

Mathematik I

J. Hellmich

Stuttgart
2023 / 24

Autor: Dr.Jürgen Hellmich

*Ich höre und vergesse,
ich sehe und erinnere mich,
ich tue es und verstehe.*
(Konfuzius)

Lineare Algebra, Analysis

© Dr. Jürgen Hellmich

Alle Rechte vorbehalten, auch die der fotomechanischen Wiedergabe und der Speicherung in elektro-nischen Medien. Die elektronische Speicherung und fotomechanische Wiedergabe ist den Hörern der Lehrveranstaltung *Lineare Algebra, Analysis* an der DHBW Stuttgart im Semester 2023/24 ausschließ-lich zur Begleitung der Vorlesung gestattet.

Stand: 13. 9. 2023

Textsatz: X_\ET_EX

Bilder: Asymptote, Dia, GeoGebra, POV-Ray, PStricks

Inhaltsverzeichnis

1	Grundlagen	1
1.1	Aussagenlogik	1
1.2	Mengen	7
2	Relationen und Funktionen	19
2.1	Allgemeine Eigenschaften von Relationen	19
2.2	Klassifikation von Relationen	21
2.3	Äquivalenzrelationen	21
2.4	Verkettung von Relationen und die inverse Relation	24
2.5	Funktionen	25
3	Zahlentheorie	35
3.1	Teilbarkeitstheorie ganzer Zahlen	35
3.2	Rechnen Modulo p	40
3.3	RSA-Verschlüsselung	45
3.4	Chinesischer Restsatz	47
4	Kombinatorik	51
4.1	Die Urnenmodelle	51
5	Vektorräume	61
5.1	\mathbb{R}^2 und \mathbb{R}^3 als Vektorraum	61
5.2	Die komplexen Zahlen	76
5.3	\mathbb{C}^n als Vektorraum	89
5.4	Der allgemeine Vektorraumbegriff*	90
5.5	Vektorräume mit Norm und Skalarprodukt	92
6	Matrizen	99
6.1	Lineare Gleichungssysteme und das GAUSS-Verfahren	99
6.2	Die Matrix zum LGS	105
6.3	Das Schema zum GAUSS-Verfahren	107
6.4	Lineare Unabhängigkeit	114
6.5	Basis und Dimension	120
6.6	Matrizen als lineare Abbildungen	132
6.7	Direkte Zerlegung eines Vektorraums	135
6.8	Die Dimensionsformel	138
6.9	Die inverse Matrix	140

6.10 Die adjungierte Matrix	144
6.11 Koordinatentransformation	151
6.12 Determinanten	156
7 Eigenwerttheorie	173
7.1 Spektrum und Eigenvektoren	173
7.2 Selbstadjungierte lineare Abbildungen	175
7.3 Funktionalkalkül	180
7.4 Normale Abbildungen	181
8 Hauptachsentransformation	185
8.1 Kegelschnitte	185
8.2 Quadratische Formen	192
8.3 Beispiele für Quadriken	195
9 Rekurrenzgleichungen	201
9.1 Lineare Rekurrenzgleichungen 1. Ordnung	202
9.2 Lineare Rekurrenzgleichungen 2. Ordnung	204
10 Folgen und Reihen	209
10.1 Der Grenzwertbegriff	211
10.2 Reihen	240
11 Funktionen	263
11.1 Stetige Funktionen	263
11.2 Differentialrechnung	283
11.3 Ableitung von Potenzreihen	298
11.4 Ableitung von Umkehrfunktionen	305
11.5 Stammfunktionen	313
11.6 Die Taylor-Entwicklung	318
11.7 Das Newton-Verfahren	331
11.8 Polynome	337
11.9 Kurvendiskussion	363
12 Integralrechnung	377
12.1 Das Flächenproblem	377
12.2 Integrationstechniken	385
12.3 Anwendungen	391
12.4 Schaubilder der elementaren Funktionen	401
13 Lösungen	1-L
13.1 Aussagenlogik	1-L
13.2 Mengen	3-L
13.3 Relationen	5-L
13.4 Zahlentheorie	6-L

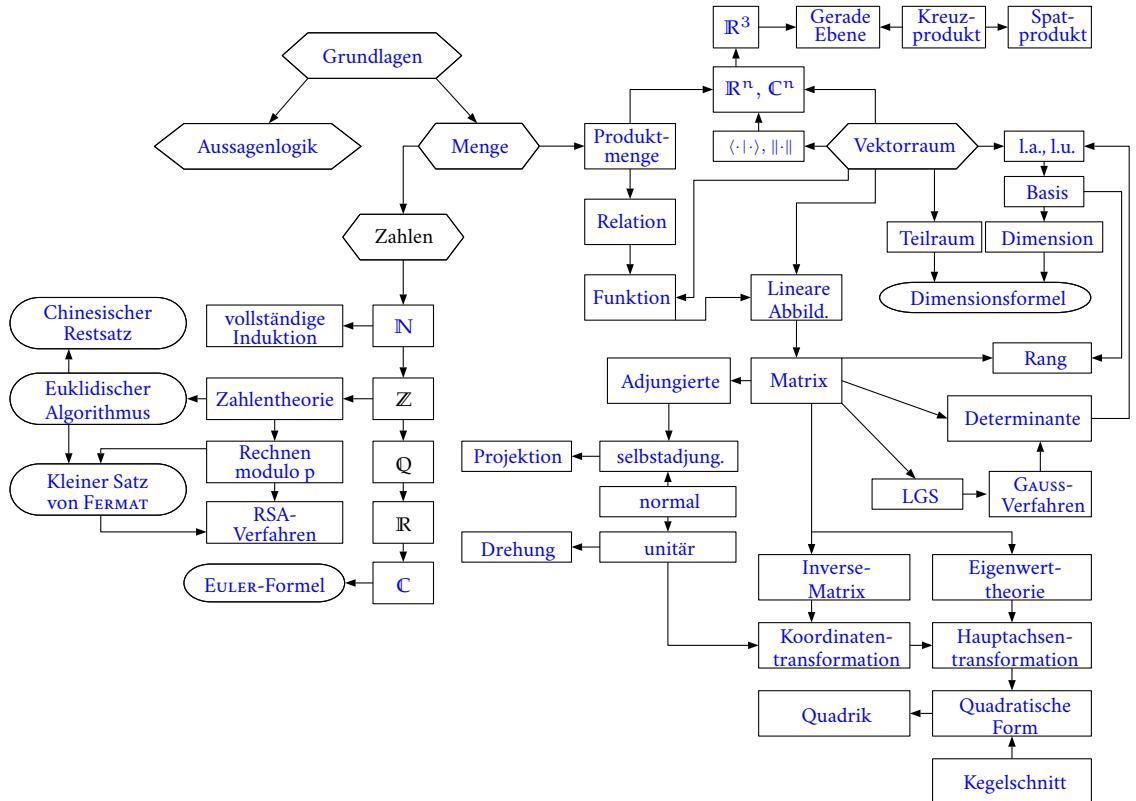
13.5 \mathbb{R}^2 und \mathbb{R}^3 als Vektorraum	8-L
13.6 Vektorräume	10-L
13.7 Matrizen	17-L
13.8 Eigenwerttheorie	33-L
13.9 Kegelschnitte	41-L
13.10 Quadratische Formen	44-L
13.11 Rekurrenzgleichungen	51-L
13.12 Folgen und Reihen	54-L
13.13 Funktionen	63-L
13.14 Kurvendiskussion	77-L

Nomenclature [1-Sy](#)

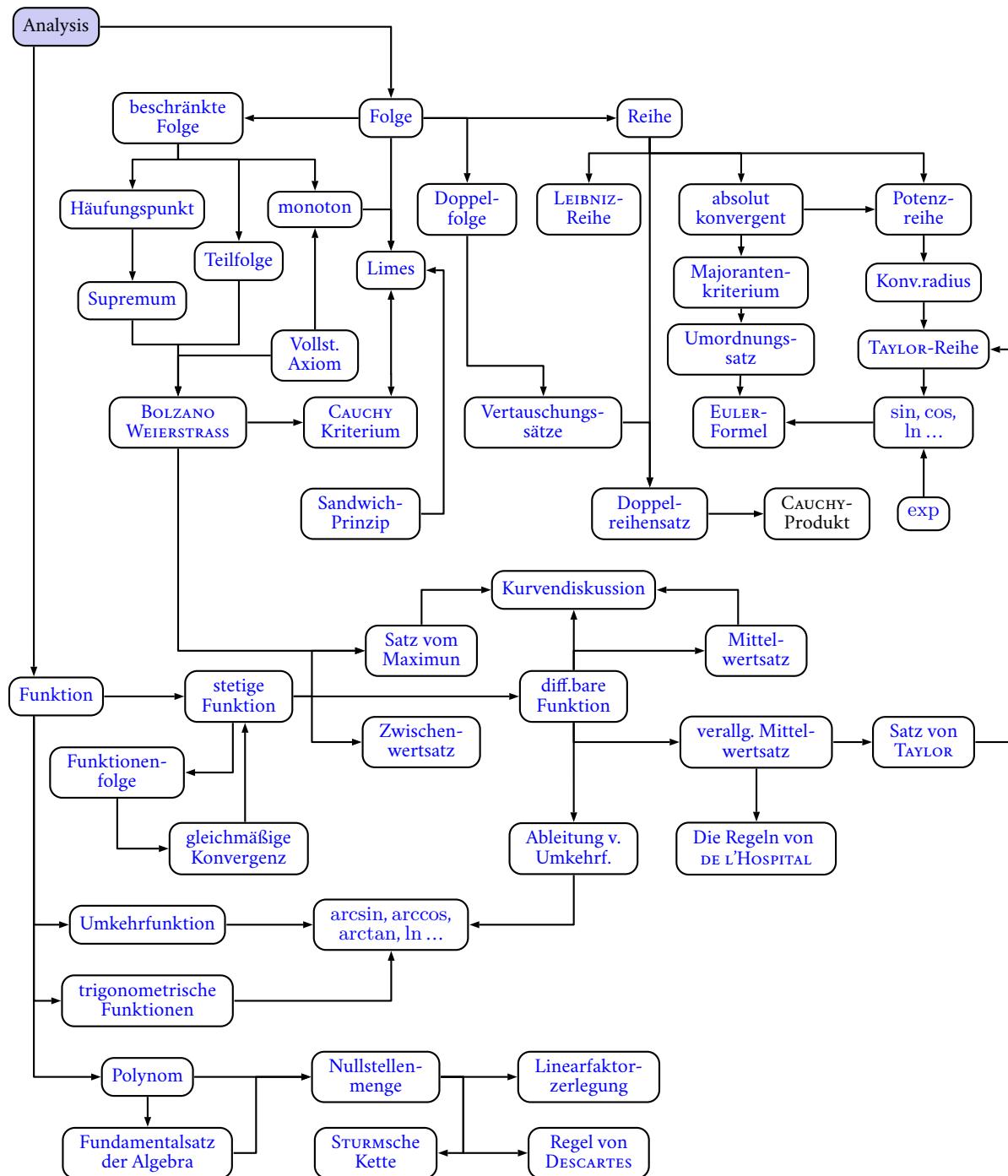
Index [1-In](#)

Themenübersicht

Lineare Algebra



Analysis



Zusammenfassung Ein Vektorraum (V, \mathbb{K}) über \mathbb{K} besteht aus einer Menge V , deren Elemente x, y, z, \dots als *Vektoren* bezeichnet werden und einem *Körper* \mathbb{K} , den sogenannten *Skalaren* t, s, r, \dots Darüber hinaus gibt es eine *Addition* $V \times V \rightarrow V, [x, y] \mapsto x + y \in V$ für zwei Vektoren x und y , sowie eine *Skalarmultiplikation* $\mathbb{K} \times V \rightarrow V, [t, x] \mapsto t \cdot x \in V$ zwischen einem Skalar t und einem Vektor x , die den gewohnten Rechenregeln folgen (5.4.1).

Beispiele:

- i) $(\mathbb{K}^n, \mathbb{K}), \mathbb{K} = \mathbb{R}, \mathbb{C}, \mathbb{F}_2$. Die Vektoren sind, wenn nichts anderes gesagt wird, Spaltenvektoren $x = [x_1, x_2, \dots, x_n]^t$ und $y = [y_1, y_2, \dots, y_n]^t$, (p. 105)

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \pm \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 \pm y_1 \\ x_2 \pm y_2 \\ \vdots \\ x_n \pm y_n \end{bmatrix}, \quad t \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} tx_1 \\ tx_2 \\ \vdots \\ tx_n \end{bmatrix}.$$

- ii) $(\mathcal{P}, \mathbb{C}), (\mathcal{P}_n, \mathbb{C})$, der Vektorraum der komplexen Polynome \mathcal{P} , bzw. der Polynome mit einem Grad $\leq n$. Hier ist die Addition zweier Vektoren p und q und die Skalarmultiplikation von p mit einer Zahl $t \in \mathbb{C}$, die Addition von Funktionen: $(p + q)(x) := p(x) + q(x)$, bzw. die Multiplikation einer Funktion mit einer Zahl: $(tp)(x) := tp(x)$.
- iii) (M_n, \mathbb{K}) die Menge der $n \times n$ -Matrizen über dem Körper \mathbb{K} . Die Addition und Skalarmultiplikation ist komponentenweise definiert, wie bei Vektoren aus \mathbb{K}^n :

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} \pm \begin{bmatrix} b_{11} & \dots & b_{1n} \\ b_{21} & \dots & b_{2n} \\ \vdots & & \vdots \\ b_{n1} & \dots & b_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} \pm b_{11} & \dots & a_{1n} \pm b_{1n} \\ a_{21} \pm b_{21} & \dots & a_{2n} \pm b_{2n} \\ \vdots & & \vdots \\ a_{n1} \pm b_{n1} & \dots & a_{nn} \pm b_{nn} \end{bmatrix},$$

$$t \begin{bmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix} = \begin{bmatrix} ta_{11} & \dots & ta_{1n} \\ ta_{21} & \dots & ta_{2n} \\ \vdots & & \vdots \\ ta_{n1} & \dots & ta_{nn} \end{bmatrix}.$$

Eine *Metrik* auf dem Vektorraum (V, \mathbb{K}) ist eine Abbildung $d: V \times V \rightarrow \mathbb{R}$, mit den Eigenschaften

- i) $d(x, y) \geq 0, d(x, y) = 0 \Leftrightarrow x = y$ (Definitheit)
- ii) $d(x, y) = d(y, x)$ (Symmetrie)
- iii) $d(x, z) \leq d(x, y) + d(y, z)$ (Dreiecksungleichung)

Ein solcher Raum heißt *metrischer Raum*. Über die Metrik kann man in V Abstände messen.

Eine *Norm* auf dem Vektorraum (V, \mathbb{K}) ist eine Abbildung $\| \cdot \| : V \rightarrow \mathbb{R}_0^+$, mit den Eigenschaften (5.5.1)

- i) $\|x\| \geq 0, \|x\| = 0 \Leftrightarrow x = \mathbf{0}$ (Definitheit)

- ii) $\|tx\| = |t|\|x\|$
- iii) $\|x + y\| \leq \|x\| + \|y\|$ (Dreiecksungleichung)

Ein Vektorraum V mit einer Norm heißt *normierter Raum*, ist er endlichdimensional, auch *Banachraum*. Durch die Norm kann man auf V Längen messen. Durch jede Norm auf V wird über $d(x, y) := \|x - y\|$ eine kanonisch zugeordnete Metrik erzeugt.

Eine *Skalarprodukt* auf einem Vektorraum (V, \mathbb{C}) ist eine Abbildung $\langle \cdot | \cdot \rangle : V \times V \rightarrow \mathbb{C}$, mit den Eigenschaften (5.5.2)

- i) $\langle x|x \rangle \geq 0, \langle x|x \rangle = 0 \Leftrightarrow x = 0$ (Definitheit)
- ii) $\langle x|y \rangle = \overline{\langle y|x \rangle}$ (Antisymmetrie)
- iii) $\langle sx + ty|z \rangle = s\langle x|z \rangle + t\langle y|z \rangle$ (Sesquilinearität)

Ein solcher Raum heißt *Hilbertraum*, wenn V endlichdimensional ist. Ein Skalarprodukt auf (V, \mathbb{R}) hat, bis auf die Eigenschaft ii), die durch $\langle x|y \rangle = \langle y|x \rangle$ zu ersetzen ist, dieselben Eigenschaften. Ein Skalarprodukt erlaubt es auf V über

$$\langle x|y \rangle = \|x\|\|y\| \cos(\alpha)$$

Winkel zu messen. Dabei ist $\|x\| := \sqrt{\langle x|x \rangle}$ die vom Skalarprodukt kanonisch erzeugte Norm, deren Dreiecksungleichung aus der CAUCHY-SCHWARZSchen Ungleichung folgt (5.5.3):

$$|\langle x|y \rangle| \leq \|x\|\|y\|.$$

Zwei Vektoren x und y heißen *orthogonal*, $x \perp y$, falls $\langle x|y \rangle = 0$ gilt.

Beispiele:

- i) Auf $(\mathbb{C}^n, \mathbb{C})$ bzw. $(\mathbb{R}^n, \mathbb{R})$ wird durch

$$\langle x|y \rangle = \sum_{k=1}^n x_k \overline{y_k}, \quad \|x\| = \sqrt{\sum_{k=1}^n |x_k|^2}, \quad \text{bzw. } \langle x|y \rangle = \sum_{k=1}^n x_k y_k, \quad \|x\| = \sqrt{\sum_{k=1}^n x_k^2}$$

das übliche Skalarprodukt mit der kanonisch zugeordneten Norm definiert.

- ii) Auf $(\mathbb{C}^n, \mathbb{C})$ wird durch

$$\|x\|_p = \sqrt[p]{\sum_{k=1}^n |x_k|^p}, \quad p \geq 1, \quad \|x\|_\infty = \max \{ |x_k| \mid k = 1, \dots, n \},$$

die Familie der L^p -Normen definiert, die außer für $p = 2$, nicht von einem Skalarprodukt stammen. (5.5.8, 5.5.14)

Ein *Teilraum* T eines Vektorraums (V, \mathbb{K}) ist eine Teilmenge von V , so daß (T, \mathbb{K}) ein Vektorraum ist (5.4.2):

$$\forall x, y \in T \forall s, t \in \mathbb{K} \quad sx + ty \in T.$$

Beispiele:

- i) V und $\{0\}$ sind (die trivialen) Teilräume eines Vektorraums (V, \mathbb{K}) .
- ii) Für zwei Teilräume S, T von (V, \mathbb{K}) ist der Schnitt $S \cap T$ wieder ein Teilraum (während die Vereinigung $S \cup T$ normalerweise keiner ist).
- iii) Für eine Menge $\mathcal{B} \subseteq V$ ist die *lineare Hülle* von \mathcal{B} , (6.5.1)

$$\text{lh } \mathcal{B} = \{ t_1 \mathbf{b}_1 + t_2 \mathbf{b}_2 + \cdots + t_n \mathbf{b}_n \mid t_1, \dots, t_n \in \mathbb{K}, \mathbf{b}_1, \dots, \mathbf{b}_n \in \mathcal{B}, n \in \mathbb{N} \},$$

ein Teilraum von V . Für zwei Teilräume S, T von (V, \mathbb{K}) bezeichnet

$$S \vee T = \text{lh } S \cup T$$

den *Aufspann* von S und T (manchmal bezeichnet man diesen Teilraum auch durch $S + T$ (6.7.1)). Sind S und T zwei Ursprungsgeraden, mit den Richtungsvektoren u und v , dann ist $S \vee T$ die Ursprungsebene, die von u und v erzeugt wird.

- iv) Für eine lineare Abbildung $A : V \rightarrow W$ sind $\ker A$ und $\text{im } A$ Teilräume von V bzw. von W . Eine Teilmenge \mathcal{B} eines Vektorraums V heißt *linear unabhängig*, wenn jede Nullkombination aus Vektoren von \mathcal{B} trivial ist. D.h., für jede Wahl von Vektoren $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n$ aus \mathcal{B} folgt aus $t_1 \mathbf{b}_1 + t_2 \mathbf{b}_2 + \cdots + t_n \mathbf{b}_n = \mathbf{0}$ immer $t_1 = \cdots = t_n = 0$. Andernfalls heißt \mathcal{B} *linear abhängig*. (6.4.1)

Eine Teilmenge \mathcal{B} eines Vektorraums V heißt *Basis* von V wenn sie linear unabhängig und erzeugend ist. (6.5.1)

Beispiele:

- i) Die kanonischen Basisvektoren für \mathbb{K}^n : (6.5.9)

$$\mathcal{E} = \left\{ \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ n \end{bmatrix} \right\} = \{ \mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n \}.$$

- ii) Die Menge $\mathcal{M} = \{ p_k \mid k \in \mathbb{N}_0 \}$ der Monome $p_k(x) = x^k$, bilden eine Basis für den Vektorraum \mathcal{P} der Polynome. (6.4.4)
- iii) In einem Vektorraum der Dimension n ist jede linear unabhängige Menge aus n Vektoren eine Basis für V .

Jeder Vektorraum (V, \mathbb{K}) hat eine Basis \mathcal{B} . Enthält diese nur endlich viele Elemente, so heißt V *endlichdimensional*, andernfalls *unendlichdimensional*. Die Anzahl der Elemente von \mathcal{B} heißt *Dimension* von V : $\dim V$. (6.5.6)

Jeder Vektor $x \in V$ hat genau eine Entwicklung

$$x = \sum_{i=1}^n x_i \mathbf{b}_i$$

bzgl. einer Basis \mathcal{B} . Der Spaltenvektor $\mathbf{x}_{\mathcal{B}} = [x_1, \dots, x_n]_{\mathcal{B}}^t$ der Koeffizienten von \mathbf{x} heißt *Basisdarstellung* von \mathbf{x} in der Basis \mathcal{B} . (6.5.2) Die Bestimmung der Koeffizienten x_i erfordert normalerweise das Lösen eines inhomogenen linearen Gleichungssystems, meist mittels GAUSS-Verfahren (6.1). Handelt es sich bei $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ um eine *Orthonormalbasis* (ONB), d. h., gilt $\langle \mathbf{b}_i | \mathbf{b}_j \rangle = \delta_{ij}$, so gewinnt man die Koeffizienten einfach durch $x_i = \langle \mathbf{x} | \mathbf{b}_i \rangle$ (6.5.15)

$$\mathbf{x} = \sum_{k=1}^n \langle \mathbf{x} | \mathbf{b}_k \rangle \mathbf{b}_k, \quad \mathbf{x}_{\mathcal{B}} = \mathcal{B}^* \mathbf{x} = [\langle \mathbf{x} | \mathbf{b}_1 \rangle, \langle \mathbf{x} | \mathbf{b}_2 \rangle, \dots, \langle \mathbf{x} | \mathbf{b}_n \rangle]_{\mathcal{B}}^t,$$

mit der Transformationsmatrix $\mathcal{B} := [\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n]$. (p. 154)

Eine *lineare Abbildung* $A : V \rightarrow W$ von dem Vektorraum (V, \mathbb{K}) in den Vektorraum (W, \mathbb{K}) ist eine Funktion mit Definitionsbereich V , die mit der Vektorraumstruktur verträglich ist, d. h., mit folgender Eigenschaft: Für alle $\mathbf{x}, \mathbf{y} \in V, s, t \in \mathbb{K}$ gilt (6.6)

$$A(s\mathbf{x} + t\mathbf{y}) = sA(\mathbf{x}) + tA(\mathbf{y}).$$

Meist schreibt man $A\mathbf{x}$ statt $A(\mathbf{x})$. Ist $\mathcal{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ eine Basis von V , so ist A durch die Bilder der Basisvektoren $A\mathbf{b}_k$, $k = 1, \dots, n$, bereits vollständig bestimmt:

$$A(x_1\mathbf{b}_1 + x_2\mathbf{b}_2 + \dots + x_n\mathbf{b}_n) = x_1A\mathbf{b}_1 + x_2A\mathbf{b}_2 + \dots + x_nA\mathbf{b}_n.$$

Wichtige Kenngrößen von A sind der *Kern* $\ker A$ und das *Bild* $\text{im } A$ von A :

$$\ker A = \{ \mathbf{x} \in V \mid A\mathbf{x} = \mathbf{0} \}, \quad \text{im } A = \{ \mathbf{y} \in W \mid \exists_{\mathbf{x} \in V} A\mathbf{x} = \mathbf{y} \}.$$

Für einen endlichdimensionalen Vektorraum V sind sie über die *Dimensionsformel* verknüpft (6.8.4):

$$\dim V = \dim \ker A + \dim \text{im } A.$$

Für $V = \mathbb{C}^n$ und $W = \mathbb{C}^m$ wird eine lineare Abbildung $A : V \rightarrow W$ meist durch eine $m \times n$ -Matrix $A = [a_{ij}]_{i=1, \dots, m}^{j=1, \dots, n}$ beschrieben:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n].$$

Dabei sind die *Spaltenvektoren* $\mathbf{a}_j = [a_{1j}, a_{2j}, \dots, a_{mj}]^t$ die Bilder Ae_j der kanonischen Basisvektoren e_j ($j = 1, \dots, n$). Für einen Vektor $\mathbf{x} = [x_1, x_2, \dots, x_n]^t$ ist

$$A\mathbf{x} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{j=1}^n a_{1j}x_j \\ \sum_{j=1}^n a_{2j}x_j \\ \vdots \\ \sum_{j=1}^n a_{mj}x_j \end{bmatrix} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \sum_{j=1}^n x_j\mathbf{a}_j.$$

Die Hintereinanderausführung $B \circ A$ zweier linearer Abbildungen $A : V \rightarrow W$ und $B : W \rightarrow X$, normalerweise einfach BA geschrieben, ist eine lineare Abbildung von V nach X .

Für die $k \times n$ -Matrix $A = [a_1, \dots, a_n] : \mathbb{K}^n \rightarrow \mathbb{K}^k$ und die $m \times k$ -Matrix $B = [b_1, \dots, b_k] : \mathbb{K}^k \rightarrow \mathbb{K}^m$ ist $BA : \mathbb{K}^n \rightarrow \mathbb{K}^m$ die $m \times n$ -Matrix (p. 134)

$$BA = [B\mathbf{a}_1, \dots, B\mathbf{a}_n].$$

Beispiele:

- i) Die identische Abbildung $\text{id} : \mathbb{K}^n \rightarrow \mathbb{K}^n, x \mapsto x$ ist durch die sogenannte *Einheitsmatrix* $\mathbb{1}$ gegeben und die *Nullabbildung* $0 : \mathbb{K}^n \rightarrow \mathbb{K}^n, x \mapsto 0$ durch die *Nullmatrix* 0:

$$\mathbb{1} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}, \quad 0 = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}.$$

- ii) Die ebene Drehung D_α um den Winkel α ist

$$D_\alpha = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}.$$

- iii) Die Projektion P_n auf die Richtung des normierten Vektors $n = [n_1, n_2, \dots, n_k]^t$ wird durch $P_n x = \langle x | n \rangle n$ beschrieben und hat die Matrix

$$P_n = \begin{bmatrix} n_1 n_1 & n_1 n_2 & \dots & n_1 n_k \\ n_2 n_1 & n_2 n_2 & \dots & n_2 n_k \\ \vdots & \vdots & \ddots & \vdots \\ n_k n_1 & n_k n_2 & \dots & n_k n_k \end{bmatrix}.$$

Die Projektion auf den zu n orthogonalen Teilraum ist $\mathbb{1} - P_n$.

- iv) Ein *lineares Gleichungssystem* (LGS)

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= y_2 \\ \vdots &= \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= y_m \end{aligned}$$

wird durch die Koeffizientenmatrix $A = [a_{ij}]_{\substack{i=1, \dots, m \\ j=1, \dots, n}}$, durch den Lösungsvektor $x = [x_1, x_2, \dots, x_n]^t$ und durch die *Inhomogenität* $y = [y_1, y_2, \dots, y_m]^t$ zu der handlichen

Vektorgleichung $Ax = y$.

Das *homogene* LGS $Ax = \mathbf{0}$ hat $\ker A$ als Lösungsmenge, und das inhomogene

$$L = \{x_1 + x_0 \mid x_0 \in \ker A\}.$$

Dabei ist x_1 eine Lösung der inhomogenen Gleichung: $Ax_1 = y$. Daher ist auch $x_1 + x_0$ eine Lösung:

$$A(x_1 + x_0) = Ax_1 + Ax_0 = y + \mathbf{0} = y.$$

Für eine lineare Abbildung $A = [a_1, \dots, a_n] : V \rightarrow V$ auf dem n -dimensionalen Vektorraum (V, \mathbb{K}) sind die folgenden Aussagen äquivalent:

- i) A ist invertierbar.
- ii) A ist injektiv.
- iii) A ist surjektiv.
- iv) $\ker A = \{\mathbf{0}\}$.
- v) Die Menge $\{a_1, \dots, a_n\}$ der Spaltenvektoren von A ist linear unabhängig.
- vi) $\det A := \det(a_1, \dots, a_n) \neq 0$.

Die *inverse Matrix* A^{-1} von A lässt sich gegebenenfalls durch das erweiterte GAUSS-Verfahren als Lösung B der Matrixgleichung $AB = \mathbb{1}$ berechnen (p. 141):

$$\left[\begin{array}{cccc|ccccc} a_{11} & a_{12} & \cdots & a_{1n} & 1 & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & a_{2n} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & 0 & 0 & \cdots & 1 \end{array} \right] \xrightarrow[\text{Verf.}]{\text{GAUSS}} \left[\begin{array}{cccc|ccccc} 1 & 0 & \cdots & 0 & b_{11} & b_{12} & \cdots & b_{1n} \\ 0 & 1 & \cdots & 0 & b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & b_{n1} & b_{n2} & \cdots & b_{nn} \end{array} \right]$$

Die *Determinante* \det auf dem Vektorraum $(\mathbb{K}^n, \mathbb{K})$ ist eine *normierte, alternierende Multilinearform*, d. h., eine Abbildung von $\mathbb{K}^n \times \mathbb{K}^n \times \dots \times \mathbb{K}^n \rightarrow \mathbb{K}$, so daß $[a_1, \dots, a_i, \dots, a_n] \mapsto \det(a_1, \dots, a_i, \dots, a_n)$ in jeder Komponente a_i linear ist, die Vertauschung zweier Komponenten a_i und a_j für $i \neq j$ zu einem Vorzeichenwechsel führt und $\det(e_1, \dots, e_n) = 1$ gilt. Durch $\det(A) := \det(a_1, \dots, a_n)$, für $A = [a_1, \dots, a_n]$, führen wir die Determinante für quadratische Matrizen ein (6.12.1). Durch die Koordinatendarstellung $a_i = \sum_{k=1}^n a_{ki} e_k$ lässt sich die Summendarstellung der Determinante gewinnen ((6.42))

$$\det(A) = \det(a_1, a_2, \dots, a_n) = \sum_{\pi \in S_n} \text{sgn}(\pi) \cdot a_{\pi(1)1} a_{\pi(2)2} \dots a_{\pi(n)n}.$$

Sie hat für konkrete Anwendungen weniger Bedeutung, aber sie stellt über das *Pfadbild* (p. 159) alternative Berechnungsmöglichkeiten bereit. Die alternierende Eigenschaft der Determinante sorgt dafür, daß sie bei elementaren Spaltenenumformungen $a_i \mapsto a_i + t a_j$ ihren Wert nicht ändert. Wegen $\det(A) = \det(A^t)$ (6.12.7), gilt das auch für Zeilenumformungen. Dadurch lässt sich eine Determinante werterhaltend in eine Determinante mit einer oberen (oder unteren)

Dreiecksmatrix umformen, so daß ihr Wert (gemäß dem Pfadbild) einfach als Produkt der Diagonalelemente abgelesen werden kann. Das ist das GAUSS-Verfahren zur Berechnung von Determinanten (6.12.8), das bei großen Matrizen angewendet wird. Ein anderes Verfahren liefert der LAPLACESche Entwicklungssatz, den man am besten an einem Beispiel versteht (6.12.10):

Wir entwickeln nach der dritten Zeile und dann nach der dritten Spalte:

$$\begin{aligned} \det \begin{bmatrix} 3^+ & 2^- & 1^+ & 4^- \\ 3^- & 3^+ & 2^- & 0^+ \\ 3^+ & 2^- & 0^+ & 0^- \\ 2^- & 3^+ & 7^- & 4^+ \end{bmatrix} &= +3 \cdot \det \begin{bmatrix} 3 & 2 & 1 & 4 \\ 3 & 3 & 2 & 0 \\ 3 & 2 & 0 & 0 \\ 2 & 3 & 7 & 4 \end{bmatrix} - 2 \cdot \det \begin{bmatrix} 3 & 2 & 1 & 4 \\ 3 & 3 & 2 & 0 \\ 3 & 2 & 0 & 0 \\ 2 & 3 & 7 & 4 \end{bmatrix} \\ &= 3 \cdot \det \begin{bmatrix} 2^+ & 1^- & 4^+ \\ 3 & 2 & 0^- \\ 3 & 7 & 4^+ \end{bmatrix} - 2 \cdot \det \begin{bmatrix} 3 & 1 & 4 \\ 3 & 2 & 0 \\ 2 & 7 & 4 \end{bmatrix} \\ &= 3 \cdot \left(4 \cdot \det \begin{bmatrix} 3 & 2 \\ 3 & 7 \end{bmatrix} + 4 \cdot \det \begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix} \right) - 2 \cdot \left(4 \cdot \det \begin{bmatrix} 3 & 2 \\ 2 & 7 \end{bmatrix} + 4 \cdot \det \begin{bmatrix} 3 & 1 \\ 3 & 2 \end{bmatrix} \right) \\ &= 3 \cdot (4 \cdot (21 - 6) + 4 \cdot (4 - 3)) - 2 \cdot (4 \cdot (21 - 4) + 4 \cdot (6 - 3)) = 32. \end{aligned}$$

Das Vorzeichen $(-1)^{i+j}$ der Unterdeterminante M_{ij} , (die aus $\det(A)$ durch Streichen der i-ten Zeile und der j-ten Spalte entsteht) findet man, in dem man in der Matrix A die linke obere Position in Gedanken mit einem + markiert und dann die restlichen Felder des Quadrats schachbrettartig abwechselnd mit – und + versieht:

$$\begin{bmatrix} + & - & + & - \\ - & + & - & + \\ + & - & + & - \\ - & + & - & + \end{bmatrix}$$

An der Position (i, j) dieser *Vorzeichenmatrix* steht das Vorzeichen von M_{ij} .

Aus der Eindeutigkeit alternierender Multilinearformen, bis auf eine multiplikative Konstante (6.12.2), erhält man sehr elegant den *Determinanten-Produktsatz* (6.12.4):

$$\det(AB) = \det(A) \det(B).$$

Die zu einer $m \times n$ -Matrix $A = [a_1, a_2, \dots, a_n]$ adjungierte $n \times m$ -Matrix A^* ist durch

$$A^* := \overline{A}^t = \begin{bmatrix} a_1^* \\ a_2^* \\ \vdots \\ a_n^* \end{bmatrix}$$

definiert (6.10.1). D.h., sie entsteht aus A durch Transponierung und Übergang zu konjugiert komplexen Einträgen $\overline{a_{ji}}$. Für reelle Matrizen gilt natürlich $A^* = A^t$.

Die zentrale Eigenschaft von A^* zeigt sich im Skalarprodukt (6.10.2):

$$\langle x | A y \rangle = \langle A^* x | y \rangle$$

gilt für alle $\mathbf{y} \in \mathbb{K}^n$ und alle $\mathbf{x} \in \mathbb{K}^m$, aus der man leicht $(AB)^* = B^*A^*$ gewinnen kann.

Wichtige Klassen quadratischer Matrizen sind die *selbstadjungierten Matrizen*, also Matrizen A , die mit A^* übereinstimmen (7.2.1)

$$A = A^*,$$

die *unitären* (6.10.4)

$$A^* = A^{-1}$$

und die *normalen Matrizen*, die durch die Bedingung

$$A^*A = AA^*$$

charakterisiert werden (7.4.1). Selbstadjungierte und unitäre Matrizen sind insbesondere normal.

Eigenvektoren $\mathbf{0} \neq \mathbf{x} \in V$ einer linearen Abbildung $A : V \rightarrow V$ sind Vektoren, die durch A lediglich gestreckt werden, die also der sogenannten *Eigenwertgleichung*

$$Ax = \lambda x$$

genügen (7.1). Der Streckungsfaktor λ heißt *Eigenwert* von A zum Eigenvektor x . Die Menge $\text{sp}(A)$ aller Eigenwerte heißt *Spektrum* von A . Eigenvektoren zu verschiedenen Eigenwerten sind linear unabhängig (7.1.2).

Um für einen n -dimensionalen Vektorraum V zu einem Lösungsverfahren der Eigenwertgleichung zu kommen, in der ja sowohl x , als auch λ unbekannt sind, wird die Bedingung $Ax = \lambda x$ so lange äquivalent umgeformt, bis man eine Bedingung an die Eigenwerte allein gefunden hat (p. 173):

$$\exists_{\mathbf{0} \neq \mathbf{x} \in V} Ax = \lambda x \Leftrightarrow \exists_{\mathbf{0} \neq \mathbf{x} \in V} (A - \lambda \mathbb{1})\mathbf{x} = \mathbf{0} \Leftrightarrow \ker A - \lambda \mathbb{1} \neq \{\mathbf{0}\} \Leftrightarrow \det(A - \lambda \mathbb{1}) = 0.$$

Bei der Funktion $\lambda \mapsto \det(A - \lambda \mathbb{1}) =: \chi(\lambda)$ handelt es sich um das sogenannte *charakteristische Polynom* von A (ein Polynom vom Grade n). Zumindest für komplexe Vektorräume hat es immer n Lösungen (von denen einige durchaus gleich sein können). Hat man die Eigenwerte λ einer Abbildung A gefunden, dann geschieht das Berechnen der zugehörigen Eigenvektoren x normalerweise durch Lösen des linearen Gleichungssystems $(A - \lambda \mathbb{1})x = \mathbf{0}$ mit Hilfe des GAUSS-Verfahrens.

Für jede normale Matrix, insbesondere also auch für jede unitäre und jede selbstadjungierte Matrix, gibt es immer eine ONB aus Eigenvektoren (7.2.4, 7.4.2). Das Spektrum einer normalen Matrix kann eine beliebige Menge aus maximal n verschiedenen komplexen Zahlen sein. Für unitäre Matrizen besteht das Spektrum aus komplexen Zahlen vom Betrag 1 (daher der Name *unitär*) (7.4.4), während das Spektrum einer selbstadjungierten Matrix immer aus reellen Zahlen besteht, selbst wenn A komplexe Einträge hat (7.2.2).

1 Grundlagen

1.1 Aussagenlogik

In der Aussagenlogik befassen wir uns mit Aussagen, für die es nur die beiden Wahrheitswerte *wahr* (w) und *falsch* (f) gibt. Aussagen können dabei durchaus umgangssprachlich ausgedrückt werden, müssen sich aber auf Objekte beziehen, die sich einer objektiven Überprüfung unterziehen lassen. Die Aussage *Mir geht es gut* ist zur Zeit der Niederschrift möglicherweise wahr, aber das könnte sich von Tag zu Tag ändern. Sie läßt sich nicht verlässlich in den Kontext der Mathematik einbetten. Die Aussage *Im Jahre 2013 haben 70% aller Menschen braune Augen* ist ebenfalls möglicherweise wahr, aber zumindest können wir davon ausgehen, daß sie einen eindeutigen Wahrheitswert hat — auch wenn wir den voraussichtlich nicht erfahren werden, da wir schwerlich alle Menschen auf ihre Augenfarbe hin untersuchen können. Dagegen sind Aussagen wie

- A := 2 ist die einzige gerade Primzahl
- B := Alle ungeraden Zahlen haben die Form $2n + 1$
- G := Jede gerade Zahl größer als 2 ist die Summe zweier Primzahlen

präzise in der Sprache der Mathematik formuliert. Aussage A ist offensichtlich wahr, und durch kurzes Nachdenken erkennt man auch die Wahrheit von B. Aussage G dagegen, obwohl einfach zu verstehen, erschließt sich nicht sofort als wahr oder falsch. Testet man sie an einigen Zahlen, ($20 = 13 + 7$, $200 = 163 + 37 = 197 + 3$, $55698022500 = 22367000459 + 33331022041$, ...) dann wird man vermutlich nur Beispiele finden, die der Wahrheit von G nicht widersprechen. Um den Wahrheitswert von G als *falsch* zu identifizieren, wäre ja nur eine gerade Zahl nötig, die sich nicht durch zwei Primzahlen darstellen läßt. Eine solche Zahl ist noch nicht gefunden worden. Bei G handelt es sich um die sogenannte *GOLDBACHSche Vermutung* (1742), die bisher nicht entschieden ist.

Unsere Aufgabe besteht im Moment nicht darin, solche Aussagen zu beweisen, sondern darin Werkzeuge bereitzustellen, die es gestatten, Verknüpfungen zwischen Aussagen zu bilden und zu untersuchen. Wir gehen dabei von elementaren Aussagen A, B, ... aus, die, was die Regeln der Aussagenlogik angeht, als atomar, also ohne weitergehende innere Struktur angesehen werden. Dabei stellen wir jede Aussage durch ihre Wahrheitstabelle dar. Für eine Aussage A hat sie die simple Form

A	(1.1)
w	
f	

und gibt einfach die möglichen Wahrheitswerte von A wieder. Für Verknüpfungen von Aussagen erweitern wir sie durch die Wahrheitswerte der beteiligten Aussagen. Eine einstellige Verknüpfung ist der Übergang von A zur *Negation* $\neg A$ (wir verwenden auch die Schreibweise \bar{A} , die manchmal eine kompaktere Darstellung erlaubt). Dabei ist $\neg A$ genau dann wahr, wenn A falsch ist und genau dann falsch, wenn A wahr ist. In der Wahrheitstabelle lässt sich das übersichtlich darstellen:

A	$\neg A$
w	f
f	w

(1.2)

Für zwei Aussagen A und B stehen uns die elementaren Verknüpfungen *und*, *oder* und *folgt* zur Verfügung, die durch die sog. *Junkturen* \wedge , \vee und \Rightarrow dargestellt werden: $A \wedge B$, $A \vee B$, $A \Rightarrow B$. Die Wahrheitstabelle zeigt die Definitionen:

A	B	$A \wedge B$	$A \vee B$	$A \Rightarrow B$
w	w	w	w	w
w	f	f	w	f
f	w	f	w	w
f	f	f	f	w

(1.3)

Die *Konjunktion* A und B, oder kurz $A \wedge B$, ist genau dann wahr, wenn A und wenn B wahr ist. Die *Disjunktion* A oder B, oder kurz $A \vee B$, ist genau dann wahr, wenn A wahr ist, oder wenn B wahr ist, oder wenn beide wahr sind. Es handelt sich bei diesem *oder* also nicht um das umgangssprachliche, ausschließende oder, das in der Mathematik als *entweder oder* bezeichnet würde. Die *Implikation* A folgt B, A impliziert B, A ist Voraussetzung für B, oder kurz $A \Rightarrow B$, entspricht nicht genau dem alltäglichen Sprachgebrauch dieser Wendungen. Wenn aus A die Aussage B folgen soll, gehen wir normalerweise davon aus, daß zwischen A und B ein kausaler Zusammenhang besteht. Das behauptet $A \Rightarrow B$ nicht, auch wenn das in der Anwendung üblicherweise der Fall ist. Wenn wir die heute als falsch erkannte Aussage A := *Der Mond besteht aus grünem Käse* und für B die oben verwendete über die ungeraden Zahlen wählen, hat die Implikation A \Rightarrow B den Wahrheitswert *wahr*, auch wenn sicher kein erkennbarer Zusammenhang zwischen A und B besteht. Auf den ersten Blick mag auch befremdlich erscheinen, daß ein Wahrheitswert f für A und w für B den Wert w für A \Rightarrow B nach sich zieht. Zusammen mit den Werten f für A, f für B und w für A \Rightarrow B kann man das als die Formalisierung dafür ansehen, daß aus etwas Falschem beliebige Wahrheitswerte gefolgert werden können. Es ist der Grund dafür, daß ein Beweis mit etwas unbestreitbar Wahrem starten und bei der zu beweisenden Aussage enden muß — eine Tatsache, die einem die Schwierigkeit des Beweisens vor Augen führt, da es keine allgemeingültigen Regeln dafür gibt, bei welcher wahren Aussage man beginnen soll. Folgert man dagegen aus der Behauptung etwas Wahres, dann kann man nicht sicher sein, daß man dabei nicht doch von einer falschen Aussage ausgegangen ist und bei anderer Schlußweise auf etwas Falsches gestoßen wäre. Anders sieht es aus, wenn man in den Folgerungen auf eine falsche Aussage trifft, denn dann kann die Behauptung nicht wahr gewesen sein.

Eine weitere Bemerkung zu den verwendeten Begriffen *folgern*, *schließen*, *beweisen* ist angebracht. Wir haben sie in unseren bisherigen Argumenten benutzt, als wäre es klar, was genau darunter zu verstehen ist. Wollten wir das aber präzisieren, dann müßten wir den Vorgang des

Schließens formalisieren, um ihn auf eine solide Grundlage zu stellen. Wollten wir das hier in Angriff nehmen, so würden wir uns aber schnell in den Abgründen der Grundlagenmathematik verlieren (formale Sprachen, Prädikatenlogik ...). Wir beschäftigen uns sozusagen nur mit der globalen Logik des Schließens, ohne dabei auf dessen Mechanismus Bezug zu nehmen, der erst bei einer konkreten Belegung durch Aussagen ins Spiel kommt. Für $A := p$ ist eine Primzahl und $B := \text{Der größte gemeinsame Teiler von } p \text{ mit jeder natürlichen Zahl ist } 1 \text{ oder } p$ kann das z. B. folgendermaßen aussehen: p sei eine Primzahl und $n \in \mathbb{N}$ eine beliebige natürliche Zahl. Da der größte gemeinsame Teiler t von p und n insbesondere ein Teiler von p ist und p nach der Definition einer Primzahl nur die Teiler 1 und p besitzen kann, muß t entweder 1 oder p sein. Dabei sind wir von der Wahrheit von A ausgegangen und haben auf die Wahrheit von B geschlossen. Das ist ein *direkter Beweis*, der durch folgende aussagenlogische Formel wiedergegeben werden kann:

$$A \wedge (A \Rightarrow B) \Rightarrow B. \quad (1.4)$$

Sehen wir uns ihre Wertetabelle an:

A	B	$A \Rightarrow B$	$A \wedge (A \Rightarrow B)$	$A \wedge (A \Rightarrow B) \Rightarrow B$
w	w	w	w	w
w	f	f	f	w
f	w	w	f	w
f	f	w	f	w

(1.5)

Formel (1.4) hat eine besondere Eigenschaft, nämlich, daß sie bei jeder Belegung von A und B mit Wahrheitswerten immer den Wert w ergibt. Eine solche Aussage wird als *Tautologie* bezeichnet. Tautologien beschreiben logische Strukturen, von denen einige wichtige Beweistechniken wiederspiegeln. Bevor wir weitere Tautologien kennenlernen, führen wir den Begriff *Äquivalenz von Aussagen* ein:

$$A \Leftrightarrow B := (A \Rightarrow B) \wedge (B \Rightarrow A). \quad (1.6)$$

Die Wahrheitstabelle zeigt, daß $A \Leftrightarrow B$ genau dann w ergibt, wenn A und B denselben Wahrheitswert aufweisen.

A	B	$A \Rightarrow B$	$B \Rightarrow A$	$A \Leftrightarrow B$
w	w	w	w	w
w	f	f	w	f
f	w	w	f	f
f	f	w	w	w

(1.7)

In obigem Beispiel für A und B gilt sogar $A \Leftrightarrow B$: $A \Rightarrow B$ haben wir schon gezeigt. Bleibt die Umkehrung $B \Rightarrow A$: Dafür sei t ein Teiler von p , d. h. insbesondere gilt $t \leq p$. Der größte gemeinsame Teiler von t und p ist also t selbst. Daher kann t nur 1 oder p sein. Das bedeutet, daß p eine Primzahl ist.

Die Tautologie $A \vee \overline{A}$, oder $A \vee \neg A$, bezeichnet man als *Satz vom ausgeschlossenen Dritten*:

A	\overline{A}	$A \vee \overline{A}$
w	f	w
f	w	w

(1.8)

Satz von der Kontraposition:

A	B	$A \Rightarrow B$	$\bar{B} \Rightarrow \bar{A}$	$(A \Rightarrow B) \Leftrightarrow (\bar{B} \Rightarrow \bar{A})$
w	w	w	w	w
w	f	f	f	w
f	w	w	w	w
f	f	w	w	w

(1.9)

Setzen wir das in die Formel (1.4) für den direkten Beweis ein, so erhalten wir die Formel für den *Widerspruchsbeweis*

$$A \wedge (\bar{B} \Rightarrow \bar{A}) \Rightarrow B. \quad (1.10)$$

Das ist folgendermaßen zu verstehen: Wir gehen von der Aussage A aus und zeigen die Implikation $\bar{B} \Rightarrow \bar{A}$. Sie kann nur dann w liefern, wenn \bar{B} und \bar{A} beide den Wert w haben, oder wenn \bar{B} den Wert f hat (vergl. Tabelle (1.3)). Der erste Fall kann nicht eintreten, denn er würde $A \wedge \bar{A}$ mit dem Wert w ausstatten, was nie der Fall sein kann. Dann bleibt nur f für \bar{B} , also w für B. Oder, etwas weniger formal: Anstatt direkt die Aussage B aus A zu folgern ist es manchmal einfacher, das Gegenteil von B anzunehmen und daraus das Gegenteil von A zu folgern. Da wir von der wahren Aussage A ausgegangen sind, kann \bar{B} nicht gelten, denn A und \bar{A} können nicht beide wahr sein. Also muß B wahr sein. Mit unserem Beispiel für A und B könnte das folgendermaßen ablaufen. Wir wollen $A \Rightarrow B$ zeigen. Wir gehen von \bar{B} aus, d. h. wir nehmen an, es gäbe eine natürliche Zahl n, die mit p einen größten gemeinsamen Teiler t hat, der von 1 und p verschieden ist. Da t dann ein echter Teiler von p ist, kann p keine Primzahl sein, im Widerspruch zur vorausgesetzten Wahrheit von A. Also muß B wahr sein, denn \bar{B} führt zu einem Widerspruch.

