} .$$

Genauso lassen sich auch die Eigenwerte -4 und 18 behandeln; -4 führt auf das lineare Gleichungssystem

$$\begin{aligned} 5x_1 + 2x_2 + 3x_3 + 4x_4 &= 0 \\ 5x_1 + 10x_2 + 7x_3 + 8x_4 &= 0 \\ 8x_1 + 7x_2 + 10x_3 + 5x_4 &= 0 \\ 4x_1 + 3x_2 + 2x_3 + 5x_1 &= 0 , \end{aligned}$$

dessen sämtliche Lösungen gerade die Vielfachen von $(-1, -1, 1, 1)$ sind, und 18 führt auf

$$\begin{aligned} -17x_1 + 2x_2 + 3x_3 + 4x_4 &= 0 \\ 5x_1 - 12x_2 + 7x_3 + 8x_4 &= 0 \\ 8x_1 + 7x_2 - 12x_3 + 5x_4 &= 0 \\ 4x_1 + 3x_2 + 2x_3 - 17x_4 &= 0 \end{aligned}$$

mit den Vielfachen von $(5, 13, 13, 5)$ als Lösungsraum.

Es ist kein Zufall, daß in diesem 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 der linearen Abbildung φ oder, wie wir auch sagen werden, die Eigenwerte von A, gleich den Nullstellen des charakteristischen Polynoms von A.

Ist x eine Nullstelle eines Polynoms $f(X)$, so kann $f(X)$ bekanntlich durch $(X - x)$ geteilt werden, und x heißt *n-fache Nullstelle* von $f(X)$, wenn $f(X)$ durch $(X - x)^n$ teilbar ist, nicht aber durch $(X - x)^{n+1}$.

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

$$\begin{aligned} 0x_1 + 1x_2 &= 0 \\ 0x_1 + 0x_2 &= 0\,, \end{aligned}$$

also gerade die Menge aller Vektoren der Form $\binom{x}{0}$ und damit eindimensional. Die geometrische Vielfachheit des Eigenwerts eins ist daher nur eins.

Allgemein gilt

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 von φ 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 A die Abbildungsmatrix von φ.

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 A steht also jeweils in der Diagonalen das

Element λ und ansonsten überall die Null, d.h. A hat die Form

$$A_{\mathcal{B}} = \begin{pmatrix} \lambda & 0 & \dots & 0 & * & \dots & * \\ 0 & \lambda & \dots & 0 & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda & * & \dots & * \\ 0 & 0 & \dots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & \mathsf{M} & \\ 0 & 0 & \dots & 0 & & & \end{pmatrix},$$

wobei uns weder die mit $*$ bezeichneten Körperelemente noch die $(n-r) \times (n-r)$-Matrix M weiter zu interessieren brauchen.

Für $A - xE$ gilt dasselbe, nur daß jetzt $\lambda - x$ in der Diagonalen steht, d.h. diese Matrix hat die Form

$$\begin{pmatrix} \lambda - x & 0 & \dots & 0 & * & \dots & * \\ 0 & \lambda - x & \dots & 0 & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda - x & * & \dots & * \\ 0 & 0 & \dots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & \mathsf{M} - \mathsf{x}\mathsf{E}_{n-r} & \\ 0 & 0 & \dots & 0 & & & \end{pmatrix},$$

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

$$\det(A - xE) = (\lambda - x)^r \det(M - xE_{n-r}) \, .$$

Somit ist $\det(A - xE)$ durch $(x - \lambda)^r$ teilbar; die algebraische Vielfachheit von λ ist also mindestens r.

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

$$r_1 + \cdots + r_\ell \leq n ,$$

denn der Grad eines Teilers kann nicht größer sein als der des Polynoms selbst.

■

§8: Symmetrische und Hermitesche Matrizen

Wie schon mehrfach erwähnt, sind viele in physikalisch-technischen Anwendungen wichtige reelle Matrizen symmetrisch. In diesem Paragraphen sollen deren besondere Eigenschaften untersucht werden; beispielsweise werden wir sehen, daß es zu einer symmetrischen Matrix stets eine Basis aus Eigenvektoren gibt und daß alle Eigenwerte reell sind.

Im Komplexen ist der Begriff der symmetrischen Matrix weniger nützlich; dort werden wir stattdessen eine Variante der Symmetrie betrachten, die sogenannte HERMITEsche Symmetrie.

Symmetrische und HERMITEsche Matrizen sind eng verbunden mit Skalarprodukten auf reellen und komplexen Vektorräumen; wir wollen deshalb mit diesen beginnen.

a) Euklidische Vektorräume

Zu Beginn dieses Kapitels hatten wir für Vektoren in $\mathbb{R}^n$ ein Skalarprodukt definiert, das uns gestattete, Längen und Winkel zu berechnen. In diesem Paragraphen sollen diese Begriffe und ihr Umfeld für beliebige reelle Vektorräume erklärt werden. Da es hier um die Grundbegriffe der klassischen EUKLIDischen Geometrie geht, sprechen wir von EUKLIDischen Vektorräumen:

Definition: Ein EUKLIDischer Vektorraum ist ein Paar $(V, \cdot)$ bestehend aus einem $\mathbb{R}$-Vektorraum V und einer Abbildung

$$\cdot : V \times V \to \mathbb{R}; \quad (\vec{v}, \vec{w}) \mapsto \vec{v} \cdot \vec{w}$$

mit folgenden Eigenschaften:

a) $\cdot$ ist linear in beiden Argumenten oder, wie wir kurz sagen, *bilinear,* d.h.

$$(\lambda\vec{u} + \mu\vec{v}) \cdot \vec{w} = \lambda(\vec{u} \cdot \vec{w}) + \mu(\vec{v}\vec{w}) \quad \text{und} \quad \vec{u} \cdot (\lambda\vec{v} + \mu\vec{w}) = \lambda(\vec{u}\vec{v}) + \mu(\vec{u}\vec{w})$$

für alle $\vec{u}, \vec{v}, \vec{w} \in V$ und $\lambda, \mu \in \mathbb{R}$.

b) $\cdot$ ist symmetrisch, d.h. $\vec{v} \cdot \vec{w} = \vec{w} \cdot \vec{v}$ für alle $\vec{v}, \vec{w} \in V$.

c) $\cdot$ ist positiv definit, d.h. $\vec{v} \cdot \vec{v} > 0$ für alle Vektoren $\vec{v} \neq \vec{0}$ aus V.

Die bilineare Abbildung $\cdot$ heißt *Skalarprodukt;* wie bei Produkten üblich, werden wir den Malpunkt oft auch weglassen.

Typisches Beispiel eines EUKLIDischen Vektorraums ist natürlich der $\mathbb{R}^n$ mit seinem üblichen Skalarprodukt; ein anderes Beispiel ist der von den Funktionen $\sin nx$ und $\cos nx$ und den Konstanten aufgespannte Untervektorraum V von $C^0(\mathbb{R}, \mathbb{R})$, d.h. der Vektorraum aller trigonometrischer Polynome

$$f = c + \sum_{n=1}^{N} a_n \sin nx + \sum_{m=1}^{M} b_m \cos mx$$

mit beliebigem N, M zusammen mit dem Skalarprodukt

$$f \cdot g = \int_0^{2\pi} f(x)g(x)\, dx\,.$$

Die Eigenschaften *a)* und *b)* aus der Definition sind klar, und für *c)* müssen wir nur die Orthogonalitätsrelationen für trigonometrische Funktionen aus Kapitel I, §*6b* beachten und erhalten, daß für

$$f = c + \sum_{n=1}^{N} a_n \sin nx + \sum_{m=1}^{M} b_m \cos mx$$

gilt

$$f \cdot f = \int_0^{2\pi} f(x)^2 \, dx = 2\pi c^2 + \pi \sum_{n=1}^{N} a_n^2 + \pi \sum_{m=1}^{M} b_m^2 ,$$

was in der Tat positiv ist, wenn auch nur einer der Koeffizienten und damit ganz f ungleich Null ist.

EUKLIDische Vektorräume haben ihren Namen davon, daß man dort so etwas wie EUKLIDische Geometrie betreiben kann; dies wird für uns zwar nur am Rande von Bedeutung sein, aber zumindest die beiden wesentlichen Grundbegriffe wollen wir kurz definieren:

Definition: *a)* Die Länge eines Vektors $\vec{v}$ aus einem EUKLIDischen Vektorraum V ist $\sqrt{\vec{v} \cdot \vec{v}}$.

b) Der Winkel zwischen zwei Vektoren $\vec{v}, \vec{w} \in V \setminus \{\vec{0}\}$ ist

$$\measuredangle(\vec{v}, \vec{w}) = \arccos \left(\frac{\vec{v} \cdot \vec{w}}{\sqrt{(\vec{v} \cdot \vec{v})(\vec{w} \cdot \vec{w})}} \right) .$$

Da der Arkuscosinus nur Werte zwischen Null und π annimmt, im Gradmaß also zwischen Null und 180°, ist dies ein *unorientierter* Winkel. Im $\mathbb{R}^3$ konnten wir durch Auszeichnung von Rechtssystemen über die Dreifingerregel oder Korkenzieherregel orientierte Winkel zwischen Null und 2π definieren; das ist in höherdimensionalen EUKLIDischen Vektorräumen nicht mehr möglich: Falls man dort einen orientierten Winkel definieren möchte, muß man zunächst eine Orientierung des Vektorraums auszeichnen. Eine entsprechende Theorie orientierter EUKLIDischer Vektorräume gibt es, jedoch brauchen wir nicht darauf einzugehen.

Wenigstens kurz sollten wir uns allerdings die Frage stellen, warum die obige Definition überhaupt einen Winkel erklärt: Sie ist offensichtlich nur sinnvoll, wenn das Argument des Arkuskosinus höchstens den Betrag eins hat.

Für das Standardskalarprodukt im $\mathbb{R}^2$ und $\mathbb{R}^3$ ist das kein Problem: In §1 dieses Kapitels hatten wir das Skalarprodukt im $\mathbb{R}^3$ *definiert* als

Produkt aus den Längen der beiden Vektoren und dem Cosinus des eingeschlossenen Winkels; wir hatten dann nachgerechnet, daß dies mit der hier verwendeten Koordinatendefinition übereinstimmt.

Nun spannen aber zwei beliebige Vektoren im $\mathbb{R}^n$ höchstens eine Ebene auf; für $n > 3$ können wir daher stets einen zweidimensionalen Untervektorraum finden, in dem beide Vektoren liegen. Am Ende von Abschnitt *c)* werden wir sehen, daß wir diesen identifizieren können mit einem $\mathbb{R}^2$ mit Standardskalarprodukt und daß wir das Skalarprodukt von V dort ausrechnen können; damit erhalten wir für das Argument des Arkuscosinus auf jeden Fall einen Wert vom Betrag höchstens eins.

b) Hermitesche Vektorräume

Für Vektorräume über den komplexen Zahlen sind die Definitionen aus dem vorigen Abschnitt nicht so gut geeignet; beispielsweise ist mit dem üblichen Skalarprodukt in $\mathbb{C}^2$

$$\begin{pmatrix} 1 \\ i \end{pmatrix} \cdot \begin{pmatrix} 1 \\ i \end{pmatrix} = 1^2 + i^2 = 0\,.$$

Solche Vektoren mit Quadrat Null sind zwar in einigen Anwendungen (wie etwa der speziellen Relativitätstheorie) durchaus sinnvoll, meist möchte man aber das Skalarprodukt eines Vektors mit sich selbst als Quadrat seiner Länge interpretieren, die für Vektoren ungleich dem Nullvektor positiv sein soll.

Wir haben bereits jeder komplexen Zahl ungleich Null eine positive reelle Zahl zugeordnet, nämlich ihren Betrag

$$|z| = \sqrt{z\overline{z}}\,,$$

und dies können wir auch auf Vektoren verallgemeinern, indem wir den beiden Vektoren

$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \quad \text{und} \quad \vec{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}$$

aus $\mathbb{C}^n$ das Produkt

$$\vec{v} \cdot \vec{w} = \sum_{\ell=1}^{n} v_\ell \overline{w_\ell}$$

zuordnen. Auch hier ist offensichtlich $\vec{v} \cdot \vec{v}$ für jeden Vektor außer dem Nullvektor eine positive reelle Zahl, denn wir summieren ja über die Betragsquadrate der Komponenten des Vektors.

Dieses Produkt erfüllt nun allerdings nicht mehr die Bedingungen *a)* und *b)* aus dem vorigen Abschnitt: Es ist zwar noch linear im ersten Argument, aber

$$\vec{v} \cdot (\lambda\vec{w}) = \sum_{\ell=1}^{n} v_\ell \overline{\lambda w_\ell} = \overline{\lambda} \sum_{\ell=1}^{n} v_\ell \overline{w_\ell} = \overline{\lambda}(\vec{v} \cdot \vec{w})\,.$$

Auch ist das Produkt nicht mehr symmetrisch, denn schon für zwei komplexe Zahlen z und w ist $z\overline{w} = \overline{w\overline{z}}$ im allgemeinen verschieden von $w\overline{z}$.

Der Begriff des HERMITEschen Vektorraums formalisiert die Eigenschaften des oben eingeführten Produkts; er ist benannt nach dem französischen Mathematiker CHARLES HERMITE (1822–1901).