1.1.1 A Zeigen Sie, daß die Formeln $(A \Rightarrow B) \Leftrightarrow \bar{A} \vee B$ und $A \wedge \bar{A} \Rightarrow B$ Tautologien sind. Wie läßt sich die zweite interpretieren?

1.1.2 A DE MORGANSche Regeln: Zeigen Sie

- i) $\overline{A \wedge B} \Leftrightarrow \bar{A} \vee \bar{B}$
- ii) $\overline{A \vee B} \Leftrightarrow \bar{A} \wedge \bar{B}$
- iii) $(A \wedge B) \vee C \Leftrightarrow (A \vee C) \wedge (B \vee C)$
- iv) $(A \vee B) \wedge C \Leftrightarrow (A \wedge C) \vee (B \wedge C)$

1.1.3 A Nach Vernehmung der Verdächtigen Mark, Robert und John erklärt Holmes seinem Freund Dr. Watson: Wenn Robert der Täter ist oder wenn John der Täter ist, so ist das Alibi von Mark echt. Sind aber Mark oder John unschuldig, so ist Robert der Täter. Wenn John schuldig ist, so war auch Mark an dem Verbrechen beteiligt. Dr. Watson kombiniert und lächelt. Er weiß jetzt, wer schuldig ist.

1.1.4 Aussageformen Für jedes x einer Menge sei eine Aussage $A(x)$ gegeben. Dann heißt A eine *Aussageform*.

Z. B. sei P die Aussageform, die durch $P(x) := x \text{ ist eine Primzahl}$ auf der Menge $X = \mathbb{N}$ der natürlichen Zahlen definiert ist. Dann ist $P(2)$ die wahre Aussage *2 ist eine Primzahl* und $P(4)$ die falsche *4 ist eine Primzahl*.

Aussageformen lassen sich quantifizieren. Dafür gibt es die *Quantoren*

\exists	Existenzquantor
\forall	Allquantor

Sie lassen sich auf Aussageformen anwenden:

$$\begin{array}{ll} \exists_{x \in X} P(x) & \text{Es gibt ein } x \text{ aus } X, \text{ für das } P(x) \text{ gilt} \\ \forall_{x \in X} P(x) & \text{Für alle } x \text{ aus } X \text{ gilt } P(x) \end{array}$$

Auf diese Weise lassen sich aus Aussageformen neue Aussagen gewinnen.

1.1.5 Beispiel

i) $X = \mathbb{N}, D(n) := \text{Die Summe aller natürlichen Zahlen bis } n \text{ ist } \frac{1}{2}n(n+1).$

Meist wird das kurz durch $1+2+\dots+n = \frac{1}{2}n(n+1)$ ausgedrückt, ohne die Aussageform D explizit anzugeben. Sie wird sozusagen mitgedacht. Damit lässt sich die Aussage

$$A := \forall_{n \in \mathbb{N}} D(n) = \forall_{n \in \mathbb{N}} 1 + 2 + \dots + n = \frac{1}{2}n(n+1)$$

bilden. Die Aussage A als wahr zu erkennen bedeutet dann, die Gleichung(en) $1 + 2 + \dots + n = \frac{1}{2}n(n+1)$ zu beweisen.

ii) $X = \mathbb{N}, R(n) := \sqrt{n} \text{ ist eine rationale Zahl}$

Wir können $B := \forall_{n \in \mathbb{N}} R(n)$, oder $C := \exists_{n \in \mathbb{N}} \overline{R(n)}$ bilden.

B behauptet, daß die Wurzel aus allen natürlichen Zahlen als Bruch geschrieben werden kann. C ist wahr, wenn es (wenigstens) eine natürliche Zahl gibt, deren Wurzel kein Bruch ist. C gilt offensichtlich genau dann, wenn B falsch ist. Das bedeutet $C \equiv \overline{B}$. Wir werden sehen, daß B falsch ist, denn schon $\sqrt{2}$ ist kein Bruch. An diesem Beispiel erkennt man aber, wie quantifizierte Aussagen zu verneinen sind:

$$\overline{\forall_{x \in X} P(x)} \Leftrightarrow \exists_{x \in X} \overline{P(x)}, \quad (1.11)$$

$$\overline{\exists_{x \in X} P(x)} \Leftrightarrow \forall_{x \in X} \overline{P(x)}, \quad (1.12)$$

nämlich einfach dadurch, daß die Quantoren vertauscht und die quantifizierten Aussagen $P(x)$ negiert werden.

iii) Quantoren lassen sich auch kombinieren:

$$\text{Pr} := \forall_{n \in \mathbb{N}} \exists_{\substack{m \in \mathbb{N} \\ m > n}} P(m)$$

liest sich folgendermaßen: Für jede natürliche Zahl n gibt es eine natürliche Zahl m , die größer als n ist und für die m eine Primzahl ist. Oder kurz gesagt: Es gibt unendlich viele Primzahlen.

$$\overline{\text{Pr}} = \exists_{n \in \mathbb{N}} \overline{\exists_{\substack{m \in \mathbb{N} \\ m > n}} P(m)} = \exists_{n \in \mathbb{N}} \forall_{\substack{m \in \mathbb{N} \\ m > n}} \overline{P(m)}$$

behauptet, daß es eine natürliche Zahl gibt, so daß alle größeren keine Primzahlen sind, d. h., daß es nur eine endlich Anzahl von Primzahlen gibt. Wir werden sehen, daß diese Aussage falsch ist.

1.2 Mengen

1.2.1 Definition (CANTOR 1895) Unter einer Menge verstehen wir jede Zusammenfassung von bestimmten, wohlunterschiedenen Objekten unserer Anschauung oder unseres Denkens zu einem Ganzen.

Die einfachsten Mengen lassen sich durch Angabe ihrer Elemente wiedergeben:

$W_1 := \{1, 2, 3, 4, 5, 6\}$, die möglichen Ergebnisse beim Würfeln mit einem Würfel.

$F := \{\text{rot, grün, blau}\}$, die drei Grundfarben.

$$\begin{aligned} W_2 := & \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \\ & (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\ & (3, 3), (3, 4), (3, 5), (3, 6), \\ & (4, 4), (4, 5), (4, 6), \\ & (5, 5), (5, 6), \\ & (6, 6)\}, \end{aligned}$$

die möglichen Ergebnisse beim Werfen zweier gleichfarbiger Würfel.

Wenn eine Menge zu viele Elemente hat, als daß man sie aufzählen könnte, sie aber einfach genug gebaut ist, verwendet man die Aufzählung mitunter trotzdem. Man zählt die Elemente soweit auf, bis das Bildungsgesetz deutlich wird (natürlich nur, wenn man es kennt) und ersetzt den Rest durch Auslassungspunkte.

$G := \{2, 4, 6, 8, 10, \dots\}$ ist, die Menge aller geraden Zahlen.

$H := \left\{1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \dots\right\}$ ist die Menge aller Stammbrüche, etc.

Diese Methode stößt natürlich schnell an ihre Grenzen. Bevor wir gleich leistungsfähigere kennlernen, legen wir die übliche Notation fest:

$m \in M$	Das Element m gehört zur Menge M ,
	m ist ein Element von M
\in	gehört zu ..., ist ein Element von ...
\notin	gehört nicht zu ..., ist kein Element von ...

1.2.2 Definition Eine Menge A ist Teilmenge einer Menge B , wenn jedes Element von A auch ein Element von B ist. Dafür schreiben wir $A \subseteq B$, oder $B \supseteq A$. Zwei Mengen A und B heißen gleich, wenn A und B dieselben Elemente enthalten. Wir schreiben $A \subset B$, oder $B \supset A$, wenn A eine Teilmenge von B ist und die beiden Mengen nicht gleich sind.

1.2.3 Lemma Für zwei Mengen A und B gilt genau dann $A = B$, wenn $A \subseteq B$ und $B \subseteq A$ gilt.

Beweis. Ist $A = B$, so ist nichts zu zeigen, denn $A \subseteq B$ und $B \subseteq A$ ist auch erfüllt, wenn die beiden Mengen gleich sind.

Gehen wir also von $A \subseteq B$ und $B \subseteq A$ aus und nehmen $A \neq B$ an, d.h., wir machen einen Widerspruchsbeweis. Da die Situation symmetrisch in A und B ist, können wir, ohne die Allgemeinheit unserer Argumentation einzuschränken, davon ausgehen, daß es ein Element a in

A gibt, das nicht zu B gehört. Das ergibt aber sofort einen Widerspruch dazu, daß laut Voraussetzung $A \subseteq B$ gilt, d. h., daß jedes Element aus A , insbesondere auch a , ein Element von B ist. Deshalb ist unsere Annahme zu verwerfen, und es gilt $A = B$. \square

Bemerkung: Es wird sich in kommenden Beweisen immer wieder die Situation ergeben, daß eine Ausgangslage symmetrisch bzgl. zweier Größen ist, so daß die folgende Argumentation nicht unzulässig eingeschränkt wird, wenn einer von beiden eine bestimmte Eigenschaft zugewiesen wird, die die andere genauso gut haben könnte. Dafür hat sich die Redewendung *ohne Beschränkung der Allgemeinheit* und deren Abkürzung *o. B. d. A.* durchgesetzt, die wir künftig auch verwenden werden.

Viele Mengen A sind Teilmengen gegebener Mengen B . Oft zeichnen sich die Elemente von A durch eine bestimmte Eigenschaft P aus, die (normalerweise) nicht allen Elementen von B zukommt. P ist eine *Aussageform*, die sich auf die Elemente von B bezieht. Für jedes $a \in B$ ist $P(a)$ eine Aussage. Die Menge A enthält dann alle Elemente a aus B , für die $P(a)$ wahr ist. Das wird durch die Schreibweise

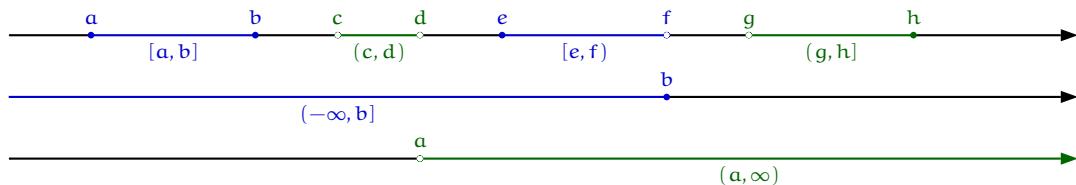
$$A = \{ a \in B \mid P(a) \} \quad (1.13)$$

wiedergegeben. Um Beispiele anführen zu können, nehmen wir vorerst die natürlichen Zahlen \mathbb{N} , die ganzen Zahlen \mathbb{Z} , die rationalen Zahlen \mathbb{Q} und die reellen Zahlen \mathbb{R} als gegeben an. Die geraden und die ungeraden Zahlen in \mathbb{N} lassen sich dann folgendermaßen darstellen:

$$G := \{ n \in \mathbb{N} \mid \exists_{k \in \mathbb{N}} n = 2k \}, \quad U := \{ n \in \mathbb{N} \mid \exists_{k \in \mathbb{N}} n = 2k - 1 \}.$$

Eine Teilmenge von \mathbb{R} der Art $\{ x \in \mathbb{R} \mid a \leq x \leq b \}$ heißt *Intervall*. Man vereinbart dafür die suggestive Schreibweise $[a, b]$. Es gibt die folgenden Intervalle:

$$\begin{array}{ll} [a, b] := \{ x \in \mathbb{R} \mid a \leq x \leq b \}, & (a, b) := \{ x \in \mathbb{R} \mid a < x < b \}, \\ [a, b) := \{ x \in \mathbb{R} \mid a \leq x < b \}, & (a, b] := \{ x \in \mathbb{R} \mid a < x \leq b \}, \\ (-\infty, b] := \{ x \in \mathbb{R} \mid x \leq b \}, & [a, \infty) := \{ x \in \mathbb{R} \mid a \leq x \}, \\ (-\infty, b) := \{ x \in \mathbb{R} \mid x < b \}, & (a, \infty) := \{ x \in \mathbb{R} \mid a < x \}. \end{array}$$



Natürlich gibt es auch $(-\infty, \infty)$, aber das ist \mathbb{R} . $[a, b]$ heißt *abgeschlossenes* und (a, b) *offenes Intervall*. $[a, b]$ und $(a, b]$ werden *halboffen* genannt. Diese Intervalle sind *beschränkt*, im Gegensatz zu $[a, \infty)$ und $(-\infty, b]$, die man *unbeschränkt* nennt. $[a, b]$ ist die Menge aller Punkte zwischen a und b , einschließlich der Randpunkte a und b . Dagegen enthält (a, b) die Randpunkte nicht und $[a, b)$ enthält zwar den unteren Randpunkt a , nicht aber den oberen b , usw. Insbesondere muß $[a, \infty)$ halboffen sein, denn ∞ ist kein Punkt aus \mathbb{R} .

Bemerkung: Für offene und halboffene Intervalle sind auch die Schreibweisen $]a, b[$, $[a, b]$, $[a, b[$ usw. gebräuchlich. Wir bevorzugen die Version mit runden Klammern, da das nach Ansicht des Autors die Lesbarkeit deutlich erhöht. Man vergleiche etwa $(-\infty, a] \cup (b e^{\mu((a,b])}, b]) \cap (3, 10^3)^c$ mit $(]-\infty, a] \cup]b e^{\mu([a,b])}, b]) \cap]3, 10^3[^c$.

Die sog. *Dreieckszahlen* sind durch $D := \{ n \in \mathbb{N} \mid \exists_{k \in \mathbb{N}} n = 1 + 2 + \dots + k \}$ definiert, die MERSENNE-Zahlen durch $M := \{ n \in \mathbb{N} \mid \exists_{k \in \mathbb{N}} n = 2^k - 1 \}$ und die FERMAT-Zahlen durch $F := \{ n \in \mathbb{N} \mid \exists_{k \in \mathbb{N}} n = 2^{2^k} + 1 \}$. Das Beispiel der letzten beiden Mengen zeigt, daß eine etwas schlankere Notation wünschenswert ist. Die Menge M wird durch die Form $2^k - 1$ seiner Elemente eindeutig festgelegt. Daher sollte man diese Menge einfach durch Angabe seiner Konstruktionsvorschrift wiedergeben können. Dafür vereinbart man die Schreibweise

$$M = \{ 2^k - 1 \mid k \in \mathbb{N} \}.$$

Man gibt die Elemente einer Menge im ersten Teil durch eine Formel an und legt nach dem senkrechten Strich ihre Variablen fest. In dieser Allgemeinheit geht diese Vorschrift über die Beschreibung von Teilmengen hinaus. Man kann sie auch dafür verwenden, um neue Mengen aus schon gegebenen zu bilden. Der Formelteil muß dabei nicht mehr Elemente einer bereits gegebenen Menge erzeugen. Wir machen das an einem Beispiel klar. Stellen wir uns vor, wir wollen die rationalen Zahlen \mathbb{Q} definieren und können dabei auf die ganzen und die natürlichen Zahlen zurückgreifen. Das Symbol $\frac{p}{q}$ für einen Bruch können wir für die Definition nicht verwenden, denn es ist ja gerade unsere Aufgabe, diesem Symbol einen mathematischen Sinn zu verleihen. Trotzdem werden wir uns davon leiten lassen, was wir über Brüche wissen. Da wir uns im Moment nicht mit den Rechenregeln der rationalen Zahlen befassen, sondern nur mit der Menge an sich, ist das Einzige, was wir beachten müssen, daß unsere Definition keinen Bruch ergibt, der sich durch Kürzen in einen anderen verwandeln läßt. Gemäß der Mengendefinition darf \mathbb{Q} nämlich nur verschiedene Elemente enthalten (CANTOR spricht von *wohlunterschiedenen Objekten*). Wir müssen daher die Situation vermeiden, in der p und q einen gemeinsamen Teiler haben, der von 1 verschieden ist. Das können wir leicht erreichen, wenn wir den Begriff *größter gemeinsamer Teiler ggT(p, q)* von p und q verwenden (siehe S. 36). Dann lautet unsere Bedingung einfach $ggT(p, q) = 1$. Als Definition der Menge der rationalen Zahlen erhalten wir

$$\mathbb{Q} := \{ [p, q] \mid p \in \mathbb{Z} \wedge q \in \mathbb{N} \wedge ggT(p, q) = 1 \}. \quad (1.14)$$

Mit dem neutralen $[p, q]$ haben wir uns bewußt von der üblichen Schreibweise $\frac{p}{q}$ abgegrenzt. Auf dieser Menge gibt es nämlich noch keinerlei Rechengesetze, die man sich bei dem Ausdruck $\frac{p}{q}$ meist dazudenkt. Diese Rechengesetze sind der zweite Schritt, nachdem die Menge als solche vorhanden ist. Sie prägen ihr eine mathematische Struktur auf, die sie erst zu dem macht, was wir unter den rationalen Zahlen verstehen.

Die Menge

$$\mathbb{P} := \{ p \in \mathbb{N} \mid p \geq 2, \ n \mid p \Rightarrow n = 1 \vee n = p \} \quad (1.15)$$

ist leicht als die Menge aller Primzahlen zu identifizieren.

1.2.4 Definition Für zwei Mengen A und B wird die Menge

$$A \times B := \{ [a, b] \mid a \in A \wedge b \in B \} \quad (1.16)$$

als Produktmenge von A und B bezeichnet. Allgemeiner ist für Mengen A_1, A_2, \dots, A_n die Produktmenge folgendermaßen definiert:

$$A_1 \times A_2 \times \dots \times A_n := \{ [a_1, a_2, \dots, a_n] \mid a_1 \in A_1, a_2 \in A_2, \dots, a_n \in A_n \}. \quad (1.17)$$

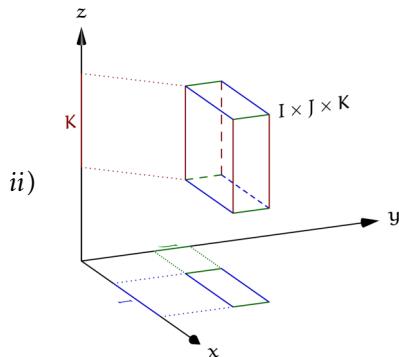
Bei der Definition (1.17) haben wir von einer gängigen Übereinkunft Gebrauch gemacht, gemäß der die Aussage $a_1 \in A_1 \wedge a_2 \in A_2 \wedge \dots \wedge a_n \in A_n$ durch eine einfache Aufzählung der einzelnen Bedingungen $a_1 \in A_1, a_2 \in A_2, \dots, a_n \in A_n$ abgekürzt werden darf.

Bemerkung: Normalerweise besteht keine Gefahr, das Tupel $[a, b]$ mit dem Intervall $[a, b] \subset \mathbb{R}$ zu verwechseln, da aus dem Kontext heraus klar wird, was gemeint ist.

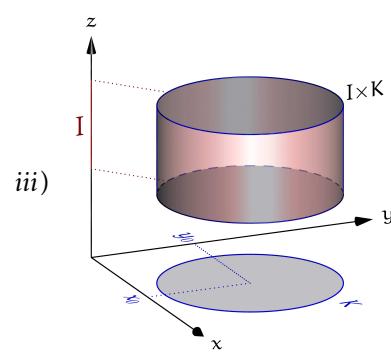
1.2.5 Beispiel

i) Für $A_1 := \mathbb{R}, \dots, A_n := \mathbb{R}$ ergibt sich $\mathbb{R}^n := \underbrace{\mathbb{R} \times \dots \times \mathbb{R}}_{n \text{ mal}}$ als die Menge

$$\mathbb{R}^n = \{ [x_1, x_2, \dots, x_n] \mid x_1, x_2, \dots, x_n \in \mathbb{R} \}. \quad (1.18)$$



I, J und K sind Intervalle in \mathbb{R} .
Dann ist $I \times J \times K \subseteq \mathbb{R}^3$ ein Quader.



$I \subseteq \mathbb{R}$ ist ein Intervall, K die Kreisscheibe
 $K := \{[x, y] \mid (x - x_0)^2 + (y - y_0)^2 \leq r^2\} \subset \mathbb{R}^2$ mit Radius r und Zentrum $[x_0, y_0]$. Dann ist $I \times K$ ein Zylinder.

1.2.6 Definition A und B seien Teilmengen einer Menge X. Wir definieren die leere Menge $\emptyset := \{ \}$ als die Menge, die kein Element enthält. Dann lassen sich die folgenden Mengenoperationen uneingeschränkt einführen.

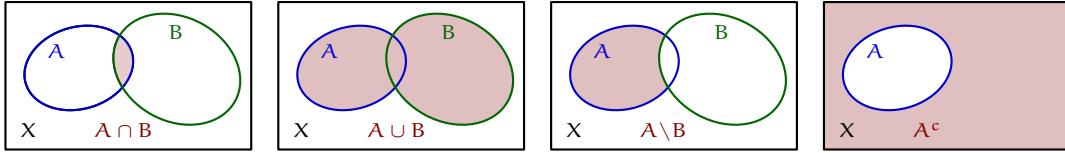
$$A \cap B := \{ x \in X \mid x \in A \wedge x \in B \}, \quad \text{Durchschnitt} \quad (1.19)$$

$$A \cup B := \{ x \in X \mid x \in A \vee x \in B \}, \quad \text{Vereinigung} \quad (1.20)$$

$$A^c := \{ x \in X \mid \neg x \in A \}, \quad \text{Komplement} \quad (1.21)$$

$$A \setminus B := \{ x \in X \mid x \in A \wedge x \notin B \}. \quad \text{Mengendifferenz} \quad (1.22)$$

A und B heißen disjunkt, falls $A \cap B = \emptyset$ gilt. Wir schreiben $x \notin A$ statt $\neg x \in A$.



Offensichtlich ist X^c die leere Menge. Die enge Verwandtschaft von \cap , \cup und c mit den logischen Junktoren \wedge , \vee und \neg spiegelt sich in den DE MORGANSchen Regeln für Mengen wieder:

1.2.7 Satz Für Teilmengen A , B und C einer Menge X gelten folgende Rechenregeln:

$$(A \cap B)^c = A^c \cup B^c \quad (1.23)$$

$$(A \cup B)^c = A^c \cap B^c \quad (1.24)$$

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C) \quad (1.25)$$

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C) \quad (1.26)$$

Beweis. Wir verwenden die DE MORGANSchen Regeln der Logik (Übung 1.1.2) und die Tatsache, daß für zwei Mengen M und N Gleichheit gilt, falls $x \in M \Leftrightarrow x \in N$ erfüllt ist.

$$\begin{aligned} x \in (A \cap B)^c &\Leftrightarrow \overline{x \in A \cap B} \Leftrightarrow \overline{x \in A \wedge x \in B} \Leftrightarrow \overline{x \in A} \vee \overline{x \in B} \Leftrightarrow x \in A^c \vee x \in B^c \\ &\Leftrightarrow x \in A^c \cup B^c. \end{aligned}$$

Das zeigt bereits die erste Gleichung. Wir verwenden $M^{cc} := (M^c)^c = M$, um aus ihr die zweite auszurechnen:

$$(A \cup B)^c = (A^{cc} \cup B^{cc})^c = ((A^c)^c \cup (B^c)^c)^c = ((A^c \cap B^c)^c)^c = A^c \cap B^c.$$

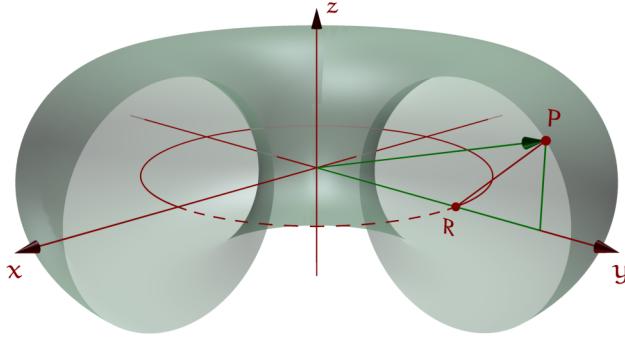
Zur dritten Gleichung:

$$\begin{aligned} x \in (A \cap B) \cup C &\Leftrightarrow (x \in A \wedge x \in B) \vee x \in C \Leftrightarrow (x \in A \vee x \in C) \wedge (x \in B \vee x \in C) \\ &\Leftrightarrow x \in A \cup C \wedge x \in B \cup C \Leftrightarrow x \in (A \cup C) \cap (B \cup C). \end{aligned}$$

Die letzte Gleichung kann man auf dieselbe Weise ableiten, oder mit Hilfe der bisherigen Ergebnisse berechnen:

$$\begin{aligned} (A \cup B) \cap C &= (A^{cc} \cup B^{cc}) \cap C^{cc} = (A^c \cap B^c)^c \cap C^{cc} = ((A^c \cap B^c) \cup C^c)^c \\ &= ((A^c \cup C^c) \cap (B^c \cup C^c))^c = (A^c \cup C^c)^c \cup (B^c \cup C^c)^c \\ &= (A \cap C) \cup (B \cap C). \end{aligned} \quad \square$$

1.2.8 A Finden Sie für den Torus eine Mengendarstellung. (Ein Torus ist ein Ring, der entsteht, wenn ein Kreis mit Radius r , der in der xz -Ebene liegt und dessen Mittelpunkt sich auf der x -Achse in der Entfernung $R > r$ vom Ursprung befindet, einmal um die z -Achse herumgedreht wird.)



Finden Sie für $R = 6$ und $r = 4$ einen Punkt auf dem Torus.

1.2.9 Die natürlichen Zahlen Wir haben bisher zur Illustration des Mengenbegriffs Beispiele herangezogen, die uns in einem strengen Aufbau der Mathematik noch gar nicht zur Verfügung stehen würden. Eine einführende Vorlesung ist jedoch nicht der geeignete Ort, um eine formale Grundlegung der Mathematik vorzunehmen. Deshalb werden wir uns auch weiterhin die Freiheit herausnehmen, die Mengen \mathbb{Z} , \mathbb{Q} und \mathbb{R} als gegeben anzusehen. Mit derselben Begründung könnten wir das auch mit den natürlichen Zahlen \mathbb{N} tun. Weil der Aufwand jedoch nicht allzu groß ist, wollen wir hier einmal demonstrieren, wie die Mengentheorie dazu eingesetzt werden kann, etwas scheinbar Vertrautes, wie die Zahlen $1, 2, 3, \dots$ einzuführen. Diesen Abschnitt kann man aber auch getrost überschlagen. Wer auf sein bisheriges Verständnis der natürlichen Zahlen $\mathbb{N} = \{1, 2, 3, 4, \dots\}$ vertraut, kann nach diesem Abschnitt unbeschadet weiterlesen. Wir stellen von NEUMANNS Idee zur Konstruktion der natürlichen Zahlen vor. Wir starten mit der einzigen Menge, die keine weiteren Voraussetzungen benötigt, mit der leeren Menge \emptyset . Sie wird die Rolle der 0 übernehmen (auch wenn wir die natürlichen Zahlen ohne die Null einführen werden). Das nächste Element ist die Menge, die die leere Menge enthält $\{\emptyset\}$. Für sie vereinbaren wir die Abkürzung 1. Es ist klar, wie das weitergehen könnte. Als 2 wählen wir die Menge $\{\{\emptyset\}\}$, als 3 dann $\{\{\{\emptyset\}\}\}$ usw. Das ist tatsächlich eine Möglichkeit, doch von NEUMANN wählte einen anderen Weg:

$$0 := \emptyset, \quad 1 := \{\emptyset\}, \quad 2 := \{\emptyset, \{\emptyset\}\}, \quad 3 := \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}, \quad \dots$$

Bevor wir das allgemeine Schema vorstellen, sehen wir uns erst einmal an, wie die 3 aus der 2 entstanden ist:

$$3 = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\} = \{\emptyset, \{\emptyset\}\} \cup \{\{\emptyset, \{\emptyset\}\}\} = 2 \cup \{2\}.$$

Das lässt ein allgemeines Verfahren erkennen: Haben wir das Element n bereits konstruiert, dann definieren wir für seinen Nachfolger n' :

$$n' := n \cup \{n\}. \quad (1.27)$$

Wir führen nun die *natürlichen Zahlen* \mathbb{N} als die Menge aller Nachfolger von 1 ein. Dieser Schritt ist nicht so harmlos, wie er scheint. Unsere Konstruktion stellt uns immer nur endlich viele natürliche Zahlen zur Verfügung, während die Vorstellung sie zu einer Menge zu vereinigen erfordert, daß wir sie in einem geeigneten Sinne alle vorliegen haben. Wenn wir

uns einmal damit abgefunden haben, daß Zahlen durch Mengen beschrieben werden, dann ist die Konstruktion eigentlich ganz überzeugend: Die Anzahl der Elemente in den Mengen $1, 2, 3, \dots$ stimmt mit unserer Vorstellung dieser Zahlen überein. Es gilt $1 = \{0\}$, $2 = \{0, 1\}$, $3 = \{0, 1, 2\}$, $n = \{0, 1, 2, \dots, n - 1\}$. Eine Zahl ist immer eine Teilmenge ihres Nachfolgers, denn $n \subset n \cup \{n\} = n'$. Die Addition auf \mathbb{N} müssen wir *rekursiv* definieren. D. h., wir können nicht sofort sagen, was wir unter $n + m$ verstehen wollen, sondern müssen schrittweise vorgehen, indem wir erst einmal $n + 1$ definieren und dann, unter der Voraussetzung, daß wir schon $n + k$ eingeführt haben, sagen was wir unter $n + k + 1$ (genau genommen unter $n + (k + 1)$) verstehen wollen. Konkret bedeutet das: $n + 1 := n'$ und $n + k' := (n + k)'$. Die Idee dahinter ist natürlich, daß nach der Festlegung von $n + 1$, nach k Schritten auch $n + k$ festgelegt ist. Genauso führen wir die Multiplikation ein: $n \cdot 1 := n$, $n \cdot k' = n \cdot (k + 1) := n \cdot k + n$ (diese Konstruktion ist ein Spezialfall des sog. Induktionsprinzips, das weiter unten vorgestellt wird). Die Zahlen $n + k$ bezeichnen wir ebenfalls als Nachfolger von n . Eine Zahl m ist demnach genau dann ein Nachfolger von n , wenn $n \subset m$ gilt. Das bedeutet, daß die \subseteq -Relation hier das gewohnte \leq auf \mathbb{N} modelliert. \subset ist dann mit $<$ zu übersetzen. Damit haben wir die natürlichen Zahlen nur mit Hilfe der Mengentheorie eingeführt, auch wenn dabei einige formale Begründungen unterschlagen wurden. Wir haben die natürlichen Zahlen ohne die Null definiert. Mitunter ist es aber wünschenswert die Menge \mathbb{N} um die Null zu erweitern. Wir schreiben dafür $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$.

1.2.10 Das Induktionsprinzip Für eine Teilmenge M von \mathbb{N} seien die folgenden beiden Eigenschaften erfüllt:

$$1 \in M, \tag{I-1}$$

$$n \in M \Rightarrow n + 1 \in M. \tag{I-2}$$

Dann besagt das *Induktionsprinzip*, daß daraus bereits $M = \mathbb{N}$ folgt. Anschaulich besagt es die einsichtige Eigenschaft, daß M mit 1 ihren Nachfolger 2, dann dessen Nachfolger 3 usw. enthalten muß und daß auf diese Weise eben alle Elemente von \mathbb{N} erreicht werden.

Dieses Prinzip läßt sich, je nach Einführung von \mathbb{N} , beweisen, oder als Teil des Axiomensystems der natürlichen Zahlen ansehen (dahinter stecken die sog. PEANO-Axiome, die wir hier aber nicht weiter erörtern werden, da wir die natürlichen Zahlen ab jetzt als intuitiv gegeben ansehen wollen). Wir nehmen den zweiten Standpunkt ein, da das Prinzip einsichtig ist.

1.2.11 Das Prinzip der vollständigen Induktion Wir stellen uns vor, daß nach einer bestimmten Vorschrift für jede natürliche Zahl $n \in \mathbb{N}$ eine Aussage A_n gegeben ist. Ein typisches Beispiel lautet etwa: $A_n : \text{Die Summe der ersten } n \text{ Zahlen ist } \frac{1}{2}n(n + 1)$. Unsere Aufgabe besteht darin, diese Aussagen auf ihren Wahrheitsgehalt hin zu untersuchen, im besten Fall also, diese unendlich vielen Aussagen zu beweisen. Daß das nicht durch viel Fleiß zu bewältigen ist, indem man sich die Aussagen der Reihe nach vornimmt, ist einsichtig. In unserem Leben können wir nur endlich viele Aufgaben erledigen, so daß wir auf diese Weise niemals *alle* Aussagen beweisen können. Hier hilft uns das Induktionsprinzip weiter. Wir definieren zu diesem Zweck die Menge

$$M := \{ n \in \mathbb{N} \mid A_n \text{ ist wahr} \} \subseteq \mathbb{N}.$$

Alle Aussagen A_n sind sicher dann wahr, wenn $M = \mathbb{N}$ gilt. Wir müssen daher nur (I-1) und (I-2) des Induktionsprinzips für M nachweisen, um auf $M = \mathbb{N}$ und daraus auf die Wahrheit aller

Aussagen A_n schließen zu können. Im Einzelnen bedeutet das: Wir müssen A_1 beweisen, denn dann gilt $1 \in M$. Dann müssen wir $n \in M \Rightarrow n+1 \in M$ zeigen. Das bedeutet, wir nehmen die Aussage A_n als wahr an ($n \in M$) und müssen mit Hilfe dieser Information auf die Wahrheit von A_{n+1} schließen ($n+1 \in M$). Entscheidend ist, daß wir nun nicht jedes einzelne A_n beweisen müssen, sondern nur noch den Mechanismus, der die Wahrheit von A_n an A_{n+1} weiterreicht, sicherzustellen haben. Ist das nämlich gelungen, dann brauchen wir uns um die Wahrheit von A_2 keine Gedanken mehr machen, da aus der (bewiesenen) Wahrheit von A_1 sofort die von A_2 , dann die von A_3 usw. folgt.

1.2.12 Satz (Vollständige Induktion) *Für jedes $n \in \mathbb{N}$ sei eine Aussage A_n gegeben. Alle Aussagen sind wahr, falls die folgenden beiden Eigenschaften erfüllt sind.*

- i) A_1 ist wahr. (Induktionsanfang)
- ii) Ist A_n wahr, dann auch A_{n+1} . (Induktionsschritt)

Den Beweis haben wir oben über das Induktionsprinzip bereits erbracht. Führen wir das Verfahren einmal am obigen Beispiel vor. Die Behauptung A_n lautet

$$1 + 2 + \dots + n = \frac{1}{2}n(n+1). \quad (1.28)$$

i) Der Induktionsanfang (meist einfach durch „ $n = 1$ “ gekennzeichnet):

A_1 behauptet $1 = \frac{1}{2} \cdot 1 \cdot 2$, was offensichtlich wahr ist.

ii) Der Induktionsschritt (durch „ $n \rightarrow n+1$ “ abgekürzt):

Wir dürfen (und müssen) A_n verwenden, d. h. wir nehmen die Aussage $1 + 2 + \dots + n = \frac{1}{2}n(n+1)$ für bare Münze. Wir kennzeichnen ihre Verwendung im Beweis an geeigneter Stelle durch IV (für Induktionsvoraussetzung). Mit ihrer Hilfe müssen wir auf A_{n+1} : $1 + 2 + \dots + n + 1 = \frac{1}{2}(n+1)(n+2)$ schließen. Es ist mitunter hilfreich, wenn man sich die Aussage A_{n+1} einmal explizit aufschreibt, um zu sehen, was denn eigentlich gezeigt werden muß. Bisher ist allerdings noch nichts passiert, wir haben sozusagen nur die Werkzeuge bereitgelegt. Starten wir:

$$\begin{aligned} 1 + 2 + \dots + n + 1 &= \color{blue}{1 + 2 + \dots + n} + n + 1 \stackrel{\text{IV}}{=} \color{blue}{\frac{1}{2}n(n+1)} + n + 1 \\ &= (n+1)\left(\frac{1}{2}n + \frac{2}{2}\right) = \frac{1}{2}(n+1)(n+2). \end{aligned}$$

Wir beginnen auf der linken Seite der Behauptung und formen sie soweit um, daß die Induktionsvoraussetzung erkennbar wird. Nachdem wir diese angewandt haben, ist der Rest nur noch einfaches Rechnen, um zur rechten Seite von A_{n+1} zu gelangen. Man kann sich denken, daß vollständige Induktion beileibe nicht immer nach diesem simplen Schema ablaufen kann. Dazu ist dieses Werkzeug zu allgemein angelegt. Trotzdem sind die Grundzüge eines typischen Induktionsbeweises erkennbar. Die Schwierigkeit im Induktionsschritt besteht meist darin, herauszufinden, wie A_n in der Aussage A_{n+1} angewendet werden kann. Dafür gibt es kein Schema, man muß die Idee normalerweise für jede Induktion neu entwickeln.

1.2.13 Der binomische Lehrsatz

1.2.14 Definition Für $n, k \in \mathbb{N}_0$ und $n \geq k$ werden durch

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} \quad (1.29)$$

die sogenannten Binomialkoeffizienten (gesprochen: n über k) definiert. Dabei ist

$$n! := 1 \cdot 2 \cdots (n-1) \cdot n, \quad 0! := 1, \quad (1.30)$$

die Fakultät von n (gesprochen: n Fakultät).

1.2.15 A Zeigen Sie, daß für $1 \leq k \leq n$ die Gleichung

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k} \quad (1.31)$$

gilt. Beweisen Sie damit, daß alle Binomialkoeffizienten natürliche Zahlen sind.

1.2.16 Satz (Binomischer Lehrsatz) Es gilt

$$(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}. \quad (1.32)$$

Mit der Notation, die uns bisher zur Verfügung steht, hätten wir die Behauptung des Satzes eigentlich folgendermaßen aufschreiben müssen:

$$\begin{aligned} (a+b)^n &= \binom{n}{0} b^n + \binom{n}{1} a b^{n-1} + \binom{n}{2} a^2 b^{n-2} + \binom{n}{3} a^3 b^{n-3} + \dots \\ &\quad \dots + \binom{n}{k} a^k b^{n-k} + \dots + \binom{n}{n-1} a^{n-1} b + \binom{n}{n} a^n. \end{aligned}$$

Diese Schreibweise ist auf die Dauer zu schwerfällig. Dabei steckt ein einfacher Mechanismus hinter dieser Summe: Alle Summanden entstehen aus einer einzigen Formel, nämlich $\binom{n}{k} a^k b^{n-k}$, indem der Reihe nach der *Summationsindex* k durch $0, 1, 2, \dots, n-1$ und n ersetzt wird und die dadurch entstandenen Ausdrücke aufsummiert werden. Genau das beschreibt das Summenzeichen \sum :

$$\sum_{k=m}^n a_k := a_m + a_{m+1} + a_{m+2} + \dots + a_{n-1} + a_n. \quad (1.33)$$

Dieselbe Aufgabe übernimmt das Produktzeichen \prod für Produkte

$$\prod_{k=m}^n a_k := a_m \cdot a_{m+1} \cdot a_{m+2} \cdots a_{n-1} \cdot a_n. \quad (1.34)$$

Wir werden gleich sehen, daß das Summenzeichen nicht nur eine bequeme Abkürzung ist, sondern daß man damit wirklich rechnen kann.

Beweis. Der Beweis wird mittels vollständiger Induktion geführt. Für $n = 1$ gilt, wegen $\binom{1}{0} = \binom{1}{1} = 1$:

$$\sum_{k=0}^1 \binom{1}{k} a^k b^{1-k} = \binom{1}{0} b + \binom{1}{1} a = (a+b)^1.$$

Das zeigt die Behauptung für $n = 1$.

Der Induktionsschritt $n \rightarrow n+1$:

$$\begin{aligned} (a+b)^{n+1} &= (a+b)(a+b)^n \stackrel{\text{IV}}{=} (a+b) \sum_{k=0}^n \binom{n}{k} a^k b^{n-k} \\ &= \sum_{k=0}^n \binom{n}{k} a^{\overbrace{k+1}^{\stackrel{=: \ell}} b^{n-k}} + \sum_{k=0}^n \binom{n}{k} a^k b^{n+1-k} \\ &= \sum_{\ell=1}^{n+1} \binom{n}{\ell-1} a^{\ell} b^{n-(\ell-1)} + \sum_{k=0}^n \binom{n}{k} a^k b^{n+1-k} \\ &= \binom{n}{n} a^{n+1} + \sum_{\ell=1}^n \binom{n}{\ell-1} a^{\ell} b^{n+1-\ell} + \sum_{k=1}^n \binom{n}{k} a^k b^{n+1-k} + \binom{n}{0} b^{n+1} \\ &= a^{n+1} + \sum_{k=1}^n \left[\binom{n}{k-1} + \binom{n}{k} \right] a^k b^{n+1-k} + b^{n+1} \\ &\stackrel{(1.31)}{=} \binom{n+1}{n+1} a^{n+1} b^0 + \sum_{k=1}^n \binom{n+1}{k} a^k b^{n+1-k} + \binom{n+1}{0} a^0 b^{n+1} \\ &= \sum_{k=0}^{n+1} \binom{n+1}{k} a^k b^{n+1-k}. \end{aligned} \quad \square$$

1.2.17 Bemerkung Zur konkreten Berechnung eines Binoms wird der Satz meist nicht verwendet. Dazu bietet sich ein Verfahren an, das mit dem sogenannten PASCALSchen Dreieck verbunden ist. Stellen wir uns vor, wir benötigen tatsächlich den Ausdruck $(a+b)^6$. Nach dem binomischen Lehrsatz (1.32) müßten wir dafür

$$(a+b)^6 = \binom{6}{6} a^6 + \binom{6}{5} a^5 b + \binom{6}{4} a^4 b^2 + \binom{6}{3} a^3 b^3 + \binom{6}{2} a^2 b^4 + \binom{6}{1} a b^5 + \binom{6}{0} b^6$$

berechnen. In dieser Formel erfordern nur die Binomialkoeffizienten $\binom{6}{k}$ wirkliche Rechenarbeit, denn die Abfolge der Potenzen $a^k b^{6-k}$ gehorcht einer starren Regel:

$$(a+b)^6 = a^6 + ? \cdot a^5 b + ? \cdot a^4 b^2 + ? \cdot a^3 b^3 + ? \cdot a^2 b^4 + ? \cdot a b^5 + b^6.$$

Wenn wir die Koeffizienten auf andere Weise, als durch Berechnung von $\binom{6}{k}$, erhalten könnten, ließe sich das Ergebnis ohne großen Aufwand hinschreiben. Das PASCALSche Dreieck ist dafür das Mittel der Wahl. Das Bildungsgesetz läßt sich leicht erraten. Jede neue Zeile entsteht aus

der vorhergehenden, indem man jeweils zwei benachbarte Einträge addiert und das Ergebnis in die Lücke darunter einträgt:

Das ist die Umsetzung der zentralen Gleichung (1.31) für die Binomialkoeffizienten. In der siebten Zeile stehen die Vorfaktoren für $(a + b)^6$:

$$(a+b)^6 = 1 \cdot a^6 + 6 \cdot a^5 b + 15 \cdot a^4 b^2 + 20 \cdot a^3 b^3 + 15 \cdot a^2 b^4 + 6 \cdot a b^5 + 1 \cdot b^6.$$

1.2.18 A Wir haben bei der Einführung des Mengenprodukts $A \times B$ zweier Mengen A und B die Schreibweise $[a, b]$ ($a \in A$ und $b \in B$), ein sogenanntes *Tupel*, verwendet, jedoch kein Wort darüber verloren, um was es sich dabei eigentlich genau handelt. Können Sie eine überzeugende mathematische Definition für das Objekt $[a, b]$ finden, die sich nur auf den Mengenbegriff stützt?

2 Relationen und Funktionen

2.1 Allgemeine Eigenschaften von Relationen

2.1.1 Definition Gegeben seien zwei Mengen A und B . Dann ist eine (zweistellige) Relation R eine Teilmenge von $A \times B$:

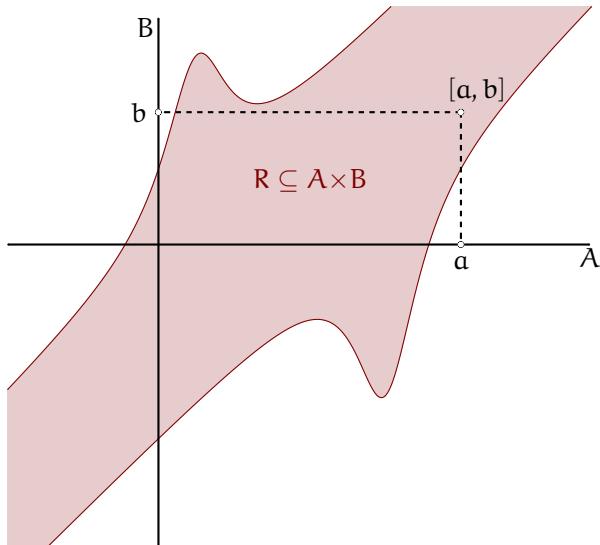
$$R \subseteq A \times B.$$

Wir sagen, ein Element $a \in A$ steht in der Relation R zu einem Element $b \in B$, wenn $[a, b] \in R$ gilt. Dafür wird aRb , manchmal auch $R(a, b)$ geschrieben.

Sind allgemeiner n Mengen A_1, A_2, \dots, A_n gegeben, so ist eine n -stellige Relation R eine Teilmenge von $A_1 \times A_2 \times \dots \times A_n$:

$$R \subseteq A_1 \times A_2 \times \dots \times A_n.$$

Wir sagen, die Elemente $a_1 \in A_1, a_2 \in A_2, \dots, a_n \in A_n$ stehen in der Relation R , falls $[a_1, a_2, \dots, a_n] \in R$ gilt. Dafür schreiben wir $R(a_1, a_2, \dots, a_n)$.



Dieser Ansatz erscheint zunächst sehr abstrakt. Tatsächlich bildet er aber auf einfache Weise das ab, was wir machen, wenn wir Dinge in eine Beziehung (Relation) zueinander setzen. Stellen wir uns als Beispiel den Fuhrpark einer (kleinen) Autovermietung vor. Praktischerweise werden wir den Bestand an Fahrzeugen in Form einer Tabelle darstellen. Dann wird das Fahrzeug mit

Fahrz.Nr.	Fabrikat	Kennzeichen	Farbe	Baujahr
1	BMW	X - Y 123	blau	2008
2	VW	Z - U 456	rot	2009
3	Fiat	V - W 100	gelb	2003

Tabelle 2.1 *Fuhrpark*

der Fahrzeugnummer 2 durch den Datensatz [2, VW, Z - U 456, rot, 2009] dargestellt. Das ist ein Element der Produktmenge

$$\mathbb{N} \times \text{Autos} \times \text{Kennzeichen} \times \text{Farben} \times \mathbb{N}_{>1999},$$

die aus den natürlichen Zahlen \mathbb{N} für die Fahrzeugnummer, der Menge aller lieferbaren Autos, der Menge der von der Straßenverkehrsbehörde zu vergebenden Autokennzeichen, der Menge der gängigen Farben und den natürlichen Zahlen größer als 1999 gebildet wird. Wir können den Fuhrpark also auch als eine Teilmenge *Fuhrpark* dieser Produktmenge auffassen:

$$[2, \text{VW}, \text{Z - U 456, rot, 2009}] \in \mathbb{N} \times \text{Autos} \times \text{Kennzeichen} \times \text{Farben} \times \mathbb{N}_{>1999}.$$

[3, Mercedes, A - BC 1, pink, 2001] ist zwar auch ein Element der Produktmenge, gehört aber offensichtlich nicht zur Relation *Fuhrpark*.

Wenn wir uns das Beispiel weiterdenken, bieten sich noch andere Relationen an, z. B. die Relationen *Kunden* und *Verleih*:

Kunden.Id.	Name	Vorname
001	Lustig	Peter
002	Riese	Adam
007	Bond	James

Tabelle 2.2 *Kunden*

Kunden.Id.	Fahrz.Nr.
007	1
007	3
001	2

Tabelle 2.3 *Verleih*

Die Relation *Verleih*(007,3) $\in \mathbb{N}_0^3 \times \mathbb{N}$ bedeutet dann, daß der gelbe Fiat an den Kunden James Bond ausgeliehen wurde.

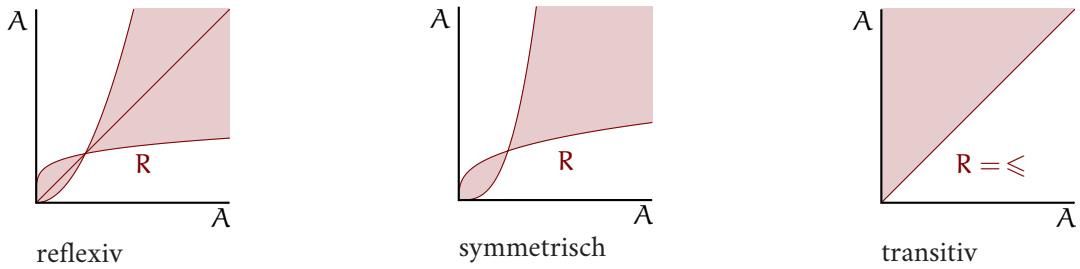
2.2 Klassifikation von Relationen

2.2.1 Definition Eine Relation $R \subseteq A \times B$ heißt

- homogen* $\Leftrightarrow A = B,$
- reflexiv* $\Leftrightarrow R \text{ ist homogen und f. a. } a \in A \text{ ist } [a, a] \in R,$
- symmetrisch* $\Leftrightarrow R \text{ ist homogen und aus } [a, b] \in R \text{ folgt } [b, a] \in R,$
- antisymmetrisch* $\Leftrightarrow R \text{ ist homogen und aus } [a, b] \in R \text{ und } [b, a] \in R \text{ folgt } a = b,$
- transitiv* $\Leftrightarrow R \text{ ist homogen und aus } [a, b] \in R, [b, c] \in R \text{ folgt } [a, c] \in R,$
- linkstotal* $\Leftrightarrow \text{f. a. } a \in A \text{ gibt es ein } b \in B, \text{ so daß } [a, b] \in R,$
- rechtstotal* $\Leftrightarrow \text{f. a. } b \in B \text{ gibt es ein } a \in A, \text{ so daß } [a, b] \in R,$
- funktional* $\Leftrightarrow \text{aus } [a, b] \in R \text{ und } [a, c] \in R \text{ folgt } b = c,$
- injektiv* $\Leftrightarrow \text{aus } [a, b] \in R \text{ und } [c, b] \in R \text{ folgt } a = c,$
- bijektiv* $\Leftrightarrow \text{f. a. } b \in B \text{ gibt es genau ein } a \in A, \text{ so daß } [a, b] \in R.$

Rechtstotale Relationen werden meist als surjektiv bezeichnet.

Das Adjektiv *bijektiv* wird üblicherweise nur für funktionale Relationen gebraucht. Man überlegt sich leicht: Eine Relation ist genau dann bijektiv, wenn sie sowohl injektiv als auch surjektiv ist.



Die für uns wichtigsten Relationen sind die *Ordnungsrelationen*, die *Äquivalenzrelationen* und natürlich die funktionalen Relationen, denn sie definieren *Funktionen*. Erstere ist eine homogene, antisymmetrische, reflexive und transitive Relation auf einer Menge A . Sie ist nach der gewohnten \leqslant -Relation auf \mathbb{R} modelliert. Deshalb schreibt man auch meist $a \leqslant b$ für eine solche Relation und nicht $[a, b] \in \leqslant$. Die abstrakten Eigenschaften sind in \mathbb{R} sofort einsichtig: Die Reflexivität bedeutet $a \leqslant a$, die Antisymmetrie, daß $a \leqslant b$ und $b \leqslant a$ nur für $a = b$ möglich ist und die Transitivität, daß aus $a \leqslant b$ und $b \leqslant c$ auch $a \leqslant c$ folgen muß.

2.3 Äquivalenzrelationen

Das Rechnen modulo p wird oft als Gelegenheit genutzt, um Äquivalenzrelationen einzuführen. Dabei werden zwei ganze Zahlen a und b als *gleich modulo einer natürlichen Zahl p* angesehen, wenn sich b von a um ein ganzzahliges Vielfaches von p unterscheidet: $b = a + t \cdot p$, $t \in \mathbb{Z}$, geschrieben als $a = b \pmod{p}$ (vergl. Seite 40). Wenn wir diese Relation vorübergehend mit R bezeichnen, dann ist $R \subseteq \mathbb{Z} \times \mathbb{Z}$ eine homogene, reflexive, symmetrische und transitive Relation. Die Reflexivität ist wegen $a = a + 0 \cdot p$ erfüllt, die Symmetrie, weil aus $a = b + t \cdot p$

natürlich $b = a + (-t) \cdot p$ folgt. Die Transitivität ist erfüllt, da aus $a = b + t \cdot p$ und $b = c + s \cdot p$ sofort $a = c + (s + t) \cdot p$ folgt. Eine Relation auf einer beliebigen Menge nach diesem Muster heißt Äquivalenzrelation. Für sie wird oft das Symbol \sim verwendet.

2.3.1 Definition Eine Relation $\sim \subseteq A \times A$ auf einer Menge A mit den Eigenschaften

- i) $a \sim a$ f. a. $a \in A$, (Reflexivität)
- ii) $a \sim b \Rightarrow b \sim a$, (Symmetrie)
- iii) $a \sim b \wedge b \sim c \Rightarrow a \sim c$, (Transitivität)

heißt Äquivalenzrelation auf A . Die Menge $[a]_\sim := \{ b \in A \mid a \sim b \}$ heißt Äquivalenzklasse der Relation \sim zum Repräsentant a . Für $a \sim b$ sagen wir: a ist zu b äquivalent.

Meist schreiben wir für eine Äquivalenzklasse einfach $[a]$, wenn es klar ist, welche Äquivalenzrelation gemeint ist. Der Repräsentant a einer Äquivalenzklasse ist normalerweise keineswegs eindeutig. Es ist $[a] = [a']$ für zwei verschiedene Elemente a und a' möglich – und zwar genau dann, wenn $a \sim a'$ gilt, also wenn die Repräsentanten äquivalent sind. Das wollen wir uns jetzt überlegen. Zunächst zeigen wir, daß aus $a \sim a'$ die Gleichheit von $[a]$ und $[a']$ folgt: Wegen $a \sim a'$ liegt a' in $[a]$. Für ein $b \in [a']$ muß $b \sim a'$ erfüllt sein. Wegen der Transitivität folgt dann aus $a' \sim a$ auch $b \sim a$, also $b \in [a]$. Jedes Element b von $[a']$ liegt daher in $[a]$. Damit haben wir $[a'] \subseteq [a]$ gezeigt. Da unsere Argumentation bezüglich a und a' symmetrisch ist, haben wir auch $[a] \subseteq [a']$ und daher $[a] = [a']$.

Für die andere Richtung gehen wir von $[a] = [a']$ aus. Das bedeutet aber, da a' wegen der Reflexivität in $[a']$ liegt, daß a' ein Element von $[a]$ ist, was per Definition $a \sim a'$ bedeutet. Damit ist alles gezeigt.

Eine unmittelbare Folgerung ist, daß zwei verschiedene Äquivalenzklassen $[a]$ und $[b]$ disjunkt sein müssen. Gäbe es nämlich ein $c \in [a] \cap [b]$, dann hieße das $a \sim c$ und $c \sim b$, woraus wegen der Transitivität sofort $a \sim b$ und, nach unseren Überlegungen, der Widerspruch $[a] = [b]$ entstehen würde.

Die Menge aller Äquivalenzklassen besteht aus lauter disjunkten Mengen, deren Vereinigung ganz A ergibt. Denn jedes Element $a \in A$ gehört zu $[a]$ und damit zu $\bigcup_{b \in A} [b] \subseteq A$. Das bedeutet $A \subseteq \bigcup_{b \in A} [b]$, oder $A = \bigcup_{b \in A} [b]$ (man beachte, daß in dieser Vereinigung viele Mengen gleich sind). Die Äquivalenzklassen bilden eine sogenannte *disjunkte Überdeckung von A* . Darunter verstehen wir folgendes: Es gibt eine Indexmenge I und für jedes $i \in I$ eine Teilmenge A_i von A , so daß $A_i \cap A_j = \emptyset$ für $i \neq j$ und $\bigcup_{i \in I} A_i = A$ gilt. Eine solche Überdeckung schreiben wir kurz in der Form $\{A_i \mid i \in I\}$. Als Indexmenge I wählen wir in unserem Kontext die Menge der verschiedenen Äquivalenzklassen $I := \bigcup_{a \in A} [a]$ (warum definieren wir nicht einfach $I := \{[a] \mid a \in A\}$?) und für die disjunkten Mengen die Elemente von I (also die Äquivalenzklassen) selbst. Jetzt können wir unsere obigen Überlegungen etwas verfeinern: Jedes $a \in A$ gehört zu $[a]$ und damit zu einer der Äquivalenzklassen aus I . Das bedeutet $A \subseteq \bigcup_{[b] \in I} [b] \subseteq A$, also $A = \bigcup_{[b] \in I} [b]$.

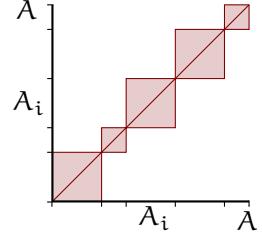
Damit haben wir die Hälfte des folgenden Satzes bereits bewiesen.

2.3.2 Satz Für jede Äquivalenzrelation \sim auf einer Menge A bilden die verschiedenen Äquivalenzklassen $[a]_\sim$, $a \in A$, eine disjunkte Überdeckung von A . Umgekehrt definiert jede disjunkte Überdeckung $\{A_i \mid i \in I\}$ von A auf folgende Weise eine Äquivalenzrelation: $a \sim b : \Leftrightarrow \exists_{i \in I} a, b \in A_i$. Die zugehörigen Äquivalenzklassen sind dabei durch die Mengen A_i gegeben.

Beweis. Wir gehen von einer disjunkten Überdeckung $\{A_i \mid i \in I\}$ von A aus und müssen zeigen, daß durch $a \sim b : \Leftrightarrow \exists_{i \in I} a, b \in A_i$ eine Äquivalenzrelation definiert wird. Die Reflexivität folgt aus der Tatsache, daß jedes $a \in A$ in genau einer der Mengen A_i liegen muß, da diese Mengen ja ganz A überdecken und paarweise disjunkt sind. Die Symmetrie ist offensichtlich. Bleibt die Transitivität: Aus $a \sim b$ und $b \sim c$ folgt zunächst, daß es genau ein $i \in I$ mit der Eigenschaft $a, b \in A_i$ gibt. Dann bedeutet aber $b \sim c$, daß auch c zu A_i gehören muß. Es folgt $a, c \in A_i$, also $a \sim c$. Die Bedingungen von 2.3.1 sind damit nachgewiesen. Die Äquivalenzklassen $[a]$ sind laut Definition $[a] = \{b \in A \mid a \sim b\}$. Andererseits liegt a in genau einer Menge A_i . Alle zu a äquivalenten Elemente b liegen gemäß der Definition der Äquivalenzrelation aber ebenfalls in A_i . Das zeigt $[a] \subseteq A_i$. Andererseits gilt für jedes $b \in A_i$ natürlich $a, b \in A_i$, also $a \sim b$, oder $b \in [a]$. Das hat die fehlende Inklusion $A_i \subseteq [a]$ zur Folge. \square

Im Lichte dieses Satzes sollte die symbolische Darstellung einer Äquivalenzrelation wie in der nebenstehenden Skizze aussehen.

Meist will man mit einer Äquivalenzrelation Eigenschaften der Ausgangsmenge A in die Menge der Äquivalenzklassen übertragen. Etwa bei dem Beispiel $x \sim y : \Leftrightarrow x = y \bmod p$ auf der Menge $A := \mathbb{Z}$ will man die Grundrechenarten \pm und \cdot auf der Menge der Äquivalenzklassen zur Verfügung haben. Dafür schreibt man dann $A/$ und spricht vom *Schnitt der Menge A nach der Relation \sim* . Es haben sich aber auch andere Schreibweisen für Konstruktionen dieser Art eingebürgert, wie wir gleich sehen werden. In unserem Beispiel haben wir die Rechenoperationen \pm und \cdot auf den Äquivalenzklassen $[x]$ und $[y]$ zu definieren. Das machen wir folgendermaßen:



$$[x] + [y] := [x + y], \quad (2.1)$$

$$[x] - [y] := [x - y], \quad (2.2)$$

$$[x] \cdot [y] := [x \cdot y]. \quad (2.3)$$

Diese Definitionen haben wir mit Hilfe der Repräsentanten x und y von $[x]$ bzw. $[y]$ formuliert. Da diese durch die Äquivalenzklassen aber nicht eindeutig festgelegt sind, müssen wir uns davon überzeugen, daß die Definition nicht in sich widersprüchlich ist. Man sagt, wir müssen prüfen, daß die Rechenoperationen *wohldefiniert* sind. Wir wissen, daß die Äquivalenzklassen $[x]$ und $[y]$ auch durch die zu x bzw. y äquivalenten Zahlen x' bzw. y' erzeugt werden: $[x] = [x']$, $[y] = [y']$. Für die linke Seite von (2.1) könnten wir auch $[x'] + [y']$ schreiben. Wir müssen also zeigen, daß die rechte Seite $[x + y]$ nicht von der speziellen Wahl von x und y abhängt. Das ist sicher dann der Fall, wenn wir $[x + y] = [x' + y']$ nachweisen können. Jetzt verwenden wir die Definition der Äquivalenzrelation: $x \sim x' \Leftrightarrow \exists_{s \in \mathbb{Z}} x' = x + sp$, $y \sim y' \Leftrightarrow \exists_{t \in \mathbb{Z}} y' = y + tp$. Es folgt $x' \pm y' = x \pm y + (s \pm t)p \sim x \pm y$ und $x' \cdot y' = x \cdot y + (sy + tx + st)p \sim x \cdot y$.

Damit haben wir die Wohldefiniertheit gleich für alle Rechenoperationen (2.1) – (2.3) gezeigt. Natürlich lässt sich jetzt auch $[x]^n := [x^n]$ für $n \in \mathbb{N}$ definieren. In der Zahlentheorie auf Seite 40 führen wir letztlich diese Rechengesetze für das Rechnen modulo p ein, auch wenn wir dort nicht diesen formalen Aufwand treiben (wir vereinbaren die Schreibweise $x =_p y$ für $x = y \pmod p$ und zeigen wie oben, daß sich diese neue Gleichheit mit den Grundrechenarten \pm und \cdot verträgt). Die Äquivalenzklasse $[0]$ besteht einfach aus allen ganzzahligen Vielfachen von p . Das ist eine Menge, die man suggestiv durch $p\mathbb{Z}$ bezeichnet. Sie legt die Äquivalenzrelation fest, denn $x \sim y \Leftrightarrow x - y \in p\mathbb{Z}$. Deshalb hat sich für die Menge der Äquivalenzklassen mit den eingeführten Rechenoperationen das aussagekräftige Symbol $\mathbb{Z}/p\mathbb{Z}$ durchgesetzt.

2.4 Verkettung von Relationen und die inverse Relation

2.4.1 Definition Für zwei Relationen $R \subseteq A \times B$ und $S \subseteq B \times C$ definiert

$$S \circ R := \{ [a, c] \in A \times C \mid \exists_{b \in B} [a, b] \in R \wedge [b, c] \in S \} \quad (2.4)$$

die Verkettung der Relationen R und S .

$$R^{-1} := \{ [b, a] \in B \times A \mid [a, b] \in R \} \quad (2.5)$$

heißt inverse Relation von R .

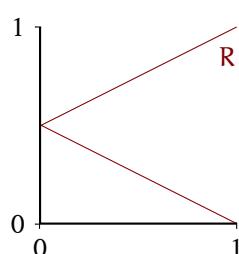
In dieser Allgemeinheit ist es durchaus möglich, daß $S \circ R$ die sogenannte *Nullrelation* $O := \emptyset \subseteq A \times C$ ist. Da für $[b, c] \in S$ nicht alle Elemente aus B vorkommen müssen, kann es sein, daß für kein Paar $[a, c] \in A \times C$ ein gemeinsames $b \in B$ mit der Eigenschaft $[a, b] \in R$ und $[b, c] \in S$ gefunden werden kann.

Die Rolle von R^{-1} zeigt sich bei der Verkettung mit R :

$$\begin{aligned} R^{-1} \circ R &= \{ [a, a'] \in A \times A \mid \exists_{b \in B} [a, b] \in R \wedge [b, a'] \in R^{-1} \} \\ &= \{ [a, a'] \in A \times A \mid \exists_{b \in B} [a, b] \in R \wedge [a', b] \in R \}. \end{aligned}$$

Offensichtlich ist mit jedem $[a, b] \in R$ das Paar $[a, a]$ in $R^{-1} \circ R$. Diese Menge kann aber weitaus mehr Elemente als diese sogenannten *Diagonalelemente* enthalten. Etwa für $A = B = \mathbb{R}$ und der \leq -Relation R ist R^{-1} die \geq -Relation. Daher ist $R^{-1} \circ R = \{ [a, a'] \in \mathbb{R} \times \mathbb{R} \mid \exists_{b \in \mathbb{R}} a \leq b \wedge a' \leq b \} = \mathbb{R} \times \mathbb{R}$, denn für je zwei reelle Zahlen a und a' gibt es immer eine reelle Zahl b , die beide übertrifft.

Man sollte also nicht zu viel hinter der Bezeichnung *Inverse* vermuten, wie z. B., daß $R^{-1} \circ R$ eine Teilmenge der *identischen Relation* (oder *Gleichheitsrelation*) $I_A := \{ [a, a] \in A \times A \mid a \in A \}$ ist. Dafür muß aus $[a, b] \in R$ und $[a', b] \in R$ immer $a = a'$ folgen,



d. h., R hat dafür injektiv zu sein. Soll auch noch $R \circ R^{-1} \subseteq I_B$ gelten, so muß aus $[b, a] \in R^{-1}$ und $[b', a] \in R^{-1}$, also aus $[a, b] \in R$ und $[a, b'] \in R$ immer $b = b'$ folgen. Das bedeutet, daß R eine funktionale Relation ist. Beide Eigenschaften sind daher nur für injektive Funktionen (s.u.) erfüllt. Man könnte meinen, daß eine injektive Relation automatisch auch funktional sein muß, aber die Relation R aus der nebenstehenden Skizze zeigt, daß das nicht der Fall sein muß.

2.5 Funktionen

Für eine funktionale Relation $\emptyset \neq f \subseteq A \times B$ definieren wir die Mengen

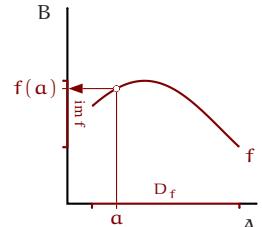
$$D_f := \{ a \in A \mid \exists_{b \in B} [a, b] \in f \}, \quad (2.6)$$

$$\text{im } f := \{ b \in B \mid \exists_{a \in A} [a, b] \in f \}. \quad (2.7)$$

$D_f \subseteq A$ heißt *Definitionsbereich* und $\text{im } f$ *Bild* von f . Solche Relationen bezeichnen wir als *Funktionen* und vereinbaren statt $[a, b] \in f$ oder $a \in f$ die üblichen Schreibweisen $f : A \rightarrow B$, $a \mapsto f(a) = b$, (streng genommen ist diese Schreibweise dem Fall $D_f = A$ vorbehalten) oder ausführlicher $f : A \supseteq D_f \rightarrow B$, $a \mapsto f(a)$. Das Bild von f ist jetzt $\text{im } f = \{ b \in B \mid \exists_{a \in A} b = f(a) \}$.

Wieso beschreiben funktionale Relationen genau das, was wir unter einer Funktion verstehen?

Die gängige Vorstellung von einer Funktion ist doch, daß aus einem Vorrat von Werten a aus einer Menge A , der *Urbildmenge*, einer geeigneten Vorschrift folgend, Bildelemente $f(a)$ aus einer Zielmenge B gebildet werden. In die Sprache der Relationen übersetzt heißt das: Jedes $a \in A$ darf mit nicht mehr als einem Element $b \in B$ in Relation f stehen. Das ist genau der Inhalt der Definition einer funktionalen Relation: Aus $[a, b] \in f$ und $[a, c] \in f$ folgt $b = c$. In der üblichen Notation heißt das: Aus $f(a) = b$ und $f(a) = c$ folgt $b = c$. Die Teilmenge f von $A \times B$ können wir jetzt durch $f = \{ [a, f(a)] \mid a \in D_f \}$ wiedergeben. f ist demnach die Menge, die man den *Graphen* der Funktion f nennt. Vom Standpunkt der Relationen aus identifiziert man also den Graphen einer Funktion mit der Funktion selbst. Aus diesem Grund werden wir oft von einer Funktion f sprechen und dabei den Definitionsbereich und die Abbildungsvorschrift als gegeben annehmen.



Wenn die Mengen A und B aus dem Kontext klar ersichtlich sind, gestatten wir es uns aber auch, einfach von *der Funktion* $f(x)$ zu sprechen, obwohl das natürlich streng genommen falsch ist, denn $f(x)$ ist der sogenannte Funktionswert von f und nicht die Funktion. Was damit gemeint ist, zeigt folgendes Beispiel: Für die Funktion f auf \mathbb{R} , die ausführlich durch $f : \mathbb{R} \supseteq \mathbb{R} \setminus \{1\} \rightarrow \mathbb{R}$, $x \mapsto \frac{x}{1-x}$ definiert ist, können wir dann einfach von der Funktion $f(x) = \frac{x}{1-x}$ sprechen. Wir vertrauen dabei darauf, daß der Leser aus dem Kontext heraus schon erkennen wird, daß es sich hier um ein reelle Funktion handelt, deren Definitionsbereich, mit Ausnahme der 1, alle reellen Zahlen umfasst, und die durch die Rechenvorschrift $\frac{x}{1-x}$ eindeutig festgelegt ist. Sollte dabei ein kleinerer Definitionsbereich D_f , als der maximal mögliche $\mathbb{R} \setminus \{1\}$ gewünscht sein, so muß dieser natürlich noch explizit angegeben werden. Trotz der augenscheinlichen Bequemlichkeit dieser Konvention, werden wir aber versuchen, sie wirklich nur in so übersichtlichen Situationen anzuwenden, wie in diesem Beispiel.

2.5.1 Rechnen mit Funktionen Für zwei Funktionen $f : X \supseteq D_f \rightarrow Y$ und $g : Y \supseteq D_g \rightarrow Z$ läßt sich, falls das Bild im f im Definitionsbereich D_g von g enthalten ist, durch

$$g \circ f : X \supseteq D_f \rightarrow Z, x \mapsto g(f(x)) \quad (2.8)$$

eine Funktion gewinnen, die als *Verkettung*, oder *Hintereinanderausführung* von g und f bezeichnet wird. Für $f(x) := x^2 + 1$ und $g(y) := \frac{1}{y}$ ist $g \circ f(x) = g(f(x)) = \frac{1}{x^2+1}$. Gemäß unserer

Übereinkunft ist der Definitionsbereich dieser Funktionen jeweils der maximal mögliche, da nichts anderes bestimmt wurde. Das heißt $D_f = \mathbb{R}$ und $D_g = \mathbb{R} \setminus \{0\}$. Wenn wir $f(x)$ zu $x^2 - 1$ abändern, lassen sich g und f nicht mehr verketten, denn $0 \in D_f$, so daß im f nicht mehr in D_g enthalten ist. Die Funktionsvorschrift $g(f(x)) = \frac{1}{x^2 - 1}$ ergibt aber durchaus eine vernünftige Funktion, mit dem natürlichen Definitionsbereich $\mathbb{R} \setminus \{-1, 1\}$. Unsere Version der Verkettung zweier Funktionen ist also zu restriktiv. Wenn wir im $f \subseteq D_g$ zu im $f \cap D_g \neq \emptyset$ abschwächen, läßt sich $g(f(x))$ für x aus einer möglicherweise echten Teilmenge von D_f bilden. Genauer gesagt läßt sich dieser Funktionswert für alle x aus der Menge $\{x \in X \mid f(x) \in D_g\} \subseteq D_f$ bilden.

2.5.2 Definition Für eine Funktion $f : X \supseteq D_f \rightarrow Y$ und eine Teilmenge $B \subseteq Y$ heißt

$$f^{-1}(B) := \{x \in X \mid f(x) \in B\} \quad (2.9)$$

Urbild(menge) von B unter der Funktion f . Für eine Teilmenge $A \subseteq D_f$ heißt die Funktion $f|_A : X \supseteq A \rightarrow Y$, $A \ni x \mapsto f(x)$ die Einschränkung von f auf A . Die Funktion $\text{id} : X \rightarrow X$, $x \mapsto x$ wird identische Funktion genannt.

Natürlich ist $f^{-1}(Y) = D_f$ und $f^{-1}(\emptyset) = \emptyset$. Die etwas unglücklich gewählte Notation $f^{-1}(A)$ hat normalerweise nichts mit einer Umkehrfunktion zu tun (die es beileibe nicht für alle Funktionen gibt, siehe Abschnitt 2.5.5), oder gar mit der Funktion $\frac{1}{f}$. Aus dem Kontext heraus ist aber meist klar, was gemeint ist, denn f^{-1} wird auf eine Menge angewandt und nicht auf das Element einer Menge.

Für die Verkettung zweier Funktionen f und g verlangen wir also noch, daß im $f \cap D_g \neq \emptyset$ gilt und haben dadurch den natürlichen Definitionsbereich $D_{g \circ f} := f^{-1}(D_g)$ für $g \circ f$.

Wenn die Zielmenge Y für zwei Funktionen $f : X \supseteq D_f \rightarrow Y$, $g : X \supseteq D_g \rightarrow Y$ eine Addition und eine Multiplikation, gegebenenfalls auch eine Division zur Verfügung stellt, dann lassen sich zu jeder dieser Rechenoperationen auch Funktionen bilden:

$$(f \pm g)(x) := f(x) \pm g(x), \quad D_{f \pm g} = D_f \cap D_g, \quad (2.10)$$

$$(f \cdot g)(x) := f(x) \cdot g(x), \quad D_{f \cdot g} = D_f \cap D_g, \quad (2.11)$$

$$\left(\frac{f}{g}\right)(x) := \frac{f(x)}{g(x)}, \quad D_{\frac{f}{g}} = D_f \cap (D_g \setminus g^{-1}(\{0\})), \quad (2.12)$$

$$f^n(x) := (f(x))^n, \quad D_{f^n} = D_f, \quad (n \in \mathbb{N}_0), \quad (2.13)$$

$$f^{-n}(x) := \frac{1}{(f(x))^n}, \quad D_{f^{-n}} = D_f \setminus f^{-1}(\{0\}), \quad (n \in \mathbb{N}). \quad (2.14)$$

Im Lichte dieser Definition sollte man jetzt mit der Übersetzung der Gleichung $\sin^2 + \cos^2 = 1$ keine Probleme haben. Es gilt $\sin^2(x) + \cos^2(x) = 1$, oder ganz ausgeschrieben: $(\sin(x))^2 + (\cos(x))^2 = 1$. Da das aber sehr umständlich aussieht, schreibt man auch $\sin(x)^2 + \cos(x)^2 = 1$ (vergl. (11.14)). Nebenbei gesagt sind Gleichungen dieser Art der Grund dafür, daß wir auch für die bekannten Funktionen \sin , \cos , \tan , \ln etc. strikt an der Funktionsschreibweise festhalten, also $\sin(x)$, $\cos(x)$ usw. statt $\sin x$, $\cos x$ schreiben. So müssen wir nie erklären, daß wir mit $\sin(x)^2$ natürlich $(\sin(x))^2$ meinen und nicht etwa $\sin(x^2)$. In der kürzeren Notation kann

$\sin x^2$ nur $\sin(x^2)$ bedeuten, $\sin(x)^2$ müßte immer durch die schwerfällige Version $(\sin x)^2$ ersetzt werden. Spätestens wenn man Konstruktionen wie $\sin(x+1)$ benötigt, läßt sich die Kurznotation nicht mehr durchhalten, da sonst $\sin x+1$ nicht von $(\sin x)+1$ zu unterscheiden wäre.

2.5.3 A Zeigen Sie: Für die Urbildmengen einer Funktion $f : X \rightarrow Y$, $D_f = X$, gelten die Rechenregeln:

$$\begin{aligned} f^{-1}(A \cap B) &= f^{-1}(A) \cap f^{-1}(B), & f^{-1}(A \cup B) &= f^{-1}(A) \cup f^{-1}(B), \\ f^{-1}(A^c) &= f^{-1}(A)^c, & f^{-1}(A \setminus B) &= f^{-1}(A) \setminus f^{-1}(B). \end{aligned}$$

2.5.4 A Prüfen Sie nach, daß die Verkettung der funktionalen Relationen g und f wieder eine funktionale Relation ist, die mit der Verkettung der zu g und f gehörenden Funktionen übereinstimmt.

2.5.5 Die Umkehrfunktion

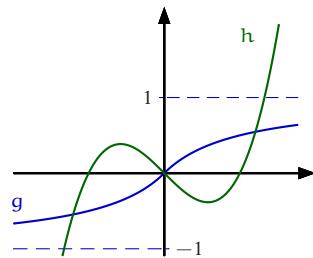
2.5.6 Definition Eine Abbildung $f : X \supseteq D_f \rightarrow Y$ von der Menge X in die Menge Y heißt injektiv, falls aus $f(x_1) = f(x_2)$ immer $x_1 = x_2$ folgt. Sie heißt surjektiv, falls das Bild im $f := \{y \in Y \mid \exists_{x \in X} y = f(x)\} =: f(X)$ die ganze Menge Y ist. f heißt bijektiv, falls f injektiv und surjektiv ist.

Wir haben mit dieser Definition die Eigenschaften injektiv, surjektiv und bijektiv einer funktionalen Relation in die gängige Sprache der Funktionen übertragen.

Die Injektivität einer Abbildung $f : X \rightarrow Y$ wird von Studenten gerne in der Form f ist injektiv, wenn jedem x aus X genau ein y aus Y zugeordnet wird wiedergegeben. Das hört sich überzeugend an, ist aber falsch. Diese Eigenschaft muß eine injektive Abbildung sicher haben, denn diese Eigenschaft muß jede Abbildung aufweisen – andernfalls wäre sie nämlich keine (sondern?). Die Injektivität verlangt aber darüber hinaus, daß jedes der Bilder $y \in f(X)$ von jeweils genau einem $x \in X$ stammt. Diese Eigenschaft wird z. B. von $f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto x^3$ erfüllt, nicht aber von $g : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto x^2$. Aus $f(x_1) = f(x_2)$, also $x_1^3 = x_2^3$ folgt nämlich $0 = x_1^3 - x_2^3 = (x_1 - x_2)(x_1^2 + x_1 x_2 + x_2^2)$. Das ist nur für $x_1^2 + x_1 x_2 + x_2^2 = 0$, oder für $x_1 = x_2$ möglich. Im ersten Fall läßt sich $x_1^2 + 2x_1 x_2 + x_2^2 = (x_1 + x_2)^2 = x_1 x_2 \geq 0$ folgern. Damit ist $x_1^2 + x_1 x_2 + x_2^2$ eine Summe aus nicht negativen Zahlen, die nur für $x_1 = x_2 = 0$ Null ergeben kann. In jedem Fall muß $x_1 = x_2$ gelten. g ist wegen $g(2) = 4 = g(-2)$ nicht injektiv.

Die Surjektivität einer Abbildung $f : X \rightarrow Y$ bedeutet, daß alle Elemente von Y Bilder sind: $\text{im } f = Y$. Die Funktion f von oben ist surjektiv, denn für jedes $y \in \mathbb{R}$ ist $f(\sqrt[3]{y}) = y$. Die Surjektivität läßt sich für eine einzelne Abbildung erzwingen, wenn sie nicht erfüllt sein sollte, indem man die Zielmenge Y gegebenenfalls verkleinert und durch $\text{im } f$ ersetzt.

Meistens ist das aber gar nicht wünschenswert. Für die Funktion $g : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto \frac{x}{1+|x|}$ würde es z. B. bedeuten, sie als surjektive



Funktion von \mathbb{R} nach $(-1, 1)$ anzusehen. Für mehrere Funktionen ist das meist sowieso nicht möglich, da sie verschiedene Bildmengen in Y haben können. Von den beiden Funktionen g und $h : \mathbb{R} \rightarrow \mathbb{R}$, $x \mapsto x^3 - x$ ist g injektiv, aber nicht surjektiv und h surjektiv, aber nicht injektiv.

Für eine injektive Funktion $f : X \supseteq D_f \rightarrow Y$ gibt es für jedes Bildelement $y \in \text{im } f$ genau ein $x \in D_f$, für das $y = f(x)$ gilt. Daher wird durch die Teilmenge $\{[y, x] \in Y \times X \mid [x, y] \in f\}$ von $Y \times X$ wieder eine funktionale Relation, also eine Funktion von Y nach X definiert. Für zwei Elemente $[y, x]$ und $[y, w]$ dieser Menge gilt nämlich $[x, y] \in f$ und $[w, y] \in f$, d. h. $y = f(x)$ und $y = f(w)$. Da f injektiv ist, muß $x = w$ folgen. Das zeigt die funktionale Eigenschaft der Relation. Die zugehörige Funktion wird *Umkehrfunktion* von f genannt und mit f^{-1} bezeichnet. Diese Schreibweise ist für jemanden, der die Potenzrechengesetze kennt, etwas irreführend. Eigentlich verstehen wir unter f^{-1} , gemäß (2.14), die Funktion $\frac{1}{f}$. Aber genau die ist hier *nicht* gemeint. Die Schreibweise röhrt daher, daß f^{-1} die Inverse von f bezüglich der Verkettung von Funktionen ist: $f^{-1} \circ f(x) = f^{-1}(f(x)) = x$ für alle $x \in D_f$. $f^{-1} \circ f$ ist demnach die identische Funktion $\text{id} : X \rightarrow X$, $x \mapsto x$, eingeschränkt auf D_f . Die Gleichung $f^{-1} \circ f = \text{id}$ erinnert an die Inversenbildung bei der Multiplikation gewöhnlicher Zahlen und ist der Grund für die Schreibweise f^{-1} . Trotz der Mehrdeutigkeit hat sich diese Notation für die Umkehrfunktion durchgesetzt, weil aus dem Zusammenhang normalerweise erkennbar ist, ob $\frac{1}{f}$, oder die Umkehrfunktion gemeint ist. An einem einfachen Beispiel wird das klar: Für $f(x) := x^3$ haben wir oben erklärt, daß es sich um eine injektive und surjektive, also eine bijektive Funktion von \mathbb{R} nach \mathbb{R} handelt. Um ihre Umkehrfunktion zu finden, müssen wir die Zuordnung $x \mapsto x^3 = y$ von f umkehren. Das bedeutet aber nichts weiter, als die Gleichung $y = x^3$ nach x aufzulösen. Wir erhalten natürlich $x = \sqrt[3]{y}$. Daher ist $f^{-1}(y) = \sqrt[3]{y}$ und $f^{-1} \circ f(x) = f^{-1}(x^3) = \sqrt[3]{x^3} = x = \text{id}(x)$, während $\frac{1}{f(x)} = \frac{1}{x^3}$ in der Verkettung mit f die Abbildung $x \mapsto \frac{1}{x^9}$ ergäbe. Die Existenz der n -ten Wurzeln überlegen wir uns in Beispiel 2.5.9.

Wie sehen der Definitionsbereich und das Bild von f^{-1} aus? Gemäß (2.6) und (2.7) haben wir

$$\begin{aligned} D_{f^{-1}} &= \{y \in Y \mid \exists_{x \in X} [y, x] \in f^{-1}\} = \{y \in Y \mid \exists_{x \in X} [x, y] \in f\} = \text{im } f, \\ \text{im } f^{-1} &= \{x \in X \mid \exists_{y \in Y} [y, x] \in f^{-1}\} = \{x \in X \mid \exists_{y \in Y} [x, y] \in f\} = D_f. \end{aligned}$$

Für die Umkehrfunktion von $f : X \supseteq D_f \rightarrow \text{im } f \subseteq Y$ gilt daher, wie erwartet, $f^{-1} : Y \supseteq \text{im } f \rightarrow D_f \subseteq X$, denn f^{-1} ordnet ja jedem Bildelement $y = f(x) \in \text{im } f$ sein Urbildelement $x \in D_f$ zu.

2.5.7 A $f : X \supseteq D_f \rightarrow \text{im } f \subseteq Y$ sei eine injektive Funktion, die nach unseren Überlegungen die Umkehrfunktion $f^{-1} : Y \supseteq \text{im } f \rightarrow D_f \subseteq X$ hat. Skizzieren Sie den Graph von f und von f^{-1} als Teilmenge von $X \times Y$. Was fällt Ihnen dabei auf?

2.5.8 A $f : X \supseteq D_f \rightarrow \text{im } f \subseteq Y$ sei eine injektive Funktion. Zeigen Sie, daß $f^{-1} \circ f = \text{id}_{|D_f}$ und $f \circ f^{-1} = \text{id}_{|\text{im } f}$ gilt.

2.5.9 Beispiel Die identische Funktion $\text{id} : \mathbb{R} \rightarrow \mathbb{R}$ ist offensichtlich bijektiv. Die Potenzfunktion $\text{id}^n : \mathbb{R} \rightarrow \mathbb{R}$, $x \mapsto \text{id}^n(x) = x^n$ ist nur für ungerade n ebenfalls bijektiv ($D_{\text{id}^n} := \mathbb{R}$). Für gerades n müssen wir den Definitionsbereich von \mathbb{R} auf $D_{\text{id}^n} := \mathbb{R}_0^+ := \mathbb{R}^+ \cup \{0\}$ einschränken, um die Bijektivität zu gewährleisten: $\text{id}^n : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$. Diese Behauptungen wollen

wir jetzt zeigen. Dafür benötigen wir eine Verallgemeinerung der dritten binomischen Formel $x^2 - y^2 = (x - y)(x + 1)$:

$$\begin{aligned} x^n - y^n &= (x - y)(x^{n-1} + x^{n-2}y + x^{n-3}y^2 + \cdots + x^2y^{n-3} + xy^{n-2} + y^{n-1}) \\ &= (x - y) \sum_{k=1}^n x^{n-k}y^k. \end{aligned} \quad (2.15)$$

Wir rechnen die Gleichung einfach nach:

$$\begin{aligned} (x - y)(x^{n-1} + x^{n-2}y + x^{n-3}y^2 + \cdots + x^2y^{n-3} + xy^{n-2} + y^{n-1}) \\ = x^n + x^{n-1}y + x^{n-2}y^2 + \cdots + x^3y^{n-3} + x^2y^{n-2} + xy^{n-1} \\ - x^{n-1}y - x^{n-2}y^2 - \cdots - x^3y^{n-3} - x^2y^{n-2} - xy^{n-1} - y^n = x^n - y^n. \end{aligned}$$

Die Idee für (2.15): Das Polynom $p(x) := x^n - y^n$ hat die Nullstelle $x = y$. Daher muß die Polynomdivision von $p(x)$ mit $x - y$ ohne Rest aufgehen (vergl. 11.8.11).

Wir können $n > 1$ annehmen, da für $n = 1$ nichts zu zeigen ist. Für die Funktion $x \mapsto id^n(x) = x^n$ müssen wir beweisen, daß aus $x \neq y \in D_{id^n}$ immer $id^n(x) \neq id^n(y)$, also $x^n \neq y^n$ folgt, oder, was dazu äquivalent ist: Aus $x^n = y^n$ muß $x = y$ folgen (vergl. (1.9)).

Aus $x^n = y^n$ ergibt sich also $x - y = 0$, oder $q := x^{n-1} + x^{n-2}y + \cdots + xy^{n-2} + y^{n-1} = 0$. Falls $x = y$ gelten sollte, sind wir fertig. Wir nehmen $x \neq y$ an. Bei geradem n gilt $x \geq 0$ und $y \geq 0$. Daher ist q eine Summe aus nicht negativen Zahlen, die nur dann 0 ergeben kann, wenn jeder einzelne Summand verschwindet. Insbesondere muß demnach $x^{n-1} = 0$ sein. Wäre $x \neq 0$, so könnten wir diese Gleichung auf beiden Seiten wiederholt durch x teilen, bis wir doch bei $x = 0$ angekommen wären. Also muß $x = 0$ gelten. Genauso können wir dann von $y^{n-1} = 0$ auf $y = 0$ schließen, also auf $x = y$, im Widerspruch zur Annahme. Für gerades n haben wir die Behauptung demnach gezeigt.

Jetzt sei n ungerade und $x \neq y$. Da $x^n = y^n$ gilt, können x und y nicht verschiedene Vorzeichen haben (bei geradem n war dieser Schluß nicht möglich, weshalb in den Voraussetzungen für diesen Fall eben $x, y \geq 0$ zu fordern war). Da $n - 1$ gerade ist, ist $x^{n-1} \geq 0$. $x^{n-2}y$ ist aber ebenfalls nicht negativ, denn, falls $x < 0$ sein sollte, gilt das auch für y , so daß $x^{n-2} < 0$ und $y < 0$ zu $x^{n-2}y > 0$ führt. Auf diese Weise können wir alle Summanden von q als nicht negativ nachweisen. Daher läßt sich, wie oben vorgeführt, auf $x = y = 0$ schließen, im Widerspruch zur Annahme.

Nun, da die Injektivität nachgewiesen ist, wissen wir nach Übung 2.5.8, daß es eine Umkehrfunktion $(id^n)^{-1} : im id^n \rightarrow D_{id^n}$ gibt, die durch Auflösen der Gleichung $x^n = y$ nach x bestimmt werden kann. Für gerades n ist $im id^n = \mathbb{R}_0^+$, für ungerades ist $im id^n = \mathbb{R}$. Das scheint offensichtlich zu sein, können wir aber im Moment noch nicht zeigen. Dafür fehlt uns noch der Stetigkeitsbegriff 11.1.1 und der Zwischenwertsatz 11.1.10. Die Umkehrfunktion ist die n -te Wurzel: $(id^n)^{-1}(y) =: \sqrt[n]{y}$. Näheres dazu in 11.1.12.

Wir könnten jetzt der Meinung sein, daß wir bei der Berechnung der Umkehrfunktion auf eine bekannte Funktion, nämlich die n -te Wurzel, gestoßen sind. Tatsächlich haben wir sie aber eben erst definiert, als Ergebnis der allgemeinen Überlegungen zur Existenz von Umkehrfunktionen.

2.5.10 Permutationen Eine bijektive Abbildung π der Menge $\mathbb{N}_n := \{1, \dots, n\}$ heißt *Permutation* der Zahlen $1, 2, \dots, n$. Wir schließen den trivialen Fall $n = 1$ aus und nehmen ab jetzt $n > 1$ an. Eine Funktion auf der endlichen Menge \mathbb{N}_n lässt sich vollständig durch eine Aufzählung ihrer Funktionswerte $[\pi(1), \pi(2), \dots, \pi(n)]$ wiedergeben. So ist durch $[2, 9, 3, 4, 1, 8, 6, 7, 5]$ die Permutation π mit den Funktionswerten $\pi(1) = 2, \pi(2) = 9, \pi(3) = 3, \dots, \pi(8) = 7$ und $\pi(9) = 5$ bestimmt. Die Inverse π^{-1} wird dann durch $[5, 1, 3, 4, 9, 7, 8, 6, 2]$ beschrieben. Wir identifizieren eine Permutation π mit der geordneten Aufzählung $[\pi(1), \pi(2), \dots, \pi(n)]$ ihrer Bilder. Für die angegebenen Beispiele schreiben wir also $\pi = [2, 9, 3, 4, 1, 8, 6, 7, 5]$ und $\pi^{-1} = [5, 1, 3, 4, 9, 7, 8, 6, 2]$. Wenn mit Permutationen gerechnet werden muß, ist diese Notation etwas knapp. Dann verwendet man gerne die ausführlichere Schreibweise $\pi = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 2 & 9 & 3 & 4 & 1 & 8 & 6 & 7 & 5 \end{bmatrix}$. So lässt sich etwa die Inverse leicht angeben, indem man dieses Schema von unten nach oben in aufsteigender Reihenfolge durchgeht: Man sucht die 1 und kann darüber ihr Urbild 5 ablesen. Damit weiß man $\pi^{-1}(1) = 5$, dann $\pi^{-1}(2) = 1$ usw. Genauso einfach lässt sich die Verkettung, also die Hintereinanderausführung zweier Permutationen bilden. Für $\sigma := \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 2 & 8 & 1 & 9 & 6 & 3 & 4 & 7 \end{bmatrix}$ ist $\sigma \circ \pi = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 2 & 7 & 8 & 1 & 5 & 4 & 6 & 3 & 9 \end{bmatrix}$. Aus diesen Darstellungen ist $1 \xrightarrow{\pi} 2 \xrightarrow{\sigma} 2$, oder $\sigma(\pi(1)) = \sigma(2) = 2$ leicht abzulesen, dann $2 \xrightarrow{\pi} 9 \xrightarrow{\sigma} 7$, also $\sigma(\pi(2)) = 7$ usw.

Die Menge \mathcal{S}_n aller Permutationen auf \mathbb{N}_n heißt *symmetrische Gruppe*. Mit der Hintereinanderausführung \circ als Verknüpfung zweier Permutationen hat sie nämlich die folgenden *Gruppeneigenschaften*: Für alle $\pi, \sigma, \tau \in \mathcal{S}_n$ gilt $\pi \circ \sigma \in \mathcal{S}_n, \pi \circ (\sigma \circ \tau) = (\pi \circ \sigma) \circ \tau$ (Assoziativität), die identische Abbildung id auf \mathbb{N}_n ist das neutrale Element der Verknüpfung, d. h., es gilt $\pi \circ \text{id} = \text{id} \circ \pi = \pi$ für alle $\pi \in \mathcal{S}_n$, und für jedes $\pi \in \mathcal{S}_n$ gibt es die Inverse $\pi^{-1} \in \mathcal{S}_n$, also eine Abbildung mit der Eigenschaft $\pi \circ \pi^{-1} = \pi^{-1} \circ \pi = \text{id}$.

Normalerweise lassen wir das Verknüpfungszeichen \circ weg, d. h., wir schreiben statt $\pi \circ \sigma$ einfach $\pi\sigma$. Auch sprechen wir lieber von dem Produkt $\pi\sigma$ der Permutationen π und σ , als von der Hintereinanderausführung.

Eine *Transposition* ist eine Permutation, bei der genau zwei Elemente vertauscht werden. Wird das Element i mit $j > i$ vertauscht, so bezeichnen wir die zugehörige Transposition mit t_{ij} . Für t_{ii+1} schreiben wir einfach t_i . Es ist eine *Transposition nächster Nachbarn*, die durch den kleineren der beiden vollständig festgelegt ist. Jede Transposition t_{ij} lässt sich durch eine ungerade Anzahl, nämlich $2(j - i) - 1$ Transpositionen nächster Nachbarn gewinnen. Denn wir brauchen $j - i$ solcher Transpositionen, um i mit $i + 1$, dann mit $i + 2$ usw. und schließlich i mit j zu vertauschen. Anschließend benötigen wir noch einmal $j - i - 1$, um j mit $j - 1$, dann mit $j - 2$ usw. und endlich mit $i + 1$ zu vertauschen:

$$[i, i+1, \dots, j-1, j] \rightarrow [i+1, i+2, \dots, j-1, j, i] \rightarrow [j, i+1, \dots, j-1, i].$$

Eine *Inversion* bei einer Permutation π ist ein Zahlenpaar $[\pi(i), \pi(i+k)]$, mit den Eigenschaften $k \geq 1$ und $\pi(i) > \pi(i+k)$. Das sind also Zahlen, die nach der Permutation nicht mehr in aufsteigender Reihenfolge vorliegen. Die Permutation $[3, 2, 1, 4, 5, 7, 9, 8, 6]$ etwa hat die Inversionen $[3, 2], [3, 1], [2, 1], [7, 6], [9, 8], [9, 6]$ und $[8, 6]$. Man erhält systematisch alle Inversionen, indem

man die Inversionen mit dem ersten Element $\pi(1)$ bildet, wenn es welche gibt, dann die mit $\pi(2)$ als erstem Element und auf diese Weise bis zu $\pi(n - 1)$ fortfährt.

Jede Transposition t_i erhöht oder vermindert die Anzahl der Inversionen einer Permutation um 1, denn aus $[\pi(1), \dots, \pi(i-1), \pi(i), \pi(i+1), \pi(i+2), \dots, \pi(n)]$ wird $[\pi(1), \dots, \pi(i-1), \pi(i+1), \pi(i), \pi(i+2), \dots, \pi(n)]$. Dabei ändert sich die Anzahl der Inversionen in $[\pi(1), \dots, \pi(i-1), \pi(i+2), \dots, \pi(n)]$ nicht. Auch Inversionen mit den ersten Elementen an Positionen zwischen 1 und $i-1$ und den zweiten bei i oder $i+1$ bleiben gleich, da sich ihre Größenbeziehungen nicht ändern. Dasselbe gilt für die erste Position bei i oder $i+1$ und die zweite zwischen $i+2$ und n . Einzig die Größenbeziehung zwischen $\pi(i)$ und $\pi(i+1)$ ändert sich. War sie ursprünglich $\pi(i) < \pi(i+1)$, dann kommt eine Inversion hinzu, andernfalls verschwindet eine.

2.5.11 Satz *Jede Permutation $\pi \in S_n$ lässt sich als Produkt von Transpositionen darstellen. Diese Darstellung ist nicht eindeutig, aber sie besteht entweder immer aus einer geraden, oder immer aus einer ungeraden Anzahl von Transpositionen.*

Beweis. Auf eine von id verschiedene Permutation $[\pi(1), \pi(2), \dots, \pi(n)]$ wenden wir die Transposition $\tau_1 := t_{\pi(k)k}$ an. Dabei ist k die größte Zahl mit $\pi(k) \neq k$, also die letzte Position, die nicht mit ihrem Eintrag übereinstimmt. Daher gilt $\pi(k) < k$, denn aus $\pi(k) = k + r$ mit einem $r > 0$ würde $k + r \neq \pi(k + r)$ folgen, im Widerspruch zur Maximalität von k . Als Ergebnis erhalten wir eine Permutation π_1 , die ab k die identische Permutation ist: $\pi_1(\ell) = \ell$ für $k \leq \ell \leq n$. Diesen Vorgang wiederholen wir für π_1 und erhalten eine Permutation π_2 , die jetzt ab einer Position $k_1 < k$ mit der identischen übereinstimmt. Nach einer endlichen Anzahl von Wiederholungen haben alle Zahlen ihre Ausgangsposition eingenommen. Wir sind daher bei der identischen Permutation angekommen: $\text{id} = \tau_\ell \tau_{\ell-1} \cdots \tau_1 \pi$. Wegen $\tau_i \tau_i = \text{id}$ für $i = 1, \dots, \ell$, erhalten wir aus dieser Gleichung durch Multiplikation mit τ_ℓ , dann mit $\tau_{\ell-1}$ usw.: $\pi = \tau_1 \cdots \tau_{\ell-1} \tau_\ell$. Damit ist $\pi \neq \text{id}$ ein Produkt aus Transpositionen. Für id lässt sich das natürlich ebenfalls erreichen, etwa durch $\text{id} = t_1 t_1$.