Definition: Ein HERMITEscher Vektorraum ist ein Paar $(V, \cdot)$ bestehend aus einem $\mathbb{C}$-Vektorraum V und einer Abbildung

$$\cdot : V \times V \to \mathbb{C}; \quad (\vec{v}, \vec{w}) \mapsto \vec{v} \cdot \vec{w}$$

mit folgenden Eigenschaften:

a) $\cdot$ ist linear im ersten Argument, d.h.

$$(\lambda\vec{u} + \mu\vec{v}) \cdot \vec{w} = \lambda(\vec{u} \cdot \vec{w}) + \mu(\vec{v} \cdot \vec{w})$$

für alle $\vec{u}, \vec{v}, \vec{w} \in V$ und $\lambda, \mu \in \mathbb{C}$.

b) $\cdot$ ist HERMITEsch symmetrisch, d.h. $\vec{v} \cdot \vec{w} = \overline{\vec{w} \cdot \vec{v}}$ für alle $\vec{v}, \vec{w} \in V$.

c) $\cdot$ ist positiv definit, d.h. $\vec{v} \cdot \vec{v}$ ist eine positive reelle Zahl für alle Vektoren $\vec{v} \neq \vec{0}$ aus V.

Die bilineare Abbildung $\cdot$ heißt HERMITE*sches Skalarprodukt,* und auch hier werden wir den Malpunkt nicht immer hinschreiben.

Durch Kombination der Eigenschaften *a)* und *b)* kann man leicht ausrechnen, was anstelle der Linearität für das zweite Argument gilt:

$$\begin{aligned}\vec{u} \cdot (\lambda\vec{v} + \mu\vec{w}) &= \overline{(\lambda\vec{v} + \mu\vec{w}) \cdot \vec{u}} = \overline{\lambda(\vec{v}\vec{u}) + \mu(\vec{w}\vec{u})} \\ &= \overline{\lambda}\,\overline{(\vec{v}\vec{u})} + \overline{\mu}\,\overline{(\vec{w}\vec{u})} = \overline{\lambda}\,(\vec{u} \cdot \vec{v}) + \overline{\mu}\,(\vec{u} \cdot \vec{w})\,.\end{aligned}$$

Diese Eigenschaft, zusammen mit der Linearität im ersten Argument, bezeichnet man gelegentlich als *Sesquilinearität,* d.h. „anderthalbfache Linearität", im Gegensatz zur echten Linearität in zwei Argumenten, der *Bilinearität.*

Typisches Beispiel eines HERMITEschen Vektorraums ist natürlich der $\mathbb{C}^n$ mit dem eingangs definierten Produkt, aber auch hier gibt es wieder ein wichtiges Beispiel eines Funktionenraums: Dieses Mal sei V der Vektorraum aller Funktionen der Form

$$f(t) = \sum_{n=-N}^{M} c_n e^{int}$$

mit $c_n \in \mathbb{C}$ und beliebigen natürlichen Zahlen N, M; wir setzen

$$f \cdot g \underset{\text{def}}{=} \int_0^{2\pi} f(t)\overline{g(t)}\,dt\,.$$

Wieder sind *a)* und *b)* klar, und *c)* folgt aus den Orthogonalitätsrelationen für die Funktionen e^{int} aus Kapitel I, §*6b.*

Auch in HERMITEschen Vektorräumen werden wir $\sqrt{\vec{v}\cdot\vec{v}}$ gelegentlich als *Länge* des Vektors $\vec{v}$ bezeichnen; auf die Definition von Winkeln dagegen wollen wir verzichten, da die Übernahme der entsprechenden Definition für EUKLIDische Vektorräume hier auf geometrisch nicht sinnvoll interpretierbare komplexe Winkel führen würde.

c) Orthonormalbasen

Im Rest dieses Paragraphen werden wir EUKLIDische und HERMITEsche Vektorräume gleichzeitig behandeln: Ersetzt man in der Definition eines HERMITEschen Vektorraums überall $\mathbb{C}$ durch $\mathbb{R}$, so erhält man einen EUKLIDischen Vektorraum. Zwar stehen noch an vielen Stellen die Querstriche für komplexe Konjugation, aber da eine reelle Zahl keinen Imaginärteil hat, bleibt sie unter Konjugation invariant, so daß für reelle Zahlen die Konjugationsstriche zwar überflüssig, aber nicht schädlich sind.

Für den $\mathbb{R}^3$ hatten wir zu Beginn dieses Kapitels gesehen, daß man das (Standard-)Skalarprodukt auch ausrechnen kann als Produkt der Längen der beiden Vektoren mal dem Cosinus des eingeschlossenen Winkels. Insbesondere ist also das Skalarprodukt zweier von $\vec{0}$ verschiedener Vektoren genau dann gleich Null, wenn der Cosinus des eingeschlossenen Winkels verschwindet, d.h. wenn der Winkel gleich $\pi/2$ oder $3\pi/2$ ist. Geometrisch gesehen heißt dies, daß die beiden Vektoren aufeinander senkrecht stehen oder, vornehmer ausgedrückt, *orthogonal* sind.

Diese Sprechweise wollen wir auch für beliebige EUKLIDische und HERMITEsche Vektorräume verwenden: Zwei Vektoren $\vec{v}, \vec{w} \in V$ heißen *orthogonal,* wenn $\vec{v} \cdot \vec{w} = 0$ ist. (Implizit haben wir diesen Begriff bereits verwendet bei den Orthogonalitätsrelationen für trigonometrische Funktionen und komplexe Exponentialfunktionen in Kapitel 1, §*6b*).

Definition: Eine Basis $\mathcal{B}$ eines EUKLIDischen oder HERMITEschen Vektorraums heißt *Orthogonalbasis,* wenn jedes Element von $\mathcal{B}$ orthogonal zu allen übrigen ist.

Eine Orthogonalbasis $\mathcal{B}$ heißt *Orthonormalbasis,* wenn zusätzlich für jeden Vektor $\vec{b} \in \mathcal{B}$ gilt: $\vec{b} \cdot \vec{b} = 1$.

Beispielsweise bilden die Einheitsvektoren

$$\vec{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \vec{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \ldots, \quad \vec{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

eine Orthonormalbasis von $\mathbb{R}^n$ und auch von $\mathbb{C}^n$. Genauso bilden die Funktionen $\sin nt$ und $\cos nt$ für $n \in \mathbb{N}$ zusammen mit der konstanten Funktion Eins eine Orthonormalbasis des von ihnen erzeugten Untervektorraums von $C^0(\mathbb{R}, \mathbb{R})$, und die Funktionen e^{int} für $n \in \mathbb{Z}$ sind eine Orthonormalbasis des von ihnen erzeugten HERMITEschen Vektorraum.

Tatsächlich hat *jeder* EUKLIDische oder HERMITEsche Vektorraum eine Orthonormalbasis; für den Beweis werden wir uns allerdings, wie immer, wenn von Basen die Rede ist, zur Vermeidung von logischen Schwierigkeiten auf den endlichdimensionalen Fall beschränken:

Satz: Jeder endlichdimensionale EUKLIDische oder HERMITEsche Vektorraum V hat (mindestens) eine Orthonormalbasis.

Beweis: Wir wissen bereits, daß V *irgendeine* Basis hat; eine solche Basis sei $\{\vec{b}_1, \ldots, \vec{b}_n\}$. Hieraus konstruieren wir zunächst eine Orthogonalbasis nach dem GRAM-SCHMIDTschen Orthogonalisierungsverfahren. Dieses Verfahren ist benannt nach dem dänischen Mathematiker JØRGAN PEDERSEN GRAM (1850–1916) und dem deutschen Mathematiker ERHARD SCHMIDT (1876–1959), obwohl es schon lange vorher dem französischen Mathematiker PIERRE-SIMON DE LAPLACE (1749–1827) bekannt war.

Das Verfahren besteht aus n Schritten, wobei nach dem r-ten Schritt r Basisvektoren $\vec{c}_1, \ldots, \vec{c}_r$ für den von $\vec{b}_1, \ldots, \vec{b}_r$ aufgespannten Untervektorraum konstruiert sind derart, daß jeder der r Vektoren $\vec{c}_i$ auf den übrigen $r-1$ senkrecht steht.

Der erste Schritt ist der einfachste: Da es hier noch keine Orthogonalitätsbedingung für $\vec{c}_1$ gibt, können wir einfach $\vec{c}_1 = \vec{b}_1$ setzen.

Nachdem wir $r \geq 1$ Schritte durchgeführt haben, haben wir r linear unabhängige Vektoren $\vec{c}_1, \ldots, \vec{c}_r$ mit $\vec{c}_i \cdot \vec{c}_j = 0$ für $i \neq j$ aus dem von $\vec{b}_1, \ldots, \vec{b}_r$ aufgespannten Untervektorraum. Ist $r = n$, haben wir eine Orthogonalbasis; andernfalls muß ein auf den bisher konstruierten $\vec{c}_i$ senkrecht stehender Vektor $\vec{c}_{r+1}$ gefunden werden, der zusammen mit diesen den von $\vec{b}_1$ bis $\vec{b}_{r+1}$ erzeugten Untervektorraum erzeugt.

Da $\vec{c}_1, \ldots, \vec{c}_r$ und $\vec{b}_1, \ldots, \vec{b}_r$ denselben Untervektorraum erzeugen, gilt dasselbe für $\vec{c}_1, \ldots, \vec{c}_r, \vec{b}_{r+1}$ und $\vec{b}_1, \ldots, \vec{b}_{r+1}$; das Problem ist, daß $\vec{b}_{r+1}$ im allgemeinen nicht orthogonal zu den $\vec{c}_i$ sein wird. Wir dürfen $\vec{b}_{r+1}$ aber abändern durch einen beliebigen Vektor aus dem von $\vec{c}_1, \ldots, \vec{c}_r$ aufgespannten Untervektorraum; also setzen wir

$$\vec{c}_{r+1} = \vec{b}_{r+1} + \lambda_1 \vec{c}_1 + \cdots + \lambda_r \vec{c}_r$$

und versuchen, die λ_i so zu bestimmen, daß dieser Vektor orthogonal zu $\vec{c}_1, \ldots, \vec{c}_r$ wird.

Wegen der Orthogonalität der $\vec{c}_i$ ist

$$\vec{c}_{r+1} \cdot \vec{c}_i = \vec{b}_{r+1} \cdot \vec{c}_i + \sum_{j=1} \lambda_j (\vec{c}_j \cdot \vec{c}_i) = \vec{b}_{r+1} \cdot \vec{c}_i + \lambda_i (\vec{c}_i \cdot \vec{c}_i) \,;$$

setzen wir also

$$\lambda_i = -\frac{\vec{b}_{r+1} \cdot \vec{c}_i}{\vec{c}_i \cdot \vec{c}_i} \,,$$

so ist $\vec{v} \cdot \vec{c}_i = 0$ für alle $i = 1, \ldots, r$.

Auf diese Weise können wir die $\vec{b}_i$ sukzessive so abändern, daß wir schließlich eine Orthogonalbasis $\{\vec{c}_1, \ldots, \vec{c}_n\}$ von V erhalten; das einzige verbleibende Problem ist, daraus eine *Orthonormalbasis* zu machen. Das ist nun aber einfach: Da eine Basis nicht den Nullvektor enthalten kann, ist jedes $\vec{c}_i \neq \vec{0}$, und damit ist $\alpha_i \underset{\text{def}}{=} \vec{c}_i \cdot \vec{c}_i$ eine positive reelle Zahl. Setzt man nun

$$\vec{e}_i = \frac{1}{\sqrt{\alpha_i}} \vec{c}_i \,,$$

so ist $\{\vec{e}_1, \ldots, \vec{e}_n\}$ offensichtlich eine Orthonormalbasis von V. ■

Als *Beispiel* dazu betrachten wir den Untervektorraum V des $\mathbb{R}^5$, der von den drei Vektoren

$$\vec{b}_1 = \begin{pmatrix} 2 \\ 2 \\ 3 \\ 4 \\ 4 \end{pmatrix}, \quad \vec{b}_2 = \begin{pmatrix} 4 \\ 4 \\ 3 \\ 2 \\ 4 \end{pmatrix} \quad \text{und} \quad \vec{b}_3 = \begin{pmatrix} 3 \\ 3 \\ 3 \\ 6 \\ 1 \end{pmatrix}$$

aufgespannt wird.

Im ersten Schritt des GRAM-SCHMIDTschen Orthogonalisierungsverfahrens setzen wir $\vec{c}_1 = \vec{b}_1$.

Im zweiten Schritt müssen wir einen Vektor

$$\vec{c}_2 = \vec{b}_2 + \lambda_1 \vec{c}_1$$

finden, so daß $\vec{c}_1 \cdot \vec{c}_2 = 0$ ist. Da $\vec{b}_2 \cdot \vec{c}_1 = 49$ und $\vec{c}_1 \cdot \vec{c}_1 = 49$ beide gleich sind, müssen wir nach GRAM-SCHMIDT $\lambda_1 = -1$ setzen, d.h.

$$\vec{c}_2 = \vec{b}_2 - \vec{c}_1 = \vec{b}_2 - \vec{b}_1 = \begin{pmatrix} 2 \\ 2 \\ 0 \\ -2 \\ 0 \end{pmatrix} .$$

Im dritten Schritt suchen wir einen Vektor

$$\vec{c}_3 = \vec{b}_3 + \mu_1 \vec{c}_1 + \mu_2 \vec{c}_2$$

mit $\vec{c}_3 \cdot \vec{c}_1 = \vec{c}_2 \cdot \vec{c}_1 = 0$. Da $\vec{b}_3 \cdot \vec{c}_1 = \vec{c}_1 \cdot \vec{c}_1 = 49$ ist, setzen wir $\mu_1 = -1$; der zweite Koeffizient μ_2 ist Null, da $\vec{b}_3$ und $\vec{c}_2$ bereits orthogonal sind. Damit wird

$$\vec{c}_3 = \vec{b}_3 - \vec{c}_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 2 \\ -3 \end{pmatrix} .$$

Schließlich müssen wir diese Orthogonalbasis noch zur Orthonormalbasis machen; wegen

$$\vec{c}_1 \cdot \vec{c}_1 = 49, \quad \vec{c}_2 \cdot \vec{c}_2 = 12 \quad \text{und} \quad \vec{c}_3 \cdot \vec{c}_3 = 15$$

setzen wir

$$\vec{e}_1 = \frac{1}{7}\vec{c}_1 = \begin{pmatrix} 2/7 \\ 2/7 \\ 3/7 \\ 4/7 \\ 4/7 \end{pmatrix}, \quad \vec{e}_2 = \frac{1}{\sqrt{12}}\vec{c}_2 = \begin{pmatrix} \sqrt{3}/3 \\ \sqrt{3}/3 \\ 0 \\ -\sqrt{3}/3 \\ 0 \end{pmatrix}$$

und

$$\vec{e}_3 = \frac{1}{\sqrt{15}}\vec{c}_3 = \begin{pmatrix} \sqrt{15}/15 \\ \sqrt{15}/15 \\ 0 \\ 2\sqrt{15}/15 \\ -\sqrt{15}/5 \end{pmatrix} .$$

Damit ist eine Orthonormalbasis von V gefunden.

Ist $\vec{e}_1, \ldots, \vec{e}_n$ eine Orthonormalbasis eines EUKLIDischen oder HERMITEschen Vektorraums V, und sind

$$\vec{v} = a_1\vec{e}_1 + \cdots + a_n\vec{v}_n \quad \text{und} \quad \vec{w} = b_1\vec{e}_1 + \cdots + b_n\vec{v}_n$$

zwei Vektoren aus V, so ist

$$\vec{v} \cdot \vec{w} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \overline{b_j}\; \vec{e}_i \cdot \vec{e}_j = \sum_{i=1}^{n} a_i \overline{b_j}\,;$$

bezüglich einer Orthonormalbasis sieht also jedes (HERMITEsche) Skalarprodukt aus wie das Standardskalarprodukt – und das unabhängig von der gewählten Orthonormalbasis. Ein (relativ kompliziertes) Beispiel hierfür sollte aus der Elektrotechnik I bekannt sein: Die Leistung eines Signals ist proportional zum HERMITEschen Skalarprodukt des Signals mit sich selbst, und nach dem Satz von PARSEVAL kann dieses auch als Summe der Betragsquadrate der FOURIER-Koeffizienten berechnet werden. (Tatsächlich ist die Situation nicht ganz mit der hier betrachteten vergleichbar, da es sich um eine unendliche Summe handelt, während wir hier von endlichen Summen ausgehen; die entsprechende Verallgemeinerung der Theorie HERMITEscher Vektorräume zu HILBERTräumen werden wir im nächsten Semester behandeln.)

Hier wollen wir uns mit einer viel einfacheren Anwendung begnügen: Aus Abschnitt *a)* ist noch die Frage offen, warum in einem EUKLIDischen Vektorraum die Winkeldefinition

$$\measuredangle(\vec{v}, \vec{w}) = \arccos\left(\frac{\vec{v} \cdot \vec{w}}{\sqrt{(\vec{v} \cdot \vec{v})(\vec{w} \cdot \vec{w})}}\right)$$

sinnvoll ist, d.h. also, warum das Argument des Arkuscosinus immer höchstens Betrag eins hat.

Wir wissen bereits, daß dies für $\mathbb{R}^3$ mit dem Standardskalarprodukt gilt, und damit gilt es natürlich erst recht für $\mathbb{R}^2$ mit dessen Standardskalarprodukt. Wir betrachten nun einen zweidimensionalen Untervektorraum, der $\vec{v}$ und $\vec{w}$ enthält, und wählen eine Orthonormalbasis dieses Untervektorraums; dies macht diesen Untervektorraum isomorph zum $\mathbb{R}^2$ mit dem Standardskalarprodukt.

Nun ergänzen wir diese Basis zu einer Orthonormalbasis des gesamten Vektorraums. Bei der Berechnung der Skalarprodukte $\vec{v}\cdot\vec{w}$, $\vec{v}\cdot\vec{v}$ und $\vec{w}\cdot\vec{w}$ nach obiger Formel spielen die neu hinzugekommenen Basisvektoren offensichtlich keine Rolle, so daß die drei Skalarprodukte tatsächlich in $\mathbb{R}^2$ berechnet werden, und damit ist wie dort

$$\left|\frac{\vec{v}\cdot\vec{w}}{\sqrt{(\vec{v}\cdot\vec{v})(\vec{w}\cdot\vec{w})}}\right| \leq 1 \,.$$

Also ist die Winkeldefinition aus Abschnitt *a)* sinnvoll.

d) 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.

Zunächst aber wollen wir sehen, was symmetrische und HERMITEsche Matrizen überhaupt mit den bislang diskutierten Skalarprodukten zu tun haben.

Dazu sei V ein endlichdimensionaler EUKLIDischer Vektorraum, und $\{\vec{b}_1, \ldots, \vec{b}_n\}$ sei eine Basis von V. Für zwei Vektoren

$$\vec{v} = \sum_{i=1}^{n} v_i\vec{b}_i \quad \text{und} \quad \vec{w} = \sum_{i=1}^{n} w_i\vec{b}_i$$

ist dann 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, die das Skalarprodukt eindeutig festlegt: Für zwei beliebige Vektoren $\vec{v}, \vec{w}$ wie oben ist

$$\vec{v} \cdot \vec{w} = \sum_{i=1}^{n} \sum_{j=1}^{n} v_i w_j \cdot c_{ij} \, .$$

Umgekehrt definiert diese Formel für eine symmetrische Matrix (c_{ij}) 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 $\{\vec{b}_1, \dots, \vec{b}_n\}$, so ist nun für zwei Vektoren

$$\vec{v} = \sum_{i=1}^{n} v_i \vec{b}_i \quad \text{und} \quad \vec{w} = \sum_{i=1}^{n} w_i \vec{b}_i$$

mit $v_i, w_i \in \mathbb{C}$

$$\vec{v} \cdot \vec{w} = \left(\sum_{i=1}^{n} v_i \vec{b}_i \right) \cdot \left(\sum_{j=1}^{n} w_j \vec{b}_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} v_i \overline{w_j} \, \vec{b}_i \cdot \vec{b}_j \, ,$$

und 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 ${}^t\!A = \overline{A}$ ist.

c) Zu einem Vektor

$$\vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \quad \text{sei} \quad \overline{\vec{v}} = \begin{pmatrix} \overline{v_1} \\ \vdots \\ \overline{v_n} \end{pmatrix}$$

der *konjugiert komplexe Vektor.*

(Letztere Schreibweise sieht zwar grausam aus, vor allem wann man sie mit der sehr viel schöneren Alternative $\overline{\mathfrak{v}}$ vergleicht, sie läßt sich aber nicht vermeiden, wenn man Vektoren mit Pfeilen kennzeichnen will.)

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 sämtliche Rechenoperationen respektiert, ist dann auch

$$ \overline{A}\,\overline{\vec{v}} = \overline{\lambda}\,\overline{\vec{v}} \,, $$

also ist

$$ {}^t\vec{v}\overline{A}\,\overline{\vec{v}} = \overline{\lambda}\,{}^t\vec{v}\overline{\vec{v}} \,. $$

Andererseits ist ${}^t(A\vec{v}) = {}^t\vec{v}\;{}^tA$ und da A als HERMITEsch vorausgesetzt wird, ist ${}^tA = \overline{A}$. Also ist ${}^t\vec{v}\,\overline{A} = \lambda\,{}^t\vec{v}$ und

$${}^t\vec{v}\,\overline{A}\,\overline{\vec{v}} = \lambda\,{}^t\vec{v}\overline{\vec{v}}\,.$$

Somit ist

$$\lambda\,{}^t\vec{v}\overline{\vec{v}} = \overline{\lambda}\,{}^t\vec{v}\overline{\vec{v}}\,,$$

und da ${}^t\vec{v}\overline{\vec{v}}$ für einen Vektor $\vec{v} \neq \vec{0}$ nicht verschwindet, folgt $\lambda = \overline{\lambda}$, d.h. λ ist reell.

b) $\vec{v}$ sei Eigenvektor zum Eigenwert λ, und $\vec{w}$ sei Eigenvektor zum davon verschiedenen Eigenwert μ, d.h.

$$A\vec{v} = \lambda\vec{v} \quad \text{und} \quad A\vec{w} = \mu\vec{w}\,.$$

Dann ist

$$\lambda\,{}^t\vec{v}\overline{\vec{w}} = {}^t(\lambda\vec{v})\overline{\vec{w}} = {}^t(A\vec{v})\overline{\vec{w}} = {}^t\vec{v}\,{}^tA\overline{\vec{w}} = {}^t\vec{v}\overline{A}\,\overline{\vec{w}} = {}^t\vec{v}\overline{A\vec{w}} = {}^t\vec{v}\overline{\mu\vec{w}} = \overline{\mu}\,{}^t\vec{v}\overline{\vec{w}}\,,$$

und da wir bereits wissen, daß alle Eigenwerte reell sind, ist $\overline{\mu} = \mu \neq \lambda$. 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.

c) λ sei ein Eigenwert von A mit geometrischer Vielfachheit r, der zugehörige Eigenraum habe also die Dimension r. Wir wählen eine Basis $\{\vec{b}_1, \ldots, \vec{b}_r\}$ davon; 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.

Diese Basis ergänzen wir nun (nach dem Basisergänzungssatz aus §5d) zu einer Basis $\mathcal{B} = \{\vec{b}_1, \ldots, \vec{b}_n\}$ des gesamten Vektorraums V; auch hier können wir wieder annehmen, daß es sich 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 natürlich A als Abbildungsmatrix; für uns interessanter ist aber die Abbildungsmatrix $A_{\mathcal{B}}$ bezüglich der neuen Basis $\mathcal{B}$. 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 also

$$ {}^tB\overline{B} = E \quad \text{und damit} \quad {}^tB = \overline{B}^{\,-1} = \overline{B^{-1}}\,. $$

Aus dieser Formel folgt, daß mit A auch $A_{\mathcal{B}}$ eine HERMITEsche Matrix ist, denn

$$ {}^tA_{\mathcal{B}} = {}^t(B^{-1}AB) = {}^tB\;{}^tA\;{}^tB^{-1} = \overline{B^{-1}}\,\overline{A}\,\overline{B} = \overline{B^{-1}AB} = \overline{A_{\mathcal{B}}}\,. $$

Da die ersten r Basisvektoren $\vec{b}_i$ Eigenvektoren von A zum Eigenwert λ sind, gilt für $i \le r$, daß $\varphi(\vec{b}_i) = \lambda\vec{b}_i$ ist, d.h. in der i-ten Spalte von $A_{\mathcal{B}}$ steht an der i-ten Stelle die reelle Zahl λ und ansonsten überall die Null. Insbesondere steht also in den ersten i Spalten unterhalb der Stelle r überall Null; da $A_{\mathcal{B}}$ eine HERMITEsche Matrix ist, steht damit auch in den ersten i Zeilen rechts von der Stelle r überall Null; die Matrix $A_{\mathcal{B}}$ hat also die Form

$$ A_{\mathcal{B}} = \begin{pmatrix} \lambda & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \lambda & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & M & \\ 0 & 0 & \dots & 0 & & & \end{pmatrix}, $$

wobei M eine $(n - r) \times (n - r)$-Matrix ist, über die wir relativ wenig wissen.

Nach diesen vielen Vorbereitungen können wir endlich daran gehen, die algebraische Vielfachheit des Eigenwerts λ zu berechnen. Dazu brauchen wir das charakteristische Polynom von A; nach dem Multiplikationssatz für Determinanten ist

$$ \begin{aligned} \det(A_{\mathcal{B}} - xE) &= \det(B^{-1}AB - xE) = \det\bigl(B^{-1}(A - xE)B\bigr) \\ &= \det(B)^{-1}\det(A - xE)\det B = \det(A - xE)\,, \end{aligned} $$

denn die Diagonalmatrix xE ist mit jeder anderen Matrix vertauschbar. Somit haben A und $A_{\mathcal{B}}$ dasselbe charakterische Polynom; da wir die Matrix $A_{\mathcal{B}}$ besser kennen, rechnen wir mit ihr.

Die Matrix $A_{\mathcal{B}} - xE$ hat die Form

$$\begin{pmatrix} \lambda - x & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \lambda - x & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda - x & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & & & \\ \vdots & \vdots & \ddots & \vdots & & M - xE_{n-r} & \\ 0 & 0 & \dots & 0 & & & \end{pmatrix},$$

wobei E_{n-r} die $(n-r) \times (n-r)$-Einheitsmatrix bezeichnet.

Eine ähnliche Matrix haben wir bereits am Ende von §*7h* betrachtet beim Beweis, daß die algebraische Vielfachheit eines Eigenwerts immer größer oder gleich der geometrischen ist; genau wie dort folgt aus dem LAPLACEschen Entwicklungssatz, daß

$$\det(A - xE) = \det(A_{\mathcal{B}} - xE) = (\lambda - x)^r \det(M - xE_{n-r}),$$

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 $A_{\mathcal{B}}$ 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 $A_{\mathcal{B}}$ 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 $\vec{b}_1, \dots, \vec{b}_r$ 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: Wir wissen, daß die Summe der algebraischen Vielfachheiten aller Eigenwerte gleich der Dimension n des Vektorraums ist; da die algebraischen gleich den geometrischen Vielfachheiten sind, gibt es also insgesamt n Eigenvektoren, die eine Basis von V bilden.