Wir nehmen an, π sei eine Permutation, für die es eine Darstellung $\pi = \tau_1 \cdots \tau_{2k}$ mit einer geraden Anzahl von Transpositionen und eine Darstellung $\pi = \sigma_1 \cdots \sigma_{2\ell+1}$ mit einer ungeraden Anzahl gibt. Das hieße, daß die identische Permutation ein Produkt aus einer ungeraden Anzahl von Transpositionen ist: $\text{id} = \sigma_{2\ell+1} \cdots \sigma_1 \tau_1 \cdots \tau_{2k}$. id wäre dann sogar das Produkt einer ungeraden Zahl von Transpositionen nächster Nachbarn, denn wir wissen, daß das für jede der beteiligten Transpositionen σ_i und τ_i gilt. Natürlich ist die Anzahl der Inversionen von id gleich Null. Andererseits erzeugt oder vernichtet jede Transposition nächster Nachbarn eine Inversion. Zählen wir also die durch die Transpositionen erzeugten Inversionen, so erhalten wir eine Summe mit einer ungeraden Anzahl Summanden, die nur 1 oder -1 sind. Als Ergebnis müßte sie 0 ergeben, was nicht möglich ist. \square

Es ist ganz instruktiv, das Verfahren einmal an einem konkreten Beispiel durchzuführen. Wir wählen $\pi := \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 6 & 3 & 7 & 2 & 5 & 1 & 8 \end{bmatrix}$ und starten mit $k = 7$ und $\pi(7) = 1$. Das führt auf $\pi_1 := t_{1,7}\pi = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 6 & 3 & 1 & 2 & 5 & 7 & 8 \end{bmatrix}$. Jetzt sind die Positionen ab $k = 7$ wieder in ihrer ursprünglichen Reihenfolge. Nun ist $k = 6$ und $\pi_1(6) = 5$.

Daher ist $\pi_2 := t_{5,6}t_{1,7}\pi = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 5 & 3 & 1 & 2 & 6 & 7 & 8 \end{bmatrix}$. Als nächste Permutation erhalten wir $\pi_3 := t_{2,5}t_{5,6}t_{1,7}\pi = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 4 & 2 & 3 & 1 & 5 & 6 & 7 & 8 \end{bmatrix}$ und schließlich $t_{1,4}t_{2,5}t_{5,6}t_{1,7}\pi = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{bmatrix} = \text{id.}$ Damit ist π eine gerade Permutation, die durch $\pi = t_{1,7}t_{5,6}t_{2,5}t_{1,4}$ aus Transpositionen aufgebaut wird.

2.5.12 Definition Eine Permutation $\pi \in S_n$ heißt gerade, wenn sie das Produkt einer geraden Zahl von Transpositionen ist, sonst ungerade. Das Vorzeichen $\text{sgn}(\pi)$ einer Permutation ist 1, falls π gerade und -1 , falls π ungerade ist.

Satz 2.5.11 zeigt, daß $\text{sgn}(\pi)$ wohldefiniert ist. Darüber hinaus ist klar, daß $\text{sgn}(\pi) = (-1)^{t(\pi)}$ gilt, wenn $t(\pi)$ die Anzahl der Transpositionen einer Darstellung von π als Produkt von Transpositionen ist. Diese ist zwar nicht eindeutig, aber zwei verschiedene Versionen müssen sich immer um eine gerade Anzahl von Transpositionen unterscheiden.

Die zentrale Eigenschaft

$$\text{sgn}(\pi\sigma) = \text{sgn}(\pi) \text{sgn}(\sigma) \quad (2.16)$$

läßt sich jetzt leicht einsehen. Sind π und σ gerade, so ist auch $\pi\sigma$ gerade, denn die gerade Anzahl von Transpositionen für π und die gerade Anzahl für σ ergibt hintereinander ausgeführt eine gerade Anzahl für $\pi\sigma$. Hier stimmt (2.16) offensichtlich: $\text{sgn}(\pi) \text{sgn}(\sigma) = 1 \cdot 1 = \text{sgn}(\pi\sigma)$. Ist eine der Permutationen π oder σ gerade, die andere aber ungerade, so läßt sich $\pi\sigma$ durch eine ungerade Anzahl von Transpositionen darstellen. Es folgt $\text{sgn}(\pi) \text{sgn}(\sigma) = 1 \cdot (-1) = \text{sgn}(\pi\sigma)$. Schließlich ist $\pi\sigma$ gerade, wenn π und σ ungerade sind, so daß (2.16) auch in diesem Fall stimmt.

$\text{sgn} : S_n \rightarrow \{-1, 1\}$ ist ein sogenannter *Gruppenhomomorphismus* zwischen den beiden Gruppen S_n und der multiplikativen Gruppe $\{-1, 1\}$ (mit den gewohnten Rechenregeln der Multiplikation, aus denen natürlich $1^{-1} = 1$ und $(-1)^{-1} = -1$ folgt). Die Homomorphismus-Eigenschaft wird durch (2.16) wiedergegeben. Er hat die Eigenschaft $\text{sgn}(t_i) = -1$, $i = 1, \dots, n-1$, durch die er bereits eindeutig festgelegt ist. Denn jeder Homomorphismus λ mit dieser Eigenschaft erfüllt $\lambda(\pi) = \lambda(t_{i_1}t_{i_2} \cdots t_{i_k}) = \lambda(t_{i_1})\lambda(t_{i_2}) \cdots \lambda(t_{i_k}) = (-1)^k = (-1)^{t(\pi)} = \text{sgn}(\pi)$, für jede Zerlegung von π in Transpositionen t_{i_ℓ} nächster Nachbarn.

Es gibt eine Formel für $\text{sgn}(\pi)$. Dafür definieren wir

$$\begin{aligned} \Delta &:= \prod_{k=1}^{n-1} \prod_{\ell>k} (\ell - k) \\ &= (2-1)(3-1)(4-1)(5-1)(6-1) \cdots (n-1) \cdot \\ &\quad (3-2)(4-2)(5-2)(6-1) \cdots (n-2) \cdot \\ &\quad (4-3)(5-3)(6-3) \cdots (n-3) \cdot \\ &\quad \cdots \\ &\quad (n-(n-1)). \end{aligned}$$

In diesem Produkt bestehen die Faktoren aus allen möglichen positiven Differenzen der Zahlen aus \mathbb{N}_n in ihrer natürlichen Reihenfolge. Jetzt bilden wir

$$\Delta(\pi) := \prod_{k=1}^{n-1} \prod_{\ell>k} (\pi(\ell) - \pi(k)).$$

Auch hier kommen alle möglichen Differenzen vor, nur sind die Faktoren in einer anderen Reihenfolge. Außerdem sind nicht mehr alle Differenzen positiv. $\pi(\ell) - \pi(k)$ ist genau dann negativ, wenn $[\pi(\ell), \pi(k)]$ eine Inversion von π ist. Die Anzahl der negativen Faktoren in $\Delta(\pi)$ ist demnach die Anzahl $i(\pi)$ der Inversionen von π . Daher ist

$$\frac{\Delta(\pi)}{\Delta} = (-1)^{i(\pi)}.$$

Wir nennen diesen Ausdruck vorläufig $\lambda(\pi)$. Offensichtlich ist für jede Transposition t_i nächster Nachbarn $\lambda(t_i) = -1$, denn sie hat genau eine Inversion. Wenn wir jetzt noch die Homomorphismus-Eigenschaft $\lambda(\pi\sigma) = \lambda(\pi)\lambda(\sigma)$ zeigen können, wissen wir $\lambda = \text{sgn}$.

$$\Delta(\pi\sigma) = \prod_{k=1}^{n-1} \prod_{\ell>k} (\pi\sigma(\ell) - \pi\sigma(k)) = \frac{\prod_{k=1}^{n-1} \prod_{\ell>k} (\pi(\sigma(\ell)) - \pi(\sigma(k)))}{\prod_{k=1}^{n-1} \prod_{\ell>k} (\sigma(\ell) - \sigma(k))} \cdot \Delta(\sigma)$$

Der Nenner des Bruchs ist nicht Δ . Da die Differenzen $\sigma(\ell) - \sigma(k)$ im Zähler und im Nenner aber in genau derselben Form auftreten, kann man den Nenner durch Umkehrung negativer Faktoren in Δ verwandeln. Dabei ändert sich der Bruch nicht, wenn man das mit dem Argumenten von π im Zähler genauso macht:

$$\Delta(\pi\sigma) = \frac{\prod_{k=1}^{n-1} \prod_{\ell>k} (\pi(\ell) - \pi(k))}{\prod_{k=1}^{n-1} \prod_{\ell>k} (\ell - k)} \cdot \Delta(\sigma) = \frac{\Delta(\pi)}{\Delta} \cdot \Delta(\sigma) = \lambda(\pi) \cdot \Delta(\sigma).$$

Teilt man beide Seiten durch Δ , so gelangt man zu $\lambda(\pi\sigma) = \lambda(\pi)\lambda(\sigma)$. Damit haben wir

$$\text{sgn}(\pi) = \frac{\Delta(\pi)}{\Delta} = (-1)^{i(\pi)} \tag{2.17}$$

abgeleitet. Das Vorzeichen einer Permutation ist also auch durch die Anzahl der Inversionen bestimmt. Insbesondere hat eine gerade bzw. ungerade Permutation auch immer eine gerade bzw. ungerade Anzahl von Inversionen.

3 Zahlentheorie

3.1 Teilbarkeitstheorie ganzer Zahlen

3.1.1 Definition Eine Zahl $0 \neq a \in \mathbb{Z}$ heißt Teiler einer Zahl $b \in \mathbb{Z}$, wenn es eine Zahl $c \in \mathbb{Z}$ mit der Eigenschaft

$$b = c \cdot a$$

gibt. Wir schreiben dafür $a | b$.

Für $b = 0$ gibt es natürlich unendlich viele Teiler, denn für $c = 0$ und alle $a \in \mathbb{Z}$ gilt $0 = 0 \cdot a$. Für $b \neq 0$ gibt es bei gegebenem Teiler a nur ein Element $c \in \mathbb{Z}$, mit $b = ca$, denn für ein weiteres $c' \in \mathbb{Z}$ mit $b = c'a$ würde $ca - c'a = (c - c')a = 0$ folgen. Da $b \neq 0$ ist, muß auch $a \neq 0$ gelten, so daß nur $c = c'$ bleibt.

3.1.2 Satz (Teilen mit Rest) Für jede ganze Zahl p und jede natürliche Zahl q gibt es eindeutig bestimmte ganze Zahlen $r, t \in \mathbb{Z}$ mit der Eigenschaft

$$p = tq + r, \quad 0 \leq r < q. \quad (3.1)$$

Ein $r > 0$ heißt Rest der Teilung von p durch q .

Beweis. Wir behandeln zunächst die Eindeutigkeit. Aus $0 \leq r' < q$ und

$$p = tq + r = t'q + r'$$

folgt, wenn wir o. B. d. A. $r' \geq r$ annehmen, $(t - t')q = r' - r \geq 0$. Damit ist $t - t' \geq 0$. Wäre $t - t' > 0$, so würde $(t - t')q = r' - r \geq q$ folgen, im Widerspruch dazu, daß sowohl r' als auch r echt kleiner als q sind, was für die Differenz $r' - r$ erst recht zutrifft. Also muß $t - t' = 0$ und damit auch $r' - r = 0$ gelten.

Zur Existenz: Dazu betrachten wir die Menge

$$M := \{ \ell \in \mathbb{Z} \mid \ell q \leq p \}.$$

In M muß es ein größtes Element ℓ_0 geben, denn andernfalls würde ℓq über alle Grenzen wachsen und dabei schließlich größer als p werden. Wir setzen $r := p - \ell_0 q (\geq 0)$ und haben damit $p = \ell_0 q + r$ mit einem $r \geq 0$ erreicht. Es muß noch $r < q$ nachgewiesen werden, um die Forderungen des Satzes zu erfüllen. Wir nehmen das Gegenteil $r \geq q$ an, d. h. $r = q + d$, $d \geq 0$. Das bedeutet

$$d = p - (\ell_0 + 1)q \geq 0 \quad \text{oder} \quad (\ell_0 + 1)q \leq p.$$

Also ist auch $\ell_0 + 1 \in M$, im Widerspruch dazu, daß ℓ_0 das größte Element von M ist. $r \geq q$ ist demnach nicht möglich, es muß $r < q$ gelten. $t := \ell_0$ und r sind die Zahlen, von denen im Satz die Rede ist. \square

Beispiel: Für $p = 82$ und $q = 7$, bzw. $p = -82$ und $q = 7$, bzw. $p = 82$ und $q = 83$ gilt:

$$82 = 11 \cdot 7 + 5, \quad -82 = -12 \cdot 7 + 2, \quad 82 = 0 \cdot 83 + 82.$$

3.1.3 Definition Der größte gemeinsame Teiler ganzer Zahlen $p_1 \neq 0, p_2 \neq 0, \dots, p_n \neq 0$ ist eine natürliche Zahl t mit den Eigenschaften

$$t | p_1 \wedge t | p_2 \wedge \dots \wedge t | p_n, \quad (3.2)$$

$$\forall s \in \mathbb{N} \ s | p_1 \wedge s | p_2 \wedge \dots \wedge s | p_n \Rightarrow s \leq t. \quad (3.3)$$

Für den größten gemeinsamen Teiler von p_1, \dots, p_n schreiben wir $\text{ggT}(p_1, \dots, p_n)$. Zahlen p_1, \dots, p_n mit größtem gemeinsamen Teiler 1 heißen teilerfremd.

Den größten gemeinsamen Teiler von 84 und 18 findet man durch wiederholtes Teilen mit Rest:

$$84 = 4 \cdot 18 + 12$$

$$18 = 1 \cdot 12 + 6$$

$$12 = 2 \cdot 6.$$

$$\text{ggT}(84, 18) = \text{ggT}(84, -18) = \text{ggT}(-84, 18) = \text{ggT}(-84, -18) = 6.$$

84													
18			18			18			18			12	
6	12	6	12	6	12	6	12	6	12	6	12	6	12
6	6	6	6	6	6	6	6	6	6	6	6	6	6

3.1.4 Satz (Euklidischer Algorithmus) Für zwei natürliche Zahlen p_0 und p_1 , $p_0 \geq p_1$, ergibt sich der größte gemeinsame Teiler durch wiederholtes Teilen mit Rest:

$$p_0 = t_1 \cdot p_1 + p_2, \quad 0 \leq p_2 < p_1$$

$$p_1 = t_2 \cdot p_2 + p_3, \quad 0 \leq p_3 < p_2$$

$$p_2 = t_3 \cdot p_3 + p_4, \quad 0 \leq p_4 < p_3$$

⋮

$$p_{n-2} = t_{n-1} \cdot p_{n-1} + p_n, \quad 0 \leq p_n < p_{n-1}$$

$$p_{n-1} = t_n \cdot p_n.$$

p_n ist der größte gemeinsame Teiler von p_0 und p_1 .

Beweis. Da die Reste p_2, p_3, \dots in jedem Schritt kleiner werden, muß das Verfahren schließlich enden. Die letzte Zeile besagt, daß p_n jedenfalls ein Teiler von p_{n-1} ist. Aus der vorletzten folgt dann, daß p_n auch p_{n-2} teilt. Auf diese Weise fortlaufend gelangt man zur zweiten Zeile und folgert, daß p_n auch p_1 teilt. $p_n \mid p_0$ folgt schließlich aus der ersten Zeile. Damit ist p_n ein gemeinsamer Teiler von p_0 und p_1 .

Sei $0 < q$ ein weiterer Teiler von p_0 und p_1 . Aus der ersten Zeile ergibt sich dann sofort, daß q den Rest p_2 teilt, aus der zweiten, daß auch p_3 von q geteilt wird usw., bis man bei der letzten Zeile angelangt ist und $q \mid p_n$ folgern kann. Das bedeutet insbesondere $q \leq p_n$. Damit ist p_n der größte gemeinsame Teiler von p_0 und p_1 . \square

3.1.5 Lemma (EUKLID) Für den größten gemeinsamen Teiler (p, q) zweier natürlicher Zahlen p und q gibt es immer eine Darstellung der Form

$$\text{ggT}(p, q) = x p + y q, \quad x, y \in \mathbb{Z}. \quad (3.4)$$

Es gilt

$$\text{ggT}(p, q) = \min \{ x p + y q \mid x, y \in \mathbb{Z} \wedge x p + y q \geq 1 \}. \quad (3.5)$$

Die Darstellung (3.4) ist keineswegs eindeutig.

Beweis. Wir zeigen (3.5). Dafür sei $r := \min \{ x p + y q \mid x, y \in \mathbb{Z}, x p + y q \geq 1 \} = x_0 p + y_0 q$ für geeignete $x_0, y_0 \in \mathbb{Z}$ und $r_0 := \text{ggT}(p, q)$. Dann ist $r_0 = r$ zu zeigen.

Als gemeinsamer Teiler von p und q teilt r_0 natürlich auch $r = x_0 p + y_0 q$. Das bedeutet $r_0 \leq r$.

Nun muß die umgekehrte Abschätzung $r \leq r_0$ gezeigt werden. Diese ist sicher erfüllt, wenn r ein gemeinsamer Teiler von p und q ist (denn r_0 ist der größte dieser Teiler). Wir gehen vom Gegenteil aus und nehmen o. B. d. A. an, daß p nicht von r geteilt wird. Nach Satz 3.1.2 gibt es eine Zahl $0 < t < r$ mit der Eigenschaft

$$p = \ell r + t,$$

für ein geeignetes $\ell \in \mathbb{Z}$. Setzen wir darin $r = x_0 p + y_0 q$,

$$p = \ell(x_0 p + y_0 q) + t$$

und lösen nach t auf:

$$0 < t = (1 - \ell x_0)p - \ell y_0 q < r.$$

Mit t haben wir ein Element der Menge $\{ x p + y q \mid x, y \in \mathbb{Z}, x p + y q > 0 \}$ gefunden, das kleiner als das Minimum r dieser Menge ist. Dieser offensichtliche Widerspruch zeigt, daß unsere Annahme falsch sein muß. Also ist r ein gemeinsamer Teiler von p und q und insbesondere $\leq r_0$. Das zeigt $r = r_0$ und damit die behauptete Darstellung $r = \text{ggT}(p, q)$.

Daß die Darstellung $\text{ggT}(p, q) = x_0 p + y_0 q$ nicht eindeutig ist, sieht man am einfachsten an einem Zahlenbeispiel:

$$\text{ggT}(5, 7) = 1 = 3 \cdot 7 - 4 \cdot 5 = -7 \cdot 7 + 10 \cdot 5.$$

\square

Übrigens erhält man alle Darstellungen systematisch auf folgende Weise:

$$1 = 3 \cdot 7 - 4 \cdot 5 + t \cdot 5 \cdot 7 - t \cdot 7 \cdot 5 = (3 + 5t) \cdot 7 - (4 + 7t) \cdot 5, \quad t \in \mathbb{Z}.$$

Lemma 3.1.5 gibt keine Vorschrift an, wie die Darstellung (3.5) zu finden ist. Wir stellen das Verfahren dafür an zwei ausführlichen Beispielen vor.

3.1.6 Beispiel $\text{ggT}(1314, 396) = 18$:

$$\begin{array}{lll} 1314 = 3 \cdot 396 + 126 & \Rightarrow & 126 = 1314 - 3 \cdot 396 \\ 396 = 3 \cdot 126 + 18 & \Rightarrow & 18 = 396 - 3 \cdot 126 \\ 126 = 7 \cdot 18. & & \end{array}$$

Die Gleichungen der letzten Spalte werden nun von unten nach oben ineinander eingesetzt und zusammengefaßt:

$$\begin{aligned} 18 &= 396 - 3 \cdot (1314 - 3 \cdot 396) = 396 - 3 \cdot 1314 + 9 \cdot 396 \\ &= 10 \cdot 396 - 3 \cdot 1314. \end{aligned}$$

3.1.7 Beispiel $\text{ggT}(5610, 637) = 1$:

$$\begin{array}{lll} 5610 = 8 \cdot 637 + 514 & \Rightarrow & 514 = 5610 - 8 \cdot 637 \\ 637 = 1 \cdot 514 + 123 & \Rightarrow & 123 = 637 - 514 \\ 514 = 4 \cdot 123 + 22 & \Rightarrow & 22 = 514 - 4 \cdot 123 \\ 123 = 5 \cdot 22 + 13 & \Rightarrow & 13 = 123 - 5 \cdot 22 \\ 22 = 1 \cdot 13 + 9 & \Rightarrow & 9 = 22 - 13 \\ 13 = 1 \cdot 9 + 4 & \Rightarrow & 4 = 13 - 9 \\ 9 = 2 \cdot 4 + 1 & \Rightarrow & 1 = 9 - 2 \cdot 4 \\ 4 = 4 \cdot 1. & & \end{array}$$

Wir setzen die Gleichungen der letzten Spalte, mit der letzten beginnend, nacheinander und nach oben fortschreitend ineinander ein. Dabei werden die Zahlen 4, 9, 13, ..., 514, 637 und 5610 wie Variablen behandelt, die sukzessive durch die nachfolgenden ersetzt werden, bis zuletzt nur noch die letzten beiden, nämlich 637 und 5610, vorhanden sind.

$$\begin{aligned} 1 &= 9 - 2 \cdot (13 - 9) & = 3 \cdot 9 - 2 \cdot 13 \\ &= 3 \cdot (22 - 13) - 2 \cdot 13 & = 3 \cdot 22 - 5 \cdot 13 \\ &= 3 \cdot 22 - 5 \cdot (123 - 5 \cdot 22) & = 28 \cdot 22 - 5 \cdot 123 \\ &= 28 \cdot (514 - 4 \cdot 123) - 5 \cdot 123 & = 28 \cdot 514 - 117 \cdot 123 \\ &= 28 \cdot 514 - 117 \cdot (637 - 514) & = 145 \cdot 514 - 117 \cdot 637 \\ &= 145 \cdot (5610 - 8 \cdot 637) - 117 \cdot 637 & = 145 \cdot 5610 - 1277 \cdot 637. \end{aligned}$$

Die gesuchte Darstellung nach EUKLID ist demnach

$$1 = 145 \cdot 5610 - 1277 \cdot 637.$$

3.1.8 Definition Zusammen mit dem euklidischen Algorithmus bezeichnet man das in den Beispielen vorgestellte Verfahren als erweiterten euklidischen Algorithmus.

3.1.9 Korollar Aus $p \mid ab$ und $\text{ggT}(p, a) = 1$ folgt $p \mid b$. Andererseits ergibt sich aus $a \mid p, b \mid p$ und $\text{ggT}(a, b) = 1$ auch $ab \mid p$.

Beweis. $\text{ggT}(p, a) = 1$ bedeutet nach Lemma 3.1.5 $1 = xp + ya$ für geeignete $x, y \in \mathbb{Z}$. Das multiplizieren wir mit b :

$$b = bxp + yab.$$

Da ab durch p teilbar ist und bxp offensichtlich auch, ist b durch p teilbar.

Die zweite Behauptung: $a \mid p$ bedeutet $p = ta$, mit einem geeigneten $t \in \mathbb{Z}$. Wir haben daher $b \mid ta$ und $\text{ggT}(b, a) = 1$, woraus $b \mid t$ und schließlich $ab \mid p$ folgt. \square

3.1.10 Korollar Für eine Primzahl p folgt aus $p \mid ab$:

$$p \mid a \text{ oder } p \mid b.$$

Beweis. Da p eine Primzahl ist, kann ein gemeinsamer Teiler von p und a nur 1 oder p sein. Es gibt daher die folgenden beiden Fälle:

1. $\text{ggT}(p, a) = p$ bedeutet, daß p ein Teiler von a ist.
2. $\text{ggT}(p, a) = 1$ bedeutet nach Korollar 3.1.9 $p \mid b$. \square

3.1.11 Korollar Für eine Primzahl p ist \sqrt{p} niemals rational.

Beweis. Wir nehmen das Gegenteil $\sqrt{p} \in \mathbb{Q}$ an, d. h. $\sqrt{p} = \frac{a}{b}$ mit geeigneten Zahlen $a, b \in \mathbb{N}$. O. B. d. A. können wir davon ausgehen, daß der Bruch $\frac{a}{b}$ vollständig gekürzt ist, d. h., daß $\text{ggT}(a, b) = 1$ gilt. Wir quadrieren $\sqrt{p} = \frac{a}{b}$ und stellen um:

$$pb^2 = a^2.$$

Das bedeutet $p \mid a^2$, nach Korollar 3.1.10 also $p \mid a$. Daher gibt es ein $t \in \mathbb{N}$, so daß $a = tp$ gilt. Eingesetzt in obige Gleichung:

$$pb^2 = t^2p^2 \Rightarrow b^2 = t^2p.$$

Wie gerade eben schließen wir daraus $p \mid b$. Damit ist p ein gemeinsamer Teiler von a und b , im Widerspruch dazu, daß $\text{ggT}(a, b) = 1$ gelten sollte. \square

3.2 Rechnen Modulo p

3.2.1 Definition Für eine Zahl $p \in \mathbb{N}$ führen wir eine neue Gleichheit zweier ganzer Zahlen x und y ein. Wir sagen

$$x =_p y \Leftrightarrow p \mid x - y. \quad (3.6)$$

Dafür ist die Schreibweise

$$x = y \pmod{p}, \quad (3.7)$$

der allgemeine Standard. Gesprochen: x gleich y modulo p .

Wir werden beide Notationen verwenden.

Zwei Zahlen x und y sind genau dann gleich modulo p , wenn sie sich um ein Vielfaches von p unterscheiden:

$$x =_p y \Leftrightarrow \exists t \in \mathbb{Z} : x = y + tp.$$

3.2.2 Beispiel $5 =_2 7$, oder $5 = 7 \pmod{2}$. $-3 =_{11} 8$, oder $-3 = 8 \pmod{11}$.

Diese Gleichheit modulo p reduziert die Anzahl der verschiedenen Elemente in \mathbb{Z} drastisch, denn es werden unendlich viele ganze Zahlen als gleich angesehen (modulo p):

$$\begin{array}{ccccccc} 0 & =_p & \pm p & =_p & \pm 2p & =_p & \pm 3p \dots \\ 1 & =_p & 1 \pm p & =_p & 1 \pm 2p & =_p & 1 \pm 3p \dots \\ \vdots & & \vdots & & \vdots & & \vdots \\ p-1 & =_p & p-1 \pm p & =_p & p-1 \pm 2p & =_p & p-1 \pm 3p \dots \end{array}$$

d. h., es gibt genau p dieser Mengen ganzer Zahlen, die modulo p nicht unterschieden werden. Offensichtlich ist jede dieser Mengen durch eine der Zahlen x zwischen 0 und $p-1$ bereits eindeutig festgelegt. Man schreibt für sie

$$[x] := \{ y \in \mathbb{Z} \mid x = y \pmod{p} \}.$$

Man kann sich leicht davon überzeugen, daß es sich bei $=_p$ um eine Äquivalenzrelation handelt und daß $[x]$ die zugehörigen Äquivalenzklassen sind. Wir werden diese Abstraktion nicht benötigen, da für unsere Zwecke die Identifikation modulo p als operativer Vorgang völlig ausreicht (wer darüber mehr wissen will, findet etwas in Abschnitt 2.3). Das bedeutet, daß wir uns vollständig innerhalb der Menge

$$\mathbb{Z}_p = \{0, 1, 2, \dots, p-1\} \quad (3.8)$$

bewegen können, wenn wir Berechnungen modulo p anstellen, auch wenn das nicht sehr praktisch ist. Tatsächlich erlauben wir es, daß in einzelnen Rechenschritten Ergebnisse durchaus außerhalb von \mathbb{Z}_p liegen dürfen, wenn wir nur das Endergebnis durch wiederholtes Teilen durch p mit Rest wieder in \mathbb{Z}_p darstellen.

Wir werden nach Satz 3.2.5 erkennen, daß sich die Multiplikation in

$$\mathbb{Z}_p^* := \{ k \in \mathbb{N} \mid k < p \wedge \text{ggT}(k, p) = 1 \} \cup \{0\} \quad (3.9)$$

besonders gut verhält, weil es zu jedem Element $0 \neq a \in \mathbb{Z}_p^*$ genau eine Element $0 \neq b \in \mathbb{Z}_p^*$ mit der Eigenschaft $ab =_p 1$ gibt. Für $p \in \mathbb{P}$ stimmt \mathbb{Z}_p mit \mathbb{Z}_p^* überein, für $p \notin \mathbb{P}$ aber nicht: $\mathbb{Z}_7^* = \{0, 1, 2, 3, 4, 5, 6\} = \mathbb{Z}_7$, aber $\mathbb{Z}_{12}^* = \{0, 1, 5, 7, 11\} \neq \mathbb{Z}_{12}$.

Bevor wir uns mit Berechnungen modulo p beschäftigen, müssen wir uns versichern, daß sie mit der neuen Gleichheit modulo p verträglich sind. Kurz gesagt vertragen sich, bis auf die Division, die ja schon in \mathbb{Z} nur eingeschränkt zur Verfügung steht, alle Grundrechenarten mit der Gleichheit modulo p. Um das einzusehen, müssen wir uns davon überzeugen, daß für modulo p gleiche Zahlen x_1 und x_2 bzw. y_1 und y_2 , das Ergebnis einer Rechnung nicht davon abhängt, ob wir sie mit x_1 und y_1 , oder mit x_2 und y_2 durchführen – vorausgesetzt, wir vergleichen diese Ergebnisse modulo p:

Aus $x_1 =_p x_2$ bzw. $y_1 =_p y_2$ folgt, daß es ganze Zahlen s und t mit $x_2 = x_1 + sp$ und $y_2 = y_1 + tp$ gilt. Dann haben wir:

$$\begin{aligned} x_2 \pm y_2 &= x_1 \pm y_1 + (s \pm t)p =_p x_1 \pm y_1, \\ x_2 y_2 &= (x_1 + sp)(y_1 + tp) = x_1 y_1 + (x_1 t + sy_2 + stp)p =_p x_1 y_1. \end{aligned}$$

Aus der letzten Gleichung ergibt sich auch $x_1^n =_p x_2^n$ für jedes $n \in \mathbb{N}$.

Als praktische Konsequenz erhalten wir: In jedem zulässigen Rechenschritt \pm und \cdot kann jedes Zwischenergebnis durch ein modulo p gleiches ersetzt werden. Das ermöglicht es, mit sehr großen Zahlen Berechnungen anzustellen, ohne tatsächlich sehr große Zahlen zu benutzen. Ein beliebtes Beispiel in diesem Zusammenhang ist folgende kleine Aufgabe:

3.2.3 Beispiel Wie lauten die letzten beiden Ziffern der Zahl 7^{1000} ?

Offensichtlich ist es nicht so ohne Weiteres möglich, diese Zahl einfach auszurechnen, um die letzten beiden Ziffern abzulesen. Erfreulicherweise ist das auch gar nicht nötig, denn die letzten beiden Ziffern von 7^{1000} können durch $7^{1000} \bmod 100$ einfach berechnet werden. Das bedeutet, daß wir nur mit Zahlen in der Größenordnung von 1000 arbeiten müssen:

$$7^{1000} = 7^{2 \cdot 500} = (7^2)^{500} = 49^{500} = (49^2)^{250} = 2401^{250} =_{100} 1^{250} = 1.$$

Also sind die letzten beiden Ziffern von 7^{1000} die 0 und die 1.

3.2.4 Beispiel (3-er, 9-er und 11-er Probe) Es gibt einen einfachen Test, mit dem eine ganze Zahl auf ihre Teilbarkeit durch 3 bzw. durch 9 geprüft werden kann: Eine Zahl ist genau dann durch 3 bzw. 9 teilbar, wenn ihre Quersumme, also die Summe ihrer Ziffern, durch 3 bzw. 9 teilbar ist. Das ist die sogenannte 3-er Probe bzw. 9-er Probe. Das liegt daran, daß $10 =_3 1$, $100 = 10^2 =_3 1^2 = 1, \dots, 10^n =_3 1^n = 1$, bzw. $10^n =_9 1$ gilt. Eine Zahl $a := a_n 10^n + a_{n-1} 10^{n-1} + \dots + a_2 10^2 + a_1 10 + a_0$ ist genau dann durch 3 teilbar, wenn $a =_3 0$ erfüllt ist. Da sich die Gleichheit modulo 3 mit der Addition und der Multiplikation verträgt, bedeutet das

$$a =_3 a_n \cdot 1 + a_{n-1} \cdot 1 + \dots + a_1 \cdot 1 + a_0 =_3 0 \Leftrightarrow 3 \mid a_n + a_{n-1} + \dots + a_1 + a_0.$$

Genauso ergibt sich auch die 9-er Probe. Die 11-er Probe:

Es gilt $1 =_{11} 1, 10 =_{11} -1, 100 = 10^2 =_{11} (-1)^2 = 1, 1000 = 10^3 =_{11} (-1)^3 = -1, \dots, 10^n =_{11} (-1)^n$. Damit folgt

$$a = \sum_{k=0}^n a_k \cdot 10^k =_{11} \sum_{k=0}^n (-1)^k a_k =_{11} 0 \Leftrightarrow 11 \mid \sum_{k=0}^n (-1)^k a_k.$$

Die Summe $\sum_{k=0}^n (-1)^k a_k$ wird *alternierende Quersumme* genannt.

Die Zahl $3\ 597\ 678\ 354$ hat die Quersumme 57 und ist daher durch 3 teilbar (denn 57 hat die Quersumme 12, die offensichtlich durch 3 geteilt werden kann), nicht aber durch 9. Überprüfen wir das: $3\ 597\ 678\ 354 = 1\ 199\ 226\ 118 \cdot 3$ und $3\ 597\ 678\ 354 = 399\ 742\ 039 \cdot 9 + 3$.

Da die alternierende Quersumme -5 ist, kann die Zahl auch nicht durch 11 geteilt werden. Tatsächlich gilt $3\ 597\ 678\ 354 = 327\ 061\ 668 \cdot 11 + 6$.

$53\ 978\ 173\ 535$ hat die alternierende Quersumme $5 - 3 + 5 - 3 + 7 - 1 + 8 - 7 + 9 - 3 + 5 = 22$. Wir haben $53\ 978\ 173\ 535 = 4\ 907\ 106\ 685 \cdot 11$, was die 11-er Probe richtig voraussagt.

3.2.5 Satz (Inverse modulo p) Für eine natürliche Zahl $p > 1$ und für eine Zahl $0 < a \in \mathbb{Z}_p^*$, also für $0 < a < p$ und $\text{ggT}(a, p) = 1$ (vergl. (3.9)), gibt es genau eine Zahl $0 < b \in \mathbb{Z}_p^*$ mit der Eigenschaft

$$ab = 1 \pmod{p}. \quad (3.10)$$

Die Zahl b wird als Inverse von a modulo p bezeichnet.

Gelegentlich schreiben wir dafür auch $a^{-1} \pmod{p}$.

Anders als in \mathbb{Z} , wo z. B. die Zahl 3 keine Inverse hat, denn wir finden keine ganze Zahl b mit der Eigenschaft $b \cdot 3 = 1$, gibt es eine solche Zahl, sagen wir modulo 7, da $5 \cdot 3 = 2 \cdot 7 + 1 =_7 1$ gilt. Die Inverse modulo 7 von 3 ist daher 5.

Beweis. $0 < a \in \mathbb{Z}_p^*$ bedeutet $a < p$ und $\text{ggT}(a, p) = 1$. Nach dem Lemma von EUKLID 3.1.5 gibt es Zahlen $x, y \in \mathbb{Z}$, so daß

$$1 = xa + yp =_p xa \quad (*)$$

gilt. Daher ist x ein Kandidat für b . $x =_p 0$ kann nicht gelten, denn sonst müßte 1 durch p teilbar sein. Ist x ein Element von \mathbb{Z}_p^* , so ist es das gesuchte b . Sollte x nicht in \mathbb{Z}_p^* liegen, so ergibt Teilen durch p mit Rest (Satz 3.1.2):

$$x = tp + r \quad (**)$$

für ein $t \in \mathbb{Z}$ und $0 < r < p$ ($r = 0$ ist nicht möglich, denn das würde $p \mid x$, also $x =_p 0$ bedeuten, s. o.). Es folgt

$$1 =_p xa = atp + ra =_p ra.$$

Für b wählen wir jetzt r , denn $r \in \mathbb{Z}_p^*$. Dafür ist nur noch $\text{ggT}(r, p) = 1$ nachzuweisen. Sei also $u \in \mathbb{N}$ ein gemeinsamer Teiler von r und p . Dann zeigt $(**)$, daß u ein Teiler von x und $(*)$, daß u auch ein Teiler von 1 ist. Das ist nur für $u = 1$ möglich. Daher sind r und p tatsächlich teilerfremd.

Die Eindeutigkeit modulo p : Für ein weiteres Element $b' \neq b$ mit den geforderten Eigenschaften $0 \neq b' \in \mathbb{Z}_p^*$ und $b'a =_p 1$ folgt

$$(b - b')a =_p 0,$$

d. h. $p \mid (b - b')a$. Wegen $\text{ggT}(a, p) = 1$ zeigt Korollar 3.1.9: $p \mid b - b'$. Das bedeutet insbesondere $p \leq |b - b'| < p$ – offensichtlich ein Widerspruch. \square

3.2.6 Beispiel Im Beispiel 3.1.7 haben wir mit dem erweiterten euklidischen Algorithmus $\text{ggT}(5610, 637) = 1$ und die Darstellung

$$1 = 145 \cdot 5610 - 1277 \cdot 637$$

gefunden. Also ist der Kandidat für die Inverse modulo 5610 von 637 die Zahl -1277 . Da diese nicht in \mathbb{Z}_{5610} liegt, addieren wir solange 5610, bis das Ergebnis das erste Mal in \mathbb{Z}_{5610} zu finden ist. Das ist bereits nach dem ersten Schritt der Fall: $-1277 + 5610 = 4333$ ist die Inverse modulo 5610 von 637: $637 \cdot 4333 =_{5610} 1$.

3.2.7 Bemerkung Neben seiner Rolle bei den Beweisen von Korollar 3.1.9 und Satz 3.2.5 ist die Darstellung des größten gemeinsamen Teilers nach EUKLID vor allem das Hilfsmittel zur Berechnung der Inversen modulo p .

3.2.8 Bemerkung Wenn wir die Multiplikation so definieren, daß die Ergebnisse mod p wieder in \mathbb{Z}_p^* liegen, dann bedeutet Satz 3.2.5, daß wir auch eine Division in \mathbb{Z}_p^* über $a \cdot b^{-1} \mod p$ zur Verfügung haben ($b^{-1} \in \mathbb{Z}_p^*$ ist die Inverse modulo p von $b \in \mathbb{Z}_p^*$). Auf diese Weise können wir in $\mathbb{Z}_p^* \setminus \{0\}$ uneingeschränkt multiplizieren und dividieren. Eine Menge mit solchen Rechenoperationen nennt man eine *Gruppe* (diese Gruppe wird *Einheitengruppe* U_p für p genannt). Ist p sogar eine Primzahl (oder die Potenz einer solchen), dann läßt sich auch zu jedem $a \in \mathbb{Z}_p^*$ genau ein $b \in \mathbb{Z}_p^*$ mit der Eigenschaft $a + b =_p 0$ finden. Jetzt ist es möglich, auf \mathbb{Z}_p^* alle Grundrechenarten durchzuführen. Eine solche Menge nennt man einen *Körper*. Bis-her kennen wir nur die Körper \mathbb{Q} und \mathbb{R} , die jeweils unendlich viele Elemente enthalten. Im Gegensatz dazu ist \mathbb{Z}_p^* ein *endlicher Körper*. Ist p nicht zu groß, dann kann man sich das an der *Additions- und Multiplikationstabelle* veranschaulichen. Etwa für $\mathbb{Z}_7^* = \{0, 1, 2, 3, 4, 5, 6\} = \mathbb{Z}_7$:

$+/7$	0	1	2	3	4	5	6	$\cdot/7$	1	2	3	4	5	6
0	0	1	2	3	4	5	6	1	1	2	3	4	5	6
1	1	2	3	4	5	6	0	2	2	4	6	1	3	5
2	2	3	4	5	6	0	1	3	3	6	2	5	1	4
3	3	4	5	6	0	1	2	4	4	1	5	2	6	3
4	4	5	6	0	1	2	3	5	5	3	1	6	4	2
5	5	6	0	1	2	3	4	6	6	5	4	3	2	1
6	6	0	1	2	3	4	5							

Dagegen $\mathbb{Z}_{12}^* = \{0, 1, 5, 7, 11\}$:

$\cdot/12$	1	5	7	11
1	1	5	7	11
5	5	1	11	7
7	7	11	1	5
11	11	7	5	1

Offensichtlich ist $1 + 5 = 6 \notin \mathbb{Z}_{12}^*$, so daß keine Hoffnung besteht, \mathbb{Z}_{12}^* zu einem Körper zu machen.

3.2.9 Satz (Kleiner Satz von FERMAT) *Für jede Primzahl p und alle $a \in \mathbb{N}$ gilt*

$$a^p =_p a. \quad (3.11)$$

Ist darüber hinaus noch $\text{ggT}(a, p) = 1$ erfüllt, so folgt sogar

$$a^{p-1} =_p 1. \quad (3.12)$$

Beweis. Den Beweis führen wir mit Hilfe vollständiger Induktion nach $a \in \mathbb{N}$.

Für $a = 1$ gibt es nichts zu zeigen.

$a \rightarrow a + 1$: Wir gehen von $a^p =_p a$ aus und müssen $(a + 1)^p =_p a + 1$ nachweisen.

$$(a + 1)^p = \sum_{k=0}^p \binom{p}{k} a^k = 1 + \sum_{k=1}^{p-1} \binom{p}{k} a^k + a^p =_p 1 + a + \sum_{k=1}^{p-1} \binom{p}{k} a^k.$$

Wenn wir zeigen können, daß die letzte Summe modulo p gleich Null ist, sind wir fertig. Dazu müssen wir uns nur klar machen, daß die Binomialkoeffizienten $\binom{p}{k}$ für $1 \leq k \leq p - 1$ durch p teilbar sind: Es gilt $\binom{p}{k} \in \mathbb{N}$ und

$$\begin{aligned} k! \binom{p}{k} &= \frac{p!}{(p-k)!} = \frac{p(p-1)(p-2)\cdots(p-k+1)\cdot(p-k)!}{(p-k)!} \\ &= p(p-1)(p-2)\cdots(p-k+1). \end{aligned}$$

Das bedeutet insbesondere $p \mid k! \binom{p}{k}$, oder $p \mid k(k-1)(k-2)\cdots 3 \cdot 2 \cdot \binom{p}{k}$. Wegen $\text{ggT}(p, k) = \text{ggT}(p, k-1) = \cdots = \text{ggT}(p, 3) = \text{ggT}(p, 2) = 1$, folgt durch wiederholte Anwendung von Korollar 3.1.10: $p \mid \binom{p}{k}$.

Damit ist $a^p =_p a$ gezeigt. Gilt jetzt auch noch $\text{ggT}(a, p) = 1$, so hat a nach Satz 3.2.5 eine Inverse b modulo p . Mit deren Hilfe erhalten wir

$$a^{p-1} =_p a^{p-1} ab = a^p b =_p ab =_p 1. \quad \square$$

3.2.10 Lemma *Für zwei Primzahlen $p \neq q$ folgt aus $a = b \pmod{p}$ und $a = b \pmod{q}$:*

$$a = b \pmod{pq}.$$

Beweis. Es gibt Zahlen $k, \ell \in \mathbb{Z}$, so daß $a = b + kp$ und $a = b + \ell q$ gilt. Daraus ergibt sich sofort $kp = \ell q$, d. h. $p \mid \ell q$. Wegen $\text{ggT}(p, q) = 1$ muß p die Zahl ℓ teilen, also $\ell = sp$ gelten, mit einem geeigneten $s \in \mathbb{Z}$. Das bedeutet $a = b + spq = b \pmod{pq}$. \square

3.2.11 Satz Für zwei Primzahlen $p \neq q$ und für eine Zahl $a \in \mathbb{N}$ mit $\text{ggT}(a, pq) = 1$ gilt

$$a^{(p-1)(q-1)} = 1 \pmod{pq}. \quad (3.13)$$

Beweis. Aus $\text{ggT}(a, pq) = 1$ folgt insbesondere $\text{ggT}(a, p) = 1$ und $\text{ggT}(a, q) = 1$, also die Voraussetzung für den kleinen Satz von FERMAT 3.2.9 bzgl. p und q . Demnach gilt $a^{p-1} = 1 \pmod{p}$ und $a^{q-1} = 1 \pmod{q}$. Dann folgt

$$\begin{aligned} a^{(p-1)(q-1)} &= (a^{p-1})^{q-1} =_p 1^{q-1} = 1, \\ a^{(p-1)(q-1)} &= (a^{q-1})^{p-1} =_q 1^{p-1} = 1. \end{aligned}$$

Nach Lemma 3.2.10 ergibt sich daraus $a^{(p-1)(q-1)} = 1 \pmod{pq}$. \square

3.3 RSA-Verschlüsselung

3.3.1 Satz (RIVEST, SHAMIR, ADLEMAN)

Das RSA-Verfahren verläuft nach folgender Vorschrift:

- i) Man wähle zwei verschiedene, große Primzahlen p und q und bilde damit $n := p \cdot q$.
- ii) Man wähle eine Zahl $1 < e \in \mathbb{N}$, die zu $\varphi := (p-1)(q-1)$ teilerfremd und kleiner als φ ist. Dann ist der öffentliche Schlüssel durch das Zahlenpaar $S_o := [e, n]$ gegeben.
- iii) $d \in \mathbb{N}$, $d < \varphi$, sei die Inverse von e modulo φ . Sie wird mit Hilfe des erweiterten euklidischen Algorithmus berechnet. Dann ist der private Schlüssel durch das Zahlenpaar $S_p := [d, n]$ gegeben.
- iv) Eine Nachricht $M \in \mathbb{N}_0$, $M < n$, wird mit S_o durch

$$C := M^e \pmod{n}$$

verschlüsselt und mit S_p durch

$$M = C^d \pmod{n}$$

entschlüsselt.

Beweis. Es ist nur zu zeigen, daß die Entschlüsselung $M =_n C^d =_n M^{e \cdot d}$ funktioniert. Der Beweis erfolgt durch eine Fallunterscheidung.

- i) $\text{ggT}(M, n) = 1$: Aus $\text{ggT}(M, pq) = 1$ folgt nach Satz 3.2.11 $M^{(p-1)(q-1)} =_{pq} 1$. Wegen $ed =_\varphi 1$ gibt es eine Zahl $\ell \in \mathbb{N}$ mit $ed = 1 + \ell\varphi = 1 + \ell(p-1)(q-1)$.

$$M^{ed} = M \cdot M^{\ell(p-1)(q-1)} = M \cdot (M^{(p-1)(q-1)})^\ell =_{pq} M =_n M.$$

- ii) $\text{ggT}(M, n) > 1$: M hat mit $n = pq$ einen gemeinsamen Teiler. Das kann nur entweder p oder q sein, denn der Teiler pq würde $M \geq n$ bedeuten, im Widerspruch zur Voraussetzung $M < n$. O. B. d. A. können wir von $\text{ggT}(M, p) = p$ und $\text{ggT}(M, q) = 1$, d. h. $M = tp$ für ein geeignetes $t \in \mathbb{N}$ ausgehen. Damit haben wir nach dem kleinen Satz von FERMAT $M^{q-1} \equiv_q 1$ und durch Potenzieren mit $\ell(p-1)$ auch $(M^{q-1})^{\ell(p-1)} \equiv_q 1$. Es gibt daher eine ganze Zahl r mit der Eigenschaft $(M^{q-1})^{\ell(p-1)} = 1 + rq$, so daß

$$M^{ed} = M \cdot (M^{q-1})^{\ell(p-1)} = M(1 + rq) = M + rMq = M + rtqp =_n M. \quad \square$$

3.3.2 Beispiel $p := 101, q := 17, n = 1717, \varphi = 1600, M := 405$. Wir wählen $e := 411$. Der erweiterte euklidische Algorithmus:

$$\begin{array}{ll} 1600 = 3 \cdot 411 + 367 & 367 = 1600 - 3 \cdot 411 \\ 411 = 1 \cdot 367 + 44 & 44 = 411 - 1 \cdot 367 \\ 367 = 8 \cdot 44 + 15 & 15 = 367 - 8 \cdot 44 \\ 44 = 2 \cdot 15 + 14 & 14 = 44 - 2 \cdot 15 \\ 15 = 1 \cdot 14 + 1 & 1 = 15 - 14 \end{array}$$

Also gilt tatsächlich $\text{ggT}(\varphi, 411) = 1$. Um die Inverse d von e modulo φ zu bestimmen, berechnen wir für den größten gemeinsamen Teiler 1 die Darstellung nach EUKLID:

$$\begin{array}{llll} 1 = 15 - 44 + 2 \cdot 15 & = & 3 \cdot 15 - 44 \\ = 3 \cdot 367 - 24 \cdot 44 - 44 & = & 3 \cdot 367 - 25 \cdot 44 \\ = 3 \cdot 367 - 25 \cdot 411 + 25 \cdot 367 & = & 28 \cdot 367 - 25 \cdot 411 \\ = 28 \cdot 1600 - 84 \cdot 411 - 25 \cdot 411 & = & 28 \cdot 1600 - 109 \cdot 411. \end{array}$$

Wir erhalten $-109 \cdot 411 = 1 - 28 \cdot 1600 =_{\varphi} 1$. Die Inverse von e ist demnach -109 . Das RSA-Verfahren erfordert jedoch ein positives d . Das kann immer erreicht werden, indem ein geeignetes Vielfaches von φ zum Kandidaten für d addiert wird: $-109 + 1600 = 1491$.