Für jeden einzelnen Eigenraum können wir die Eigenvektoren so wählen, daß sie eine Orthonormalbasis bilden; da Eigenvektoren zu verschiedenen Eigenwerten stets orthogonal sind, ist die Vereinigungsmenge dieser Basen Orthonormalbasis von V.

■

e) Quadriken im $\mathbb{R}^n$

Symmetrische Matrizen sind auch ein nützliches Werkzeug für die Geometrie (und haben über diese viele weiterreichende Anwendungen beispielsweise in der mehrdimensionalen Analysis). Aus der Schule ist, zumindest für die Ebene und vielleicht auch den dreidimensionalen Raum, bekannt, wie sich Geraden und Ebenen durch lineare Gleichungssysteme beschreiben lassen. In diesem Abschnitt wollen wir die einfachste nichtlineare Verallgemeinerung solcher Fragen betrachten, die *Quadriken:*

Definition: Unter einer Quadrik im $\mathbb{R}^n$ verstehen wir die Menge aller Punkte $(x_1, \dots, x_n)$, die einer quadratischen Gleichung

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{k=1}^{n} b_k x_k = c$$

genügen mit $a_{ij}, b_k, c \in \mathbb{R}$.

Nachdem wir uns nun schon rund ein halbes Semester mit linearer Algebra beschäftigen, ist hoffentlich jedermann klar, daß diese Definition geradezu schreit nach einer Umformulierung mit Vektoren und

Matrizen: Mit der Matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ und den Vektoren

$$\vec{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \quad \text{und} \quad \vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

ist

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j + \sum_{k=1}^{n} b_kx_k = {}^t\vec{x}A\vec{x} + \vec{b}\cdot\vec{x}\,,$$

wobei wir auch hier wieder Vektoren mit $1 \times n$-Matrizen identifizieren und $\cdot$ für das Standardskalarprodukt im $\mathbb{R}^n$ steht.

Durch diese kompaktere Schreibweise hat sich natürlich noch nichts wesentliches geändert; die wirkliche Vereinfachung kommt erst, wenn wir uns klarmachen, daß wir hier *Geometrie* betreiben und daß es für einen geometrischen Sachverhalt völlig irrelevant ist, ob und gegebenenfalls wo wir Koordinatenachsen ins Bild zeichnen.

Bevor wir ein günstiges Koordinatensystems wählen, sollten wir uns zuerst überlegen, welche Freiheiten wir haben, um die Gleichung der Quadrik zu vereinfachen.

Wegen der Kommutativität der Multiplikation in $\mathbb{R}$ ist

$$a_{ij}x_ix_j + a_{ji}x_jx_i = (a_{ij} + a_{ji})x_ix_j\;;$$

wir können daher nach Belieben einen der beiden Koeffizienten um einen gewissen Betrag vergrößern und den anderen entsprechend verkleinern, ohne daß sich an der Gesamtsumme etwas ändert.

Eine denkbare Strategie wäre, einen der beiden Koeffizienten auf Null zu bringen; auf diese Weise könnten wir erreichen, daß A zu einer Dreiecksmatrix wird. Das wäre sicherlich eine Vereinfachung, allerdings wissen wir nur wenig über die Eigenvektoren einer Dreiecksmatrix. (Die *Eigenwerte* sind dafür sehr einfach zu bestimmen: Wie man sich leicht überlegt, sind es einfach die Einträge in der Diagonalen.)

Eine bessere Alternative besteht darin, daß wir sowohl a_{ij} als auch a_{ji} durch den gemeinsamen Mittelwert $\frac{1}{2}(a_{ij}+a_{ji})$ ersetzen; dadurch wird A zu einer symmetrischen Matrix, und über symmetrische Matrizen wissen wir aus dem vorigen Paragraphen ziemlich viel; insbesondere wissen wir, daß der $\mathbb{R}^n$ eine Orthonormalbasis aus Eigenvektoren von A hat. Die Richtungen dieser Vektoren bezeichnen wir als die *Hauptachsen* der Quadrik, und die Transformation der Quadrik in das entsprechende Koordinatensystems als *Hauptachsentransformation.*

In dieser Basis ausgedrückt wird A zu einer Diagonalmatrix und der quadratische Teil der Gleichung vereinfacht sich zu

$$^t\vec{x}A\vec{x} = \sum_{i=1}^{n} a_{ii}x_i^2 \,.$$

Wir können noch weiter vereinfachen. Dazu betrachten wir zunächst den zwar völlig uninteressanten, dafür aber sehr einfachen Fall $n = 1$. Hier ist die allgemeinste Quadrikgleichung

$$ax^2 + bx = c \,.$$

Die klassische (über drei Tausend Jahre alte) Lösungsmethode für quadratische Gleichungen durch quadratischen Ergänzung besteht darin, im Falle $a \neq 0$ diese Gleichung in der Form

$$a\left(x + \frac{b}{2a}\right)^2 = c + \frac{b^2}{4a}$$

oder kurz

$$\alpha y^2 = \gamma$$

zu schreiben mit der neuen Variablen $y = x + b/2a$, wobei wir die ganze Gleichung durch ein von Null verschiedenes γ dividieren können, so daß wir noch zusätzlich annehmen können, daß $\gamma = 0$ oder 1 ist.

Dieser neuen Form sieht man sofort die Lösungsmenge der Gleichung an: Für $\gamma = 1$ und positives α gibt es zwei verschiedene Lösungen, während es für negatives α (und auch den hier ausgeschlossenen Fall $\alpha = 0$) keine gibt.

Für $\gamma = 0$ gibt es im Falle $\alpha \neq 0$ die zweifache Nullstelle Null und für $\alpha = 0$ ist jede reelle Zahl Lösung der dann zu $0 = 0$ degenerierten Gleichung.

Genauso können wir die Methode der quadratischen Ergänzung auch auf die Gleichung

$$\sum_{i=1}^{n} a_{ii}x_i^2 + \sum_{i=1}^{n} b_i x_i = c$$

anwenden: Für jedes i, für das a_{ii} nicht verschwindet, können wir durch quadratische Ergänzung oder – geometrisch ausgedrückt – Verschiebung des Nullpunkts mittels der Transformation $x_i \mapsto x_i + b_i/2a_{ii}$ erreichen, daß $b_i = 0$ wird. (Der Einfachheit halber bezeichnen wir die neuen Variablen genau wie die alten wieder mit x_i.)

Weiter ändert sich nichts an der Lösungsmenge einer Gleichung, wenn wir die gesamte Gleichung mit einer von Null verschiedenen Zahl multiplizieren; daher können wir erreichen, daß c, sofern es nicht verschwindet, zu eins wird. Wir können also für das folgende davon ausgehen, daß $c = 0$ oder 1 ist.

Schließlich ist noch die Reihenfolge, in der wir die n Koordinatenachsen benennen, unserer Willkür überlassen; wir legen sie so fest, daß die ersten a_{ii} positiv sind, dann die negativen kommen und zum Schluß die Nullen. Es gibt dann also Zahlen $0 \leq s \leq r \leq n$, so daß a_{ii} für $1 \leq i \leq s$ positiv ist, für $s < i \leq r$ negativ und für $i > r$ Null. Insbesondere ist also $b_i = 0$ fir $i \leq r$. Wir setzen $\lambda_i \underset{\text{def}}{=} |a_{ii}|$ und schreiben die Gleichung in der Form

$$\sum_{i=1}^{s} \lambda_i x_i^2 - \sum_{j=s+1}^{r} \lambda_j x_j^2 + \sum_{k=s+1}^{n} b_i x_i = c$$

mit $c = 0$ oder 1.

Die hier auftretenden Zahlen r und s sind wichtige Kenngrößen der Quadrik: Wir bezeichnen r als den *Rang* der Quadrik und das Paar $(s, r - s)$ als *Signatur*.

Falls $c = 0$ ist, können wir die gesamte Gleichung mit -1 multiplizieren um aus der Signatur $(s, r - s)$ die Signatur $(r - s, s)$ zu machen; in diesem Fall können wir daher noch zusätzlich fordern, daß $s \geq r - s$ sein soll.

Als letzte Vereinfachung können wir schließlich im Fall $r < n$, falls es mindestens ein $b_i \neq 0$ gibt, stets $c = 0$ erreichen, indem wir für $c = 1$ die Variable x_i durch $x_i + 1/b_i$ ersetzen.

Nachdem wir die Gleichung nun genügend vereinfacht haben, wollen wir in den Dimensionen zwei und drei genauer untersuchen, welche Möglichkeiten es gibt.

1) Quadriken im $\mathbb{R}^2$***:*** Hier bezeichnen wir die Koordinaten wie üblich mit x und y anstelle von x_1 und x_2. Die positiven reellen Zahlen λ_1 und λ_2, sofern sie auftreten, können wir als Quadrate reeller Zahlen schreiben; um nachher bekannt aussehende Gleichungen zu bekommen, schreiben wir sie als *Kehrwerte* von Quadraten:

$$\lambda_1 = \frac{1}{a^2} \quad \text{und} \quad \lambda_2 = \frac{1}{b^2} .$$

Der Rang einer Quadrik kann höchstens gleich der Dimension des zugrundeliegenden Raums sein, also ist $r \leq 2$.

Beginnen wir mit dem Fall $r = 2$. Dazu gibt es die drei Unterfälle $s = 0$, $s = 1$ und $s = 2$.

Für $s = 2$ haben wir, je nachdem, ob $c = 0$ oder $c = 1$ ist, die beiden Normalformen

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 0 \quad \text{und} \quad \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 .$$

Die erste Gleichung hat nur die eine Lösung $x = y = 0$, die Quadrik ist also zu einem Punkt degeneriert, und im zweiten Fall haben wir die in Abbildung 37 dargestellte *Ellipse* mit Halbachsen a und b.

Nächster Fall ist $s = 1$; hier haben wir die beiden Normalformen

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 0 \quad \text{und} \quad \frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 .$$

Wegen

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = \left(\frac{x}{a} + \frac{y}{b}\right)\left(\frac{x}{a} - \frac{y}{b}\right)$$

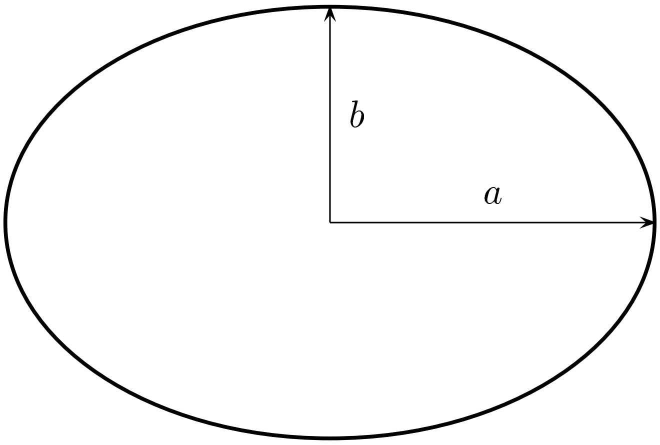


Abb. 37: Ellipse mit Halbachsen a und b

besteht die erste Quadrik aus den beiden Geraden

$$y = \pm \frac{b}{a} x \,,$$

die sich im Nullpunkt schneiden; die zweite ist die in Abbildung 38 zu sehende *Hyperbel.* Der Parameter a ist der Achsenabschnitte auf der x-Achse, und die (gestrichelt eingezeichneten) Geraden $y = \pm\frac{b}{a}x$ sind die Asymptoten.

Als letzter Fall einer Quadrik mir Rang 2 bleibt $s = 0$; hier haben alle Quadrate negatives Vorzeichen. Da hier $s < r - s$ ist, liefert dieser Fall nur für $c = 1$ etwas neues, und auch da beschreibt

$$-\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

nur die leere Menge.

Gehen wir also weiter zu Rang 1; hier gibt es die Unterfälle $s = 1$ und $s = 0$.

Für $s = 1$ gibt es die Normalformen

$$\frac{x^2}{a^2} = 1 \quad \text{und} \quad \frac{x^2}{a^2} + by = 0 \,.$$

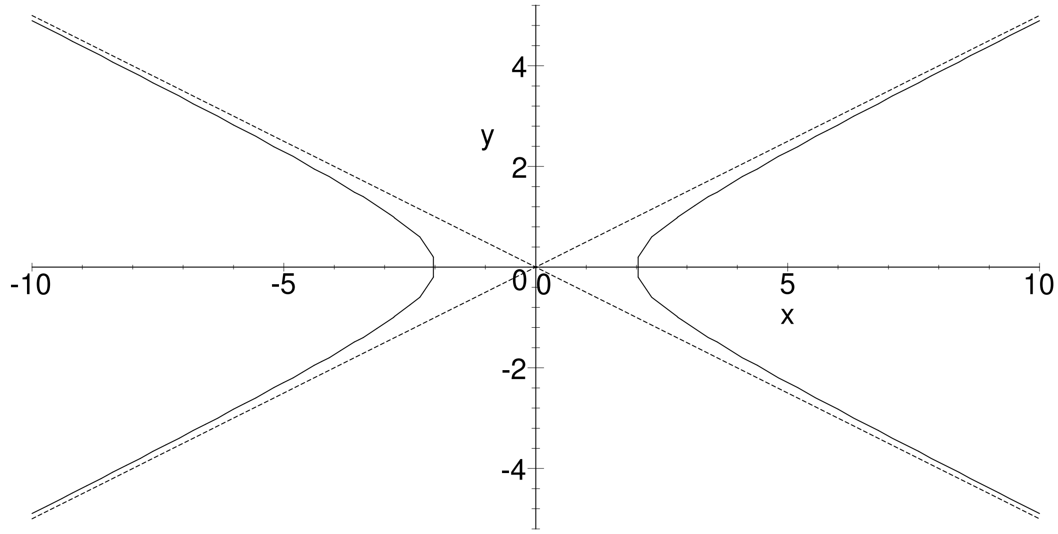


Abb. 38: Hyperbel $x^2/4 - y^2 = 1$ mit Asymptoten

Erstere definiert die beiden parallelen Geraden $x = \pm a$, letztere die *Parabel* $y = \frac{-1}{a^2 b} x^2$ aus Abbildung 39.