Ergebnis: $S_o = [411, 1717], S_p = [1491, 1717]$.

Verschlüsselung:

$$\begin{array}{llll} C =_n 405^{411} & = & 405 \cdot (405^2)^{205} \\ =_n 405 \cdot 910^{205} & = & 405 \cdot 910 \cdot (910^2)^{102} \\ =_n 405 \cdot 910 \cdot 506^{102} & = & 405 \cdot 910 \cdot (506^2)^{51} \\ =_n 405 \cdot 910 \cdot 203^{51} & = & 405 \cdot 910 \cdot 203 \cdot (203^2)^{25} \\ =_n 405 \cdot 910 \cdot 203 & =_n & 809. \end{array}$$

Die verschlüsselte Nachricht ist $C = 809$.

Entschlüsselung:

$$\begin{array}{llll} C^d =_n 809^{1491} & = & 809 \cdot (809^2)^{745} \\ =_n 809 \cdot 304^{745} & = & 809 \cdot 304 \cdot (304^2)^{372} \end{array}$$

$$\begin{aligned}
 &=_{\text{n}} 809 \cdot 304 \cdot 1415^{372} &= & 809 \cdot 304 \cdot (1415^2)^{186} \\
 &=_{\text{n}} 809 \cdot 304 \cdot 203^{186} &= & 809 \cdot 304 \cdot (203^2)^{93} \\
 &=_{\text{n}} 809 \cdot 304 &=_{\text{n}} & 405 = M.
 \end{aligned}$$

3.4 Chinesischer Restsatz

Der folgende Satz ist für das RSA-Verfahren nicht nötig. Da wir aber gerade alle Hilfsmittel für seinen Beweis beisammen haben, nehmen wir ihn mit auf.

3.4.1 Satz (Chinesischer Restsatz) *Die Zahlen $m_1, \dots, m_n \in \mathbb{N}$ seien paarweise teilerfremd. Dann hat das Gleichungssystem*

$$\begin{aligned}
 x &= a_1 \pmod{m_1} \\
 x &= a_2 \pmod{m_2} \\
 &\vdots \\
 x &= a_n \pmod{m_n}
 \end{aligned}$$

genau eine Lösung x modulo $m_1 m_2 \cdots m_n$.

Beweis. Es sei $\mu := m_1 m_2 \cdots m_n$ und $\mu_i := \prod_{k \neq i} m_k$. Dann gilt $\text{ggT}(m_i, \mu_i) = 1$, denn jeder Teiler von m_i muß laut Voraussetzung teilerfremd zu m_k sein ($k \neq i$) und nach Korollar 3.1.9 dann auch zu μ_i . Nach Satz 3.2.5 hat μ_i eine Inverse v_i modulo m_i . Also gilt $v_i \mu_i = 1 \pmod{m_i}$, aber $v_i \mu_i = 0 \pmod{m_k}$, für $k \neq i$, denn $m_k \mid \mu_i$. Wir definieren daher

$$x := \sum_{i=1}^n a_i v_i \mu_i.$$

Dann gilt $x = a_i \pmod{m_i}$ für $i = 1, \dots, n$, d. h., x ist eine Lösung. Für eine weitere Lösung x' dieses Gleichungssystems haben wir $m_i \mid (x - x')$ für $i = 1, \dots, n$. Da die m_i paarweise teilerfremd sind, folgt aus dem zweiten Teil von Korollar 3.1.9 schließlich $\mu \mid (x - x')$, also $x = x' \pmod{\mu}$. \square

3.4.2 Beispiel Ein klassisches Beispiel ist die Aufgabe, eine natürliche Zahl x zu finden, die jeweils den Rest 1 aufweist, wenn man sie durch 2, 3, 4, 5 und durch 6 teilt. Darüber hinaus soll sie auch noch durch 7 teilbar sein. Das bedeutet

$$\begin{aligned}
 x &= 1 \pmod{2}, & x &= 1 \pmod{3}, & x &= 1 \pmod{4}, \\
 x &= 1 \pmod{5}, & x &= 1 \pmod{6}, & x &= 0 \pmod{7}.
 \end{aligned}$$

Der chinesische Restsatz lässt sich nicht sofort anwenden, weil die Reste 2, 4 und 6 nicht teilerfremd sind. Allerdings können wir die ersten fünf Gleichungen mit Hilfe von Lemma 3.4.5 zu $x = 1 \pmod{60}$ zusammenfassen. Auf diese und die letzte Gleichung lässt sich Satz 3.4.1 anwenden: $m_1 := 60$, $m_2 := 7$, $\mu := 420$, $a_1 := 1$, $a_2 := 0$, $\mu_1 := 7$, $\mu_2 := 60$ ergibt $v_1 = 43$ als Inverse von 7 mod 60 und $v_2 = 2$ als die Inverse von 60 mod 7. Wir erhalten demnach $x = 1 \cdot 43 \cdot 7 + 0 \cdot 2 \cdot 60 = 301$. Alle anderen Lösungen haben die Form $301 + t \cdot 420$, für $t \in \mathbb{N}$.

Für dieses Beispiel benötigen wir noch etwas Handwerkszeug:

3.4.3 Definition Für natürliche Zahlen p_1, \dots, p_n bezeichnet $\text{kgV}(p_1, \dots, p_n)$ das kleinste gemeinsame Vielfache von p_1, \dots, p_n , also die Zahl

$$\text{kgV}(p_1, \dots, p_n) := \min \{ x \in \mathbb{N} \mid \exists_{t_1, \dots, t_n \in \mathbb{N}} x = t_1 p_1 = \dots = t_n p_n \}. \quad (3.14)$$

Natürlich ist $x = p_1 p_2 \cdots p_n$ ein gemeinsames Vielfaches der Zahlen p_1 bis p_n , so daß die Menge in (3.14) nicht leer ist. Falls p_1, \dots, p_n teilerfremd sind, ist $p_1 p_2 \cdots p_n$ auch schon das kleinste gemeinsame Vielfache dieser Zahlen.

Es gilt etwa $\text{kgV}(42, 120) = \frac{2^3 \cdot 3 \cdot 5 \cdot 2 \cdot 3 \cdot 7}{2 \cdot 3} = \frac{42 \cdot 120}{\text{ggT}(42, 120)} = 840$. Dieses Ergebnis ist kein Zufall:

3.4.4 Lemma $\text{kgV}(p, q)$ ist ein Teiler jedes gemeinsamen Vielfachen von p und q . Der Zusammenhang mit $\text{ggT}(p, q)$ ist

$$\text{kgV}(p, q) \text{ggT}(p, q) = pq. \quad (3.15)$$

Beweis. Seien $g := \text{ggT}(p, q)$ und $k := \text{kgV}(p, q)$. Dann gibt es Zahlen $s, t, m, n \in \mathbb{N}$, so daß $p = sg$, $q = tg$ und $k = mp = nq$ gilt.

Sei $\mu > 0$ ein gemeinsames Vielfaches von p und q und $k < \mu$. Dann gilt $\mu = dk + r$ für ein geeignetes $d \in \mathbb{N}$, und einen Rest $0 \leq r < k$. Insbesondere ist $r = \mu - dk$ ein gemeinsames Vielfaches von p und q , denn μ und k sind es. Wäre $r > 0$, dann müßte $r \geq k$ gelten, denn k ist das kleinste gemeinsame Vielfache von p und q : W! Also bleibt nur $r = 0$, d. h., $k \mid \mu$.

Natürlich gilt $\text{ggT}(s, t) =: u = 1$. Es gibt nämlich Zahlen $s', t' \in \mathbb{N}$, mit $s = us'$ und $t = ut'$. Dann zeigt $p = s'ug$ und $q = t'ug$, daß ug ein gemeinsamer Teiler von p und q ist. Daher gilt $ug \leq g$, was nur für $u = 1$ möglich ist.

Jetzt zu (3.15). Der Ausdruck

$$\frac{pq}{g} = sq = tp$$

ist offensichtlich ein gemeinsames Vielfaches von p und q . Das bedeutet $k \mid \frac{pq}{g}$. Es gibt daher ein $\ell \in \mathbb{N}$ mit der Eigenschaft $sq = tp = \ell k = \ell m p = \ell n q$. Daraus folgt $t = \ell m$ und $s = \ell n$. Da s und t teilerfremd sind, muß $\ell = 1$ gelten. Das zeigt $\frac{pq}{g} = k$, also (3.15). \square

3.4.5 Lemma Für natürliche Zahlen p_1, p_2, \dots, p_n gilt

$$\forall_{i \in \{1, \dots, n\}} a = b \bmod p_i \Rightarrow a = b \bmod \text{kgV}(p_1, p_2, \dots, p_n).$$

Das ist eine Verallgemeinerung von Lemma 3.2.10.

Beweis. $a = b \bmod p_1$ und $a = b \bmod p_2$ bedeutet $a = b + tp_1 = b + sp_2$ für geeignete $t, s \in \mathbb{Z}$. Also ist $tp_1 = sp_2$ ein gemeinsames Vielfaches von p_1 und p_2 und wird nach Lemma 3.4.4 von $\text{kgV}(p_1, p_2)$ geteilt: $tp_1 = \alpha \text{kgV}(p_1, p_2)$. Das heißt $a = b + \alpha \text{kgV}(p_1, p_2)$, oder $a = b \bmod \text{kgV}(p_1, p_2)$. Zusammen mit $a = b \bmod p_3$ folgt damit $a = b \bmod \text{kgV}(\text{kgV}(p_1, p_2), p_3)$, also $a = b \bmod \text{kgV}(p_1, p_2, p_3)$, nach Übung 3.4.6, v). Das kann jetzt fortgesetzt werden. \square

3.4.6 A Für $p, q \in \mathbb{N}$ sei $g := \text{ggT}(p, q)$ und $k := \text{kgV}(p, q)$. Also gilt $p = sg$ und $q = tg$ und $k = pm = nq$ für geeignete $m, n, s, t \in \mathbb{N}$.

- i) Zeigen Sie mit Hilfe von Lemma 3.1.5, daß jeder gemeinsame Teiler von p und q ein Teiler von g ist und daß s und t teilerfremd sind.
- ii) Zeigen Sie, daß auch m und n teilerfremd sind.
- iii) Zeigen Sie $\text{ggT}(p, q, r) = \text{ggT}(\text{ggT}(p, q), r)$.
- iv) Zeigen Sie, daß es auch eine Darstellung nach EUKLID für $\text{ggT}(p, q, r)$ gibt und daß sie durch eine zu (3.5) analoge Formel gegeben ist.
- v) Zeigen Sie $\text{kgV}(p, q, r) = \text{kgV}(\text{kgV}(p, q), r)$, $\text{kgV}(p, q, r, s) = \text{kgV}(\text{kgV}(p, q, r), s)$, usw.
- vi) Zeigen Sie: Aus $\text{ggT}(p, q) = 1$, $\text{ggT}(a, b) = 1$ für $a, b \in \mathbb{N}$ und $pa = qb$ folgt $a = q$ und $b = p$.
- vii) Zeigen Sie: $ap = qb$ und $\text{ggT}(a, b) = 1$ für $a, b \in \mathbb{N}$ hat $ap = \text{kgV}(p, q)$ zur Folge.

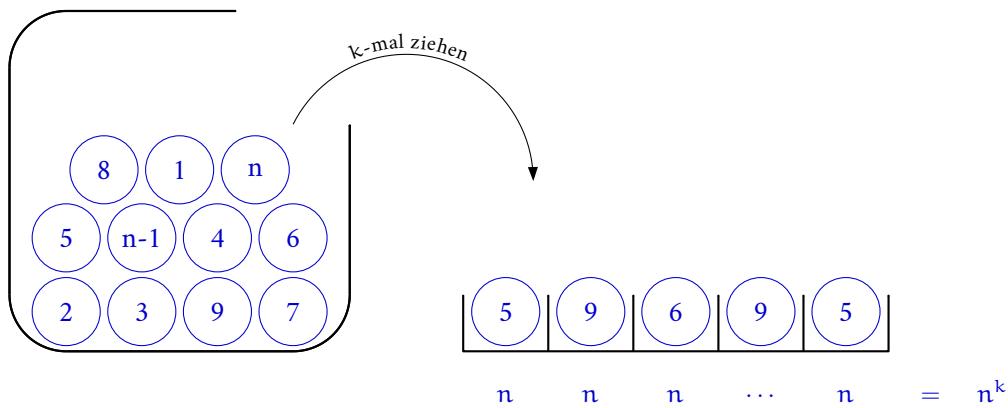
3.4.7 Beispiel Es sei $p := 30$, $q := 120$ und $r := 252$. Dann ist offensichtlich $\text{kgV}(p, q) = 120$, $\text{ggT}(120, 252) = 12$ und daher $\text{kgV}(30, 120, 252) = \frac{120 \cdot 252}{12} = 2520 = 84 \cdot p = 21 \cdot q = 10 \cdot r$. Die Zahlen 84, 21 und 10 sind teilerfremd. Damit ist nicht gemeint, daß sie paarweise teilerfremd sein müssen, sondern als Gesamtheit der drei Zahlen.

$\text{ggT}(30, 120, 252) = \text{ggT}(\text{ggT}(30, 120), 252) = \text{ggT}(30, 252) = 6$. Wegen $30 = -3 \cdot 30 + 1 \cdot 120$ und $6 = 17 \cdot 30 - 2 \cdot 252$ ist $6 = -51 \cdot 30 + 17 \cdot 120 - 2 \cdot 252$ eine mögliche Darstellung nach EUKLID für $\text{ggT}(30, 120, 252)$.

4 Kombinatorik

4.1 Die Urnenmodelle

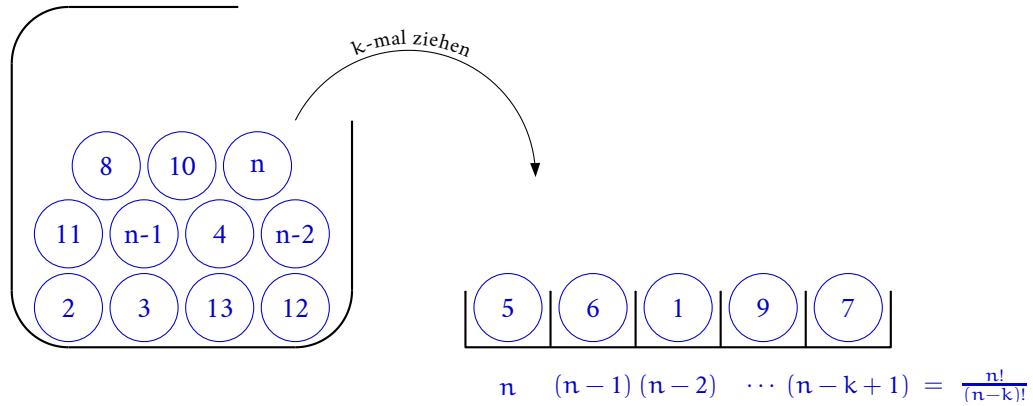
4.1.1 Ungeordnete Ziehung mit Zurücklegen Aus einer Urne mit n durchnummierten Kugeln wird k -mal gezogen. Die erhaltene Kugel wird notiert. Anschließend wird sie wieder zurückgelegt. Die Reihenfolge der Ziehung bleibt unverändert. Daher gibt es für die erste Ziehung n mögliche Ergebnisse. Für die zweite aber auch, da in der Urne alle Kugeln wieder vorhanden sind. Für die ersten beiden Ziehungen haben wir demnach $n \cdot n = n^2$ mögliche Ziehungsergebnisse. Für die ersten drei dann n^3 , usw. Für die k Ziehungen ergeben sich so n^k Möglichkeiten.



4.1.2 Ungeordnete Ziehung ohne Zurücklegen Aus einer Urne mit n durchnummierten Kugeln wird k -mal gezogen. Die erhaltenen Kugeln werden nicht zurückgelegt. Die Reihenfolge der Ziehung bleibt unverändert. Daher gibt es für die erste Ziehung n mögliche Ergebnisse. Für die zweite aber nur noch $n - 1$, da die Kugel der ersten Ziehung in der Urne fehlt. Für die ersten beiden Ziehungen haben wir demnach $n(n - 1)$ mögliche Ziehungsergebnisse. Für die ersten drei dann $n(n - 1)(n - 2)$, usw. Für die k Ziehungen ergeben sich auf diese Weise

$$n(n - 1) \cdots (n - k + 1) = \frac{n(n - 1) \cdots (n - k + 1) \cdot (n - k) \cdots 3 \cdot 2}{(n - k)!} = \frac{n!}{(n - k)!}$$

Möglichkeiten.

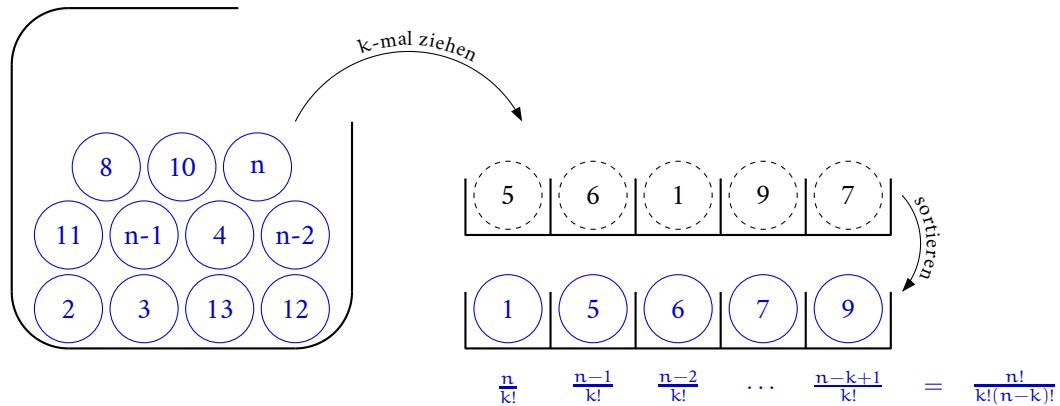


Einen wichtigen Spezialfall erhalten wir für $k = n$, also für den Fall, bei dem die Urne vollständig geleert wird. Es ergeben sich $\frac{n!}{(n-n)!} = n!$ Ziehungsergebnisse (vergl. (1.30)). Da die Ziehung ungeordnet ist, entstehen dabei alle möglichen Anordnungen der Zahlen 1 bis n , also alle *Permutationen* dieser Zahlen (vergl. Beispiel 4.1.9, 6), oder 2.5.10).

4.1.3 Satz *Mit den Elementen einer n -elementigen Menge lassen sich $n!$ Permutationen bilden. D.h., es gibt $n!$ Möglichkeiten n unterscheidbare Objekte anzugeordnen.*

Es gibt $\frac{n!}{(n-k)!}$ Möglichkeiten für ungeordnete Ziehungen ohne Zurücklegen von k Kugeln aus einer Urne mit n Kugeln.

4.1.4 Geordnete Ziehung ohne Zurücklegen Aus einer Urne mit n durchnummierten Kugeln wird k -mal gezogen. Die erhaltenen Kugeln werden nicht zurückgelegt. Die Reihenfolge der Ziehung ist unwichtig. Nach der Ziehung ordnet man daher die Kugeln nach aufsteigenden Nummern an. Dabei geht die Reihenfolge der ursprünglichen Ziehung natürlich verloren. Um die Anzahl der möglichen Ziehungsergebnissen zu erfahren, stellen wir uns zunächst alle möglichen Ausgänge vor, bevor die Kugeln sortiert werden. Das sind $\frac{n!}{(n-k)!}$ k -Tupel aus verschiedenen Zahlen, (wie etwa 56197 für $k = 5$) denn jetzt handelt es sich noch um eine ungeordnete Ziehung ohne Zurücklegen.



Wir gruppieren sie in Blöcke, die jeweils nur Tupel mit den gleichen k Zahlen, jedoch in unterschiedlicher Reihenfolge, aufweisen (also etwa 15679, 15697, 15967, ..., 97651). Jeder Block

enthält genau $k!$ Elemente, denn er enthält alle Permutationen seines ersten Tupels. Durch das Sortieren fallen für jeden Block diese $k!$ -Möglichkeiten zu einer einzigen zusammen (vergl. Beispiel 4.1.9, 2) und 3)). Die Zahl $\frac{n!}{(n-k)!}$ der ungeordneten Tupel reduziert sich daher um den Faktor $k!$ auf $\frac{n!}{k!(n-k)!} = \binom{n}{k}$.

Es gibt weitere Interpretationen dieses Ergebnisses. Die vielleicht wichtigste betrifft die Anzahl der Möglichkeiten, aus einer Menge mit n Elementen, k -elementige Teilmengen zu bilden. Das liegt daran, daß das Ziehungsergebnis ungeordnet ist. Das Ergebnis 15679 ist vollständig durch die Menge $\{1, 5, 6, 7, 9\}$ beschrieben, da die Reihenfolge ihrer Elemente definitionsgemäß nicht von Bedeutung ist. Jedes Ziehungsergebnis bestimmt eine k -elementige Teilmenge von $\{1, 2, \dots, n-1, n\}$ und jede solche Teilmenge einen möglichen Ausgang der geordneten Ziehung ohne Zurücklegen.

4.1.5 Satz $\binom{n}{k}$ ist die Anzahl

- i) der möglichen k -elementigen Teilmengen einer n -elementigen Menge,
- ii) der möglichen Anordnungen von k gleichen Objekten auf n Plätze,
- iii) der möglichen geordneten Ziehungen von k Kugeln aus einer Urne mit n Kugeln.

Eine Verallgemeinerung des Binomialkoeffizienten $\binom{n}{k}$ ist der Multinomialkoeffizient

$$\binom{n}{k_1, k_2, \dots, k_r} := \frac{n!}{k_1! k_2! \cdots k_r!}, \quad (4.1)$$

mit $k_1 + k_2 + \cdots + k_r = n$, $k_i \in \mathbb{N}_0$. Er beschreibt die Anzahl der möglichen Anordnungen von k_1, \dots, k_r jeweils gleichen Objekten auf n Plätze. Außerdem übernimmt er die Rolle des Binomialkoeffizienten in der Verallgemeinerung des binomischen Lehrsatzes 1.2.16

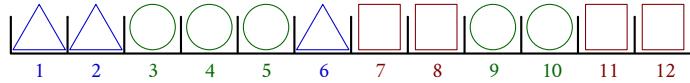
$$(a_1 + a_2 + \cdots + a_r)^n = \sum_{k_1 + \cdots + k_r = n} \binom{n}{k_1, k_2, \dots, k_r} a_1^{k_1} a_2^{k_2} \cdots a_r^{k_r}. \quad (4.2)$$

Beweis. i) Den Beweis haben wir im Wesentlichen schon erbracht. Wir müssen uns noch mit dem Sonderfall $\binom{n}{0} = 1$ beschäftigen. Auch hier stimmt die Behauptung, denn die einzige 0-elementige Menge ist die leere Menge, die Teilemenge jeder Menge ist.

ii) Die Anzahl möglicher Anordnungen machen wir uns an einem Beispiel klar: Eine Anordnung von 5 Kreisen auf 12 Plätze ist durch die Menge der Platznummern, im folgenden Beispiel $\{3, 4, 5, 9, 10\}$, eindeutig bestimmt. Daher gibt es genau so viele Anordnungen, wie 5-elementige Teilmengen der 12-elementigen Menge $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. Das sind $\binom{12}{5}$.



Die Interpretation der Multinomialkoeffizienten $\binom{n}{k_1, k_2, \dots, k_r}$ demonstrieren wir am Beispiel der $\binom{12}{3,4,5}$ möglichen Anordnungen von 3 Dreiecken, 4 Quadraten und 5 Kreisen auf 12 Plätze.



Zunächst verteilen wir die 5 Kreise auf den 12 verfügbaren Plätzen. Dafür gibt es $\binom{12}{5}$ Möglichkeiten. Bei jeder sind 7 Plätze nicht durch Kreise besetzt, die wir jetzt mit den 3 Dreiecken und 4 Quadraten belegen können. Das geht jeweils auf $\binom{7}{3}$ verschiedene Weisen. Insgesamt gibt es daher

$$\binom{12}{5} \binom{7}{3} = \binom{12}{7} \binom{7}{3} = \frac{12!}{5! 7!} \cdot \frac{7!}{3! 4!} = \frac{12!}{3! 4! 5!} = \binom{12}{3, 4, 5}$$

Möglichkeiten, die Dreiecke, die Quadrate und die Kreise zu verteilen.

Die Verallgemeinerung des binomischen Lehrsatzes zeigen wir nur für den Fall $(a_1 + a_2 + a_3)^n$. Hier lässt sich das Wesentliche verstehen.

$$\begin{aligned} (a_1 + a_2 + a_3)^n &= (a_1 + (a_2 + a_3))^n \stackrel{(1.32)}{=} \sum_{k=0}^n \binom{n}{k} a_1^{n-k} (a_2 + a_3)^k \\ &= \sum_{k=0}^n \sum_{\ell=0}^k \binom{n}{k} \binom{k}{\ell} a_1^{n-k} a_2^{k-\ell} a_3^\ell = \sum_{k=0}^n \sum_{\ell=0}^k \frac{n!}{(n-k)! k!} \frac{k!}{(k-\ell)! \ell!} a_1^{n-k} a_2^{k-\ell} a_3^\ell \\ &= \sum_{k=0}^n \sum_{\ell=0}^k \binom{n}{n-k, k-\ell, \ell} a_1^{n-k} a_2^{k-\ell} a_3^\ell = \sum_{k_1+k_2+k_3=n} \binom{n}{k_1, k_2, k_3} a_1^{k_1} a_2^{k_2} a_3^{k_3}. \quad \square \end{aligned}$$

4.1.6 Korollar Die Anzahl aller Teilmengen einer n -elementigen Menge ist 2^n .

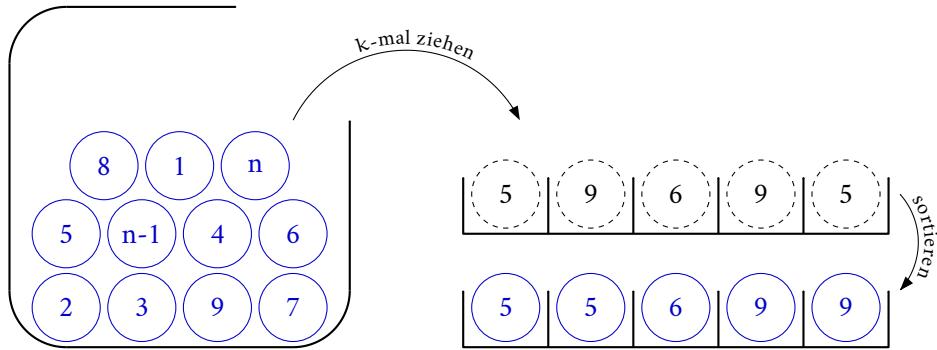
Beweis. Die Anzahl aller Teilmengen setzt sich aus der Anzahl der 0-, der 1-, der 2-elementigen usw. zusammen. Eine Anwendung des binomischen Lehrsatzes 1.2.16 ergibt dafür

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = \sum_{k=0}^n \binom{n}{k} 1^k \cdot 1^{n-k} = (1+1)^n = 2^n. \quad \square$$

4.1.7 Geordnete Ziehung mit Zurücklegen Durch geeignete Kodierung der Ziehungsergebnisse führen wir diesen Fall auf die Anordnung von Objekten zurück. Das Ergebnis 55699 etwa ist durch Angabe der Häufigkeiten k_i für das Auftreten der Zahl i eindeutig charakterisiert. Für 55699 heißt das $k_1 = k_2 = k_3 = k_4 = 0, k_5 = 2, k_6 = 1, k_7 = k_8 = 0, k_9 = 2, \dots, k_n = 0$. Da wir $k = 5$ mal ziehen, muß $k_1 + k_2 + \cdots + k_n = 5$ gelten. Jede mögliche Ziehung wird durch eine *Zerlegung*

$$k_1 + k_2 + \cdots + k_n = k$$

der Anzahl k der Ziehungen in n Summanden $k_i \in \mathbb{N}_0$ eindeutig festgelegt. Wir müssen also nur noch bestimmen, wie viele Zerlegungen dieser Art eine natürliche Zahl k haben kann.



An Beispielen, wie $0+0+0+0+0+2+1+0+0+2+0 = 5$ für $k = 5$ und $n = 10$, demonstrieren wir die Kodierung, die es gestattet, das Ziehungsergebnis durch Anordnungen von Objekten auf einer festen Anzahl von Plätzen zu beschreiben. Wir ersetzen einfach die Häufigkeiten k_i für das Ziehungsergebnis i durch die entsprechende Anzahl von Punkten \bullet . Das bedeutet

$$\begin{aligned}
 0 + 0 + 0 + 0 + 0 + 2 + 1 + 0 + 0 + 2 + 0 &\simeq + + + + \bullet \bullet + \bullet + + + \bullet \bullet + \\
 3 + 0 + 0 + 1 + 0 + 0 + 0 + 0 + 0 + 1 &\simeq \bullet \bullet \bullet + + + \bullet + + + + + + \bullet \\
 1 + 0 + 1 + 1 + 0 + 0 + 1 + 0 + 1 + 0 &\simeq \bullet + + \bullet + \bullet + + + \bullet + + \bullet + \\
 &\dots && \dots
 \end{aligned}$$

Offensichtlich handelt es sich jetzt nur noch darum, 5 Punkte auf $10 - 1 + 5 = 14$ Plätze zu verteilen. Dafür gibt es, wie wir bereits wissen, $\binom{14}{5} = 2002$ Möglichkeiten. Im allgemeinen Fall haben wir $n - 1 +$ -Zeichen, um mit den n Häufigkeiten k_1, \dots, k_n die Summe $k_1 + k_2 + \dots + k_n = k$ zu bilden. In der Kodierung bedeutet das, k Punkte auf $n - 1 + k$ Plätze zu verteilen, wofür es $\binom{n+k-1}{k}$ Möglichkeiten gibt.

4.1.8 Satz $\binom{n+k-1}{k}$ ist die Anzahl der Möglichkeiten

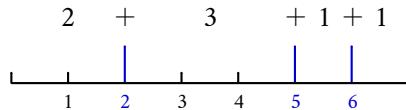
- i) einer geordneten Ziehung mit Zurücklegen von k Kugeln aus einer Urne mit n Kugeln,
- ii) die Zahl k in n nicht negative ganze Zahlen zu zerlegen.

$\binom{k-1}{n-1}$ ist die Anzahl der Möglichkeiten, die Zahl k in n natürliche Zahlen zu zerlegen.

Beweis. Es ist nur noch die Anzahl der Zerlegungen von k durch n natürliche Zahlen zu bestimmen. Offensichtlich muß jetzt $k \geq n$ gelten. Die Kodierung machen wir uns am Beispiel $k = 7$ und $n = 4$ klar:

$$\begin{aligned}
 7 &= 4 + 1 + 1 + 1 = 1 + 4 + 1 + 1 = 1 + 1 + 4 + 1 = 1 + 1 + 1 + 4 \\
 &= 3 + 2 + 1 + 1 = 3 + 1 + 2 + 1 = 3 + 1 + 1 + 2 = 1 + 3 + 1 + 2 = 1 + 1 + 3 + 2 = 1 + 3 + 2 + 1 \\
 &= 2 + 3 + 1 + 1 = 2 + 1 + 3 + 1 = 2 + 1 + 1 + 3 = 1 + 2 + 1 + 3 = 1 + 1 + 2 + 3 = 1 + 2 + 3 + 1 \\
 &= 2 + 2 + 2 + 1 = 2 + 2 + 1 + 2 = 2 + 1 + 2 + 2 = 1 + 2 + 2 + 2
 \end{aligned}$$

Die Zerlegung $7 = 2 + 3 + 1 + 1$ etwa kodieren wir durch aufeinanderfolgende Strecken mit 2 Einheiten, 3 Einheiten und zwei mit der Länge einer Einheit:



Diese Anordnung ist eindeutig durch die Positionen 2, 5 und 6 bestimmt, bei denen die Strecken aneinander stoßen. Bei einer Streckenlänge k kommen $k-1$ Positionen für die $n-1$ möglichen Trennstellen der n aufeinander folgenden Strecken in Frage. Es geht also um die Anzahl der Möglichkeiten, $n-1$ Trennstellen aus $k-1$ auszuwählen. Das sind $\binom{k-1}{n-1}$, wie wir inzwischen wissen. \square

Die vier Urnenmodelle lassen sich übersichtlich in folgender Tabelle zusammenfassen (hier ist u bzw. g die Abkürzung für *ungeordnet* bzw. *geordnet* und $m Z$ bzw. $o Z$ für *mit Zurücklegen* bzw. *ohne Zurücklegen*):

	$m Z$	$o Z$	(4.3)
u	n^k	$\frac{n!}{(n-k)!}$	
g	$\binom{n+k-1}{k}$	$\binom{n}{k}$	

4.1.9 Beispiel Wir wählen $n = 5$ und $k = 3$, um die verschiedenen Ziehungsmodelle an einem konkreten Fall vorzustellen.

1) Ungeordnete Ziehung mit Zurücklegen. Es gibt $5^3 = 125$ Ergebnisse:

111	112	113	114	115	311	312	313	314	315	511	512	513	514	515
121	122	123	124	125	321	322	323	324	325	521	522	523	524	525
131	132	133	134	135	331	332	333	334	335	531	532	533	534	535
141	142	143	144	145	341	342	343	344	345	541	542	543	544	545
151	152	153	154	155	351	352	353	354	355	551	552	553	554	555
211	212	213	214	215	411	412	413	414	415					
221	222	223	224	225	421	422	423	424	425					
231	232	233	234	235	431	432	433	434	435					
241	242	243	244	245	441	442	443	444	445					
251	252	253	254	255	451	452	453	454	455					

2) Ungeordnete Ziehung ohne Zurücklegen. Es gibt $\frac{5!}{(5-3)!} = 5 \cdot 4 \cdot 3 = 60$ Möglichkeiten:

123	124	125	134	135	145	234	235	245	345
132	142	152	143	153	154	242	253	254	354
312	412	512	413	513	514	423	523	524	534
213	214	215	314	315	415	324	325	425	435
231	241	251	341	351	451	342	352	452	453
321	421	521	431	531	541	432	532	542	543

3) Geordnete Ziehung ohne Zurücklegen. Es gibt $\binom{5}{3} = \frac{5 \cdot 4 \cdot 3!}{3! \cdot 2!} = \frac{5 \cdot 4}{2} = 10$ Ergebnisse:

123	124	125	134	135	145	234	235	245	345
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

4) Geordnete Ziehung mit Zurücklegen. Es gibt $\binom{5+3-1}{3} = 35$ Möglichkeiten:

111	112	113	114	115	333	334	335		
	122	123	124	125		344	345		
		133	134	135			355		
			144	145		444	445		
				155			455		
	222	223	224	225				555	
		233	234	235					
			244	245					
				255					

5) Es gibt $\binom{5-1}{3-1} = 6$ Möglichkeiten, die Zahl 5 in 3 Summanden aus \mathbb{N} zu zerlegen, nämlich $5 = 3 + 1 + 1 = 1 + 3 + 1 = 1 + 1 + 3 = 2 + 2 + 1 = 2 + 1 + 2 = 1 + 2 + 2$, aber $\binom{3+5-1}{5} = 21$ mit Summanden aus \mathbb{N}_0 ($n = 3, k = 5$):

$$\begin{aligned} 5 &= 5 + 0 + 0 = 0 + 5 + 0 = 0 + 0 + 5 \\ &= 4 + 1 + 0 = 4 + 0 + 1 = 0 + 4 + 1 = 1 + 4 + 0 = 1 + 0 + 4 = 0 + 1 + 4 \\ &= 3 + 2 + 0 = 3 + 0 + 2 = 0 + 3 + 2 = 2 + 3 + 0 = 2 + 0 + 3 = 0 + 2 + 3 \\ &= 3 + 1 + 1 = 1 + 3 + 1 = 1 + 1 + 3 \\ &= 2 + 2 + 1 = 2 + 1 + 2 = 1 + 2 + 2 \end{aligned}$$

6) Es gilt $(a_1 + a_2 + a_3)^4 = a_1^4 + 4a_1^3a_2 + 6a_1^2a_2^2 + 4a_1a_2^3 + a_2^4 + 4a_1^3a_3 + 12a_1^2a_2a_3 + 12a_1a_2^2a_3 + a_3^4$

7) Es gibt $4! = 24$ Möglichkeiten, die Zahlen 1, 2, 3 und 4 anzugeben:

$$\begin{array}{ccccccc} 1, 2, 3, 4 & 1, 3, 2, 4 & 3, 1, 2, 4 & 2, 1, 3, 4 & 2, 3, 1, 4 & 3, 2, 1, 4 \\ 1, 2, 4, 3 & 1, 3, 4, 2 & 3, 1, 4, 2 & 2, 1, 4, 3 & 2, 3, 4, 1 & 3, 2, 4, 1 \\ 1, 4, 2, 3 & 1, 4, 3, 2 & 3, 4, 1, 2 & 2, 4, 1, 3 & 2, 4, 3, 1 & 3, 4, 2, 1 \\ 4, 1, 2, 3 & 4, 1, 3, 2 & 4, 3, 1, 2 & 4, 2, 1, 3 & 4, 2, 3, 1 & 4, 3, 2, 1 \end{array}$$

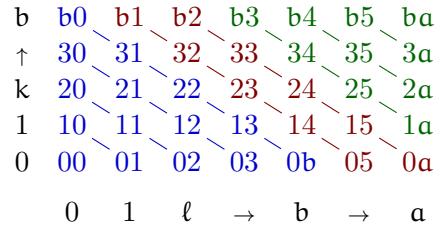
Die $5! = 120$ Möglichkeiten, die Zahlen 1, 2, 3, 4, 5 anzugeben, kann man aus den Permutationen von 1, 2, 3, 4 systematisch dadurch gewinnen, daß man die Zahl 5 an jede anhängt, sie dann jeweils an die vierte Position setzt, an die dritte, die zweite und die erste:

$$\begin{array}{ccccccc} 1, 2, 3, 4, 5 & 1, 3, 2, 4, 5 & 3, 1, 2, 4, 5 & 2, 1, 3, 4, 5 & 2, 3, 1, 4, 5 & 3, 2, 1, 4, 5 \\ 1, 2, 4, 3, 5 & 1, 3, 4, 2, 5 & 3, 1, 4, 2, 5 & 2, 1, 4, 3, 5 & 2, 3, 4, 1, 5 & 3, 2, 4, 1, 5 \\ 1, 4, 2, 3, 5 & 1, 4, 3, 2, 5 & 3, 4, 1, 2, 5 & 2, 4, 1, 3, 5 & 2, 4, 3, 1, 5 & 3, 4, 2, 1, 5 \\ 4, 1, 2, 3, 5 & 4, 1, 3, 2, 5 & 4, 3, 1, 2, 5 & 4, 2, 1, 3, 5 & 4, 2, 3, 1, 5 & 4, 3, 2, 1, 5 \\ 1, 2, 3, 5, 4 & 1, 3, 2, 5, 4 & 3, 1, 2, 5, 4 & 2, 1, 3, 5, 4 & 2, 3, 1, 5, 4 & 3, 2, 1, 5, 4 \\ 1, 2, 4, 5, 3 & 1, 3, 4, 5, 2 & 3, 1, 4, 5, 2 & 2, 1, 4, 5, 3 & 2, 3, 4, 5, 1 & 3, 2, 4, 5, 1 \\ 1, 4, 2, 5, 3 & 1, 4, 3, 5, 2 & 3, 4, 1, 5, 2 & 2, 4, 1, 5, 3 & 2, 4, 3, 5, 1 & 3, 4, 2, 5, 1 \\ 4, 1, 2, 5, 3 & 4, 1, 3, 5, 2 & 4, 3, 1, 5, 2 & 4, 2, 1, 5, 3 & 4, 2, 3, 5, 1 & 4, 3, 2, 5, 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 5, 1, 2, 3, 4 & 5, 1, 3, 2, 4 & 5, 3, 1, 2, 4 & 5, 2, 1, 3, 4 & 5, 2, 3, 1, 4 & 5, 3, 2, 1, 4 \\ 5, 1, 2, 4, 3 & 5, 1, 3, 4, 2 & 5, 3, 1, 4, 2 & 5, 2, 1, 4, 3 & 5, 2, 3, 4, 1 & 5, 3, 2, 4, 1 \\ 5, 1, 4, 2, 3 & 5, 1, 4, 3, 2 & 5, 3, 4, 1, 2 & 5, 2, 4, 1, 3 & 5, 2, 4, 3, 1 & 5, 3, 4, 2, 1 \\ 5, 4, 1, 2, 3 & 5, 4, 1, 3, 2 & 5, 4, 3, 1, 2 & 5, 4, 2, 1, 3 & 5, 4, 2, 3, 1 & 5, 4, 3, 2, 1 \end{array}$$

4.1.10 Die VANDERMONDSche Identität

Für $a, b \in \mathbb{N}$ und $b \leq a$ gilt:

$$\begin{aligned}
 (1+x)^a(1+x)^b &= (1+x)^{a+b} \\
 &= \sum_{n=0}^{a+b} \binom{a+b}{n} x^n = \sum_{k=0}^a \sum_{\ell=0}^b \binom{b}{k} \binom{a}{\ell} x^{k+\ell} \\
 &= \sum_{n=0}^b \sum_{k=0}^n \binom{b}{k} \binom{a}{n-k} x^n \\
 &\quad + \sum_{n=b+1}^a \sum_{k=0}^b \binom{b}{k} \binom{a}{n-k} x^n + \sum_{n=a+1}^{a+b} \sum_{k=n-a}^b \binom{b}{k} \binom{a}{n-k} x^n.
 \end{aligned}$$



Die Doppelsumme wird entlang der Diagonalen konstanter $k + \ell$ ausgeführt. Ein Koeffizientenvergleich ergibt

$$\begin{aligned}
 \sum_{k=0}^n \binom{b}{k} \binom{a}{n-k} &= \binom{a+b}{n}, \quad n \leq b, \quad \sum_{k=0}^b \binom{b}{k} \binom{a}{n-k} = \binom{a+b}{n}, \quad b \leq n \leq a, \\
 \sum_{k=n-a}^b \binom{b}{k} \binom{a}{n-k} &= \binom{a+b}{n}, \quad a \leq n \leq a+b.
 \end{aligned}$$

Treffen wir die Vereinbarung, daß $\binom{p}{q} := 0$ für $q < 0$, oder $q > p$ zu setzen ist, dann lassen sich diese drei Fälle für $a, b, n \in \mathbb{N}$, $n \leq a+b$, zur VANDERMONDSchen Identität zusammenfassen:

$$\sum_{k=0}^n \binom{b}{k} \binom{a}{n-k} = \binom{a+b}{n}. \quad (4.4)$$

Wir haben diese Identität rein algebraisch erhalten. Es gibt aber auch eine anschauliche Deutung. Dazu stellen wir uns eine Urne mit $a+b$ Kugeln vor, aus der n -mal gezogen wird. Dafür gibt es $\binom{a+b}{n}$ Möglichkeiten. Wir können uns aber auch vorstellen, daß wir die $a+b$ Kugeln auf zwei Urnen aufteilen, eine mit b und eine mit a Kugeln. Dann ziehen wir k mal aus der ersten und $n-k$ -mal aus der zweiten Urne, was im Ergebnis der Ziehung von n Kugeln aus der Urne mit $a+b$ Kugeln entspricht. Es gibt $\binom{b}{k}$ Ergebnisse für die Ziehung aus der ersten Urne, und für jede dieser Ziehungen jeweils $\binom{a}{n-k}$ Möglichkeiten für die zweite Urne. Das sind also $\binom{b}{k} \binom{a}{n-k}$ Möglichkeiten. Um alle Möglichkeiten zu erhalten, muß jetzt noch über k summiert werden, womit wir auch bei (4.4) angekommen sind.

Für $n = b$ erhalten wir, wegen $\sum_{k=0}^n \binom{n}{k} \binom{a}{n-k} = \sum_{k=0}^n \binom{n}{n-k} \binom{a}{n-k} = \sum_{k=0}^n \binom{n}{k} \binom{a}{k}$:

$$\sum_{k=0}^n \binom{n}{k} \binom{a}{k} = \binom{a+n}{n}. \quad (4.5)$$

Das ergibt für $a = n$:

$$\sum_{k=0}^n \binom{n}{k}^2 = \binom{2n}{n}. \quad (4.6)$$

Schreiben wir (4.4) systematisch in der Form

$$\sum_{k_1+k_2=n} \binom{a_1}{k_1} \binom{a_2}{k_2} = \binom{a_1+a_2}{n},$$

so erkennt man schnell, wie diese Beziehung verallgemeinert werden kann. Setzen wir $a_2 = a_3 + a_4$ und verwenden (4.4):

$$\begin{aligned} \binom{a_1+a_3+a_4}{n} &= \sum_{k_1+k_2=n} \binom{a_1}{k_1} \binom{a_3+a_4}{k_2} = \sum_{k_1+k_2=n} \sum_{k_3+k_4=k_2} \binom{a_1}{k_1} \binom{a_3}{k_3} \binom{a_4}{k_4} \\ &= \sum_{k_1+k_3+k_4=n} \binom{a_1}{k_1} \binom{a_3}{k_3} \binom{a_4}{k_4}. \end{aligned}$$

Diese Überlegung lässt sich wiederholen und ergibt schließlich die verallgemeinerte VANDERMONDSche Identität

$$\sum_{k_1+k_2+\dots+k_p=n} \binom{a_1}{k_1} \binom{a_2}{k_2} \dots \binom{a_r}{k_r} = \binom{a_1+a_2+\dots+a_r}{n}. \quad (4.7)$$

4.1.11 Beispiel (Lotto) Wir überlegen uns die Gewinnchancen der einzelnen Gewinnränge beim Lotto. Bekanntlich werden für ein Spiel auf einem quadratischen Feld mit den Zahlen 1 bis 49 sechs angekreuzt, sagen wir $\{1, 2, 3, 4, 5, 6\}$. Die Ziehung der Lottozahlen ist geordnet, ohne Zurücklegen: Nachdem sechs Kugeln aus einer Trommel mit 49 durchnummerierten Kugeln gezogen sind, werden die gezogenen Nummern in aufsteigender Reihenfolge als Ziehungsergebnis präsentiert. Da beim Ankreuzen der sechs Zahlen auf dem Tippzettel keine Reihenfolge festgelegt werden kann, darf sie für das Ziehungsergebnis auch keine Rolle spielen. Die Anzahl möglicher Ziehungen beträgt $\binom{49}{6} = 13983816$, also knapp 14 Millionen. Wenn man auch noch die Zusatzzahl spielt (eine Zahl zwischen 0 und 9, die auf dem Tippschein steht), multipliziert sich die Anzahl der Möglichkeiten noch einmal mit 10 und ergibt rund 140 Millionen Möglichkeiten. Da sechs Richtige nur auf eine Weise erzielt werden können, ist die Chance dafür $1 : \binom{49}{6} \approx 7.15 \cdot 10^{-8}$, und bei sechs Richtigen mit Zusatzzahl $\approx 7.15 \cdot 10^{-9}$.

Es gibt weitere Gewinnränge. 5 Richtige können auf $6 \cdot 43 = 258$ Arten getippt werden. Bei unserem fiktiven Beispiel $\{1, 2, 3, 4, 5, 6\}$ kommt man in diesen Gewinnrang, wenn genau eine der sechs Zahlen durch eine der 43 nicht gezogenen ersetzt wird, wie etwa $\{7, 2, 3, 4, 5, 6\}$, $\{1, 12, 3, 4, 5, 6\}$, ... $\{1, 2, 3, 4, 5, 49\}$. Genau 5 Richtige zu tippen hat demnach die Chance $258 : \binom{49}{6} \approx 1.84 \cdot 10^{-5} \approx 1 : 50000$.

Bei 4 Richtigen sind zwei der korrekten Zahlen durch nicht gezogene zu ersetzen. Für jede der $\binom{6}{2}$ Möglichkeiten, zwei Kugeln des tatsächlichen Ziehungsergebnisses auszuwählen, gibt es $\binom{43}{2}$ Weisen, sie durch zwei der 43 nicht gezogenen zu ersetzen. Das sind $\binom{6}{2} \cdot \binom{43}{2} = 13545$ Möglichkeiten. Die Chance dafür beträgt daher $\binom{6}{2} \cdot \binom{43}{2} : \binom{49}{6} \approx 9.69 \cdot 10^{-4} \approx 1 : 1000$.

3 Richtige sind auf $\binom{6}{3} \cdot \binom{43}{3} = 246820$ Weisen zu erhalten, mit einer Chance von $\binom{6}{3} \cdot \binom{43}{3} : \binom{49}{6} \approx 0.018 \approx 1 : 55$.

Obwohl sie nicht mehr zu den Gewinnrängen zählen, bestimmen wir noch die Möglichkeiten, genau zwei und genau eine Richtige zu tippen, nämlich $\binom{6}{4} \cdot \binom{43}{4} = 1851150$, bzw. $\binom{6}{5} \cdot \binom{43}{5} =$

5 775 588. Wie nicht anders zu erwarten, entfällt der Löwenanteil $\binom{6}{6} \cdot \binom{43}{6} = 6\,096\,454$ der Möglichkeiten darauf, alle Zahlen falsch zu tippen. Die Chance, beim Lotto nichts zu gewinnen, beläuft sich somit auf $(1\,851\,150 + 5\,775\,588 + 6\,096\,454) : \binom{49}{6} \approx 0.98 = 98 : 100$.