$s = 0$ liefert, da $s < r - s$ ist, nur für $c = 1$ etwas neues, und auch das nur, wenn keine linearen Terme auftreten, da wir sonst sofort auf $c = 0$ reduzieren können. Was bleibt ist also die Gleichung

$$-\frac{x^2}{a^2} = 1 \quad \text{oder} \quad x^2 = -a^2 ,$$

die wieder einmal die leere Menge definiert.

Damit kommen wir zum Rang Null, d.h. zu den „Quadriken", in deren Gleichung kein quadratischer Term auftritt. Die allgemeinste Gleichung ist

$$ax + by = c ,$$

d.h. also die allgemeine Geradengleichung, die in Spezialfällen auch wieder die leere Menge (setze $a = b = 0$ und $c = 1$) oder auch die gesamte Ebene (für $a = b = c = 0$) als Lösung haben kann.

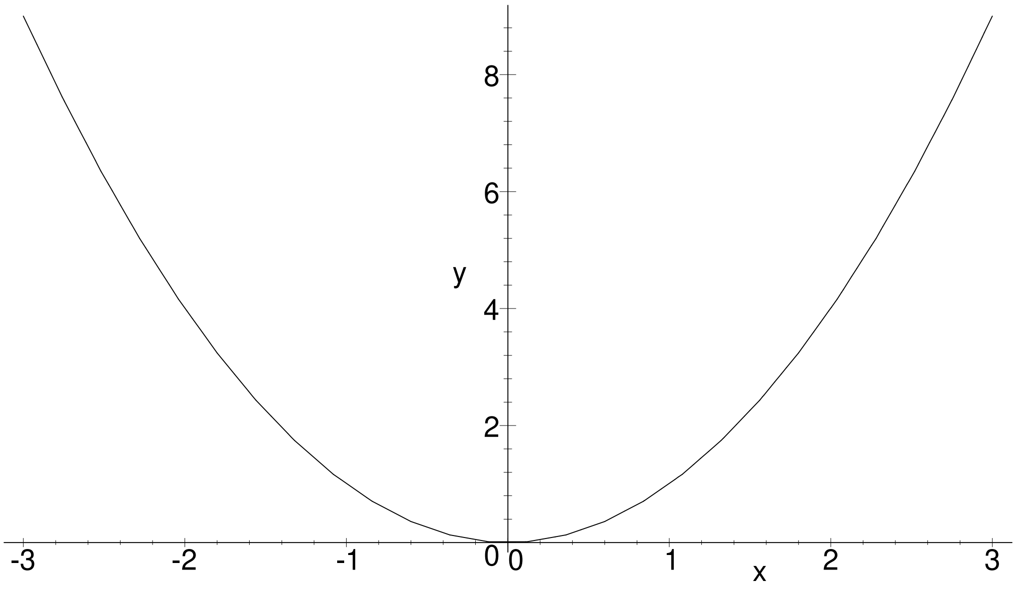


Abb. 39: Die Parabel $y = x^2$

2) Quadriken im $\mathbb{R}^3$***:*** Hier gibt es naturgemäß schon mehr Möglichkeiten. Wir bezeichnen die drei Variablen mit x, y, z und setzen in Analogie zum zweidimensionalen Fall

$$\lambda_1 = \frac{1}{a^2}, \quad \lambda_2 = \frac{1}{b^2} \quad \text{und} \quad \lambda_3 = \frac{1}{c^2}\,.$$

Auch hier beginnen wir wieder mit dem größtmöglichen Rang $r = 3$.

Falls auch $s = 3$ den größtmöglichen Wert hat, bekommen wir die beiden Normalformen

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 0 \quad \text{und} \quad \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1\,,$$

deren erstete wieder nur einen Punkt als Lösung hat. Die zweite Gleichung, deren Lösungsmenge man in Abbildung 40 sehen kann, ist die offensichtliche Verallgemeinerung der Ellipsengleichung aus dem $\mathbb{R}^2$; man spricht daher hier von einem *Ellipsoid.*

Für $s = 2$ sind die entsprechenden Normalformen

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 0 \quad \text{und} \quad \frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1\,.$$

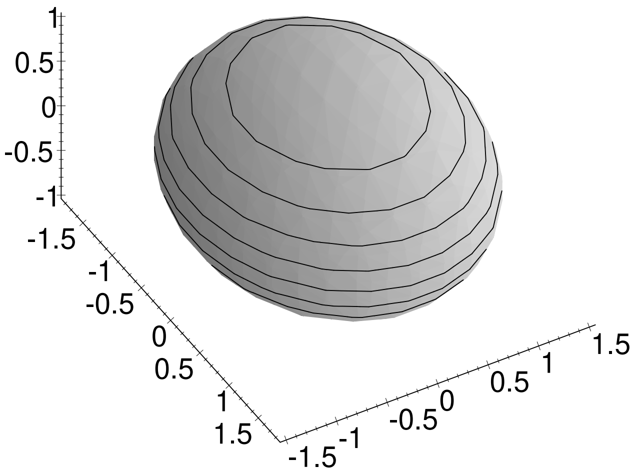


Abb. 40: Das Ellipsoid $\frac{x^2}{3} + \frac{y^2}{2} + z^2 = 1$

Die erste dieser Gleichungen können wir auch als

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = \left(\frac{z}{c}\right)^2$$

schreiben, und aus dieser Form sehen wir, daß die Lösungsmenge für konstantes z eine Ellipse ist, deren Halbachsen linear mit z ansteigen; wir haben also einen (elliptischen) *Kegel* mit Spitze im Nullpunkt. Abbildung 41 zeigt, was Maple daraus macht; man sollte also Computern trotz deren unbestreitbarer Nützlichkeit nicht zu sehr vertrauen.

Die zweite Gleichung hat für konstantes x oder y Hyperbeln als Schnittkurven und für konstantes z Ellipsen; man bezeichnet die in Abbildung 42 dargestellte zugehörige Quadrik als *einschaliges Hyperboloid.*

Der Zusatz „einschaliges“ wird klar, wenn wir den nächsten Fall $s = 1$ betrachten; da hier $s < r - s$ ist, interessiert nur die Gleichung

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$$

mit $c = 1$. Sie hat für konstantes y oder z Hyperbeln als Schnittkurven und für konstantes x entweder die leere Menge oder eine Ellipse.

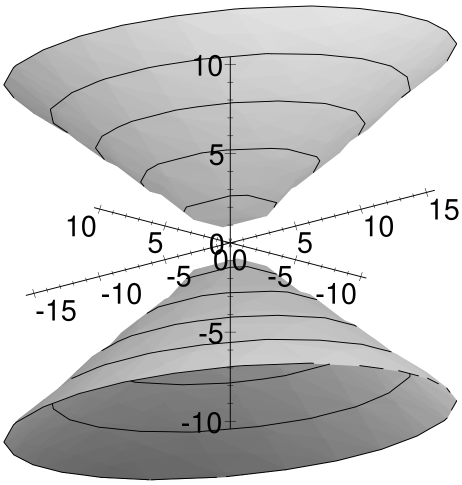


Abb. 41: Der Kegel $x^2 + y^2 = z^2$

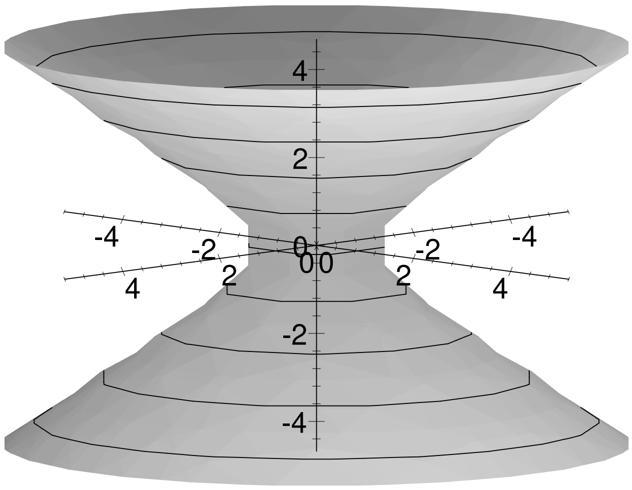


Abb. 42: Das einschalige Hyberboloid $x^2 + y^2 - z^2 = 1$

Da die Schnittmenge für betragskleine x leer ist, zerfällt die Quadrik in zwei Teile; sie heißt deshalb *zweischaliges Hyperboloid* und ist in Abbildung 43 zu sehen.

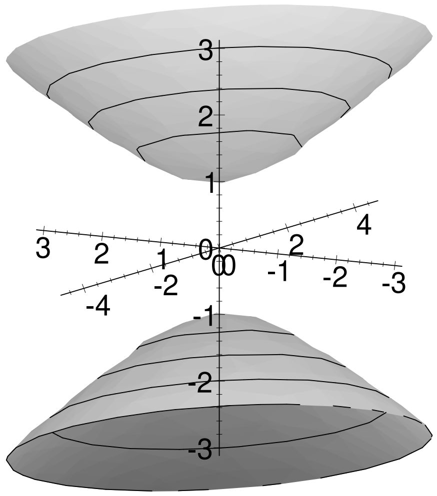


Abb. 43: Das zweischalige Hyberboloid $x^2 - y^2/2 - z^2 = 1$

Für $s = 0$ und konstanten Term eins definiert

$$-\frac{x^2}{a^2} - \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$$

natürlich wieder die leere Menge.

Damit kommen wir zum Rang zwei. Für den maximal möglichen Wert $s = 2$ haben wir die beiden Normalformen

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \beta z = 0 \quad \text{und} \quad \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1\,,$$

denn $c = 1$ ist nur möglich, wenn kein linearer Term auftritt.

Bei der ersten Gleichung müssen wir unterscheiden, ob β verschwindet oder nicht. Für $\beta \neq 0$ erhalten wir eine Quadrik, deren Schnittkurven für konstantes z Ellipsen sind und für konstantes x oder y Parabeln. Man

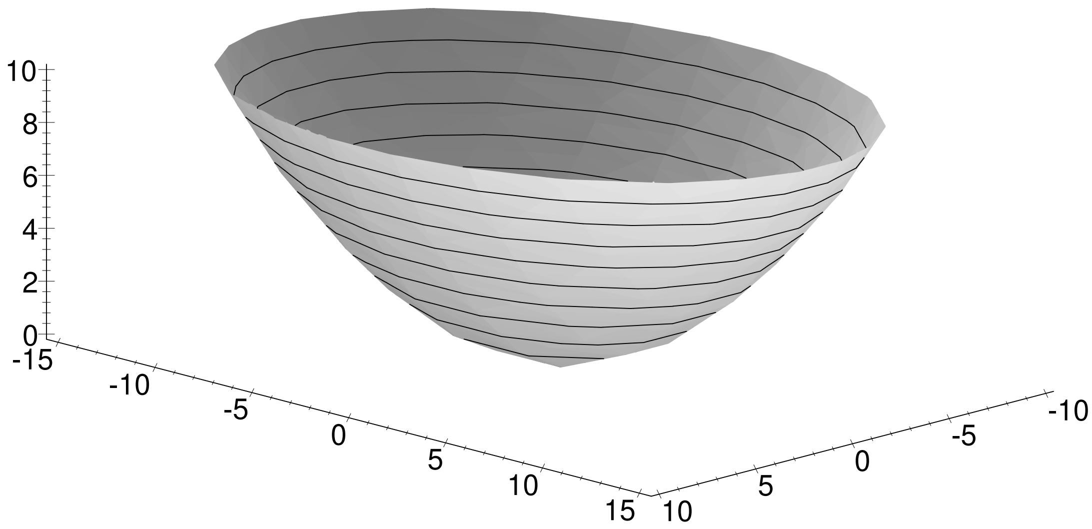


Abb. 44: Das elliptischen Paraboloid $x^2 + y^2/2 - 10z = 0$

spricht daher von einem *elliptischen Paraboloid;* es ist in Abbildung 44 dargestellt.

Für $\beta = 0$ gibt es nur die Lösung $x = y = 0$, die Gleichung ist also eine komplizierte Darstellung der z-Achse des Koordinatensystems, d.h wir haben eine (doppelt zu zählende) Gerade.