Wenn unsere Analyse der einzelnen Gewinnränge zutreffend ist, sollten sich die Anzahl ihrer Realisierungen zur Gesamtzahl $\binom{49}{6}$ aller Möglichkeiten aufsummieren. Das kann man anhand der konkreten Zahlen leicht überprüfen – oder wir vertrauen auf die Folgerung (4.5) aus der VANDERMONDSchen Identität:

$$\begin{aligned} & \binom{6}{0} \binom{43}{0} + \binom{6}{1} \binom{43}{1} + \binom{6}{2} \binom{43}{2} + \binom{6}{3} \binom{43}{3} + \binom{6}{4} \binom{43}{4} + \binom{6}{5} \binom{43}{5} + \binom{6}{6} \binom{43}{6} \\ &= \sum_{k=0}^6 \binom{6}{k} \binom{43}{k} \stackrel{(4.5)}{=} \binom{6+43}{6} = \binom{49}{6}. \end{aligned}$$

5 Vektorräume

5.1 \mathbb{R}^2 und \mathbb{R}^3 als Vektorraum

In Definition 1.2.4 haben wir die Produktmenge von zwei oder mehr Mengen eingeführt und als Beispiel dafür in 1.2.5 den \mathbb{R}^n durch

$$\mathbb{R}^n = \{ [x_1, x_2, \dots, x_n] \mid x_1, x_2, \dots, x_n \in \mathbb{R} \}$$

definiert. Die Elemente $x = [x_1, x_2, \dots, x_n] \in \mathbb{R}^n$ bezeichnen wir als *Vektoren*. Vorläufig werden wir die Schreibweise dieser Vektoren den Gegebenheiten anpassen, d.h., benötigen wir die Koordinaten eines Vektors $x \in \mathbb{R}^n$ im fließenden Text, so schreiben wir sie als Zeilen $[x_1, \dots, x_n]$, in abgesetzten Formeln dagegen als Spalten

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

Wir unterscheiden Vektoren x durch die fette Schreibweise von gewöhnlichen Zahlen x . Auf \mathbb{R}^n führen wir eine Addition für Vektoren und die sogenannte *Skalarmultiplikation* von Vektoren mit einer Zahl (Skalar) aus \mathbb{R} ein.

5.1.1 Definition Sind $x = [x_1, \dots, x_n]$ und $y = [y_1, \dots, y_n]$ Vektoren aus \mathbb{R}^n und $t \in \mathbb{R}$, so wird durch

$$x + y := \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix} \quad (5.1)$$

$$t \cdot x := \begin{bmatrix} t \cdot x_1 \\ \vdots \\ t \cdot x_n \end{bmatrix} \quad (5.2)$$

die Summe $x + y$ von x und y bzw. die Skalarmultiplikation $t \cdot x$ von x mit t definiert.

Natürlich werden wir meist tx statt $t \cdot x$ schreiben. Außerdem ist es üblich, wie beim Rechnen mit gewöhnlichen Zahlen, das Symbol $-x$ für $(-1) \cdot x$ zu vereinbaren und dann $x - y$ für $x + (-1) \cdot y$ zu schreiben. Der Vektor $\mathbf{0} := [0, \dots, 0]$ wird als *Nullvektor* bezeichnet.

Eine geometrische Interpretation der Vektoren und ihrer Rechengesetze geben wir zunächst für den Fall \mathbb{R}^2 und später auch für \mathbb{R}^3 an. Wir stellen uns einen Vektor $x = [x_1, x_2]$ als eine

gerichtete Strecke in einem zweidimensionalen Koordinatensystem vor. Ausgehend von einem beliebigen Punkt lautet die Vorschrift: Gehe x_1 Einheiten in Richtung der x_1 -Achse und x_2 Einheiten in Richtung der x_2 -Achse. Das Ende dieser Strecke markiert man durch eine Pfeilspitze, um die Richtung, in die der Vektor zeigt, kenntlich zu machen. Machen wir uns diese Interpretation von Vektoren zu eigen, so ergibt sich zwangsläufig, daß die Parallelverschiebung einer solchen gerichteten Strecke denselben Vektor darstellt, denn der Startpunkt in der oben angeführten Vorschrift ist beliebig. Wählen wir allerdings den Ursprung O des Koordinatensystems als Startpunkt, dann zeigen die Vektoren \mathbf{x} und \mathbf{y} auf Punkte $X := [x_1, x_2]$ bzw. $Y := [y_1, y_2]$, mit denselben Koordinaten x_1 und x_2 , wie die zugehörigen Vektoren.

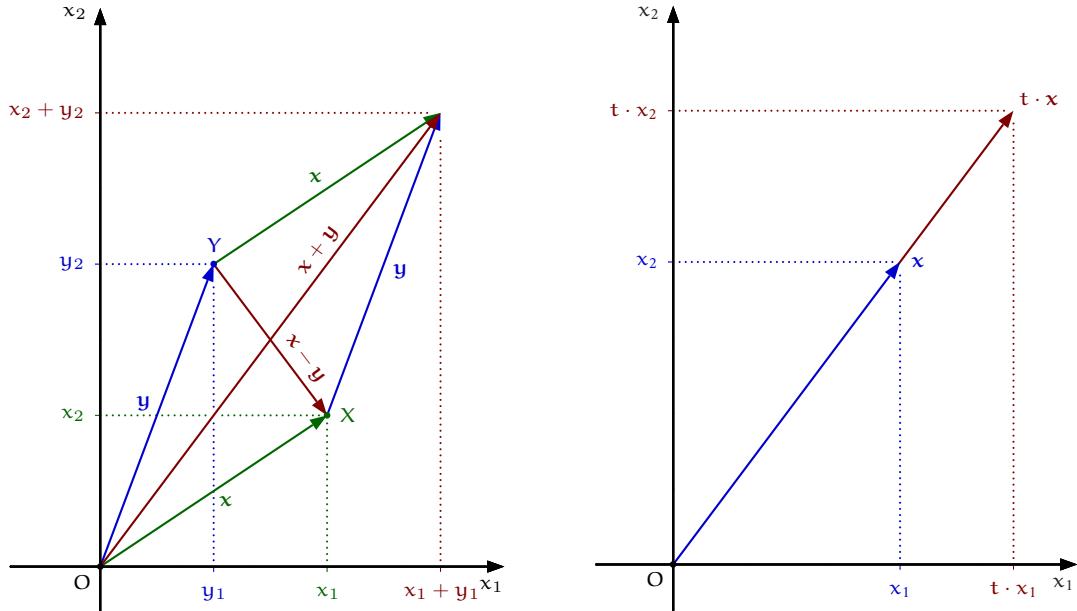
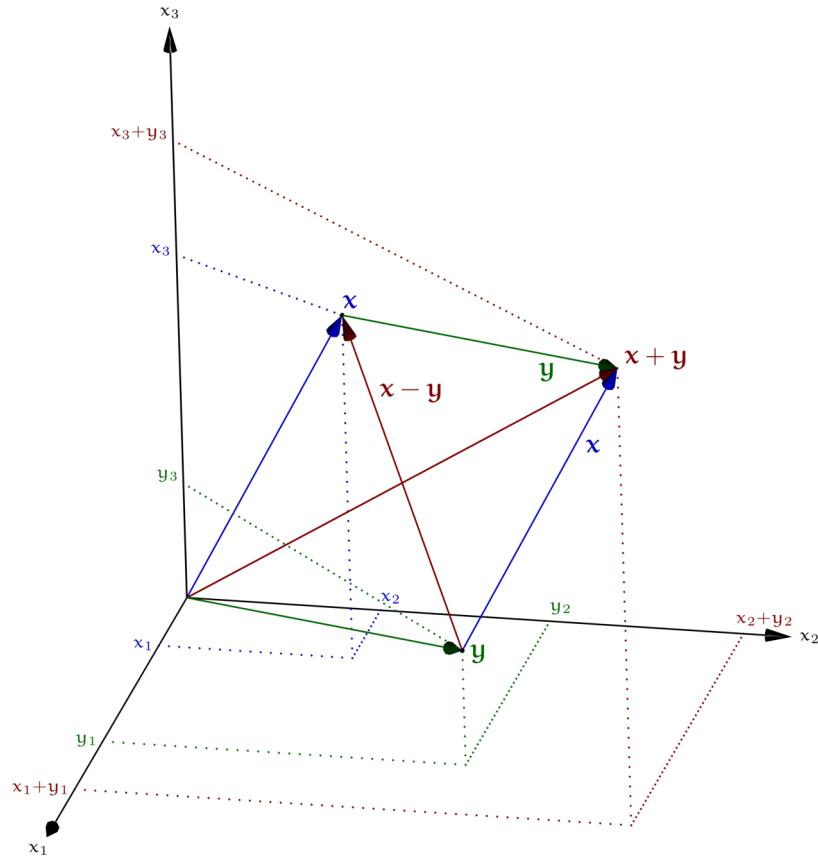


Abb. 5.1 Addition, Subtraktion und Skalarmultiplikation im \mathbb{R}^2

Die Addition $\mathbf{x} + \mathbf{y}$ von \mathbf{x} und \mathbf{y} ist die Hintereinanderausführung der Vorschriften \mathbf{x} und \mathbf{y} . Das heißt, sich x_1 Einheiten für \mathbf{x} und y_1 für \mathbf{y} in Richtung der x_1 -Richtung zu bewegen, insgesamt also um $x_1 + y_1$ Einheiten. Dasselbe gilt für die x_2 -Richtung. Es ergibt sich so die Rechenvorschrift aus (5.1). Graphisch bedeutet das, die gerichtete Strecke \mathbf{y} an das Ende der gerichteten Strecke \mathbf{x} zu fügen. Abb. 5.1 zeigt, daß dieser Vorgang nicht von der Reihenfolge abhängt, daß also $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ gilt. Den Vektor, der bei Y startet und bei X endet bezeichnen wir mit \overrightarrow{YX} . Für ihn gilt $\mathbf{y} + \overrightarrow{YX} = \mathbf{x}$, also $\overrightarrow{YX} = \mathbf{x} - \mathbf{y}$ und insbesondere $\overrightarrow{OX} = \mathbf{x}$. Da X und \overrightarrow{OX} dieselben Koordinaten haben, werden wir meist X mit dem Vektor $\mathbf{x} = \overrightarrow{OX}$ identifizieren und von dem Punkt \mathbf{x} sprechen. Die Skalarmultiplikation $t\mathbf{x}$ eines Vektors \mathbf{x} mit dem Faktor t ist eine Streckung von \mathbf{x} um das t -Fache.

Die Interpretation der Vektoroperationen Addition, Subtraktion und Skalarmultiplikation bleibt in höheren Dimensionen, aber vornehmlich in der dritten, erhalten. Im \mathbb{R}^3 hat man gegenüber \mathbb{R}^2 lediglich eine weitere Richtung zur Verfügung, in die man sich bewegen kann, beschrieben durch die dritte Koordinate x_3 eines Vektors $\mathbf{x} = [x_1, x_2, x_3] \in \mathbb{R}^3$.

Abb. 5.2 Addition und Subtraktion im \mathbb{R}^3

Die allgemeinste Operation, die mit Vektoren durchzuführen ist, besteht in einer Kombination aus Streckung (Skalarmultiplikation) und Addition. D. h., wir werden keinen allgemeineren Konstruktionen, als den sogenannten *Linearkombinationen*

$$t_1x_1 + t_2x_2 + \cdots + t_mx_m \quad (5.3)$$

aus Vektoren x_1, \dots, x_m und Skalaren t_1, \dots, t_m begegnen. In den folgenden Abschnitten stellen wir die grundlegenden Objekte *Gerade*, *Ebene* und *Determinante* im Raum \mathbb{R}^3 vor, wo sie noch sehr anschaulich sind, bevor wir sie später auf höhere Dimensionen verallgemeinern.

5.1.2 Geraden und Ebenen Eine Gerade g durch zwei verschiedene Punkte a und b wird durch diese eindeutig festgelegt. Sie lässt sich durch zwei Vektoren charakterisieren, nämlich

durch den sog. *Stützvektor* $\mathbf{q} := \mathbf{a}$ der Geraden und ihren *Richtungsvektor* $\mathbf{u} := \mathbf{b} - \mathbf{a}$:

$$g := \{ \mathbf{q} + t\mathbf{u} \mid t \in \mathbb{R} \}. \quad (5.4)$$

Man beachte dabei, daß die Gerade, d. h. die Menge g , durch \mathbf{A} und \mathbf{B} eindeutig bestimmt ist, nicht aber die Bestimmungsstücke \mathbf{q} und \mathbf{u} . Für \mathbf{q} hätte man durchaus auch \mathbf{b} wählen können und statt \mathbf{u} z. B. $2\mathbf{u}$ oder $-\mathbf{u}$ (vergl. Abb. 5.3).

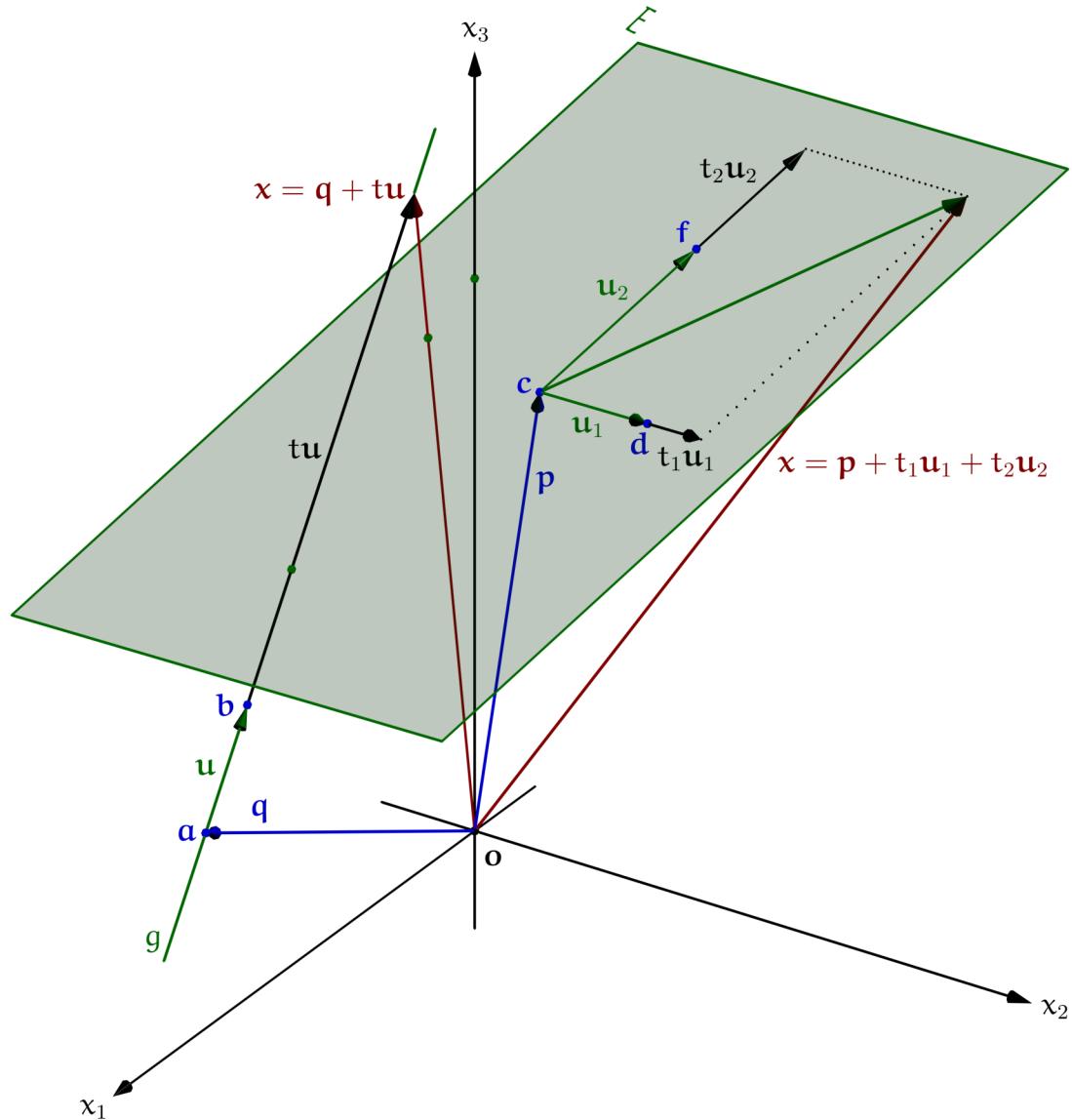


Abb. 5.3 Gerade und Ebene im \mathbb{R}^3

Eine Ebene E wird durch drei verschiedene Punkte c, d und f , die nicht auf einer Geraden liegen, eindeutig bestimmt. Wie für eine Gerade wählt man etwa $\mathbf{p} := \mathbf{c}$ als Stützvektor und die beiden

Richtungsvektoren $\mathbf{u}_1 := \mathbf{d} - \mathbf{c}$, $\mathbf{u}_2 := \mathbf{f} - \mathbf{c}$. Dann ist die sogenannte *Parameterdarstellung* einer Ebene E folgendermaßen definiert:

$$E := \{ \mathbf{p} + t_1 \mathbf{u}_1 + t_2 \mathbf{u}_2 \mid t_1, t_2 \in \mathbb{R} \}. \quad (5.5)$$

5.1.3 Die Norm Unter der *euklidischen Norm* $\|\mathbf{x}\|$ eines Vektors \mathbf{x} (im Folgenden aber einfach als *Norm* bezeichnet) aus \mathbb{R}^2 oder \mathbb{R}^3 wollen wir zunächst die Länge der gerichteten Strecke \mathbf{x} verstehen. Für $\mathbf{x} = [x_1, x_2] \in \mathbb{R}^2$ wird sie also einfach über den Satz des PYTHAGORAS bestimmt:

$$\left\| \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\| := \sqrt{x_1^2 + x_2^2}. \quad (5.6)$$

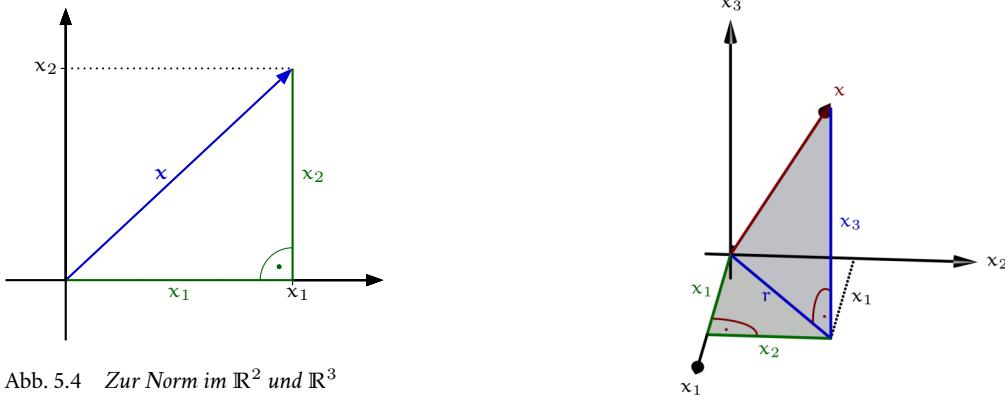


Abb. 5.4 Zur Norm im \mathbb{R}^2 und \mathbb{R}^3

Um die Norm eines Vektors $\mathbf{x} = [x_1, x_2, x_3] \in \mathbb{R}^3$ zu finden, bemühen wir den Satz des PYTHAGORAS zweimal. Für das rechtwinklige Dreieck mit den Katheten r und x_3 (vergl. Abb. 5.4) mit der Hypotenuse $\|\mathbf{x}\|$ erhalten wir $\|\mathbf{x}\|^2 = r^2 + x_3^2$. Das Dreieck mit den Katheten x_1, x_2 und der Hypotenuse r liefert uns den fehlenden Ausdruck $r^2 = x_1^2 + x_2^2$. Eingesetzt ergibt sich die Norm von \mathbf{x} zu

$$\left\| \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \right\| := \sqrt{x_1^2 + x_2^2 + x_3^2}. \quad (5.7)$$

Den Abstand zweier Punkte \mathbf{x} und \mathbf{y} bezeichnen wir mit $d(\mathbf{x}, \mathbf{y})$. Er ist offensichtlich die Länge des Verbindungsvektors $\mathbf{x} - \mathbf{y}$ (vergl. Abb. 5.2):

$$d(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}. \quad (5.8)$$

5.1.4 Skalarprodukt Der Winkel α zwischen Vektoren $\mathbf{x} = [x_1, x_2, x_3]$ und $\mathbf{y} = [y_1, y_2, y_3]$ lässt sich mit Hilfe des *Skalarprodukts*

$$\langle \mathbf{x} | \mathbf{y} \rangle := \|\mathbf{x}\| \|\mathbf{y}\| \cos(\alpha) \quad (5.9)$$

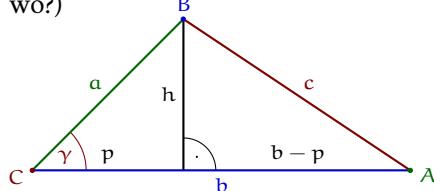
dieser beiden Vektoren bestimmen, wenn wir einen Weg finden, die rechte Seite nur unter Verwendung der Koordinaten x_i und y_i von \mathbf{x} bzw. \mathbf{y} auszurechnen. Dann lässt sich (5.9) nämlich nach $\cos(\alpha)$ auflösen und daraus dann auf α schließen. Der Weg zur Formel für (5.9) verläuft über den Kosinussatz.

5.1.5 Satz (Kosinussatz) Für ein ebenes Dreieck mit den Seiten a, b, c und dem Winkel γ zwischen a und b gilt

$$c^2 = a^2 + b^2 - 2ab \cos(\gamma). \quad (5.10)$$

Für $\gamma = \frac{\pi}{2}$ ergibt sich der Satz des PYTHAGORAS.

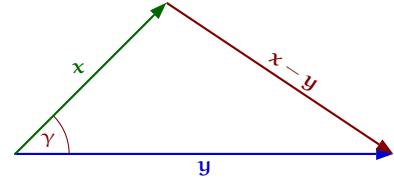
Beweis. Der Skizze entnimmt man $p = a \cos(\gamma)$ und $c^2 = h^2 + (b - p)^2 = a^2 - p^2 + b^2 + p^2 - 2bp = a^2 + b^2 - 2ab \cos(\gamma)$. (Es wurde zweimal der Satz des PYTHAGORAS verwendet, wo?)



Wegen $\cos\left(\frac{\pi}{2}\right) = 0$ ergibt sich der Satz des PYTHAGORAS als Sonderfall. Das ist natürlich nicht weiter überraschend, da wir ihn für den Beweis ja vorausgesetzt haben. \square

Jetzt können wir uns daran machen, eine Rechenmethode für das Skalarprodukt zu finden. Wir wenden den Kosinussatz auf die Situation in nebenstehender Skizze an. Dabei ist $a = \|\mathbf{x}\|$, $b = \|\mathbf{y}\|$ und $c = \|\mathbf{x} - \mathbf{y}\|$:

$$\begin{aligned} \|\mathbf{x} - \mathbf{y}\|^2 &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\|\mathbf{x}\|\|\mathbf{y}\| \cos(\gamma) \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\langle \mathbf{x} | \mathbf{y} \rangle, \end{aligned}$$



also

$$\begin{aligned} 2\langle \mathbf{x} | \mathbf{y} \rangle &= x_1^2 + x_2^2 + x_3^2 + y_1^2 + y_2^2 + y_3^2 - (x_1 - y_1)^2 - (x_2 - y_2)^2 - (x_3 - y_3)^2 \\ &= x_1^2 + x_2^2 + x_3^2 + y_1^2 + y_2^2 + y_3^2 \\ &\quad - x_1^2 + 2x_1 y_1 - y_1^2 - x_2^2 + 2x_2 y_2 - y_2^2 - x_3^2 + 2x_3 y_3 - y_3^2 \\ &= 2x_1 y_1 + 2x_2 y_2 + 2x_3 y_3. \end{aligned}$$

Wir erhalten die einfache Formel

$$\langle \mathbf{x} | \mathbf{y} \rangle = x_1 y_1 + x_2 y_2 + x_3 y_3. \quad (5.11)$$

Natürlich entfällt der dritte Summand $x_3 y_3$, wenn die Vektoren aus \mathbb{R}^2 sind. Wählen wir für \mathbf{y} den Vektor \mathbf{x} , dann zeigt (5.7) den engen Zusammenhang des Skalarprodukts mit der Norm:

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x} | \mathbf{x} \rangle}. \quad (5.12)$$

Mit Hilfe der Gleichung (5.11) ist es nun eine leichte Übungsaufgabe, die Eigenschaften

$$\text{i)} \quad \langle \mathbf{x} | \mathbf{x} \rangle \geq 0, = 0 \iff \mathbf{x} = \mathbf{0}, \quad (\text{Definitheit}) \quad (5.13)$$

$$\text{ii)} \quad \langle \mathbf{x} | \mathbf{y} \rangle = \langle \mathbf{y} | \mathbf{x} \rangle, \quad (\text{Symmetrie}) \quad (5.14)$$

$$\text{iii)} \quad \langle t\mathbf{x} + s\mathbf{y} | \mathbf{z} \rangle = t\langle \mathbf{x} | \mathbf{z} \rangle + s\langle \mathbf{y} | \mathbf{z} \rangle \quad (\text{Linearität}) \quad (5.15)$$

nachzurechnen, die den Namen Skalarprodukt rechtfertigen. Diese Rechenregeln sind in der ursprünglichen Definition (5.9) kaum zu erkennen.

5.1.6 Beispiel Es sei $\mathbf{x} := [8, -4, 1]$, $\mathbf{y} := [2, 5, 14]$ und $\mathbf{z} := [4, 7, -4]$. Dann ist gemäß (5.7) $\|\mathbf{x}\| = \sqrt{64 + 16 + 1} = \sqrt{81} = 9$, $\|\mathbf{y}\| = \sqrt{4 + 25 + 196} = \sqrt{225} = 15$ und $\|\mathbf{z}\| = 9$. Das Skalarprodukt ergibt sich nach (5.11) zu $\langle \mathbf{x} | \mathbf{y} \rangle = 16 - 20 + 14 = 10$. Nun lösen wir (5.9) nach $\cos(\alpha)$ auf:

$$\cos(\alpha) = \frac{\langle \mathbf{x} | \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{10}{9 \cdot 15} = \frac{2}{27}.$$

Wir erhalten $\alpha = \cos^{-1}(\frac{2}{27}) \approx 1.497$, oder $\alpha \approx 1.497 \cdot \frac{180^\circ}{\pi} \approx 85.75^\circ$, wenn wir das Ergebnis statt in Bogenmaß lieber in Grad angeben wollen.

Für den Winkel β zwischen \mathbf{x} und \mathbf{z} bestimmen wir

$$\langle \mathbf{x} | \mathbf{z} \rangle = 32 - 28 - 4 = 0.$$

In diesem Fall brauchen wir die Formel für $\cos(\beta)$ gar nicht mehr bemühen, denn der Kosinus wird in dem uns ausschließlich interessierenden Bereich $\beta \in [0, \pi]$ nur an der Stelle $\beta = \frac{\pi}{2}$ Null. Die beiden Vektoren \mathbf{x} und \mathbf{z} schließen also einen rechten Winkel ein, d. h. sie stehen senkrecht aufeinander. Man sagt auch \mathbf{x} und \mathbf{z} sind *orthogonal* und schreibt dafür mitunter $\mathbf{x} \perp \mathbf{z}$:

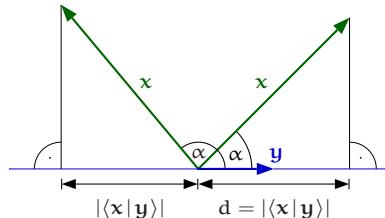
$$\mathbf{x} \perp \mathbf{z} : \iff \langle \mathbf{x} | \mathbf{z} \rangle = 0. \quad (5.16)$$

5.1.7 Geometrische Interpretation des Skalarprodukts Ist einer der beiden Vektoren \mathbf{x} oder \mathbf{y} , sagen wir \mathbf{y} , *normiert*, d. h. gilt $\|\mathbf{y}\| = 1$, dann lässt sich der Betrag $|\langle \mathbf{x} | \mathbf{y} \rangle|$ des Skalarprodukts als Länge der Projektion von \mathbf{x} auf die Richtung \mathbf{y} interpretieren.

An der Skizze sehen wir $|\cos(\alpha)| = \frac{d}{\|\mathbf{x}\|}$. Dabei ist d die Länge der Projektion von \mathbf{x} auf die Richtung \mathbf{y} . Wegen $\|\mathbf{y}\| = 1$ folgt zusammen mit (5.9)

$$d = \|\mathbf{x}\| |\cos(\alpha)| = \|\mathbf{x}\| \|\mathbf{y}\| |\cos(\alpha)| = |\langle \mathbf{x} | \mathbf{y} \rangle|.$$

Wenn wir den Betrag in $|\langle \mathbf{x} | \mathbf{y} \rangle|$ weglassen, dann sagt uns das Vorzeichen von $\langle \mathbf{x} | \mathbf{y} \rangle$ auch noch die Richtung der Projektion relativ zur Richtung von \mathbf{y} . Ist es positiv, dann liegt sie in der Richtung in die \mathbf{y} zeigt. Ist es negativ, dann liegt sie in der entgegengesetzten Richtung, denn $\cos(\alpha)$ ist für $\frac{\pi}{2} < \alpha < \pi$ negativ.



5.1.8 Satz (CAUCHY-SCHWARZ-Ungleichung) Für alle Vektoren $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$ gilt

$$|\langle \mathbf{x} | \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|. \quad (5.17)$$

Das Gleichheitszeichen gilt dabei genau dann, wenn \mathbf{x} und \mathbf{y} parallel sind.

Beweis. Die Ungleichung (5.17) ist eine einfache Konsequenz aus der Definition (5.9) des Skalarprodukts und der Tatsache $|\cos(\alpha)| \leq 1$: $|\langle \mathbf{x} | \mathbf{y} \rangle| = \|\mathbf{x}\| \|\mathbf{y}\| |\cos(\alpha)| \leq \|\mathbf{x}\| \|\mathbf{y}\|$. Falls \mathbf{x} und \mathbf{y} parallel sind, ist $\alpha = 0$ oder $\alpha = \pi$ und $\cos(\alpha) = 1$ bzw. $\cos(\alpha) = -1$, also $|\cos(\alpha)| = 1$ und damit $|\langle \mathbf{x} | \mathbf{y} \rangle| = \|\mathbf{x}\| \|\mathbf{y}\|$. Gehen wir andererseits von dieser Gleichheit aus, dann muß $|\cos(\alpha)| = 1$ gelten, was nur bei $\alpha = 0$ oder $\alpha = \pi$ vorkommt. Für diese Winkel sind \mathbf{x} und \mathbf{y} parallel. \square

5.1.9 Das Kreuzprodukt Anders als das Skalarprodukt in \mathbb{R}^3 , das aus zwei Vektoren \mathbf{x} und \mathbf{y} eine Zahl $\langle \mathbf{x} | \mathbf{y} \rangle$ aus \mathbb{R} (einen *Skalar*) macht, gibt es im \mathbb{R}^3 ein weiteres Produkt, daß als Ergebnis einen Vektor aus \mathbb{R}^3 liefert. Es wird als *Kreuzprodukt* $\mathbf{x} \times \mathbf{y}$ bezeichnet und hat die schöne Eigenschaft (neben weiteren), einen Vektor zu erzeugen, der senkrecht auf den beiden Ausgangsvektoren steht. Ein Vektor \mathbf{n} , der senkrecht auf $\mathbf{x} = [x_1, x_2, x_3]$ und senkrecht auf $\mathbf{y} = [y_1, y_2, y_3]$ steht, muß die beiden Gleichungen $\langle \mathbf{x} | \mathbf{n} \rangle = x_1 n_1 + x_2 n_2 + x_3 n_3 = 0$ und $\langle \mathbf{y} | \mathbf{n} \rangle = y_1 n_1 + y_2 n_2 + y_3 n_3 = 0$ erfüllen. Da wir das GAUSS-Verfahren zur Lösung eines solchen *linearen Gleichungssystems* noch nicht zur Verfügung haben, (siehe 6.1, Seite 99) geben wir das Ergebnis einfach an und rechnen dafür nach, daß es die gewünschten Eigenschaften aufweist:

$$\mathbf{x} \times \mathbf{y} := \begin{bmatrix} x_2 y_3 - x_3 y_2 \\ x_3 y_1 - x_1 y_3 \\ x_1 y_2 - x_2 y_1 \end{bmatrix}. \quad (5.18)$$

5.1.10 Satz Das Kreuzprodukt $\mathbf{x} \times \mathbf{y}$ hat folgende Eigenschaften:

$$(\mathbf{t}\mathbf{x} + \mathbf{s}\mathbf{y}) \times \mathbf{z} = \mathbf{t}\mathbf{x} \times \mathbf{z} + \mathbf{s}\mathbf{y} \times \mathbf{z} \quad (5.19)$$

$$\mathbf{x} \times \mathbf{y} = -\mathbf{y} \times \mathbf{x}, \quad (5.20)$$

$$\mathbf{x} \times \mathbf{y} = \mathbf{0} \iff \mathbf{x} \parallel \mathbf{y} \quad (5.21)$$

$$\mathbf{x} \times \mathbf{y} \perp \mathbf{x}, \quad \mathbf{x} \times \mathbf{y} \perp \mathbf{y}, \quad (5.22)$$

$$\|\mathbf{x} \times \mathbf{y}\| = \|\mathbf{x}\| \|\mathbf{y}\| \sin(\alpha). \quad (5.23)$$

α ist der Winkel, den \mathbf{x} und \mathbf{y} einschließen. $\|\mathbf{x} \times \mathbf{y}\|$ ist der Flächeninhalt des Parallelogramms, das von \mathbf{x} und \mathbf{y} aufgespannt wird.

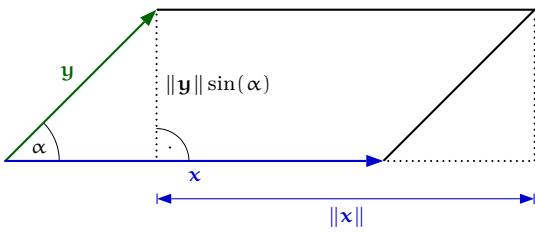
Beweis: (5.19) und (5.20) sind simple Übungen. (5.21): $\mathbf{x} \parallel \mathbf{y}$ bedeutet, daß \mathbf{x} ein Vielfaches von \mathbf{y} ist: $\mathbf{x} = t\mathbf{y}$. Es folgt $\mathbf{x} \times \mathbf{y} = t\mathbf{x} \times \mathbf{x} = -t\mathbf{x} \times \mathbf{x} = -\mathbf{x} \times \mathbf{y}$, also $\mathbf{x} \times \mathbf{y} = \mathbf{0}$. Dabei haben wir (5.20) auf $\mathbf{x} \times \mathbf{x}$ angewandt. Für die Umkehrung verwenden wir (5.23). Aus $\|\mathbf{x} \times \mathbf{y}\| = 0$ folgt dann $\sin(\alpha) = 0$, also $\alpha = 0$ und daraus $\mathbf{x} \parallel \mathbf{y}$. (5.22) ist eine einfache Rechenaufgabe:

$$\begin{aligned} \langle \mathbf{x} \times \mathbf{y} | \mathbf{x} \rangle &= (x_2 y_3 - x_3 y_2) x_1 + (x_3 y_1 - x_1 y_3) x_2 + (x_1 y_2 - x_2 y_1) x_3 \\ &= x_2 y_3 x_1 - x_3 y_2 x_1 + x_3 y_1 x_2 - x_1 y_3 x_2 + x_1 y_2 x_3 - x_2 y_1 x_3 = 0. \end{aligned}$$

$\langle \mathbf{x} \times \mathbf{y} | \mathbf{y} \rangle = -\langle \mathbf{y} \times \mathbf{x} | \mathbf{y} \rangle = 0$. Das zeigt (5.22). (5.23) ist aufwendiger (vergl. Übung 7.4.6):

$$\begin{aligned} \|\mathbf{x} \times \mathbf{y}\|^2 &= (x_2 y_3 - x_3 y_2)^2 + (x_3 y_1 - x_1 y_3)^2 + (x_1 y_2 - x_2 y_1)^2 \\ &= x_2^2 y_3^2 + x_3^2 y_2^2 - 2x_2 x_3 y_2 y_3 + x_3^2 y_1^2 + x_1^2 y_2^2 - 2x_1 x_2 y_1 y_2 \end{aligned}$$

$$\begin{aligned}
& + x_1^2 y_2^2 + x_2^2 y_1^2 - 2x_1 x_2 y_1 y_2 + x_1^2 y_1^2 + x_2^2 y_2^2 + x_3^2 y_3^2 - x_1^2 y_1^2 - x_2^2 y_2^2 - x_3^2 y_3^2 \\
& = (x_1^2 + x_2^2 + x_3^2)(y_1^2 + y_2^2 + y_3^2) - (x_1 y_1 + x_2 y_2 + x_3 y_3)^2 \\
& = \|x\|^2 \|y\|^2 - \langle x | y \rangle^2 = \|x\|^2 \|y\|^2 (1 - \cos^2(\alpha)) = \|x\|^2 \|y\|^2 \sin^2(\alpha).
\end{aligned}$$



Da $\sin(\alpha)$ in dem uns interessierenden Bereich $\alpha \in [0, \pi]$ nicht negativ ist, können wir aus obiger Gleichung auf beiden Seiten die Wurzel ziehen und erhalten (5.23). Die Interpretation von $\|x\| \|y\| \sin(\alpha)$ als Flächeninhalt des Parallelogramms, das von x und y erzeugt wird, sieht man an der Skizze. \square

Die etwas sperrige Formel (5.18) lässt sich durch das folgende grafische Verfahren leicht merken: Die Koordinaten von x und die von y werden jeweils zweimal untereinander geschrieben und dann wie folgt über Kreuz multipliziert:

$$\left[\begin{array}{cc} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \end{array} \right] : \left[\begin{array}{c} x_2 y_3 - y_2 x_3 \\ x_3 y_1 - y_3 x_1 \\ x_1 y_2 - y_1 x_2 \end{array} \right] = x \times y. \quad (5.24)$$

5.1.11 Die Normalenform der Ebene Eine Ebene $E \subset \mathbb{R}^3$ lässt sich durch einen Punkt $p \in E$ und einen Vektor n , der in eine zu E senkrechten Richtung zeigt, charakterisieren. Genauer: n steht für alle $x, y \in E$ senkrecht auf $x - y$. n ist der sogenannte *Normalenvektor* von E . Er ist bis auf ein Vielfaches durch die Ebene festgelegt. Verlangen wir, daß n die Länge 1 hat, so machen wir das gelegentlich durch die Schreibweise n_0 deutlich. Für einen weiteren Vektor $x \in E$ gilt also

$$\langle x - p | n_0 \rangle = \langle x | n_0 \rangle - \langle p | n_0 \rangle = 0.$$

Für alle Vektoren $x \in E$ ist demnach das Skalarprodukt mit dem Normalenvektor n_0 die feste Zahl $d_0 := \langle p | n_0 \rangle$. Gemäß 5.1.7 handelt es sich, bis auf ein eventuelles Vorzeichen, um die Länge der Projektion von x auf die Richtung n_0 . Laut Abb. 5.5 ist das aber gerade der Abstand der Ebene E vom Ursprung O (durch Änderung des Vorzeichens von n_0 kann immer $d_0 \geq 0$ erreicht werden). Also liegen alle Punkte von E in der Menge $\{x \in \mathbb{R}^3 \mid \langle x | n_0 \rangle = d_0\}$. Abb. 5.5 zeigt aber auch, daß alle Vektoren x mit derselben Projektion d_0 in Richtung n_0 zu E gehören (wir wollen es hier mit der Anschauung bewenden lassen, da ein strenger Beweis eine genauere Kenntnis der Begriffe *Dimension* und *Basis* erfordert, die uns im Augenblick noch nicht zur Verfügung stehen). Damit haben wir die Normalenform einer Ebene gefunden, die einen vorgegebenen Punkt p enthält und den Normalenvektor n_0 der Länge 1 hat:

$$E = \{x \in \mathbb{R}^3 \mid \langle x | n_0 \rangle = d_0\}, \quad d_0 = \langle p | n_0 \rangle, \quad \|n_0\| = 1. \quad (5.25)$$

Man kann in (5.25) auf die Forderung $\|n_0\| = 1$ verzichten, wenn man die Interpretation von d_0 als Abstand der Ebene zum Ursprung nicht benötigt. Aus Abb. 5.5 lässt sich auch eine Formel

für den Abstand $d(\mathbf{y}, E)$ eines Punktes \mathbf{y} von der Ebene E ablesen. $\langle \mathbf{y} | \mathbf{n}_0 \rangle$ ist die Projektion von \mathbf{y} auf die Richtung \mathbf{n}_0 . Ziehen wir davon den Abstand d_0 ab, so erhalten wir, bis auf ein mögliches Vorzeichen (warum?) den Abstand $d(\mathbf{y}, E)$. Wir haben damit die sogenannte **HESSE-Form** für den Abstand eines Punktes \mathbf{y} von der Ebene E gefunden. Für praktische Rechnungen möchte man sich gern von der unbequemen Forderung $\|\mathbf{n}_0\| = 1$ befreien können. Die Ebenengleichung lautet dann für einen beliebigen Normalenvektor \mathbf{n} : $\langle \mathbf{y} | \mathbf{n} \rangle = d$, mit einer festen Zahl d , die i. Allg. nicht mehr als Abstand interpretierbar ist. Um den Anschluß an (5.25) zu finden, muß \mathbf{n} normiert werden, was einfach durch $\mathbf{n}_0 = \frac{1}{\|\mathbf{n}\|} \mathbf{n}$ erreicht wird. Die zugehörige Ebenengleichung lautet jetzt $\langle \mathbf{y} | \mathbf{n}_0 \rangle = \frac{d}{\|\mathbf{n}\|} = d_0$. Setzen wir das in (5.26)

$$d(\mathbf{y}, E) = |\langle \mathbf{y} | \mathbf{n}_0 \rangle - d_0| \quad (5.26)$$

ein: $d(\mathbf{y}, E) = |\langle \mathbf{y} | \frac{\mathbf{n}}{\|\mathbf{n}\|} \rangle - \frac{d}{\|\mathbf{n}\|}|$, also

$$d(\mathbf{y}, E) = \frac{1}{\|\mathbf{n}\|} |\langle \mathbf{y} | \mathbf{n} \rangle - d|. \quad (5.27)$$

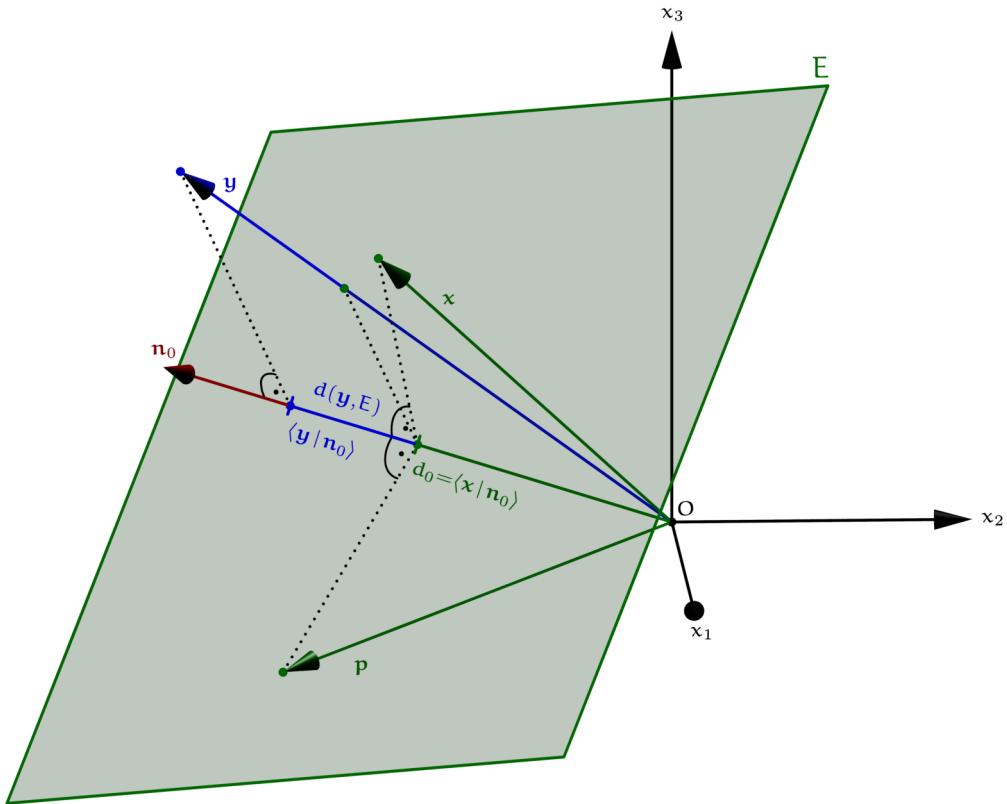


Abb. 5.5 Zur Normalenform einer Ebene

5.1.12 Beispiel Wir gehen von drei Punkten \mathbf{p} , \mathbf{q} und \mathbf{r} aus, bestimmen die Normalenform der Ebene E , die diese Punkte enthält und berechnen anschließend den Abstand eines weiteren Punktes \mathbf{s} von E .

$$\mathbf{p} := \frac{1}{4}[5, -8, -2], \mathbf{q} := \frac{1}{4}[12, 1, 2], \mathbf{r} := \frac{1}{4}[3, -6, 6], \mathbf{s} := \frac{1}{2}[2, -5, 4].$$

Wir wählen \mathbf{p} als Stützvektor, $\mathbf{u}_1 := 4(\mathbf{q} - \mathbf{p}) = [7, 9, 4]$ und $\mathbf{u}_2 := 2(\mathbf{r} - \mathbf{p}) = [-1, 1, 4]$ als Richtungsvektoren. Die Parameterdarstellung von E lautet damit

$$E = \left\{ \frac{1}{4} \begin{bmatrix} 5 \\ -8 \\ -2 \end{bmatrix} + t_1 \begin{bmatrix} 7 \\ 9 \\ 4 \end{bmatrix} + t_2 \begin{bmatrix} -1 \\ 1 \\ 4 \end{bmatrix} \mid t_1, t_2 \in \mathbb{R} \right\}.$$

Der Normalenvektor ist ein geeignetes Vielfaches des Kreuzprodukts $\mathbf{u}_1 \times \mathbf{u}_2$:

$$\begin{bmatrix} 7 & -1 \\ 9 & 1 \\ 4 & 4 \\ 4 & -1 \\ 7 & 1 \\ 9 & 1 \\ 4 & 4 \end{bmatrix} : \begin{bmatrix} 36 & - & 4 \\ -4 & - & 28 \\ 7 & + & 9 \end{bmatrix} = \begin{bmatrix} 32 \\ -32 \\ 16 \end{bmatrix}. \text{ Wir wählen also } \mathbf{n} := \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}.$$

Die Ebenengleichung lautet damit $\langle \mathbf{x} | \mathbf{n} \rangle = \langle \mathbf{q} | \mathbf{n} \rangle = 6 - 2 + 2 = 6$, oder ausgeschrieben in der sogenannten *Koordinatenform*: $2x_1 - 2x_2 + x_3 = 6$. Das ist die Version, mit der üblicherweise gerechnet wird. Wir haben jetzt $E = \{\mathbf{x} \in \mathbb{R}^3 \mid 2x_1 - 2x_2 + x_3 = 6\}$. Die HESSE-Form ergibt sich aus $|\langle \mathbf{y} | \mathbf{n} \rangle - 6|$ einfach durch Teilen mit $\|\mathbf{n}\| = 3$:

$$d(\mathbf{y}, E) = \frac{1}{3} \cdot |2y_1 - 2y_2 + y_3 - 6|.$$

Also ist $d(\mathbf{s}, E) = \frac{1}{3} \cdot |2 + 5 + 2 - 6| = 1$.

5.1.13 Spatprodukt und Determinante Wir haben für das Skalarprodukt und das Kreuzprodukt geometrische Interpretationen gefunden. Daher sollte auch eine für die Kombination $\langle \mathbf{x} \times \mathbf{y} | \mathbf{z} \rangle$ dieser beiden Produkte möglich sein. Wir werden gleich sehen, daß es sich dabei im Wesentlichen um das Volumen des *Spat* handelt, der von den Vektoren \mathbf{x} , \mathbf{y} und \mathbf{z} aufgespannt wird (Abb. 5.6). Daher röhrt der Name *Spatprodukt* für $\langle \mathbf{x} \times \mathbf{y} | \mathbf{z} \rangle$.

Wir wissen, daß $\|\mathbf{x} \times \mathbf{y}\|$ den Inhalt der Grundfläche des Spats angibt (5.1.7).

Die Projektion von \mathbf{z} auf $\mathbf{n}_0 := \frac{\mathbf{x} \times \mathbf{y}}{\|\mathbf{x} \times \mathbf{y}\|}$ hat die Länge $h := \|\mathbf{h}\| = |\langle \mathbf{n}_0 | \mathbf{z} \rangle| = \frac{1}{\|\mathbf{x} \times \mathbf{y}\|} |\langle \mathbf{x} \times \mathbf{y} | \mathbf{z} \rangle|$. Um also $\langle \mathbf{x} \times \mathbf{y} | \mathbf{z} \rangle$ interpretieren zu können, müssen wir zunächst $h \cdot \|\mathbf{x} \times \mathbf{y}\|$ verstehen. Das ist allerdings leicht, denn es handelt sich einfach um das Volumen des senkrechten Prismas über der Grundfläche des Spats (die gestrichelten Linien in Abb. 5.6). Worüber man etwas nachdenken muß, ist, daß es auch das Volumen des Spats ist.