Die zweite Gleichung ist identisch mit der Ellipsengleichung aus dem $\mathbb{R}^2$, allerdings gibt es nun im $\mathbb{R}^3$ noch eine dritte Koordinate, die in der Gleichung nicht vorkommt und die somit beliebig gewählt werden kann; die Quadrik ist also ein Zylinder mit elliptischer Grundlinie, d.h. ein *elliptischer Zylinder.*

Der Fall $s = 1$ liefert die beiden Normalformen

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} + \beta z = 0 \quad \text{und} \quad \frac{x^2}{a^2} - \frac{y^2}{b^2} = 1\,.$$

Letztere definiert einen Zylinder mit hyperbolischer Grundlinie, erstere hat für konstantes z Hyperbeln als Schnittkurven und für konstantes x oder y Parabeln; man spricht daher von einem *hyperbolischen Parabo-*

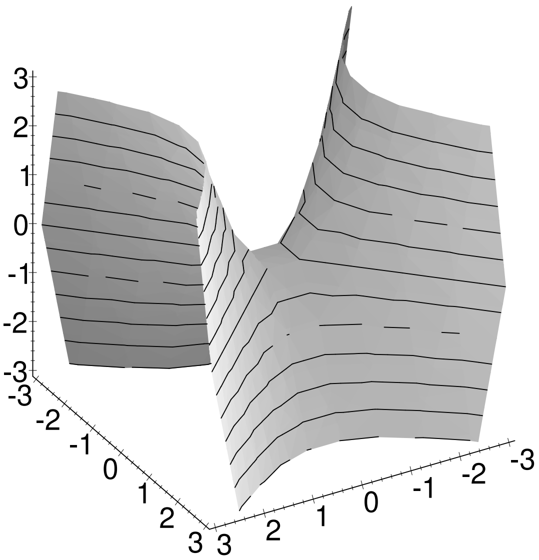


Abb. 45: Die Sattelfläche $x^2 - y^2 - z = 0$

loid oder, aus Gründen, die man in Abbildung 45 erkennen kann, einer *Sattelfläche.*

Für $s = 0$ schließlich ist entsprechend nur

$$-\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

interessant, und diese Gleichung ist unlösbar.

Damit ist der Fall $r = 2$ erledigt; wir kommen zu $r = 1$. Für $s = 1$ haben wir die Normalformen

$$\frac{x^2}{a} + by + cz = 0 \quad \text{und} \quad \frac{x^2}{a} = 1\,.$$

Letztere definiert die beiden parallelen Ebenen $x = \pm a$, erstere können wir durch den Koordinatenwechsel $y \mapsto by + cz$ vereinfachen zu $x^2 + a^2 y = 0$, d.h. wir haben einen Zylinder mit einer Parabel als Grundlinie.

Für $s = 0$ sind die Normalformen entsprechend

$$-\frac{x^2}{a} + by + cz = 0 \quad \text{und} \quad -\frac{x^2}{a} = 1\,;$$

nun ist die zweite Gleichung unlösbar, und die erste definiert weiterhin einen parabolischen Zylinder.

Damit bleibt nur noch der Fall $r = 0$, in dem wir keine echte Quadrik haben, sondern eine Ebene, die gegebenfalls auch zur leeren Menge oder zum gesamten $\mathbb{R}^3$ degenerieren kann.

§9: Komprimierung von Bilddaten

Zum Abschluß des Semesters wollen wir wenigstens noch kurz eine praktische Anwendung von Eigenvektoren und Matrizen kennenlernen, die Komprimierung von Bilddaten und Audiodaten.

In der Vorlesung wurde dieses Thema nur kurz am Ende der letzten Vorlesungsstunde behandelt; hier im Skriptum sollen die Grundlagen etwas ausführlicher dargestellt werden, wobei wir uns allerdings bei den statistischen Grundlagen auch hier eher auf Veranschaulichung als auf exakte Definitionen oder gar Beweise beschränken müssen. Deshalb und wegen der erheblich knapperen Behandlung in der Vorlesung ist dieser Paragraph für Prüfungszwecke kein Teil der Vorlesung.

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.

Bild- und Audiodaten werden typischerweise als Matrizen ganzer Zahlen zwischen 0 und 255 *bzw.* Vektoren ganzer Zahlen zwischen 0 und 65535 ($= 2^{16} - 1$) digitalisiert. (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 Gehöhr noch auf sehr feine Dynamikunterschiede 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 beispielsweise 256×256 oder 512×512 Bildpunkte abgetastet (je nach Seitenverhältnis des Bilds und den Beschränkungen des Scanners kommen auch fast beliebige andere Bildgrößen vor), bei Schwarzweißbilder nur nach Helligkeit, bei Farbbildern nach insgesamt drei physikalischen 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 und dementsprechend meist klein ist.

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 wohl 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.

Alle Verfahren, die beliebige, aber realistische Bilder komprimieren möchten, 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.

Damit kommen wir in Bereiche der Mathematik, die jenseits des Stoffs dieser Vorlesung liegen; der erste (und wichtigste) statistische Begriff, den wir zwar unbedingt brauchen, aber mit unseren bisherigen Mathematikkenntnissen nicht exakt definieren können, ist der der *Zufallsvariablen.* Unter einer (diskreten) Zufallsvariablen wollen wir uns einen Prozeß vorstellen, der uns zufällig einen Wert aus einer vorgegebenen endlichen Menge

$$\{x_0, \dots, x_N\}$$

liefert. Dieser „Zufall“ muß natürlich, falls er mathematisch faßbar sein soll, irgendwelchen Regeln genügen, und hier kommt der zweite fundamentale, aber hier nicht definierte Begriff zum Einsatz: Wir nehmen an, daß für jeden der möglichen Werte x_i feststeht, mit welcher *Wahrscheinlichkeit* p_i er angenommen wird, wobei hoffentlich jeder Leser eine gewisse Vorstellung davon hat, was der Begriff „Wahrscheinlichkeit p_i“ bedeuten soll und einige vielleicht auch in der Schule die eine oder andere Definition dafür kennengelernt haben. Hier seien nur zwei formale Konsequenzen erwähnt:

$$0 \le p_i \le 1 \quad \text{für alle } i \text{ und} \quad \sum_{i=0}^{N} p_i = 1\,.$$

Sollen beispielsweise alle x_i mit gleicher Wahrscheinlichkeit angenommen werden (was für Bilddaten nicht unbedingt ein sehr realistisches Modell ist), so sind alle $p_i = 1/(N+1)$.

In unserem Modell soll ein „Bild“ dann produziert werden durch eine Matrix (X_{ij}) von geeigneten Zufallsvariablen; eine „Audiosequenz“ entsprechend durch einen Vektor $(X_0, \dots, X_N)$.

Welche Zufallsvariablen sind geeignet? In unserem einfachen Modell wollen wir uns nicht darauf festlegen, wie die p_i gewählt werden sollen, sondern stattdessen nur einige schwächere Kenngrößen der Zufallsvariablen festlegen, die wir, modulo der vagen Vorbereitungen von oben, nun exakt definieren können; weiter hinten sind zur Veranschaulichung sechs in der Bildverarbeitung beliebte Testbilder zusammen mit den Werten dieser Kenngrößen abgedruckt. Diese Werte sind entnommen aus

P.M. FARELLE: Recursive Block Coding for Image Data Compression, *Springer,* 1990 ;

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 abgedruckten Bilder in derselben Weise verunstaltet sind.

Definition: Der *Erwartungswert* $E(X)$ einer Zufallsvariablen X ist

$$E(X) = \sum_{i=0}^{N} p_i x_i \,.$$

Falls alle x_i mit gleicher Wahrscheinlichkeit angenommen werden, ist der Erwartungswert also einfach das arithmetische Mittel

$$\frac{1}{N+1} \sum_{i=0}^{N} x_i$$

der möglichen Werte; ansonsten ist er ein sogenanntes gewichtetes arithmetisches Mittel.

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 Wert 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“.

Peppers

$$\begin{aligned} \mu &= 115{,}6 \\ \sigma^2 &= 5632 \\ \sigma &= 75{,}0 \\ \rho &= 0{,}98 \\ x_{\text{min}} &= 0 \\ x_{\text{max}} &= 237 \end{aligned}$$

Lenna

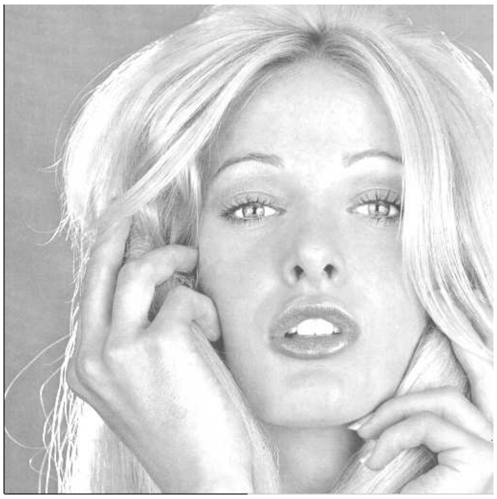
$$\begin{aligned} \mu &= 99{,}1 \\ \sigma^2 &= 2796 \\ \sigma &= 52{,}9 \\ \rho &= 0{,}97 \\ x_{\text{min}} &= 3 \\ x_{\text{max}} &= 248 \end{aligned}$$

Sailboat

$$\begin{aligned} \mu &= 124{,}3 \\ \sigma^2 &= 6027 \\ \sigma &= 77{,}6 \\ \rho &= 0{,}97 \\ x_{\text{min}} &= 0 \\ x_{\text{max}} &= 249 \end{aligned}$$

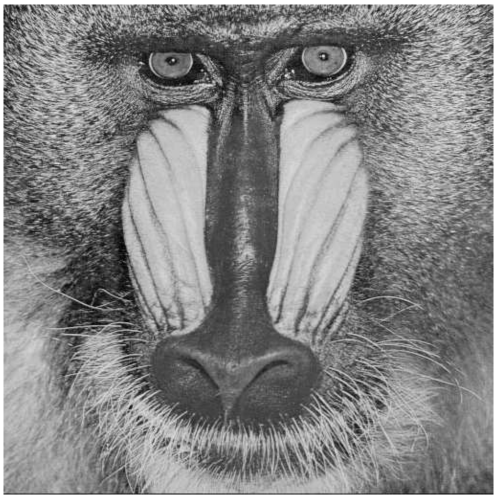
Stream

$$\begin{aligned} \mu &= 113{,}8 \\ \sigma^2 &= 2996 \\ \sigma &= 54{,}7 \\ \rho &= 0{,}94 \\ x_{\min} &= 0 \\ x_{\max} &= 255 \end{aligned}$$



Tiffany

$$\begin{aligned} \mu &= 208{,}6 \\ \sigma^2 &= 1126 \\ \sigma &= 33{,}6 \\ \rho &= 0{,}87 \\ x_{\min} &= 3 \\ x_{\max} &= 255 \end{aligned}$$



Baboon

$$\begin{aligned} \mu &= 128{,}9 \\ \sigma^2 &= 2282 \\ \sigma &= 47{,}8 \\ \rho &= 0{,}86 \\ x_{\min} &= 0 \\ x_{\max} &= 236 \end{aligned}$$

Die nächste wichtige Kenngröße ist die Varianz, welche angibt, wie stark eine Zufallsvariable um ihren Erwartungswert streut:

Definition: Die Varianz einer Zufallsvariablen X mit Erwartungswert $E(X)$ ist

$$\sigma_X^2 = E\big((X - E(X))^2\big) = \sum_{i=0}^{N} p_i\big(x_i - E(X)\big)^2 ;$$

ihre Standardabweichung ist $\sigma_X = \sqrt{\sigma_X^2}$.

Wir haben uns nicht ein halbes Semester mit linearer Algebra befaßt, um diese Definition so stehen zu lassen – insbesondere, da wir es im folgenden mit noch erheblich schlimmeren Summen zu tun bekommen werden. Wir definieren daher einen Vektor $\vec{v}_X \in \mathbb{R}^{N+1}$ mit Komponenten $v_{X0}, \dots, v_{XN}$, wobei

$$\vec{v}_{Xi} = \sqrt{p_i}\big(x_i - E(X)\big)$$

sein soll; dann ist

$$\sigma_X^2 = \vec{v}_X \cdot \vec{v}_X \quad \text{und} \quad \sigma_X = |\vec{v}_X| .$$

Auch hier wollen wir wieder davon ausgehen, daß alle Zufallsvariablen zu einem gegebenen Bild *bzw.* einer gegebenen Audiosequenz dieselbe Varianz haben; diese schätzen wir aufgrund der gegebenen Daten als

$$\sigma^2 = \frac{1}{N}\sum_{i=0}^{N}(x_i - \mu)^2 ,$$

wobei $x_0, \dots, x_N$ die Helligkeits- *bzw.* Lautstärkewerte sind. (Es ist kein Druckfehler, daß vor dieser Summe mit $N + 1$ Summanden nur N im Nenner steht; es liegt im wesentlichen daran, daß wir den Erwartungswert der Zufallsvariablen aufgrund des Bildes nur schätzen können. Im nächsten Semester, bei der Fehler- und Ausgleichsrechnung, werden wir solche Fragen genauer betrachten.)

Was die Varianz bedeutet, 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“.

Schließlich kommen wir zu der Eigenschaft, die wirklich hilft, reale Bilddaten zu komprimieren, der *Korrelation* zwischen zwei Zufallsvariablen.

Ausgangspunkt ist die Beobachtung, daß es in einem Bild oder Musikstück nur wenige aprupte Übergänge gibt. Bei einer Musik-CD etwa wird das Signal 44 100-mal pro Sekunde abgetastet; verglichen mit den Zeitspannen, innerhalb derer selbst das schrägste Musikstück große Sprünge macht, ist die 44 100-tel Sekunde so kurz, daß diese Sprünge in der Tat sehr seltene Ereignisse sind; wir können daher davon ausgehen, daß sich bei Bild- und Audiodaten die unmittelbaren Nachbarn eines Datums *im Mittel* nur wenig vom gegebenen Datum unterscheiden.

In unserem bisherigen Modell haben die Zufallsvariablen für zwei benachbarte Bildpunkte noch nichts miteinander zu tun; sie können voneinander völlig unabhängig sein. Die noch fehlende statistische Kenngröße muß also irgendwie die Abhängigkeit zwischen zwei Zufallsvariablen quantifizieren.

Um diese noch einzuführende Größe zu verstehen, betrachten wir für zwei Extremfälle den Erwartungswert

$$E\big((X - E(X))(Y - E(Y))\big) = \vec{v}_X \cdot \vec{v}_Y$$

für zwei Zufallsvariablen X und Y mit demselben Wertebereich:

a) Wenn X und Y voneinander unabhängig sind und der Wert x_i aus dem gemeinsamen Wertebereich von X mit Wahrscheinlichkeit p_i und von Y mit Wahrscheinlichkeit q_i angenommen wird, ist die Wahrscheinlichkeit dafür, daß $X = x_i$ und $Y = x_j$ ist, gleich dem Produkt $p_i q_j$. (In der Wahrscheinlichkeitstheorie betrachtet man dies als *Definition* für die

Unabhängigkeit der Ereignisse $X = x_i$ und $Y = x_j$.) Also ist

$$E\big((X - E(X))(Y - E(Y))\big) = \sum_{i=0}^{N}\sum_{j=0}^{N} p_i q_j (x_i - E(X))(x_j - E(Y))$$
$$= \sum_{i=0}^{N}\left(p_i(x_i - E(X))\sum_{j=0}^{N} q_j(x_j - E(Y))\right) = 0\,,$$

denn

$$\sum_{j=0}^{N} q_j(x_j - E(Y)) = \sum_{j=0}^{N} q_j x_j - \sum_{j=0}^{N} q_j E(Y) = \sum_{j=0}^{N} q_j x_j - 1\cdot\sum_{j=0}^{N} q_j x_j = 0\,.$$

b) Wenn X und Y grundsätzlich immer denselben Wert liefern, ist erstens stets $p_i = q_i$ und damit auch $E(X) = E(Y)$ sowie $\sigma_X = \sigma_Y$, und zweitens ist die Wahrscheinlichkeit des Ereignisses $X = x_i$ und $Y = x_j$ null, falls $i \neq j$, und p_i sonst. Also ist

$$E\big((X - E(X))(Y - E(Y))\big) = \sum_{i=0}^{N} p_i\big(x_i - E(X)\big)^2 = \sigma_X^2 = \sigma_X\sigma_Y\,.$$

Definition: *a)* Die Kovarianz zwischen den beiden Zufallsvariablen X und Y ist

$$\operatorname{cov}(X, Y) = E\big((X - E(X))(Y - E(Y))\big) = \vec{v}_X \cdot \vec{v}_Y\,.$$

b) Die Korrelation zwischen den beiden Zufallsvariablen X und Y ist

$$\rho(X, Y) = \frac{E\big((X - E(X))(Y - E(Y))\big)}{\sigma_X\sigma_Y} = \frac{\vec{v}_X \cdot \vec{v}_Y}{\sqrt{(\vec{v}_X \cdot \vec{v}_X)(\vec{v}_Y \cdot \vec{v}_Y)}}\,.$$

Damit haben zwei voneinander unabhängige Zufallsvariablen also Korrelation und Kovarianz null (man sagt auch, sie seien unkorreliert), während die Korrelation einer Zufallsvariablen mit sich selbst gleich eins ist. Allgemeiner rechnet man in völliger Analogie zu oben leicht nach, daß $\rho(X, Y)$ auch dann gleich ± 1 ist, wenn der Wert von Y stets eine lineare Funktion des Werts von X ist, wobei das Vorzeichen gleich dem des Koeffizienten von X ist.

Im Falle von Bild- oder Audiodaten erwarten wir nach dem oben Gesagten ganz sicher nicht, daß die verschiedenen Zufallsvariablen völlig unkorreliert sind, andererseits aber auch nicht, daß sie alle gleich sind, denn sonst wäre etwa das Bild eine einfarbige Fläche ohne jegliche Struktur.

Für unsere Anwendungen wollen wir vereinfachend annehmen, daß für ein vorgegebenes Bild oder Audiosignal alle Zufallsvariablen denselben Erwartungswert und dieselbe Varianz haben; für Bilder mit weißem oder schwarzem Rand oder auch für Symphonien mit grandiosem Finale ist das zwar sicherlich falsch, aber ein universelles alternatives Modell ist schwer vorstellbar, und in Einzelfällen kann man, falls man den Aufwand als lohnend betrachtet, immer noch beispielsweise den gesamten Bildrand getrennt vom Bild in einem anderen Format kodieren.

Auch für die Korrelationen zwischen den einzelnen Zufallsvariablen wollen wir eine möglichst einfache Annahme machen, allerdings ist es hier natürlich unrealistisch zu sagen, die Korrelation zwischen zwei Zufallsvariablen zu beliebigen Punkten solle immer dieselbe sein; wir nehmen daher nur an, daß die Zufallsvariablen zu zwei *benachbarten* Punkten immer dieselbe Korrelation ρ haben sollen. Diesen Wert ρ bezeichnen wir als die *Autokorrelation* des Bildes.

Die Autokorrelation und ihre Rolle für die Komprimierung von Bilddaten wird besser verständlich, wenn wir für zwei Zufallsvariablen X und Y mit Autokorrelation ρ die Differenz

$$Z = Y - \rho X$$

betrachten: Für diese ist

$$\begin{aligned} \operatorname{cov}(X, Z) &= \vec{v}_X \cdot \vec{v}_Z = \vec{v}_X \cdot (\vec{v}_Y - \rho \vec{v}_X) = \vec{v}_X \cdot \vec{v}_Y - \rho \vec{v}_X \cdot \vec{v}_X \\ &= \operatorname{cov}(X, Y) - \rho \vec{v}_X \cdot \vec{v}_X = \rho \sigma_X \sigma_Y - \rho \sigma_X^2 = 0\,, \end{aligned}$$

da nach Voraussetzung $\sigma_X = \sigma_Y$ ist. Also sind X und Z unkorreliert. Anderst ausgedrückt heißt dies, daß wir

$$Y = \rho X + Z$$

schreiben können; hier ist der erste Bestandteil vollständig bestimmt durch X, während der zweite unabhängig von X ist.

Da der Begriff der Autokorrelation wohl das 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. Da beide Bilder allerdings auch flächige Strukturen haben, ist sie immer noch ziemlich groß.

Betrachten wir als nächstes eine Folge $X_1, \ldots, X_n$ von Zufallsvariablen mit gemeinsamem Erwartungswert μ und gemeinsamer Standardabweichung σ, 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 Stochastiker handelt es sich hier um spezielle sogenannte MARKOV-Ketten, benannt nach dem russischen Mathematiker ANDREĬ ANDREEVICH MARKOV (1856–1922).)

Falls wir bei einer solchen Folge von Zufallsvariablen die Werte von $X_1, \dots, X_n$ 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

$$X_1, X_2 - X_1, \dots, X_n - X_{n-1}\,.$$

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, die hier zu weit führen würden; die Dekorrelation der Zufallsvariablen aber führt uns geradewegs zu Eigenvektoren symmetrischer Matrizen:

Wir definieren für eine Folge von Zufallsvariablen deren *Kovarianz-* und *Korrelationsmatrizen*

$$\mathrm{Cov}(X_1, \dots, X_n) \in \mathbb{R}^{n\times n} \quad \text{und} \quad \mathrm{Kor}(X_1, \dots, X_n) \in \mathbb{R}^{n\times n}$$

dadurch, daß der Eintrag an der Stelle ij dieser Matrizen jeweils die Kovarianz *bzw.* Korrelation zwischen X_i und X_j sein soll.

Das Ideal, auf das wir hinarbeiten, sind Zufallsvariablen, deren Korrelations- und Kovarianzmatrix Diagonalmatrizen sind, denn dann sind je zwei verschiedene Variablen unkorreliert. Die Korrelationsmatrix ist in diesem Fall natürlich die Einheitsmatrix.

Da die Kovarianzmatrix eine symmetrische Matrix ist, gibt es eine Orthonormalbasis des $\mathbb{R}^n$ aus reellen Eigenvektoren, bezüglich derer die Kovarianzmatrix 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

$$\begin{aligned}
\operatorname{cov}(Y_i, Y_k) = \vec{v}_{Y_i} \cdot \vec{v}_{Y_k} &= \left(\sum_{j=1}^{n} \alpha_{ij} vv_{X_j} \right) \cdot \left(\sum_{\ell=1}^{n} \alpha_{k\ell} \vec{v}_{X_\ell} \right) \\
&= \sum_{j=1}^{n} \sum_{\ell=1}^{n} \alpha_{ij} \ \vec{v}_{X_j} \cdot \vec{v}_{X_\ell} \ \alpha_{k\ell} = \sum_{j=1}^{n} \sum_{\ell=1}^{n} \alpha_{ij} \ \operatorname{cov}(X_j, X_\ell) \ \alpha_{k\ell} \\
&= \sum_{\ell=1}^{n} \left(\sum_{j=1}^{n} \alpha_{ij} \ \operatorname{cov}(X_j, X_\ell) \right) \ \alpha_{k\ell} \, .
\end{aligned}$$

Der Inhalt der großen Klammer ist offensichtlich der Eintrag an der Stelle $i\ell$ der Produktmatrix $A \cdot \operatorname{Cov}(X_1, \ldots, X_n)$, 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

$$\operatorname{Cov}(Y_1, \ldots, Y_n) = A \cdot \operatorname{Cov}(X_1, \ldots, X_n) \cdot {}^tA$$

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

$$\mathrm{Cov}(Y_1, \ldots, Y_n) = A \cdot \mathrm{Cov}(X_1, \ldots, X_n) \cdot A^{-1}\,.$$

Damit ist $\mathrm{Cov}(Y_1, \ldots, Y_n)$ eine Diagonlmatrix, 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 Kovarianzmatrix, 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 als KARHUNEN-LOÈVE-Transformation, da sie in Europa fast gleichzeitig um 1947 *bzw.* 1948 in wahrscheinlichkeitstheoretischen Arbeiten des Finnen H. KARHUNEN und des Franzosen M. LOÈVE erschien; in USA war sie schon 1933 von dem Statistiker H. HOTELLING vorgeschlagen worden.

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.

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 Kompromis 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 sonst irgendwo die KARHUNEN-LOÈVE-Transformation wirklich angewandt. Der Grund liegt an der Struktur der Eigenvektoren der Kovarianzmatrix: Betrachten wir etwa als typisches Beispiel den Fall $\rho = 0{,}95$ und $n = 8$. Die

Kovarianzmatrix für $n = 8$ ist

$$\mathrm{Cov}(X_1, \ldots, X_8) = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 & \rho^7 \\ \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^7 & \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}.$$

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 46 bis 53 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.

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-LOEÈVE-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* im nächsten Semester verwiesen.

Die diskrete Cosinustransformation ist Teil fast aller gängiger Normen zur Bildkomprimierung: Sowohl der JPEG-Standard für Photographien,