Die Idee besteht darin, den Spat über volumengleiche Körper in das Prisma zu verwandeln. Im ersten Schritt wird vom Spat entlang der Fläche $\mathbf{x} \mathbf{a} \mathbf{b} \mathbf{c}$ ein Keil abgetrennt, der an der gegenüberliegenden Seite wieder angesetzt wird (die gepunkteten Linien in Abb. 5.6).

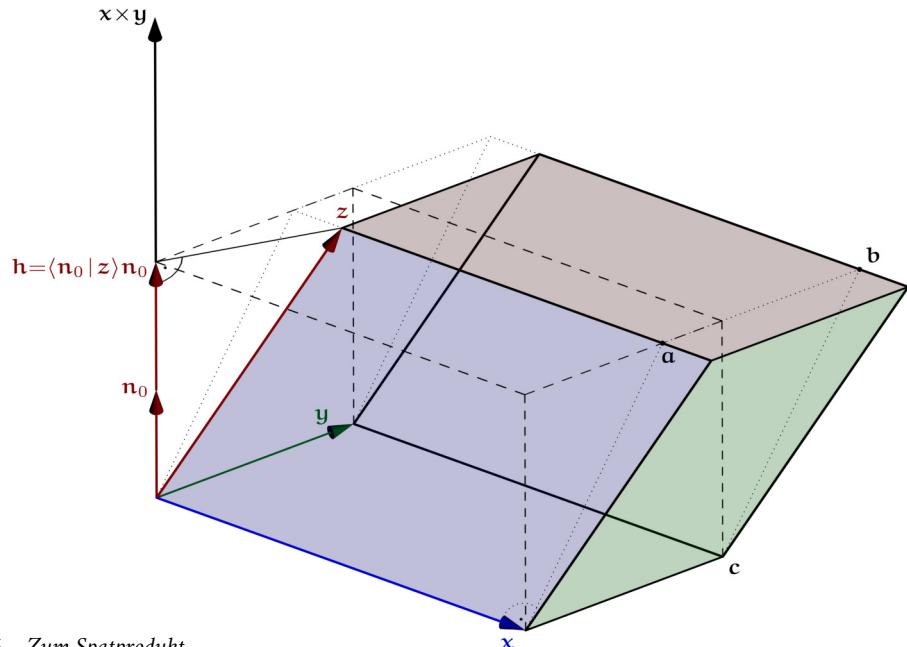


Abb. 5.6 Zum Spatprodukt

Im zweiten Schritt wiederholt man diesen Vorgang entlang der \mathbf{y} -Richtung und gelangt so zum Prisma. Man überlege sich die nötigen Zwischenschritte, falls der Spat sehr viel schiefer ist, als in Abb. 5.6.

Als Zwischenergebnis haben wir, daß $|\langle \mathbf{x} \times \mathbf{y} | \mathbf{z} \rangle|$ das Volumen des Spats wiedergibt, der von den Vektoren \mathbf{x} , \mathbf{y} und \mathbf{z} aufgespannt wird. Da das Skalarprodukt auch negativ werden kann, muß $\langle \mathbf{x} \times \mathbf{y} | \mathbf{z} \rangle$ noch eine Information enthalten, die über das Volumen hinaus geht.

In Abb. 5.6 ist das Spatprodukt positiv. Ersetzen wir \mathbf{z} durch $-\mathbf{z}$, so erhalten wir einen Spat gleichen Volumens. Der Ausdruck $\langle \mathbf{x} \times \mathbf{y} | -\mathbf{z} \rangle$ wird jetzt aber negativ. Vergleicht man die Lage der sogenannten Dreibeine $\mathbf{x}, \mathbf{y}, \mathbf{z}$ und $\mathbf{x}, \mathbf{y}, -\mathbf{z}$, so erkennt man, daß sie durch keine Bewegung im Raum zur Deckung gebracht werden können (ausprobieren). Ordnet man den Daumen, Zeigefinger und Ringfinger der rechten Hand so zueinander an, daß sie nicht in einer Ebene liegen, dann läßt sie sich so drehen, daß \mathbf{x} in Richtung des Daumens, \mathbf{y} in Richtung des Zeigefingers und \mathbf{z}

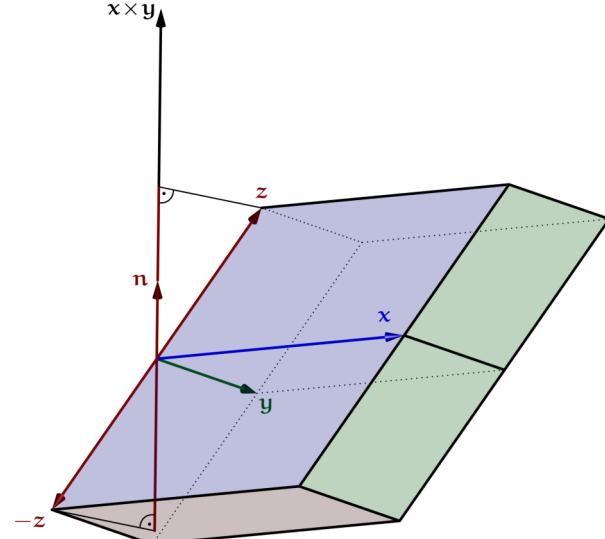


Abb. 5.7 Orientierung

in Richtung des Mittelfingers zeigt. Für $\mathbf{x}, \mathbf{y}, -\mathbf{z}$ ist das nicht möglich. Man sagt, das Dreibein $\mathbf{x}, \mathbf{y}, \mathbf{z}$ gehorcht der *Rechte-Hand-Regel*. Das Spatprodukt kann offensichtlich zwischen diesen beiden Situationen unterscheiden. Man vereinbart daher, daß drei Vektoren (*mathematisch*) *positiv orientiert* heißen, wenn ihr Spatprodukt positiv ist und andernfalls (*mathematisch*) *negativ orientiert*. Die Verallgemeinerung dieser Überlegungen auf höhere Dimensionen führt auf den Begriff der *Determinante*. Daher werden wir für das Spatprodukt auch die Schreibweise $\det(\mathbf{x}, \mathbf{y}, \mathbf{z})$ verwenden und von der Determinante der Vektoren $\mathbf{x}, \mathbf{y}, \mathbf{z}$ sprechen.

Vertauschen wir die Reihenfolge der Vektoren \mathbf{x}, \mathbf{y} und \mathbf{z} , so ändert sich dabei offensichtlich nicht das Volumen des Spats, da dieses ja nicht von der Abfolge abhängen kann, in der wir die Vektoren aufzählen. Allerdings kann sich die Orientierung ändern, was am Vorzeichen der Determinante erkennbar ist. So ist etwa $\det(\mathbf{y}, \mathbf{x}, \mathbf{z}) = \langle \mathbf{y} \times \mathbf{x} | \mathbf{z} \rangle = -\langle \mathbf{x} \times \mathbf{y} | \mathbf{z} \rangle = -\det(\mathbf{y}, \mathbf{x}, \mathbf{z})$, denn das Kreuzprodukt ändert sein Vorzeichen, wenn die Reihenfolge der Vektoren vertauscht wird (5.20). $\det(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \det(\mathbf{z}, \mathbf{x}, \mathbf{y})$ wird durch unsere Überlegungen zum Volumen nahegelegt, muß aber streng genommen nachgerechnet werden, denn diese Gleichung stellt auch die Behauptung über ein gleiches Vorzeichen auf. Wir haben oben zwar die Rechte-Hand-Regel eingeführt und $\mathbf{x}, \mathbf{y}, \mathbf{z}$ sowie $\mathbf{z}, \mathbf{x}, \mathbf{y}$ gehorchen derselben Regel (ausprobieren), wir müssen uns aber immer noch davon überzeugen, daß sich die Ausdrücke $\langle \mathbf{x} \times \mathbf{y} | \mathbf{z} \rangle$ und $\langle \mathbf{z} \times \mathbf{x} | \mathbf{y} \rangle$ auch tatsächlich daran halten.

Abb. 5.6 zeigt einen Spat mit positiver Orientierung. Die Vektoren $\mathbf{x}, \mathbf{y}, \mathbf{z}$ gehorchen der Rechte-Hand-Regel und das Skalarprodukt $\langle \mathbf{x} \times \mathbf{y} | \mathbf{z} \rangle$ ist positiv. Nach 5.9 muß also $\mathbf{x} \times \mathbf{y}$ und \mathbf{z} in dieselbe Richtung zeigen. Das bedeutet, daß \mathbf{x}, \mathbf{y} und $\mathbf{x} \times \mathbf{y}$ ebenfalls der Rechte-Hand-Regel genügen müssen.

5.1.14 Satz *Das Spatprodukt $\det(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \langle \mathbf{x} \times \mathbf{y} | \mathbf{z} \rangle$ ist, bis auf ein mögliches Vorzeichen, das Volumen des Spats, der von \mathbf{x}, \mathbf{y} und \mathbf{z} aufgespannt wird. Das Vorzeichen gibt die Orientierung der drei Vektoren an. Es gilt für alle $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{w} \in \mathbb{R}^3$ und $t, s \in \mathbb{R}$:*

$$\det \begin{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \end{bmatrix} = x_1 y_2 z_3 + x_3 y_1 z_2 + x_2 y_3 z_1 - x_3 y_2 z_1 - x_1 y_3 z_2 - x_2 y_1 z_3, \quad (5.28)$$

$$\begin{aligned} \det(\mathbf{x}, \mathbf{y}, \mathbf{z}) &= \det(\mathbf{z}, \mathbf{x}, \mathbf{y}) = \det(\mathbf{y}, \mathbf{z}, \mathbf{x}) = \\ &\quad -\det(\mathbf{x}, \mathbf{z}, \mathbf{y}) = -\det(\mathbf{z}, \mathbf{y}, \mathbf{x}) = -\det(\mathbf{y}, \mathbf{x}, \mathbf{z}), \end{aligned} \quad (5.29)$$

$$\det(\mathbf{x}, \mathbf{y}, t\mathbf{v} + s\mathbf{w}) = t \det(\mathbf{x}, \mathbf{y}, \mathbf{v}) + s \det(\mathbf{x}, \mathbf{y}, \mathbf{w}), \quad (5.30)$$

$$\det(\mathbf{x}, \mathbf{y}, \mathbf{z} + t\mathbf{x} + s\mathbf{y}) = \det(\mathbf{x}, \mathbf{y}, \mathbf{z}). \quad (5.31)$$

$\det(\mathbf{x}, \mathbf{y}, \mathbf{z}) = 0$ gilt genau dann, wenn \mathbf{x}, \mathbf{y} und \mathbf{z} in einer Ebene durch den Ursprung liegen.

Beweis. Gleichung (5.28) ist simples Ausrechnen des Spatprodukts. Mit dieser Gleichung kann man sich dann leicht von (5.29) überzeugen. Da alle Faktoren x_i, y_j und z_k in den Produkten $x_i y_j z_k$ linear vorkommen (d.h. nicht mit höheren Potenzen als 1), ist (5.30) nur eine leichte Fleißaufgabe. Man sagt, die Determinante ist in der letzten Komponente linear. Für (5.31) genügt es daher, wenn wir uns mit $\det(\mathbf{x}, \mathbf{y}, \mathbf{z} + \mathbf{x})$ beschäftigen. Wir erhalten $\det(\mathbf{x}, \mathbf{y}, \mathbf{z}) +$

$\det(\mathbf{x}, \mathbf{y}, \mathbf{x})$. Für den zweiten Summanden gilt aber nach (5.29) $\det(\mathbf{x}, \mathbf{y}, \mathbf{x}) = -\det(\mathbf{x}, \mathbf{y}, \mathbf{x})$, woraus natürlich $\det(\mathbf{x}, \mathbf{y}, \mathbf{x}) = 0$ folgt. Also haben wir $\det(\mathbf{x}, \mathbf{y}, \mathbf{z} + \mathbf{x}) = \det(\mathbf{x}, \mathbf{y}, \mathbf{z})$ gezeigt. Die letzte Behauptung ist im Lichte der Volumeninterpretation der Determinante klar. \square

Ist $\det(\mathbf{x}, \mathbf{y}, \mathbf{z}) = 0$, so nennt man $\{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ linear abhängig, andernfalls linear unabhängig. Natürlich gelten Gleichungen wie (5.30) und (5.31) auch für die erste und die zweite Variable von $\det(\mathbf{x}, \mathbf{y}, \mathbf{z})$. Das ist eine direkte Konsequenz von (5.29).

Niemand berechnet eine Determinante gemäß (5.28). Genau wie beim Kreuzprodukt gibt es eine einfache grafische Merkregel:

$$\begin{bmatrix} x_1 & y_1 & z_1 & x_1 & y_1 \\ x_2 & y_2 & z_2 & x_2 & y_2 \\ x_3 & y_3 & z_3 & x_3 & y_3 \end{bmatrix} : \quad \begin{aligned} &x_1 y_2 z_3 + y_1 z_2 x_3 + z_1 x_2 y_3 \\ &- x_3 y_2 z_1 - y_3 z_2 x_1 - z_3 x_2 y_1 = \det(\mathbf{x}, \mathbf{y}, \mathbf{z}). \end{aligned} \quad (5.32)$$

5.1.15 Die Determinante für den \mathbb{R}^2

Hier soll sie, wie im \mathbb{R}^3 , einen Inhalt messen, nur daß es sich jetzt um den Flächeninhalt eines Parallelogramms handelt, das von zwei Vektoren $\mathbf{x} = [x_1, x_2]$ und $\mathbf{y} = [y_1, y_2]$ aufgespannt wird. Mit einem kleinen Trick profitieren wir dabei von der Arbeit, die wir uns für den Fall \mathbb{R}^3 gemacht haben. Wir betten die Vektoren in den \mathbb{R}^3 ein, indem wir ihnen eine

weitere Koordinate 0 hinzufügen: $\mathbf{x} = [x_1, x_2, 0]$ und $\mathbf{y} = [y_1, y_2, 0]$. Als dritten Vektor wählen wir $\mathbf{z} := [0, 0, 1]$ und erhalten durch den Aufspann von \mathbf{x} , \mathbf{y} und \mathbf{z} ein Prisma der Höhe 1 und dem Parallelogramm als Grundfläche. Sein Volumen $\det(\mathbf{x}, \mathbf{y}, \mathbf{z})$ hat daher dieselbe Maßzahl wie der Flächeninhalt des Parallelogramms (Abb. 5.8). In (5.28) eingesetzt bleiben nur die Summanden übrig, die keinen Faktor z_1 oder z_2 aufweisen, also $x_1 y_2 - x_2 y_1$. Das ist die richtige Definition der Determinante im \mathbb{R}^2 :

$$\det \left[\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right] := x_1 y_2 - x_2 y_1. \quad (5.33)$$

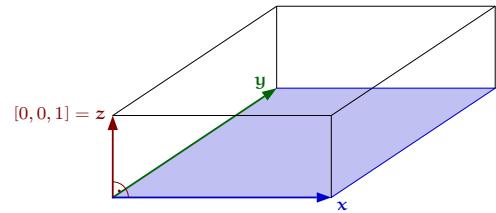


Abb. 5.8 Zur Determinante im \mathbb{R}^2

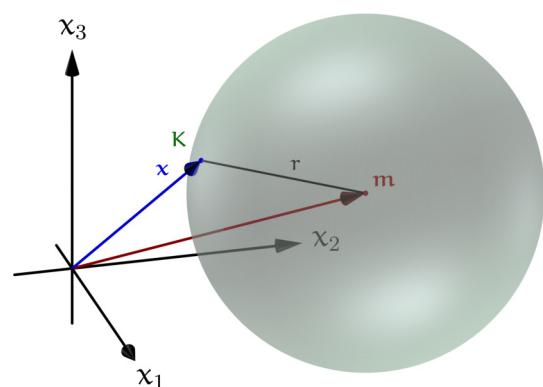
5.1.16 Die Kugel

Eine Kugeloberfläche K ist der geometrische Ort aller Punkte \mathbf{x} , die von einem gegebenen Punkt \mathbf{m} – dem Mittelpunkt – einen festen Abstand r haben. r ist der Radius von K . Das bedeutet also

$$K = \{ \mathbf{x} \in \mathbb{R}^3 \mid \| \mathbf{x} - \mathbf{m} \| = r \}. \quad (5.34)$$

In der sogenannten Koordinatenform der Kugelgleichung verwendet man üblicherweise die quadrierte Version von $\| \mathbf{x} - \mathbf{m} \| = r$:

$$(x_1 - m_1)^2 + (x_2 - m_2)^2 + (x_3 - m_3)^2 = r^2. \quad (5.35)$$



5.1.17 A Zeigen Sie, daß die Tangentialebene T an die Kugel K mit Mittelpunkt \mathbf{m} und Radius r in einem Punkt $\mathbf{p} \in K$ durch die Menge

$$T = \{ \mathbf{x} \in \mathbb{R}^3 \mid \langle \mathbf{x} - \mathbf{m} | \mathbf{p} - \mathbf{m} \rangle = r^2 \} . \quad (5.36)$$

gegeben ist. Finden Sie eine Darstellung der Kugel (also nicht nur der Sphäre). Zeigen Sie für $\mathbf{m} := [3, 2, 2]$, $r := 1.5$ und $\mathbf{p} := [4, 1.5, 3]$, daß $\mathbf{x} := [-9, 213.5, 122]$ zu T gehört.

5.1.18 A $g := \{ \mathbf{p} + t \mathbf{u} \mid t \in \mathbb{R} \}$ sei eine Gerade mit einem Richtungsvektor \mathbf{u} der Länge 1. Machen Sie sich anhand der Abbildung 5.9 klar, daß der Abstand $d(\mathbf{q}, g)$ eines Punktes \mathbf{q} von g durch

$$d(\mathbf{q}, g) = \| \mathbf{u} \times (\mathbf{q} - \mathbf{p}) \| \quad (5.37)$$

bestimmt ist. Berechnen Sie damit den Abstand von $\mathbf{q} := [15, 6, 5]$ zu der Geraden

$$g := \left\{ \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} + t \begin{bmatrix} 3 \\ 4 \\ -1 \end{bmatrix} \mid t \in \mathbb{R} \right\} .$$

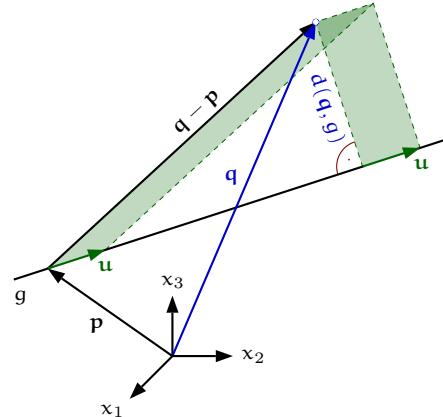


Abb. 5.9 Abstand Punkt–Gerade

5.1.19 Euklidisches Skalarprodukt und euklidische Norm im \mathbb{R}^n Für den euklidischen Raum \mathbb{R}^3 stehen uns zwei wichtige Werkzeuge zur Verfügung: Das Skalarprodukt zur Winkelmessung und die Norm zur Längen- und Abstandsmessung. Gleichung (5.12) zeigt, daß das erste das zweite schon bereitstellt. Daher müssen wir nur das Skalarprodukt auf den \mathbb{R}^n verallgemeinern und erhalten eine Norm gleich mitgeliefert. Im Hinblick auf (5.11) ist klar, wie das sinnvoll geschehen kann: Für $\mathbf{x} = [x_1, \dots, x_n] \in \mathbb{R}^n$ und $\mathbf{y} = [y_1, \dots, y_n] \in \mathbb{R}^n$ definieren wir das *euklidische Skalarprodukt* durch

$$\langle \mathbf{x} | \mathbf{y} \rangle := x_1 y_1 + x_2 y_2 + \dots + x_n y_n = \sum_{k=1}^n x_k y_k . \quad (5.38)$$

Die zugehörige *euklidische Norm* ist dann

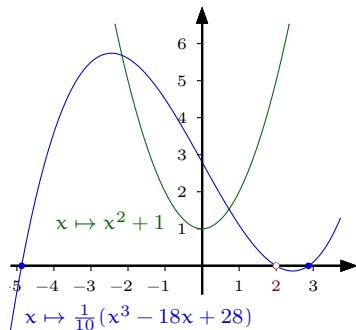
$$\| \mathbf{x} \| := \sqrt{\langle \mathbf{x} | \mathbf{x} \rangle} = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} = \sqrt{\sum_{k=1}^n x_k^2} . \quad (5.39)$$

Das Augenmerk liegt im \mathbb{R}^n nicht mehr auf der Winkelmessung, sondern darauf, mit dem Skalarprodukt den Orthogonalitätsbegriff zur Verfügung zu haben:

$$\mathbf{x} \perp \mathbf{y} : \iff \langle \mathbf{x} | \mathbf{y} \rangle = 0 . \quad (5.40)$$

Wenn nicht ausdrücklich etwas anderes gesagt wird, verstehen wir unter dem Skalarprodukt und der zugehörigen Norm auf \mathbb{R}^n immer das euklidische Skalarprodukt und die euklidische Norm.

5.2 Die komplexen Zahlen



Es gibt einfache algebraische Gleichungen, die in \mathbb{R} nicht lösbar sind. Die wichtigste Gleichung dieser Art, und wie wir sehen werden, auch die einzige, deren Lösung wir zu \mathbb{R} hinzufügen müssen, ist

$$x^2 = -1. \quad (5.41)$$

Da alle Quadrate in \mathbb{R} nicht negativ sind, kann es hier keine reelle Lösung geben. Der Graph der Funktion $x \mapsto x^2 + 1$ macht das ebenfalls deutlich. Eine Lösung von (5.41) wäre eine Nullstelle, die diese Funktion offensichtlich nicht hat.

Bei quadratischen Gleichungen $ax^2 + bx + c = 0$, mit reellen Koeffizienten a , b und c , für die $4ac > b^2$ gilt, versagt die Lösungsformel, die sogenannte *Mitternachtsformel* $x_{1/2} = \frac{1}{2a}(-b \pm \sqrt{b^2 - 4ac})$ regelmäßig, weil der Ausdruck $b^2 - 4ac$ unter der Wurzel negativ ist. Allerdings bedeutet das einfach, daß die Funktion $x \mapsto ax^2 + bx + c$ eben keine reelle Nullstelle hat. Anders sieht das bei kubischen Gleichungen, wie etwa $x^3 + px + q = 0$, aus. Auch für solche Gleichungen sind Lösungsformeln bekannt, wie die *Cardanischen Formeln* (siehe 5.2.11)

$$x = u - v, \quad \text{mit} \quad u := \sqrt[3]{\sqrt{\Delta} - \frac{q}{2}}, \quad v := \sqrt[3]{\sqrt{\Delta} + \frac{q}{2}} = \frac{p}{3u}. \quad (5.42)$$

Dabei ist $\Delta := (\frac{q}{2})^2 + (\frac{p}{3})^3$ die sogenannte *Diskriminante*. Etwa für die Gleichung

$$x^3 - 18x + 28 = 0$$

ist $\Delta = -20$. Daher kann man (5.42) nicht anwenden, obwohl $x^3 - 18x + 28 = (x^2 + 2x - 14)(x - 2)$ zeigt, daß es eine glatte Lösung, nämlich $x_1 = 2$, geben muß (tatsächlich sind $x_{2/3} = -1 \pm \sqrt{15}$ zwei weitere). Die Formel (5.42) würde die unmöglichen Zwischenergebnisse $u = \sqrt[3]{2\sqrt{-5} - 14}$ und $v = \sqrt[3]{2\sqrt{-5} + 14}$ ergeben. Will man sich damit nicht abfinden, dann kann man ja mal sehen, was passiert, wenn man sich über das Denkverbot $\sqrt{-20} = 2\sqrt{-5}$ hinwegsetzt. Bei diesem Beispiel ist das besonders einfach, weil man weiß, daß $u = 1 - \sqrt{-5}$ die formale dritte Wurzel ist:

$$(1 - \sqrt{-5})^3 = 1 - 3\sqrt{-5} + 3\sqrt{-5}^2 - \sqrt{-5}^3 = 1 - 3\sqrt{-5} - 15 + 5\sqrt{-5} = 2\sqrt{-5} - 14.$$

Genauso erhält man $v = -1 - \sqrt{-5}$. Diese formalen Manipulationen liefern überraschenderweise tatsächlich ein richtiges Endergebnis $x = 1 - \sqrt{-5} + 1 + \sqrt{-5} = 2$. Solche Rechnungen kann man als einen der Gründe vermuten, warum man angefangen hat darüber nachzudenken, ob man nicht durch Hinzunahme fiktiver Lösungen von Gleichungen des Typs (5.41) zu \mathbb{R} einen Zahlkörper erhalten kann, in dem Probleme, wie im vorgestellten Beispiel, nicht mehr auftreten. Dabei hat sich herausgestellt, daß die einzige *neue Zahl*, die benötigt wird, die Lösung von (5.41) ist. Diese Zahl wird mit i bezeichnet, was auf ihren imaginären Charakter hinweisen soll, wenn man sie vom Standpunkt reeller Zahlen aus betrachtet. Die Wurzeln u und v wären dann als $1 - \sqrt{5} i$ und $-1 - \sqrt{5} i$ zu interpretieren. Bevor wir die komplexen Zahlen formal einführen, überlegen wir uns, was wir benötigen, um konsistent mit ihnen rechnen

zu können. Dafür bilden wir zunächst einmal ganz naiv den allgemeinsten Ausdruck $z_1 + iz_2$, $z_1, z_2 \in \mathbb{R}$, den wir aus reellen Zahlen und i bilden können. Die scheinbar allgemeinere Form $z_0 + iz_1 + i^2z_2 + \dots + i^n z_n$ lässt sich immer in $\tilde{z}_1 + i\tilde{z}_2$ verwandeln, wenn wir verwenden, was die Zahl i einzig ausmacht:

$$i^2 = -1. \quad (5.43)$$

Überlegen wir uns, welche Rechenregeln wir zu definieren haben. Für Addition, Subtraktion und Multiplikation sind sie offensichtlich, einzig bei der Division müssen wir zunächst das dritte Binom einsetzen, um den Bruch in die Standardform $u_1 + iu_2$ zu bringen:

$$\begin{aligned} (z_1 + iz_2) \pm (w_1 + iw_2) &= z_1 \pm w_1 + i(z_2 \pm w_2), \\ (z_1 + iz_2)(w_1 + iw_2) &= z_1w_1 + i(z_1w_2 + z_2w_1) + i^2z_2w_2 \\ &= z_1w_1 - z_2w_2 + i(z_1w_2 + z_2w_1), \\ \frac{z_1 + iz_2}{w_1 + iw_2} &= \frac{(z_1 + iz_2)(w_1 - iw_2)}{(w_1 + iw_2)(w_1 - iw_2)} \\ &= \frac{z_1w_1 - i^2z_2w_2 + i(z_2w_1 - z_1w_2)}{w_1^2 - i^2w_2^2} \\ &= \frac{z_1w_1 + z_2w_2}{w_1^2 + w_2^2} + i \frac{z_2w_1 - z_1w_2}{w_1^2 + w_2^2}. \end{aligned}$$

5.2.1 Die Gaußsche Zahlenebene Um die komplexen Zahlen auf mathematisch solidem Grund einzuführen, überlegen wir uns, was in obiger Rechnung an Struktur wirklich benutzt wurde. Die erste Abstraktion folgt aus der Beobachtung, daß zur Festlegung von $z_1 + iz_2$ tatsächlich nur das Zahlenpaar $[z_1, z_2] \in \mathbb{R}^2$ nötig ist. Daher werden wir die komplexen Zahlen als Punktmenge mit der Ebene \mathbb{R}^2 (der sogenannten *Gaußschen Zahlenebene*) identifizieren und auf dieser die Rechenregeln einführen, die von unseren Vorerörungen nahegelegt werden.

5.2.2 Definition Die komplexen Zahlen \mathbb{C} sind die Menge \mathbb{R}^2 , versehen mit folgenden Rechenregeln

$$\begin{aligned} [z_1, z_2] \pm [w_1, w_2] &:= [z_1 \pm w_1, z_2 \pm w_2], \\ [z_1, z_2] \cdot [w_1, w_2] &:= [z_1w_1 - z_2w_2, z_1w_2 + z_2w_1]. \end{aligned}$$

Die Zahlen $[x, 0]$ werden mit den reellen Zahlen $x \in \mathbb{R}$ identifiziert. $[0, 1]$ wird mit dem Symbol i bezeichnet. Damit hat jede komplexe Zahl $z \in \mathbb{C}$ die Gestalt $z = z_1 + iz_2$, $z_1, z_2 \in \mathbb{R}$. Dabei heißen $\operatorname{Re}(z) := z_1$ und $\operatorname{Im}(z) := z_2$ Realteil bzw. Imaginärteil der Zahl z .

Die Addition und die Subtraktion in \mathbb{C} sind einfach die entsprechenden Vektoroperationen des \mathbb{R}^2 . Die Multiplikation ist eine Operation, die über die Vektorraumstruktur des \mathbb{R}^2 hinaus geht. Die Division haben wir bisher nicht definiert. Da sie die Umkehrung der Multiplikation darstellt, muß sie sich aus dieser Rechenregel gewinnen lassen.

5.2.3 A Zeigen Sie, daß die Multiplikation aus der Definition die üblichen Rechenregeln respektiert, d. h., daß $z \cdot w = w \cdot z$, $z \cdot (u + w) = z \cdot u + z \cdot w$, $z \cdot (u \cdot w) = (z \cdot u) \cdot w$ gilt.

5.2.4 Definition Für jede komplexe Zahl $z = z_1 + iz_2$ ist

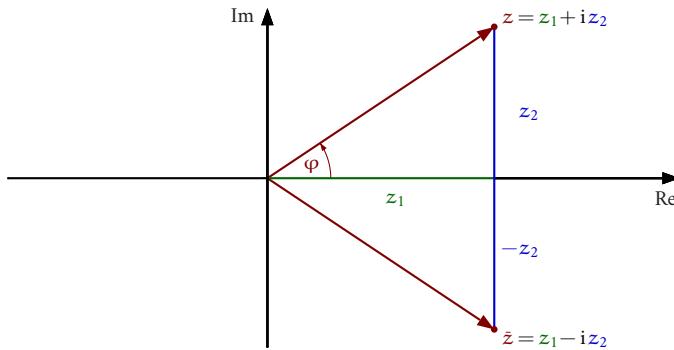
$$|z| := \sqrt{z_1^2 + z_2^2} \quad (5.44)$$

ihr Betrag und

$$\bar{z} := z_1 - iz_2 \quad (5.45)$$

die zu z konjugiert komplexe Zahl.

Der Betrag $|z|$ von z ist gerade die euklidische Norm des Vektors $[z_1, z_2] \in \mathbb{R}^2$, der zu z korrespondiert. Der Identifizierung von \mathbb{C} mit \mathbb{R}^2 , was die Vektorraumstruktur angeht, trägt auch die Darstellung komplexer Zahlen als Vektoren in der *Gaußschen Zahlebene* Rechnung:



5.2.5 A

- i) Zeigen Sie $z \cdot \bar{z} = |z|^2$ und $\bar{z} \pm w = \bar{z} \pm \bar{w}$, sowie $\bar{z} \cdot w = \bar{z} \cdot \bar{w}$. Verwenden Sie das, um nachzuweisen, daß \mathbb{C} nullteilerfrei ist, d. h., daß aus $z \neq 0$ und $z \cdot w = 0$ immer $w = 0$ folgt.
- ii) Folgern Sie aus i), daß die Gleichung $z \cdot w = u$ für jedes $z \neq 0$ eine eindeutig bestimmte Lösung w hat und daß diese durch $w = \frac{1}{|z|^2} \bar{z} \cdot u$ gegeben ist. Folgern Sie schließlich die Kürzungsregel für \mathbb{C} , indem Sie die Gleichung $z \cdot w = u$ mit einer Zahl $v \neq 0$ multiplizieren.

Die Lösung $w = \frac{1}{|z|^2} \bar{z}$ von $z \cdot w = 1$ bezeichnen wir mit $\frac{1}{z}$, $1/z$ oder z^{-1} .

- iii) Zeigen Sie: $\overline{z/w} = \bar{z}/\bar{w}$, $|\bar{z}| = |z|$,
 $\operatorname{Re}(z) = \frac{1}{2}(z + \bar{z})$, $\operatorname{Im}(z) = \frac{1}{2i}(z - \bar{z})$,
 $|z \cdot w| = |z| \cdot |w|$, $\left| \frac{z}{w} \right| = \frac{|z|}{|w|}$, $|z + w| \leq |z| + |w|$.

Verwenden Sie für die letzte Ungleichung, der sog. *Dreiecksungleichung*, daß aus $0 \leq a \leq b$ auch $0 \leq \sqrt{a} \leq \sqrt{b}$ folgt (warum eigentlich?).

Zeigen Sie die *umgekehrte Dreiecksungleichung* $|z - w| \geq ||z| - |w||$.

- iv) Zeigen Sie den *Zwei-Quadrat-Satz*: Das Produkt aus der Summe von jeweils zwei Quadranten ganzer Zahlen ist wieder eine solche Summe.
 $(a^2 + b^2)(c^2 + d^2) = e^2 + f^2$, $a, b, c, d, e, f \in \mathbb{Z}$.

Bemerkung: Ab jetzt schreiben wir, wie allgemein üblich, meist zw statt $z \cdot w$.

Aus der Skizze der GAUSSSchen Zahlenebene auf Seite 78 entnimmt man leicht die sog. *Polar-darstellung* einer komplexen Zahl:

$$z = |z| (\cos(\varphi) + i \sin(\varphi)). \quad (5.46)$$

Der Winkel $\varphi \in (-\pi, \pi]$ ist eindeutig durch

$$\varphi = \begin{cases} \arccos\left(\frac{\operatorname{Re}(z)}{|z|}\right), & \operatorname{Im}(z) \geq 0, \\ -\arccos\left(\frac{\operatorname{Re}(z)}{|z|}\right), & \operatorname{Im}(z) < 0. \end{cases} \quad (5.47)$$

bestimmt. Dabei ist \arccos , oder \cos^{-1} im Moment einfach ein Name für die Möglichkeit, vom Wert des Kosinus auf den zugehörigen Winkel zu schließen (vergl. 11.4.7).

5.2.6 Die EULER-Formel Der Polardarstellung (5.46) ist nicht ohne Weiteres anzusehen, daß sie den eigentlichen Grund dafür liefert, daß das Rechnen mit komplexen Zahlen so leistungsfähig ist. Dazu fehlt noch eine berühmte Formel, nämlich die *EULER-Formel*, die wir hier im Vorgriff auf die erst später zu behandelnden Reihenentwicklung von Funktionen angeben wollen. Sie lautet (vergl. Beispiel 11.6.11)

$$e^{i\varphi} = \cos(\varphi) + i \sin(\varphi). \quad (5.48)$$

Mit dieser Formel ist eine Verbindung zwischen den trigonometrischen Funktionen und der Exponentialfunktion geschaffen, die doch scheinbar nichts miteinander zu tun haben, wenn man von den Funktionsgraphen ausgeht. Eine vorläufige Rechtfertigung dafür, den Ausdruck $\cos(\varphi) + i \sin(\varphi)$ mit der Exponentialfunktion in Verbindung zu bringen, liefern die *Additionssätze der Trigonometrie* (siehe auch S. 135, 272)

$$\cos(\varphi + \psi) = \cos(\varphi) \cos(\psi) - \sin(\varphi) \sin(\psi), \quad (5.49)$$

$$\sin(\varphi + \psi) = \sin(\varphi) \cos(\psi) + \cos(\varphi) \sin(\psi), \quad (5.50)$$

Damit erhalten wir nämlich (das zunächst noch formale) Potenzrechengesetz

$$\begin{aligned} e^{i\varphi} e^{i\psi} &= (\cos(\varphi) + i \sin(\varphi))(\cos(\psi) + i \sin(\psi)) \\ &= \cos(\varphi) \cos(\psi) - \sin(\varphi) \sin(\psi) + i(\sin(\varphi) \cos(\psi) + \cos(\varphi) \sin(\psi)) \\ &= \cos(\varphi + \psi) + i \sin(\varphi + \psi) = e^{i(\varphi+\psi)}. \end{aligned}$$

Die Polardarstellung einer komplexen Zahl lautet jetzt

$$z = |z| e^{i\varphi}. \quad (5.51)$$

Der Fortschritt, den diese Formel gegenüber der normale Darstellung $z = z_1 + iz_2$ bereitstellt, liegt darin begründet, daß die etwas umständlich auszuführenden Rechenoperationen Multiplikation, Division und insbesondere das Potenzieren auf die einfachen Potenzrechengesetze zurückgeführt werden. Für $z = |z| e^{i\varphi}$ und $w = |w| e^{i\psi}$ gilt

$$zw = |z||w| e^{i(\varphi+\psi)}, \quad (5.52)$$

$$\frac{z}{w} = \frac{|z|}{|w|} e^{i(\varphi-\psi)}, \quad (5.53)$$

$$z^n = |z|^n e^{in\varphi}. \quad (5.54)$$

An diesen Formeln lassen sich einfache geometrische Interpretationen der Multiplikation und der Division ablesen (für die Addition haben wir ja bereits die Vektoraddition). Besonders einfach werden sie am *Einheitskreis* $\mathbb{C}_1 := \{ z \in \mathbb{Z} \mid |z| = 1 \}$. Hier bedeutet Multiplikation zw und Division $\frac{z}{w}$ einfach die Addition bzw. Subtraktion der Winkel φ und ψ der beteiligten Faktoren z und w . Außerhalb des Einheitskreises ist noch die Multiplikation bzw. Division der Beträge zu berücksichtigen.

5.2.7 Beispiel (HERONS Formel) Eine überraschende Anwendung des Rechnens mit komplexen Zahlen ist eine elegante Herleitung von HERONS Formel

$$F = \sqrt{s(s-a)(s-b)(s-c)} \quad (5.55)$$

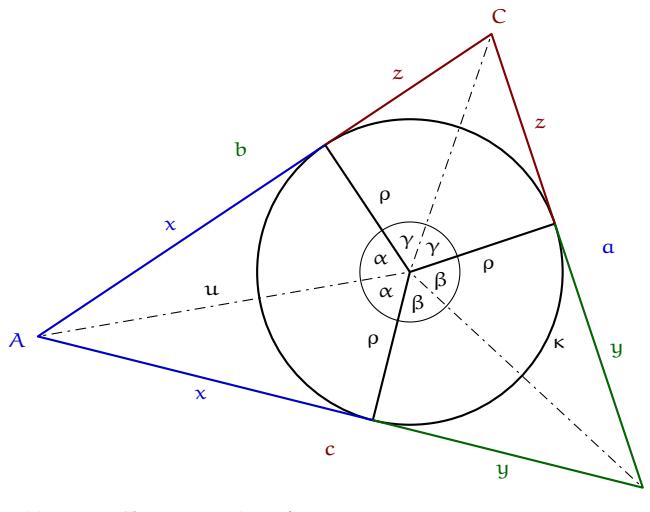


Abb. 5.10 Zu HERONS Formel

für den Flächeninhalt eines Dreiecks, bei gegebenen Seitenlängen a, b und c . Dabei ist $s := \frac{1}{2}(a+b+c) = x+y+z$ der halbe Umfang des Dreiecks (siehe Abb. 5.10).

ρ ist der Radius des *Inkreises* κ , also des größten Kreises, der ganz im Dreieck enthalten ist. Die Zahl $\rho + ix$ muß laut Skizze die Polardarstellung $u e^{i\alpha}$ haben, mit einem geeigneten $u > 0$.

Genauso erhalten wir zwei weitere Polardarstellungen:

$$\rho + iy = v e^{i\beta}, \rho + iz = w e^{i\gamma}.$$

Das Produkt $uvw e^{i(\alpha+\beta+\gamma)}$ dieser drei Zahlen ist $-uvw$, also reell, denn $\alpha + \beta + \gamma = \pi$ und $e^{i\pi} = -1$. Sein Imaginärteil muß daher verschwinden:

$$\operatorname{Im}(\rho + ix)(\rho + iy)(\rho + iz) = \rho^2(x + y + z) - xyz = \rho^2 s - xyz = 0.$$

Daraus läßt sich der *Inkreisradius* ρ bestimmen. Laut Skizze ist $y + z = a$, also $x + y + z = s = x + a$ und damit $x = s - a$. Genauso erhalten wir $y = s - b$ und $z = s - c$. Das ergibt

$$\rho = \sqrt{\frac{1}{s}(s-a)(s-b)(s-c)}. \quad (5.56)$$

Da sich das Dreieck in drei Rechtecke mit den Flächeninhalten $\rho x, \rho y$ und ρz umwandeln läßt, gilt $F = \rho(x + y + z) = \rho s$. Daraus ist (5.55) leicht zu folgern.

Zur Bestimmung des Inkreismittelpunkts siehe Aufgabe 6.4.10.

Dieser Beweis stammt von Miles Dillon Edwards (Lassiter High School, Marietta, GA 30066).

Für ein Dreieck mit den Seiten $a = 2, b = 3$ und $c = 4$ ist $\rho = \frac{1}{6}\sqrt{15}$ und $F = \frac{3}{4}\sqrt{15}$.

5.2.8 A Zeigen Sie:

$$\begin{aligned} i) \quad e^{i\frac{\pi}{2}} &= i, \quad e^{i\pi} = -1, \quad e^{2\pi i} = 1, \quad |e^{i\varphi}| = 1, \quad \overline{e^{i\varphi}} = e^{-i\varphi}. \\ e^{i\frac{\pi}{4}} &= \frac{1}{\sqrt{2}}(1+i), \quad e^{i\frac{\pi}{3}} = \frac{1}{2}(1+\sqrt{3}i). \end{aligned}$$

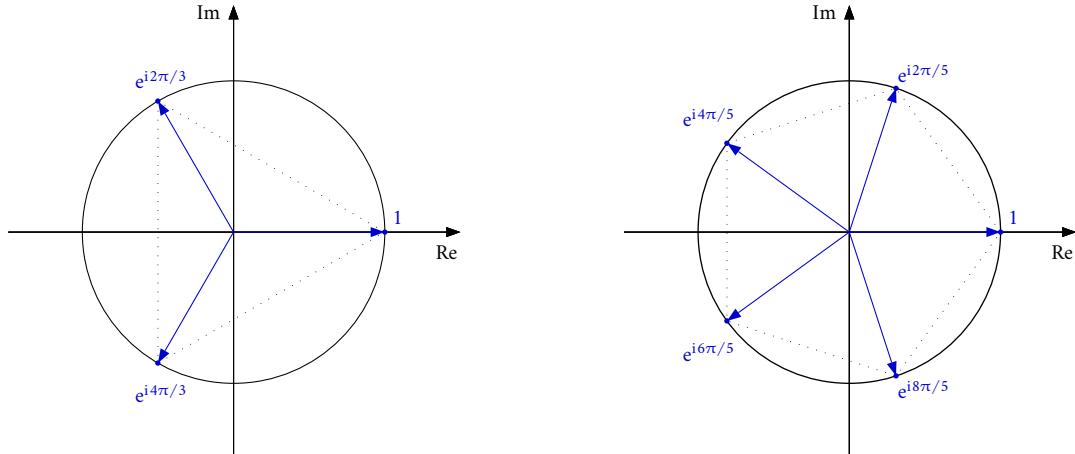
- ii) Multiplikation einer Zahl z mit $\pm i$ bedeutet eine Rotation von z um $\pm \frac{\pi}{2}$.
 Multiplikation einer Zahl z mit $e^{i\psi}$ bedeutet eine Rotation von z um ψ .

5.2.9 Die n -ten Wurzeln Die Bildung von w^n für $w = |w| e^{i\psi}$ bedeutet, den Betrag $|w|$ zur n -ten Potenz zu erheben und den Winkel ψ mit n zu multiplizieren. Eine n -te Wurzel von z ist eine Zahl w , deren n -te Potenz w^n gerade $z = |z| e^{i\varphi}$ ergibt: $w^n = |w|^n e^{in\psi} = |z| e^{i\varphi} = z$. Der Betrag von w sollte also die n -te Wurzel von $|z|$ sein. Für ψ führt sicher die Wahl $\frac{\varphi}{n}$ zum Ziel:

$$w^n = \left(\sqrt[n]{|z|} e^{i\frac{\varphi}{n}}\right)^n = \sqrt[n]{|z|}^n e^{i\frac{\varphi}{n} \cdot n} = |z| e^{i\varphi} = z.$$

Da es aber schon in \mathbb{R} , z. B. für die Quadratwurzel, zwei Lösungen gibt, ist zu erwarten, daß weitere Lösungen der Gleichung $w^n = z$ existieren. Wir werden gleich sehen, daß es sogar immer genau n verschiedene Lösungen gibt. Der Grund dafür ist die Gleichung $e^{p \cdot 2\pi i} = 1$, $p \in \mathbb{Z}$, die einfach aus der Tatsache folgt, daß $\cos(p \cdot 2\pi) = 1$ und $\sin(p \cdot 2\pi) = 0$ gilt. Bevor wir die allgemeine Lösung der Gleichung $z^n = w$ angeben, beschäftigen wir uns mit dem wichtigen Spezialfall $w^n = 1$, der den Schlüssel zur allgemeinen Lösungstheorie liefert. Eine Lösung ist natürlich $u_0 := 1$. Eine weitere ist $u_1 := e^{i\frac{2\pi}{n}}$. Allerdings sind auch $u_p := e^{ip \cdot \frac{2\pi}{n}}$ Lösungen, denn $u_p^n = e^{ip \cdot \frac{2\pi}{n} \cdot n} = e^{ip \cdot 2\pi} = 1$. Nur für $p = 1, 2, \dots, n-1$ entstehen auf diese Weise neue Lösungen. u_0, u_1, \dots, u_{n-1} sind die sogenannten n -ten Einheitswurzeln.

In der folgenden Skizze sind die dritten und die fünften Einheitswurzeln zu sehen.



5.2.10 Satz Für $z = |z| e^{i\varphi}$ gibt es immer genau n verschiedene Lösungen w_0, w_1, \dots, w_{n-1} der Gleichung $w^n = z$, die aus der Grundlösung $w_0 := \sqrt[n]{|z|} e^{i\frac{\varphi}{n}}$ durch Multiplikation mit den n -ten Einheitswurzeln u_0, u_1, \dots, u_{n-1} entstehen ($p = 0, 1, \dots, n-1$):

$$w_p = \sqrt[n]{|z|} e^{i\frac{1}{n}(\varphi + p \cdot 2\pi)} = w_0 u_p, \quad (5.57)$$

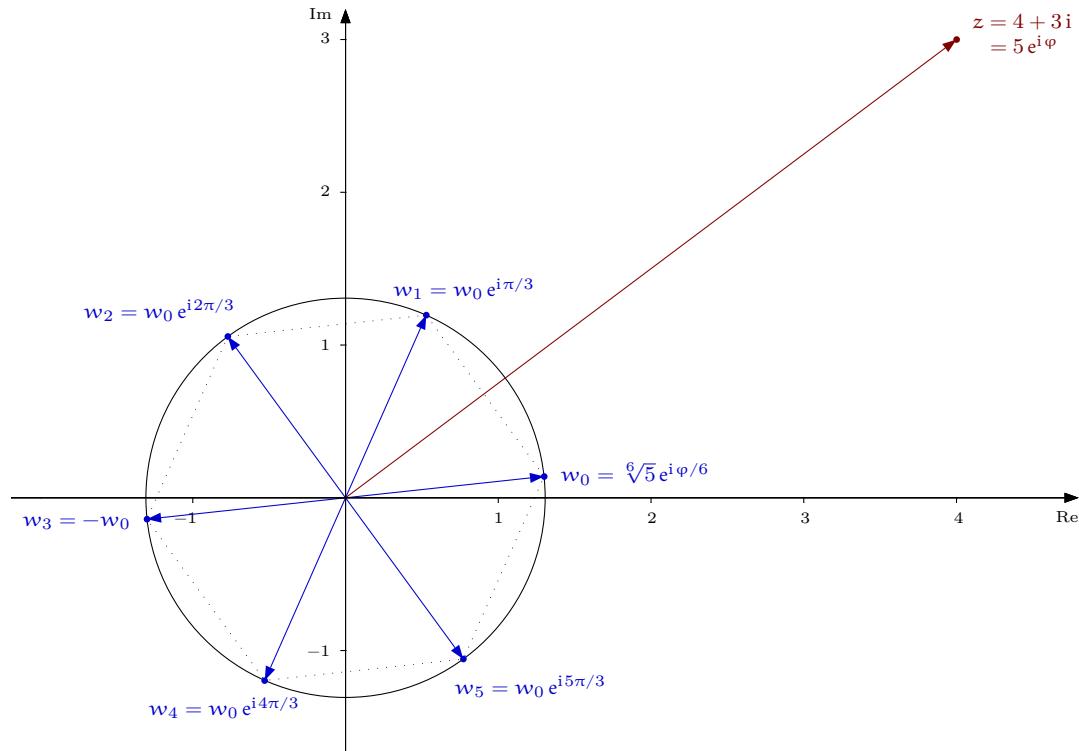
$$u_p = e^{ip \cdot \frac{2\pi}{n}}. \quad (5.58)$$

Beweis. Wir müssen nachrechnen, daß tatsächlich $w_p^n = z$ gilt. Das folgt aber leicht aus $w_0^n = z$ und $u_p^n = 1$.