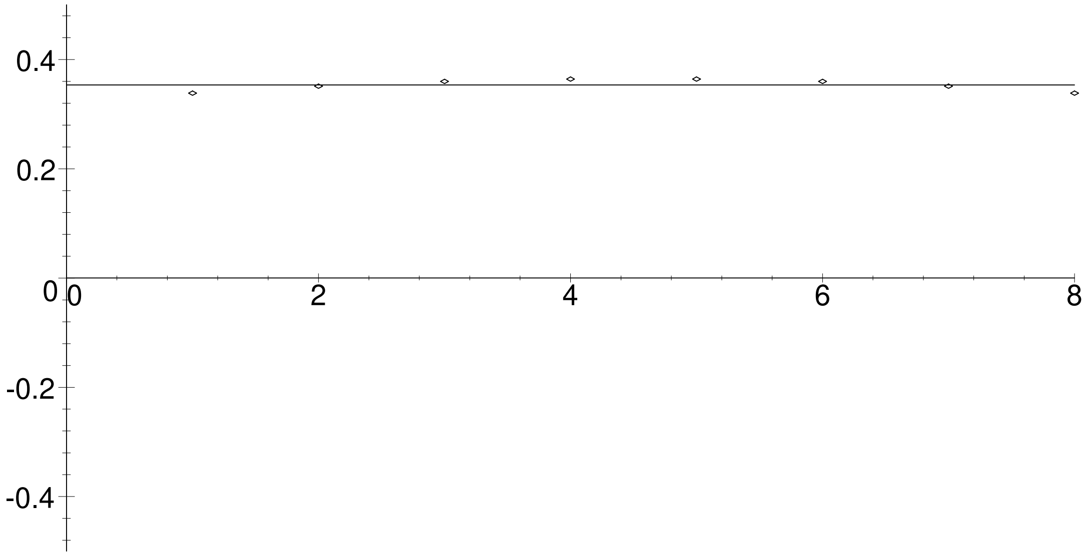


Abb. 46: Der erste Eigenvektor der Kovarianzmatrix

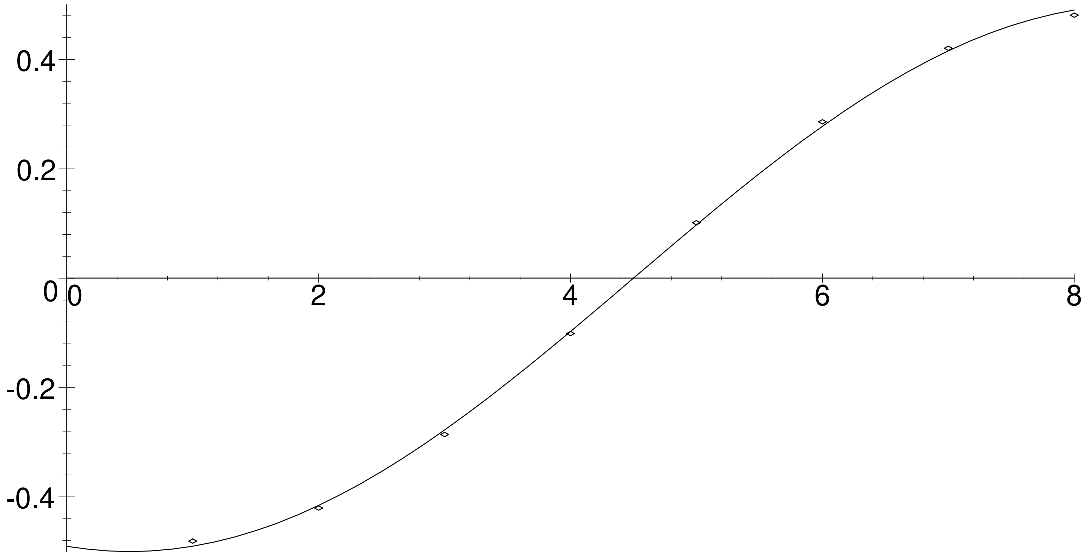


Abb. 47: Der zweite Eigenvektor der Kovarianzmatrix

die Standards MPEG 1 und 2 für digitale (Unterhaltungs-)Videos als

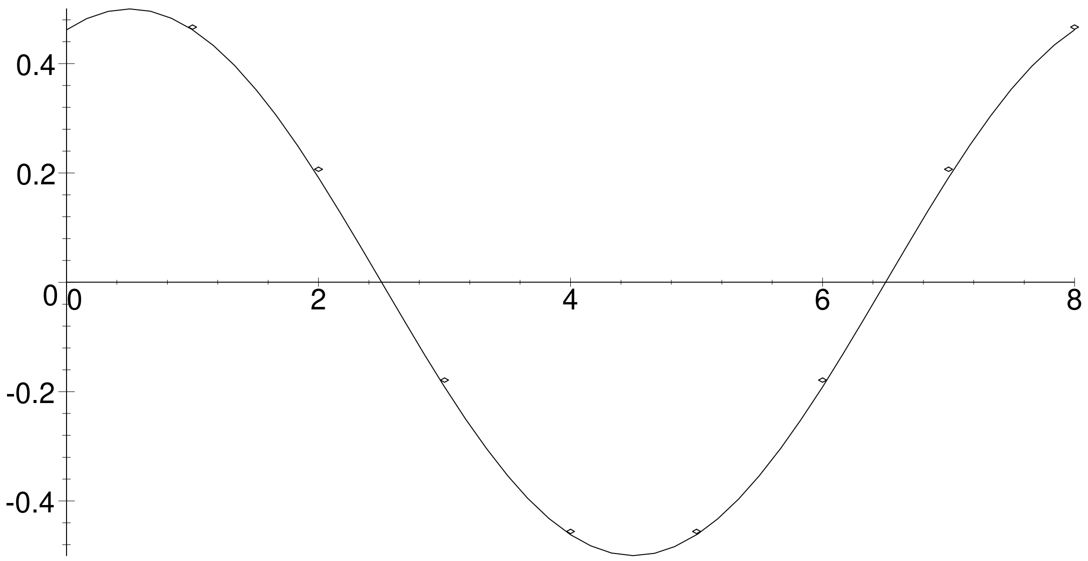


Abb. 48: Der dritte Eigenvektor der Kovarianzmatrix

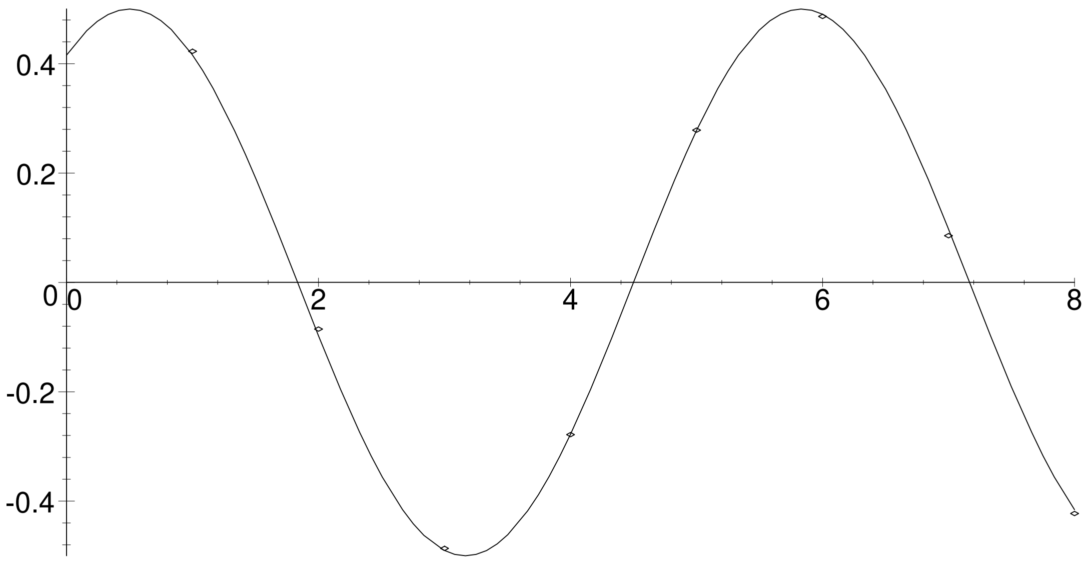


Abb. 49: Der vierte Eigenvektor der Kovarianzmatrix

auch der Standard CCITT H.261 für Videokonferenzen enthalten (ne-

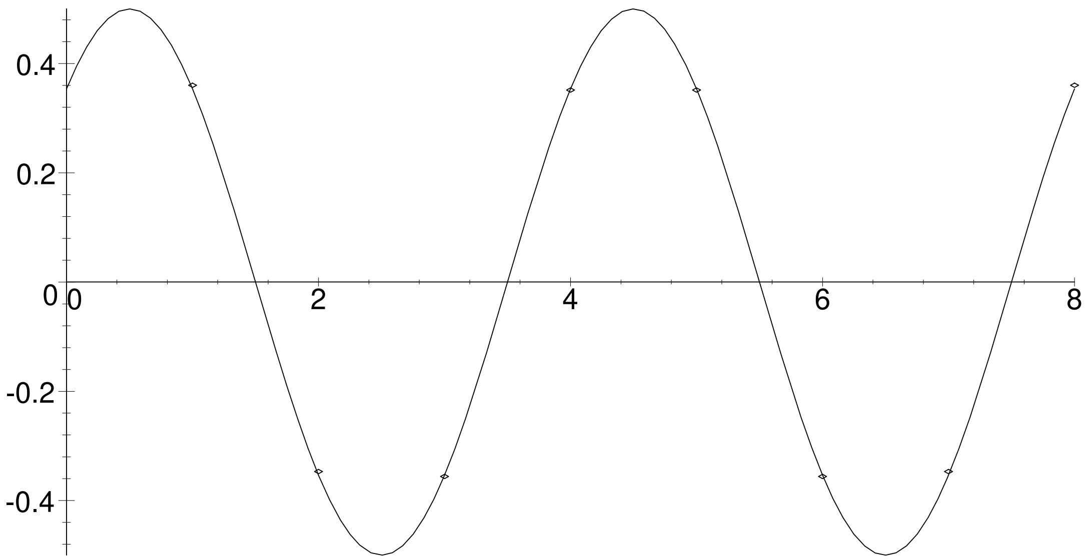


Abb. 50: Der fünfte Eigenvektor der Kovarianzmatrix

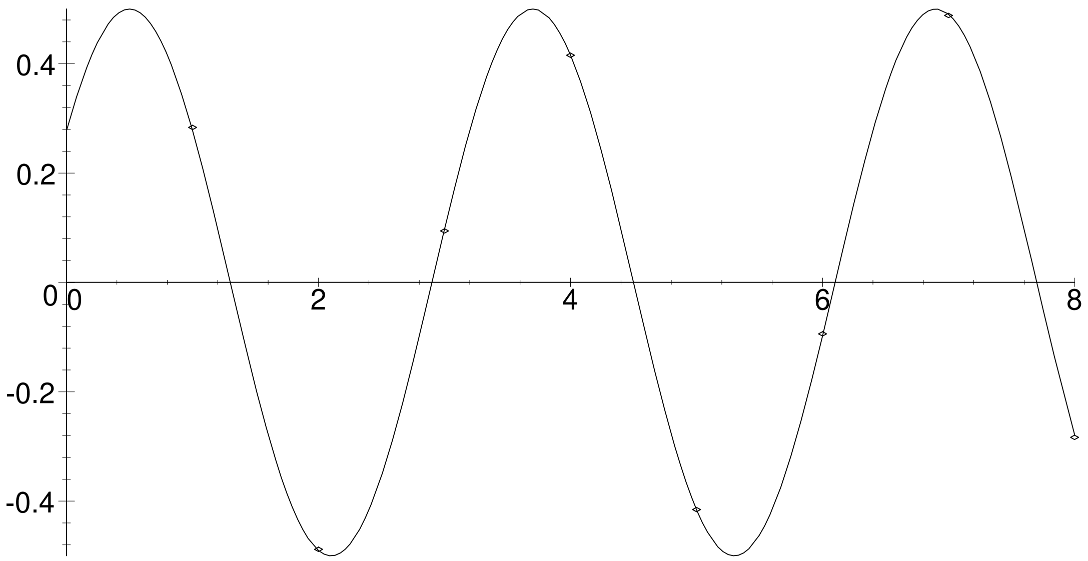


Abb. 51: Der sechste Eigenvektor der Kovarianzmatrix

ben anderen Bestandteilen) jeweils eine diskrete Cosinustransformation.

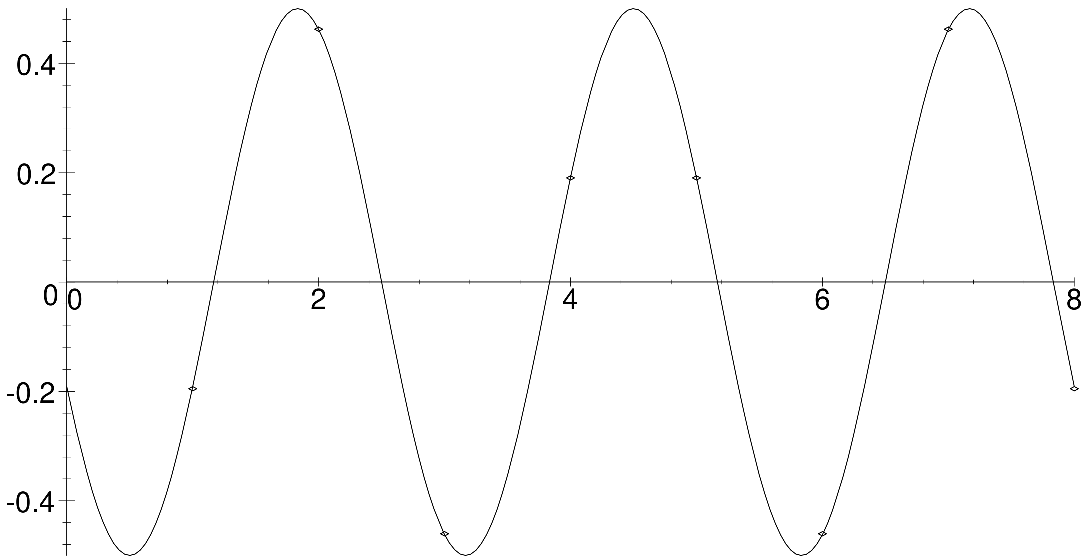


Abb. 52: Der siebte Eigenvektor der Kovarianzmatrix

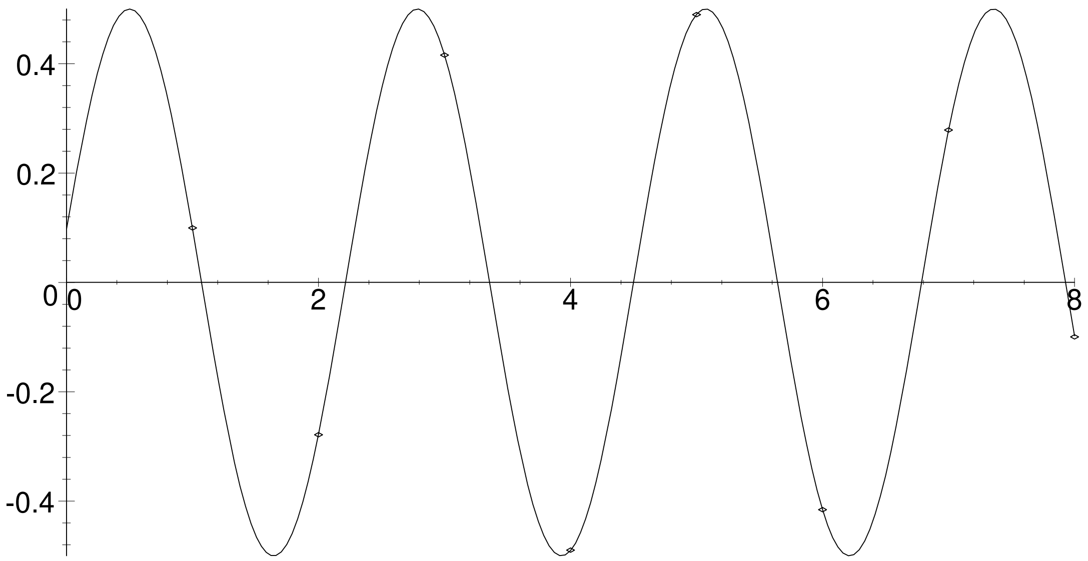


Abb. 53: Der achte Eigenvektor der Kovarianzmatrix

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 Beschreibung 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-Transformationen 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 oder in vielen Fällen sogar gleich gar nicht auftreten. Demenstsprechend 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.

E N D E