Jetzt müssen wir nur noch zeigen, daß auch alle Lösungen gefunden wurden. Dazu überlegen wir zunächst, daß für ψ in der Gleichung $e^{i\psi} = 1$ nur die Zahlen $p \cdot 2\pi$, $p \in \mathbb{Z}$ möglich sind (eine einfache Übung). Für eine Lösung $w = |w|e^{i\psi}$ von $w^n = z$ muß $|w^n| = |w|^n = |z|$, also $|w| = \sqrt[n]{|z|}$ gelten. $w^n = z$ führt dann auf $e^{in\psi} = e^{i\varphi}$ und nach Multiplikation mit $e^{-i\varphi}$ auf $e^{i(n\psi-\varphi)} = 1$. Das bedeutet $n\psi - \varphi = p \cdot 2\pi$, oder $\psi = \frac{1}{n}(\varphi + p \cdot 2\pi)$, $p \in \mathbb{Z}$. Einzig die Einschränkung von p auf den Bereich von 0 bis $n - 1$ muß noch begründet werden. Würden wir $p > n - 1$ zulassen, also $p = r \cdot n + q$ mit einer Zahl $r \in \mathbb{N}$ und $q \in \{0, \dots, n - 1\}$, so erhielten wir keine neue Lösung, sondern einfach wieder w_q aus $\{w_0, w_1, \dots, w_{n-1}\}$:

$$\sqrt[n]{|z|} e^{i(\varphi + (r \cdot n + q) \cdot 2\pi)/n} = \sqrt[n]{|z|} e^{i(\varphi + q \cdot 2\pi)/n} \cdot e^{ir \cdot n \cdot 2\pi/n} = w_q \cdot e^{ir \cdot 2\pi} = w_q.$$

Genauso überlegt man sich, daß auch $p < 0$ zu keiner neuen Lösung führt. \square



Bestimmen wir die sechsten Wurzeln von $z = 4 + 3i = 5e^{i\varphi}$ mit $\varphi = \arccos(0.8) \approx 0.6435$: $w_0 = \sqrt[6]{5} e^{i\frac{\varphi}{6}}$ und

$$w_1 = w_0 e^{i\frac{\pi}{3}}, \quad w_2 = w_0 e^{i\frac{2\pi}{3}}, \quad w_3 = -w_0, \quad w_4 = w_0 e^{i\frac{4\pi}{3}}, \quad w_5 = w_0 e^{i\frac{5\pi}{3}}.$$

Wie man sieht entstehen auch im Falle allgemeiner Wurzeln regelmäßige n -Ecke, nur daß diese jetzt um den Winkel von w_0 gedreht sind.

5.2.11 Kubische Gleichungen Eine kubische Gleichung $x^3 + bx^2 + cx + d = 0$ kann durch die Substitution $y := x - e$ in die Form

$$y^3 + py + q = 0 \quad (5.59)$$

gebracht werden, mit $p := c - \frac{1}{3}b^2$ und $q := \frac{2}{27}b^3 - \frac{1}{3}bc + d$. Für $e := -\frac{b}{3}$ erhält man nämlich

$$\begin{aligned} 0 &= x^3 + bx^2 + cx + d \\ &= y^3 + 3ey^2 + 3e^2y + e^3 + b(y^2 + 2ey + e^2) + c(y + e) + d \\ &= y^3 + (3e + b)y^2 + (3e^2 + 2eb + c)y + e^3 + e^2b + ce + d \\ &= y^3 + \left(\frac{b^2}{3} - \frac{2b^2}{3} + c\right)y - \frac{b^3}{27} + \frac{b^3}{9} - \frac{bc}{3} + d \\ &= y^3 + \left(c - \frac{1}{3}b^2\right)y + \frac{2}{27}b^3 - \frac{1}{3}bc + d. \end{aligned}$$

Der Lösungsansatz für (5.59) ist $y = u - v$, mit den beiden Unbekannten u und v :

$$\begin{aligned} 0 &= u^3 - 3u^2v + 3uv^2 - v^3 + (u - v)p + q \\ &= u^3 - v^3 + 3uv(v - u) + (u - v)p + q = u^3 - v^3 + (p - 3uv)(u - v) + q \\ &= u^3 - v^3 + q, \quad \text{falls } p = 3uv \text{ gilt.} \end{aligned}$$

Wenn also u und v so bestimmt werden können, daß die Gleichungen

$$0 = u^3 - v^3 + q, \quad (5.60)$$

$$p = 3uv \quad (5.61)$$

gelten, dann ist $y = u - v$ eine Lösung von (5.59). Multiplizieren wir (5.60) mit u^3 , so erhalten wir eine quadratische Gleichung für u^3 :

$$0 = u^6 + q u^3 - \left(\frac{p}{3}\right)^3,$$

mit den Lösungen

$$u^3 = -\frac{q}{2} \pm \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3},$$

die, wie immer bei den Lösungen quadratischer Gleichungen, richtig zu interpretieren sind, falls die sogenannte *Diskriminante* $\Delta := \left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3$ negativ, oder gar komplex sein sollte.

Wir wählen

$$u := \sqrt[3]{\sqrt{\Delta} - \frac{q}{2}} \quad \text{und} \quad v := \sqrt[3]{\sqrt{\Delta} + \frac{q}{2}}.$$

Dann gilt offensichtlich $u^3 - v^3 = -q$ und $uv = \sqrt[3]{\Delta - \frac{q^2}{4}} = \sqrt[3]{\frac{p^3}{27}} = \frac{p}{3}$ – letzteres allerdings nur, wenn die Wurzeln reell sind. Die Gleichungen (5.60) und (5.61) sind erfüllt. Damit ist durch

$$y_1 := \sqrt[3]{\sqrt{\Delta} - \frac{q}{2}} - \sqrt[3]{\sqrt{\Delta} + \frac{q}{2}} = \sqrt[3]{\sqrt{\Delta} - \frac{q}{2}} - \frac{p}{3 \sqrt[3]{\sqrt{\Delta} - \frac{q}{2}}} \quad (5.62)$$

eine Formel für eine Lösung der kubischen Gleichung gefunden. Auch hier muß gesagt werden, daß es sich im Falle komplexer Wurzel $\sqrt{\Delta}$ weniger um eine Lösungsformel handelt, als

vielmehr um eine Anweisung, aus welcher komplexen Zahl eine dritte Wurzel gezogen werden soll. Deshalb ist der zweite Teil von (5.62) dem ersten vorzuziehen, wenn in den Wurzeln komplexe Zahlen vorkommen, da sonst nicht klar ist, welche dritte Wurzel von $\sqrt{\Delta} + \frac{q}{2}$ gemeint ist. Es genügt, eine Lösung in der Form $y = u - v$ zu finden, da die beiden anderen systematisch mit Hilfe der dritten Einheitswurzeln (5.58) $u_k := \exp(k \frac{2\pi i}{3})$, $k = 0, 1, 2$, konstruiert werden können:

$$y_2 = u_1 u - u_2 v, \quad y_3 = u_2 u - u_1 v. \quad (5.63)$$

Für y_2 etwa gilt $3u_1 u \cdot u_2 v = 3uv = p$ und $(u_1 u)^3 - (u_2 v)^3 = u^3 - v^3 = -q$.

5.2.12 Satz Eine kubische Gleichung $x^3 + bx^2 + cx + d = 0$ kann mit der Substitution $x = y - \frac{b}{3}$ auf die Standardform

$$y^3 + py + q = 0 \quad (5.64)$$

gebracht werden. Dabei ist $p := c - \frac{1}{3}b^2$ und $q := \frac{2}{27}b^3 - \frac{1}{3}bc + d$. Deren Lösungen werden (für $p, q \neq 0$) durch die sogenannten Cardanischen Formeln gegeben:

$$y_1 = \sqrt[3]{\sqrt{\Delta} - \frac{q}{2}} - \frac{p}{3\sqrt[3]{\sqrt{\Delta} - \frac{q}{2}}} = u - v, \quad (5.65)$$

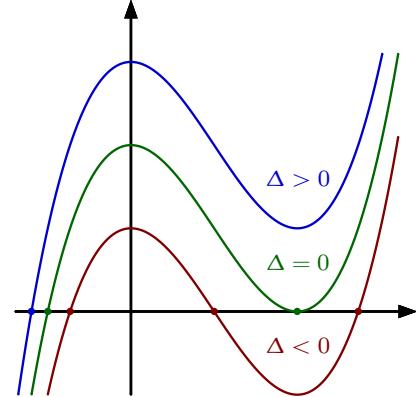
$$y_2 = u_1 u - u_2 v, \quad y_3 = u_2 u - u_1 v. \quad (5.66)$$

$u_0 = 1$, $u_{1/2} = \frac{1}{2}(-1 \pm i\sqrt{3})$ sind die dritten Einheitswurzeln. $\Delta := (\frac{q}{2})^2 + (\frac{p}{3})^3$ ist die sogenannte Diskriminante. Für reelle p und q bedeutet

$$\Delta > 0: \quad y_1 = \sqrt[3]{\sqrt{\Delta} - \frac{q}{2}} - \sqrt[3]{\sqrt{\Delta} + \frac{q}{2}} \in \mathbb{R}, \\ y_3 = y_2 \notin \mathbb{R} \text{ und } y_2 \neq y_3,$$

$$\Delta = 0: \quad y_1 = \frac{3q}{p} \in \mathbb{R} \text{ und } y_2 = y_3 = -\frac{3q}{2p} \in \mathbb{R},$$

$$\Delta < 0: \quad y_{1/2/3} = 2 \operatorname{Re} \left(u_{0/1/2} \sqrt[3]{-\frac{q}{2} + i\sqrt{|\Delta|}} \right), \text{ oder} \\ y_1 = 2 \sqrt{\frac{|p|}{3}} \cos(\varphi_0), \\ y_{2/3} = -2 \sqrt{\frac{|p|}{3}} \cos(\varphi_0 \mp \frac{\pi}{3}), \\ \varphi_0 := \frac{1}{3} \arccos \left(-\frac{q}{2} \sqrt{\frac{27}{|p|^3}} \right).$$



Beweis. $p \neq 0$ bedeutet, daß $\sqrt{\Delta} \neq \frac{q}{2}$ gilt, so daß (5.65) sinnvoll ist. Wir müssen zeigen, daß $y_1 = u - v$, mit $u = \sqrt[3]{\sqrt{\Delta} - \frac{q}{2}}$ und $v = \frac{p}{3\sqrt[3]{\sqrt{\Delta} - \frac{q}{2}}}$ die beiden Gleichungen (5.60) und (5.61) erfüllen. Letztere ist offensichtlich. Bleibt (5.60):

$$u^3 - v^3 = \sqrt{\Delta} - \frac{q}{2} - \frac{(\frac{p}{3})^3}{\sqrt{\Delta} - \frac{q}{2}} = \frac{\Delta + (\frac{q}{2})^2 - q\sqrt{\Delta} - (\frac{p}{3})^3}{\sqrt{\Delta} - \frac{q}{2}} = \frac{\frac{q^2}{2} - q\sqrt{\Delta}}{\sqrt{\Delta} - \frac{q}{2}} = -q.$$

Dabei haben wir nur die definierende Eigenschaft der beteiligten Wurzeln verwendet, also $u^3 = \sqrt{\Delta} - \frac{q}{2}$ und $\sqrt{\Delta}^2 = \Delta$. Wir können daher eine beliebige Quadratwurzel aus der (möglicherweise komplexen) Zahl Δ und eine beliebige dritte Wurzel aus $\sqrt{\Delta} - \frac{q}{2}$ wählen, um die erste Lösung y_1 zu bilden. Daß es sich dann bei y_2 und y_3 ebenfalls um Lösungen handelt, haben wir oben schon gezeigt.

Jetzt noch der Fall reeller Koeffizienten $p \neq 0$ und $q \neq 0$ und die Rolle der Diskriminante. Für $\Delta \geq 0$ haben wir es nur mit der dritten Wurzel aus reellen Zahlen zu tun, so daß für y_1 die reelle Wurzel gewählt werden kann, für die die gewohnten Rechenregeln gelten:

$$\begin{aligned} y_1 &= \sqrt[3]{\sqrt{\Delta} - \frac{q}{2}} - \frac{p}{3\sqrt[3]{\sqrt{\Delta} - \frac{q}{2}}} = \sqrt[3]{\sqrt{\Delta} - \frac{q}{2}} - \frac{p\sqrt[3]{\sqrt{\Delta} + \frac{q}{2}}}{3\sqrt[3]{(\frac{p}{3})^3 + (\frac{q}{2})^2 - (\frac{q}{2})^2}} \\ &= \sqrt[3]{\sqrt{\Delta} - \frac{q}{2}} - \sqrt[3]{\sqrt{\Delta} + \frac{q}{2}} = u - v. \end{aligned}$$

Offensichtlich sind u und v verschieden und daher $y_1 \neq 0$. Wäre $y_2 = u_1 u - u_2 v = u_1 u - \bar{u_1}v$ reell, so müsste $u_1 u - \bar{u_1}v = \bar{u_1}u - u_1v$ gelten, woraus sofort $2i \operatorname{Im}(u_1)u = -2i \operatorname{Im}(u_1)v$, also $u = -v$ folgen würde. Das wiederum wäre zu $\Delta = 0$ äquivalent. Nur in diesem Fall sind also die weiteren Lösungen y_2 und y_3 reell und darüber hinaus gleich: $v = -u$ hat $y_1 = -2\sqrt[3]{\frac{q}{2}}$ und $y_2 = (u_1 + \bar{u_1})u = 2 \operatorname{Re}(u_1)u = \sqrt[3]{\frac{q}{2}} = y_3$ zur Folge. $\Delta = 0$ bedeutet $(\frac{p}{3})^3 = -(\frac{q}{2})^2$, also $(\frac{p}{3})^3 \frac{q}{2} = -(\frac{q}{2})^3$, oder $-\frac{q}{2} = (\frac{3q}{2p})^3$, woraus $y_1 = \frac{3q}{p}$ und $y_2 = -\frac{3q}{2p}$ unmittelbar folgt.

Für $\Delta > 0$ dagegen ist $y_2 \notin \mathbb{R}$ und daher $y_2 \neq \bar{y}_2 = y_3$. Damit sind drei verschiedene Lösungen vorhanden, von denen genau eine reell ist.

Für $\Delta < 0$ wählen wir $\sqrt{\Delta} := i\sqrt{|\Delta|}$ und erhalten mit $u = \sqrt[3]{-\frac{q}{2} + i\sqrt{|\Delta|}}$ die erste Lösung $y_1 = u - \frac{p}{3u}$. Die formen wir um (vergl. Aufgabe 5.2.15):

$$\begin{aligned} y_1 &= u - \frac{p\bar{u}}{3|u|^2} = u - \frac{p\bar{u}}{3\sqrt[3]{|-\frac{q}{2} + i\sqrt{|\Delta|}|^2}} = u - \frac{p\bar{u}}{3\sqrt[3]{|-\frac{q}{2} + i\sqrt{|\Delta|}|^2}} \\ &= u - \frac{p\bar{u}}{3\sqrt[3]{(\frac{q}{2})^2 - (\frac{p}{3})^3 - (\frac{q}{2})^2}} = u + \bar{u} = 2 \operatorname{Re} \sqrt[3]{-\frac{q}{2} + i\sqrt{|\Delta|}}. \end{aligned}$$

Die Formeln für y_2 und y_3 ergeben sich jetzt mit $v = -\bar{u}$ aus einer einfachen Rechnung. u , $u_1 u$ und $u_2 u$ sind alle dritten Wurzeln von $-\frac{q}{2} + i\sqrt{|\Delta|}$. Sie bilden also die Eckpunkte eines gleichseitigen Dreiecks. Wegen $u \notin \mathbb{R}$ liegt es nicht symmetrisch zur reellen Achse, weshalb die Realteile der Eckpunkte y_1 , y_2 und y_3 paarweise verschieden sein müssen.

Die zweite Form der Lösungen ergibt sich mit Hilfe des Ansatzes $y = t \cos(\varphi)$ und der Gleichung $\cos(3\varphi) = 4 \cos^3(\varphi) - 3 \cos(\varphi)$ (vergl. Aufgabe 5.2.19). Setzen wir das in (5.64) ein, um φ und t zu bestimmen:

$$0 = t^3 \cos^3(\varphi) + pt \cos(\varphi) + q = \frac{1}{4}t^3 \cos(3\varphi) + t\left(\frac{3}{4}t^2 + p\right)\cos(\varphi) + q.$$

Wenn wir $t = \sqrt{-\frac{4}{3}p}$ wählen, verschwindet der Term mit $\cos(\varphi)$. Das ist möglich, denn $\Delta = (\frac{p}{3})^3 + (\frac{q}{2})^2 < 0$ hat $p < 0$ zur Folge. Damit führt $t = 2\sqrt{\frac{|p|}{3}}$ auf $\cos(3\varphi) = -\frac{q}{2}\sqrt{(\frac{3}{|p|})^3}$. Auch diese Gleichung ist lösbar, denn $\Delta < 0$ bedeutet auch $|(\frac{p}{3})^3| = (\frac{|p|}{3})^3 > (\frac{q}{2})^2$, also $\frac{|q|}{2} < \sqrt{(\frac{|p|}{3})^3}$ und $\frac{|q|}{2}\sqrt{(\frac{3}{|p|})^3} < 1$. Die Lösungen sind $\varphi_k := \frac{1}{3}\arccos\left(-\frac{q}{2}\sqrt{(\frac{3}{|p|})^3}\right) + k\frac{2\pi}{3}$, $k = 0, 1, 2$. Für $k = 0$ erhalten wir $y_1 := 2\sqrt{\frac{|p|}{3}}\cos(\varphi_0)$. Für $k = 1$ ergibt sich, wegen $\cos(\varphi_0 + \frac{2\pi}{3}) = \cos(\varphi_0 - \frac{\pi}{3} + \pi) = -\cos(\varphi_0 - \frac{\pi}{3})$ und für $k = 2$, wegen $\cos(\varphi_0 + \frac{4\pi}{3}) = \cos(\varphi_0 + \frac{\pi}{3} + \pi) = -\cos(\varphi_0 + \frac{\pi}{3})$ die behaupteten Ausdrücke für y_2 und y_3 . \square

5.2.13 Beispiel

1. Für die Nullstellen der Funktion $f(x) := x^3 - 18x + 28$ ist $\Delta = -20 < 0$. Daher ist u eine dritte Wurzel von $-14 + 2i\sqrt{5}$, etwa $u = 1 - i\sqrt{5}$:

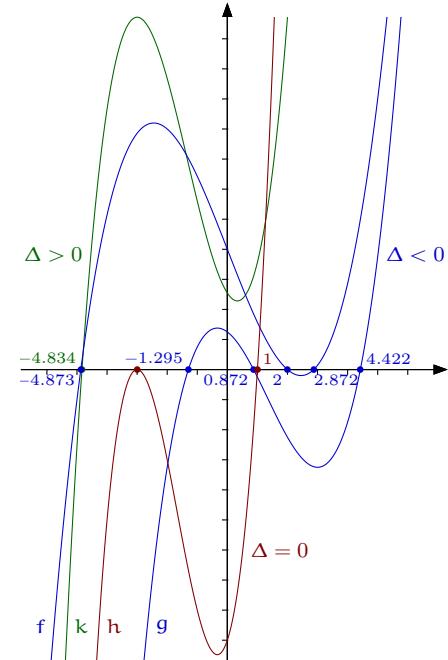
$$(1 - i\sqrt{5})^3 = 1 - 3i\sqrt{5} - 15 + 5i\sqrt{5} = -14 + 2i\sqrt{5}.$$

Die Nullstellen von f sind dann $x_1 = 2\operatorname{Re}(u) = 2$ und $x_{2/3} = 2\operatorname{Re}(u_{1/2}u) = \operatorname{Re}(-1 + i\sqrt{5} \pm \sqrt{15} \pm i\sqrt{3}) = -1 \pm \sqrt{15}$, in Übereinstimmung mit den Lösungen der quadratischen Gleichung $x^2 + 2x - 14 = 0$, die aus der kubischen durch Division mit dem Linearfaktor $x - 2$ entsteht.

Für den (seltenen) Fall, daß man die dritte Wurzel geschlossen angeben kann, ist die erste Version der Lösungsformeln für $\Delta < 0$ von Vorteil. Normalerweise aber findet man keine algebraische Ausdrücke für die Wurzeln. Dann ist die cos-Version der Lösung vorzuziehen. Für die Nullstellen von $g(x) := \frac{1}{4}(x^3 - 4x^2 - 3x + 5)$ etwa, führt die Substitution $x = y + \frac{4}{3}$ auf die Normalform $y^3 - \frac{25}{3}y - \frac{101}{27} = 0$, mit $\Delta = -\frac{13 \cdot 149}{108} < 0$. Wir erhalten $\varphi_0 = \frac{1}{3}\arccos\left(\frac{101}{250}\right)$ und daraus $x_1 = \frac{10}{3}\cos\left(\frac{1}{3}\arccos\left(\frac{101}{250}\right)\right) + \frac{4}{3} \approx 4.422697$, $x_2 = -\frac{10}{3}\cos\left(\frac{1}{3}\arccos\left(\frac{101}{250}\right) - \frac{\pi}{3}\right) + \frac{4}{3} \approx -1.295415$ und $x_3 \approx 0.872717$.

2. $h(x) := x^3 + 5x^2 + 3x - 9 = 0$ führt mit der Substitution $x = y - \frac{5}{3}$ auf $y^3 - \frac{16}{3}y - \frac{128}{27} = 0$, mit $\Delta = -(\frac{16}{3})^3 + (\frac{64}{27})^2 = -\frac{16^3}{27^2} + \frac{64^2}{27^2} = 0$. Daher ist $y_1 = \frac{3 \cdot 128 \cdot 3}{27 \cdot 16} = \frac{8}{3}$, also $x_1 = \frac{8}{3} - \frac{5}{3} = 1$ und $y_{2/3} = -\frac{4}{3}$, d.h., $x_{2/3} = -3$. Das bedeutet $h(x) = (x - 1)(x + 3)^2$ (was man mit der systematischen Methode zum Auffinden rationaler Lösungen 11.8.7, bei diesem Beispiel, viel schneller erhalten hätte).

3. Für die Funktion $k(x) := \frac{27}{53}(x^3 + 4x^2 - 3x + 5)$ liefert die Substitution $x = y - \frac{4}{3}$ die Normalform $y^3 - \frac{25}{3}y + \frac{371}{27} = 0$, mit $\Delta = \frac{11^2 \cdot 23}{6^2 \cdot 3} > 0$. Das ergibt $u = \sqrt[3]{\frac{11}{6}\sqrt{\frac{23}{3}}} - \frac{371}{54} \approx$



-1.215110 und $v = \sqrt[3]{\frac{11}{6} \sqrt{\frac{23}{3}} + \frac{371}{54}} \approx 2.286030$. Wir erhalten damit $x_1 \approx -3.501140 - \frac{4}{3} \approx -4.834473$ und $x_{2/3} = \frac{1}{2}(v - u) \pm i \frac{\sqrt{3}}{2}(u + v) - \frac{4}{3} \approx -0.417237 \pm 0.927444 i$.

4. $\ell(x) := x^3 + 6x^2 + 3(4+i)x + 9 + 7i$ führt mit der Substitution $x = y - 2$ auf $p = 3(4+i) - 12 = 3i$, $q = \frac{16-27}{27} - 6(4+i) + 9 + 7i = 1 + i$ und $\Delta = \frac{1}{4}(1+i)^2 + i^3 = \frac{1}{2}i - i = -\frac{1}{2}i$. $y^3 + 3iy + 1 + i = 0$ ist die zu $\ell(x) = 0$ gehörende Normalform. Mit (5.67) erhalten wir $\sqrt{\Delta} = \frac{1}{2}(1-i)$ als eine mögliche Quadratwurzel. Daraus ergibt sich $u = \sqrt[3]{\sqrt{\Delta} - \frac{q}{2}} = \sqrt[3]{\frac{1}{2}(1-i) - \frac{1}{2}(1+i)} = \sqrt[3]{-i} = i$ und $v = 1$, also $x_1 = i - 3$. Die verbleibenden Lösungen von $\ell(x) = 0$ sind $x_2 = u_1i - u_2 - 2 = -\frac{1}{2}(3 + \sqrt{3}) + \frac{1}{2}(\sqrt{3} - 1)i$ und $x_3 = u_2i - u_1 - 2 = -\frac{1}{2}(3 - \sqrt{3}) - \frac{1}{2}(1 + \sqrt{3})i$.

5. $x^3 + 3ix + 2 + i = 0$, also $p = 3i$ und $q = 2 + i$, sowie $\Delta = \frac{3}{4}$. Wir benötigen für u eine dritte Wurzel von $z := -\frac{q}{2} + \sqrt{\Delta} = \frac{1}{2}(\sqrt{3}-2) - \frac{1}{2}i = |z|e^{i\varphi}$. Mit $|z| = \sqrt{2-\sqrt{3}} = \frac{1}{\sqrt{2}}(\sqrt{3}-1)$ ist $\frac{\varphi}{3} = -\frac{1}{3}\arccos\left(\sqrt{2}\frac{\sqrt{3}-2}{\sqrt{3}-1}\right) = -\frac{1}{3}\arccos\left(\frac{\sqrt{2}}{4}(1-\sqrt{3})\right) \approx -0.610865$. Wir erhalten $u = \sqrt[3]{|z|}e^{i\frac{\varphi}{3}} \approx 0.802926(\cos(0.610865) - i\sin(0.610865)) \approx 0.657719 - 0.460539i$ und $v = \frac{p}{3u} \approx -0.714357 + 1.020209i$, so daß schließlich $x_1 = u - v \approx 1.372076 - 1.480748i$ und $x_2 \approx -1.170726 + 0.691323i$ sowie $x_3 \approx -0.201350 + 0.789425i$ als Ergebnis folgt.

Dieses Beispiel zeigt, worin die Schwierigkeiten in der Anwendung der Cardanischen Formeln bestehen: Ist der Ausdruck $z = a + ib$ in der dritten Wurzel komplex, so ist eben eine dritte Wurzel aus einer komplexen Zahl zu finden. Das wird nicht immer in der Form eines geschlossenen Wurzelausdrucks, wie in Beispiel 4, möglich sein können. Vielmehr wird man sich über die Polardarstellung $z = |z|e^{i\varphi}$ mit einer Wurzel in der Form $\sqrt[3]{|z|}(\cos(\frac{\varphi}{3}) + i\sin(\frac{\varphi}{3}))$ behelfen müssen, in der man die trigonometrischen Funktionen im Allgemeinen nicht loswerden kann. Im Gegensatz zu Quadratwurzeln, wo man über $\cos(\frac{\varphi}{2}) = \sqrt{\frac{1}{2}(1+\cos(\varphi))}$ und $\sin(\frac{\varphi}{2}) = \sqrt{\frac{1}{2}(1-\cos(\varphi))}$ die Berechnung der Halbwinkelfunktionen auf den Kosinus zurückspielen kann, für den man über $\cos(\varphi) = \pm \frac{a}{\sqrt{a^2+b^2}}$ einen algebraischen Ausdruck gewinnt (vergl. Aufgabe 5.2.18), ist das für $\cos(\frac{\varphi}{3})$ und $\sin(\frac{\varphi}{3})$ offenbar nicht möglich. Das liegt daran, daß $\cos(3\varphi) = 4\cos^3(\varphi) - 3\cos(\varphi)$ gilt, so daß $4\cos^3(\frac{\varphi}{3}) - 3\cos(\frac{\varphi}{3}) - \cos(\varphi) = 0$ die Lösungen einer kubische Gleichung erfordert, um $\cos(\frac{\varphi}{3})$ als Funktion von $\cos(\varphi)$ darzustellen. Man könnte auf die Idee kommen, dafür (5.62) zu verwenden. Aber dabei dreht man sich im Kreis, denn der Ausdruck unter der dritten Wurzel ergibt genau $\cos(\varphi) \pm i\sin(\varphi) = e^{\pm i\varphi}$.

5.2.14 A Bestimmen Sie die 8-ten Wurzeln von $z = 5 + 12i$.

5.2.15 A Zeigen Sie: Ist v eine n -te Wurzel von x und w eine n -te Wurzel von z , dann ist vw eine n -te Wurzel von xz . Wenn wir v durch $\sqrt[n]{x}$ und w durch $\sqrt[n]{z}$ bezeichnen würden, ließe sich das durch $\sqrt[n]{x}\sqrt[n]{z} = \sqrt[n]{xz}$ wiedergeben. Machen Sie sich an folgender Rechnung klar, daß diese vermeintliche Gleichung für komplexe Zahlen im Allgemeinen falsch ist:

$$-1 = i^2 = \sqrt{-1}^2 = \sqrt{-1}\sqrt{-1} = \sqrt{(-1)\cdot(-1)} = \sqrt{1} = 1.$$

Wo liegt der Fehler in dieser Überlegung?

Zeigen Sie: Für den Betrag der n -ten Wurzel v von x gilt $|v| = \sqrt[n]{|x|}$.

5.2.16 A Zeigen Sie, daß der Ausdruck

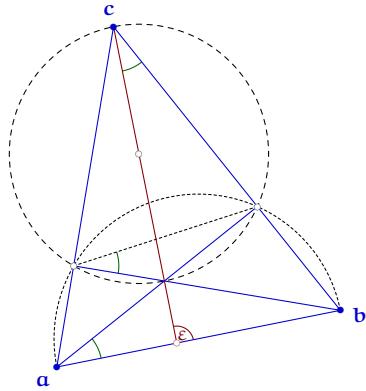
$$\frac{(e^{i\alpha} - e^{i\beta})^2 |e^{-i\alpha} - e^{i\beta}|^2}{|e^{i\alpha} - e^{i\beta}|^2 (e^{-i\alpha} - e^{i\beta})^2}$$

für $|\beta| \neq |\alpha|$ und $\alpha, \beta \in [-\pi, \pi]$ nicht von β abhängt. Geben Sie eine Interpretation dieses Ergebnisses an.

5.2.17 A

Verwenden Sie die Interpretation von Übung 5.2.16 und ergänzen Sie damit die Skizze für einen kurzen Beweis des Höhensatzes in Dreiecken:

In einem Dreieck schneiden sich die drei Höhen in einem gemeinsamen Punkt.



5.2.18 A Zeigen Sie, daß durch

$$\sqrt{z} := \sqrt{\frac{1}{2}(|z| + x)} + i \nu(y) \sqrt{\frac{1}{2}(|z| - x)} \quad (5.67)$$

eine Quadratwurzel der Zahl $z := x + iy$ gegeben ist. Eine weitere ist offensichtlich $-\sqrt{z}$. Berechnen Sie damit $\sqrt{12 - 5i}$ und $\sqrt{-4 - 5i}$.

Dabei ist $\nu(y) := 1$, falls $y \geq 0$ und $\nu(y) := -1$, falls $y < 0$ gilt. Diese Funktion unterscheidet sich von der sogenannten *Vorzeichenfunktion* sgn (auch als *Signum* bezeichnet) an der Stelle $y = 0$:

$$\text{sgn}(y) := \begin{cases} 1 & , y > 0, \\ 0 & , y = 0, \\ -1 & , y < 0. \end{cases} \quad (5.68)$$

5.2.19 A Zeigen Sie $\cos(3\varphi) = 4 \cos^3(\varphi) - 3 \cos(\varphi)$.

5.3 \mathbb{C}^n als Vektorraum

Der Schritt vom \mathbb{R}^n zum \mathbb{C}^n ist nicht sehr groß, denn wir haben in den Vektoren $\mathbf{x} = [x_1, x_2, \dots, x_n]$ nur komplexe Zahlen $x_i \in \mathbb{C}$ und in den Linearkombinationen

$$\lambda\mathbf{x} + \gamma\mathbf{y}$$

$\lambda, \gamma \in \mathbb{C}$ zuzulassen. Die Rechenregeln 5.1.1 können wir ansonsten wörtlich aus dem \mathbb{R}^n übernehmen.

5.3.1 Euklidisches Skalarprodukt und euklidische Norm im \mathbb{C}^n Das Skalarprodukt $\langle \mathbf{x} | \mathbf{y} \rangle = \sum_{k=1}^n x_k y_k$ zwischen zwei Vektoren $\mathbf{x} = [x_1, x_2, \dots, x_n]$ und $\mathbf{y} = [y_1, y_2, \dots, y_n]$ lässt sich so nicht aus dem \mathbb{R}^n übertragen, denn sonst würde der wichtige Zusammenhang $\langle \mathbf{x} | \mathbf{x} \rangle = \sum_{k=1}^n x_k^2 = \|\mathbf{x}\|^2$ zwischen Skalarprodukt und Norm verloren gehen. Der Grund ist, daß $\sum_{k=1}^n x_k^2$ nicht mehr positiv zu sein braucht. Diese Eigenschaft hat das Quadrat x_k^2 einer komplexen Zahl x_k im Allgemeinen nicht mehr. Damit würde die Norm eine komplexe Größe sein, was nicht unserer Vorstellung von einem Längenbegriff entspricht. Weil wir aber auch im \mathbb{C}^n ein Skalarprodukt und eine Norm zur Verfügung haben wollen, passen wir die Begriffe aus dem \mathbb{R}^n an die Situation im \mathbb{C}^n an. Am einfachsten ist das für die Norm. Wir ersetzen in $\sum_{k=1}^n x_k^2$ die Summanden x_k^2 durch $|x_k|^2 = \overline{x_k} x_k \geq 0$, d. h., wir definieren die *euklidische Norm* im \mathbb{C}^n durch

$$\|\mathbf{x}\| := \sqrt{\sum_{k=1}^n |x_k|^2}. \quad (5.69)$$

Den oben angesprochenen Zusammenhang zwischen Skalarprodukt und Norm erhalten wir aufrecht, wenn wir das *euklidische Skalarprodukt* im \mathbb{C}^n durch

$$\langle \mathbf{x} | \mathbf{y} \rangle := \sum_{k=1}^n \overline{x_k} y_k \quad (5.70)$$

einführen: $\langle \mathbf{x} | \mathbf{x} \rangle = \sum_{k=1}^n \overline{x_k} x_k = \sum_{k=1}^n |x_k|^2 = \|\mathbf{x}\|^2$. Die Linearität ist jetzt nur noch in der zweiten Komponente gegeben, $\langle \mathbf{x} | \lambda \mathbf{y} + \gamma \mathbf{z} \rangle = \lambda \langle \mathbf{x} | \mathbf{y} \rangle + \gamma \langle \mathbf{x} | \mathbf{z} \rangle$, während die erste Komponente *antilinear* ist: $\langle \lambda \mathbf{x} + \gamma \mathbf{y} | \mathbf{z} \rangle = \bar{\lambda} \langle \mathbf{x} | \mathbf{z} \rangle + \bar{\gamma} \langle \mathbf{y} | \mathbf{z} \rangle$. Diese Eigenschaft wird als *Sesquilinearität* bezeichnet (das meint so etwas, wie ‚halb linear‘). Wie beim \mathbb{R}^n meinen wir, wenn nicht ausdrücklich etwas anderes gesagt wird, mit dem Skalarprodukt und der Norm auf \mathbb{C}^n jeweils die euklidische Version. Fassen wir die Eigenschaften des Skalarprodukts zusammen (vergl. 5.5.6):

i) $\langle \mathbf{x} | \mathbf{x} \rangle \geq 0, \langle \mathbf{x} | \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$, *(Definitheit)*

ii) $\langle \mathbf{x} | \mathbf{y} \rangle = \overline{\langle \mathbf{y} | \mathbf{x} \rangle}$, *(Antilinearität)*

iii) $\langle \mathbf{x} | t\mathbf{y} + s\mathbf{z} \rangle = t\langle \mathbf{x} | \mathbf{y} \rangle + s\langle \mathbf{x} | \mathbf{z} \rangle$, *(Sesquilinearität)*

für alle $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n, s, t \in \mathbb{C}$.

5.3.2 Beispiel Für $\mathbf{x} := [1, 1+i, 2i]$ und $\mathbf{y} := [1-3i, 5+2i, 1+3i]$ ist $\|\mathbf{x}\| = \sqrt{1+|1+i|^2+4} = \sqrt{1+2+4} = \sqrt{7}$, $\|\mathbf{y}\| = \sqrt{1+9+25+4+1+9} = \sqrt{49} = 7$, $\langle \mathbf{x} | \mathbf{y} \rangle = 1+3i+(1+i)(5-2i)+2i(1-3i) = 14+8i$.

5.4 Der allgemeine Vektorraumbegriff*

Wir haben inzwischen einige Beispiele von Mengen gesehen, die wir jeweils als Vektorraum bezeichnet haben. Eine solche Menge, etwa \mathbb{R}^n , oder die Menge \mathcal{P} der Polynome, ist mit zwei grundlegenden Rechenoperationen ausgestattet: Zwei Elemente lassen sich addieren und jeder Vektor lässt sich strecken, d. h. lässt sich mit einer Zahl multiplizieren. Die Mengen haben darüber hinaus üblicherweise weitere Eigenschaften, aber was die Zugehörigkeit zu den Vektorräumen angeht, ist nicht mehr nötig. In der folgenden Definition geben wir allgemein an, was unter einem Vektorraum zu verstehen ist.

5.4.1 Definition Ein Vektorraum (V, \mathbb{K}) über \mathbb{K} besteht aus einer Menge V , deren Elemente x, y, z, \dots als Vektoren bezeichnet werden und einem Körper \mathbb{K} , den sogenannten Skalaren t, s, r, \dots . Darüber hinaus gibt es eine Addition $x + y \in V$ für zwei Vektoren x und y , sowie eine Skalarmultiplikation $t \cdot x \in V$ zwischen einem Skalar t und einem Vektor x , die die folgenden Eigenschaften aufweisen:

- i) $x + y = y + x$. (Kommutativität)
- ii) $(x + y) + z = x + (y + z)$. (Assoziativität)
- iii) Es gibt einen Vektor $\mathbf{0}$, den Nullvektor, mit der Eigenschaft $x + \mathbf{0} = x$.
- iv) Für alle x gibt es einen Vektor $-x$ mit $x + (-x) = \mathbf{0}$. (Existenz der Inversen)
- v) $t \cdot (x + y) = t \cdot x + t \cdot y$.
- vi) $(t + s) \cdot x = t \cdot x + s \cdot x$.
- vii) $(ts) \cdot x = t \cdot (s \cdot x)$.
- viii) Für das Einselement $1 \in \mathbb{K}$ gilt $1 \cdot x = x$.

Ziehen wir ein paar elementare Folgerungen aus dieser Definition.

- 1) Es gibt nur einen Nullvektor, denn für einen Vektor $\mathbf{0}'$ mit der Eigenschaft iii) gilt $\mathbf{0}' \stackrel{\text{iii)}}{=} \mathbf{0}' + \mathbf{0} \stackrel{\text{i)}}{=} \mathbf{0} + \mathbf{0}' \stackrel{\text{iii)}}{=} \mathbf{0}$.
- 2) Für jedes $x \in V$ gibt es nur einen Vektor $-x$, denn aus $x + y = \mathbf{0}$ folgt $-x \stackrel{\text{iii)}}{=} -x + \mathbf{0} = -x + (x + y) \stackrel{\text{ii)}}{=} (-x + x) + y \stackrel{\text{i)}}{=} (x + (-x)) + y \stackrel{\text{iv)}}{=} \mathbf{0} + y \stackrel{\text{i)}}{=} y + \mathbf{0} \stackrel{\text{iii)}}{=} y$. Daraus folgt insbesondere $-\mathbf{0} = \mathbf{0}$ und $-(-x) = x$, denn $\mathbf{0} \stackrel{\text{iii)}}{=} \mathbf{0} + \mathbf{0}$, bzw. $\mathbf{0} \stackrel{\text{iv)}}{=} x + (-x) \stackrel{\text{i)}}{=} (-x) + x$.
- 3) Für $0 \in \mathbb{K}$ gilt $0 \cdot x = \mathbf{0}$, denn $0 \cdot x = (0 + 0) \cdot x \stackrel{\text{vi)}}{=} 0 \cdot x + 0 \cdot x$. Daraus folgt $\mathbf{0} \stackrel{\text{iv)}}{=} 0 \cdot x + (-0 \cdot x) = (0 \cdot x + 0 \cdot x) + (-0 \cdot x) \stackrel{\text{ii)}}{=} 0 \cdot x + (0 \cdot x + (-0 \cdot x)) \stackrel{\text{iv)}}{=} 0 \cdot x + \mathbf{0} \stackrel{\text{iii)}}{=} 0 \cdot x$.
- 4) $-(t \cdot x) = (-t) \cdot x = t \cdot (-x)$, denn $t \cdot x + (-t) \cdot x \stackrel{\text{vi)}}{=} (t + (-t)) \cdot x = 0 \cdot x = \mathbf{0}$. Daraus folgt insbesondere auch $(-1) \cdot x = -(1 \cdot x) \stackrel{\text{viii)}}{=} -x$. Das wiederum zeigt $t \cdot (-x) = t \cdot ((-1) \cdot x) \stackrel{\text{vii)}}{=} (-t) \cdot x$.
- 5) $t \cdot \mathbf{0} = \mathbf{0}$: $t \cdot \mathbf{0} \stackrel{\text{iii)}}{=} t \cdot (\mathbf{0} + \mathbf{0}) = t \cdot (\mathbf{0} + (-\mathbf{0})) \stackrel{\text{v)}}{=} t \cdot \mathbf{0} + t \cdot (-\mathbf{0}) = t \cdot \mathbf{0} + (-t \cdot \mathbf{0}) = \mathbf{0}$.

- 6) Aus $t \cdot x = \mathbf{0}$ folgt $t = 0$, oder $x = \mathbf{0}$: Zunächst sei $t \neq 0$. Dann existiert $t^{-1} \in \mathbb{K}$, so daß $\mathbf{0} = t^{-1} \cdot \mathbf{0} = t^{-1} \cdot (t \cdot x) \stackrel{\text{vii)}}{=} (t^{-1}t) \cdot x = 1 \cdot x \stackrel{\text{viii)}}{=} x$. Ist dagegen $x \neq \mathbf{0}$, so muß $t = 0$ gelten, denn andernfalls würde, wie eben gezeigt, $x = \mathbf{0}$ folgen.

Nachdem wir jetzt die gewohnten Rechenregeln gefolgert haben, können wir es uns erlauben statt $x + (y + z)$ einfach $x + y + z$ zu schreiben. Genauso verkürzen wir $x + (-y)$ zu $x - y$, indem wir die Rechenoperation – durch $x - y := x + (-y)$ einführen. Wenn keine Gefahr für Mißverständnisse besteht, schreiben wir künftig auch einfach tx statt $t \cdot x$. Meistens ist aus dem Kontext klar, über welchen Körper \mathbb{K} ein gegebener Vektorraum zu bilden ist. Daher werden wir meist von dem Vektorraum V , statt von (V, \mathbb{K}) sprechen können.

Eigentlich vermißt man eine weitere Eigenschaft, nämlich, daß der einzige Vektor x , der mit seinem Inversen $-x$ übereinstimmt, der Nullvektor ist. In dieser Allgemeinheit ist das jedoch nicht wahr. In der Definition ist nämlich auch der Fall enthalten, daß \mathbb{K} der kleinstmögliche Körper $\mathbb{F}_2 = \{0, 1\}$ ist, für den die Rechenregeln $0 + 0 = 0$, $0 + 1 = 1$, $0 \cdot 1 = 0$, $1 \cdot 1 = 1$ und insbesondere $1 + 1 = 0$ gelten. Für einen Vektorraum V über \mathbb{F}_2 gilt daher $\mathbf{0} = 0 \cdot x = (1 + 1) \cdot x = 1 \cdot x + 1 \cdot x = x + x$. Das bedeutet $x = -x$ für alle $x \in V$.

Allerdings haben wir diese Eigenschaft sofort zur Verfügung, sobald in \mathbb{K} nicht mehr $1 + 1 = 0$ gilt. Denn dann haben wir $\mathbf{0} = x - x = x + x = 1 \cdot x + 1 \cdot x = (1 + 1) \cdot x$. Aus $1 + 1 \neq 0$ folgt jetzt $x = \mathbf{0}$.

5.4.2 Definition Für einen Vektorraum (V, \mathbb{K}) heißt eine Teilmenge $W \subseteq V$ Teilraum von V , falls (W, \mathbb{K}) mit der Addition $+$ und der Skalarmultiplikation \cdot aus (V, \mathbb{K}) wieder ein Vektorraum über \mathbb{K} ist.

5.4.3 Satz Eine Teilmenge W von V ist genau dann ein Teilraum von (V, \mathbb{K}) , wenn für alle $t, s \in \mathbb{K}$ und alle $x, y \in W$ folgendes gilt:

$$s \cdot x + t \cdot y \in W. \quad (5.71)$$

Einen Vektor der Form $t_1 \cdot x_1 + t_2 \cdot x_2 + \cdots + t_n \cdot x_n$ nennen wir *Linearkombination* der Vektoren x_1, x_2, \dots, x_n . Der Satz sagt dann, daß eine Teilmenge genau dann ein Teilraum ist, wenn keine Linearkombination seiner Elemente aus ihm hinausführt.

Beweis. Der Beweis ist sehr einfach. Wenn (W, \mathbb{K}) ein Vektorraum über \mathbb{K} ist, muß nichts gezeigt werden. Für die andere Richtung müssen wir die Vektorraumaxiome für W nachrechnen. Laut (5.71) haben wir bereits $x + y \in W$ und $t \cdot y \in W$ für alle $x, y \in W$ und $t \in \mathbb{K}$. Die Eigenschaften i) – viii) für (W, \mathbb{K}) werden jetzt von (V, \mathbb{K}) bereitgestellt. \square

Es ist an der Zeit, Beispiele für Vektorräume anzugeben. Wir fangen mit einem sehr allgemeinen an, aus dem alle für uns wichtigen mit Hilfe von Satz 5.4.3 leicht folgen.

5.4.4 Satz Sei X eine Menge und \mathbb{K} ein Körper. Dann ist die Menge $\mathcal{F}(X, \mathbb{K})$ aller Funktionen auf X mit Werten in \mathbb{K} ein Vektorraum über \mathbb{K} , wenn wir die Addition und die Skalarmultiplikation folgendermaßen definieren: Für alle $x \in X$, $t \in \mathbb{K}$, $f, g \in \mathcal{F}(X, \mathbb{K})$ gilt

$$(f + g)(x) := f(x) + g(x), \quad (5.72)$$

$$(t \cdot f)(x) := t \cdot f(x). \quad (5.73)$$

Beweis. Durch (5.72) und (5.73) werden offensichtlich wieder Funktionen von X nach \mathbb{K} definiert. Daher gilt $f + g \in V := \mathcal{F}(X, \mathbb{K})$ und $t \cdot f \in V$. Nun zu den einzelnen Punkten: Auch wenn es sich eigentlich um eine Routineaufgabe handelt, machen wir uns dieses eine Mal die Mühe, die einzelnen Punkte nachzuweisen.

Zu i): $(f + g)(x) = f(x) + g(x) = g(x) + f(x) = (g + f)(x)$ für alle $x \in X$, da die Addition auf \mathbb{K} kommutativ ist. Daher stimmen die beiden Funktionen $f + g$ und $g + f$ in allen ihren Funktionswerten überein. Das bedeutet $f + g = g + f$.

Zu ii): $((f + g) + h)(x) = (f + g)(x) + h(x) = f(x) + g(x) + h(x) = f(x) + (g + h)(x) = (f + (g + h))(x)$, für alle $f, g, h \in V$ und alle $x \in X$. Wie unter i) folgt die Behauptung.

Zu iii): Die Nullfunktion $\mathbf{0} \in V$, die durch $\mathbf{0}(x) := 0$ definiert ist, leistet das Gewünschte: $(f + \mathbf{0})(x) = f(x) + \mathbf{0}(x) = f(x)$ für alle $x \in X$ bedeutet $f + \mathbf{0} = f$.

Zu iv): Wir definieren $-f$ durch $(-f)(x) := -f(x)$. Dann gilt $(f + (-f))(x) = f(x) + (-f)(x) = f(x) - f(x) = 0 = \mathbf{0}(x)$ für alle $x \in X$. Das zeigt $f + (-f) = \mathbf{0}$.

Zu v): $(t \cdot (f + g))(x) = t \cdot (f + g)(x) = t \cdot (f(x) + g(x)) = t \cdot f(x) + t \cdot g(x) = (t \cdot f)(x) + (t \cdot g)(x) = (t \cdot f + t \cdot g)(x)$ zeigt $t \cdot (f + g) = t \cdot f + t \cdot g$.

Zu vi): $((t+s) \cdot f)(x) = (t+s) \cdot f(x) = t \cdot f(x) + s \cdot f(x) = (t \cdot f)(x) + (s \cdot f)(x) = (t \cdot f + s \cdot f)(x)$. Das zeigt $(t+s) \cdot f = t \cdot f + s \cdot f$.

Zu vii): $((ts) \cdot f)(x) = (ts) \cdot f(x) = t \cdot s \cdot f(x) = t \cdot (s \cdot f)(x) = (t \cdot (s \cdot f))(x)$. Das zeigt $(ts) \cdot f = t \cdot (s \cdot f)$.

Zu viii): $(1 \cdot f)(x) = 1 \cdot f(x) = f(x)$ ergibt schließlich $1 \cdot f = f$. □

Wir haben in diesem Satz Funktionen f ausnahmsweise in fetter Schrift wiedergegeben, um ihren Charakter als Vektor gemäß unserer bisherigen Übereinkunft hervorzuheben. Davon werden wir im Folgenden aber wieder Abstand nehmen.

Wenn es sich bei X um die endliche Menge $\{1, 2, \dots, n\}$ handelt, dann lässt sich jede Funktion $f : X \rightarrow \mathbb{K}$ durch Angabe aller ihrer Funktionswerte $f_1 := f(1), f_2 := f(2), \dots, f_n := f(n)$ eindeutig festlegen. Das bedeutet, f wird durch das Tupel $[f_1, f_2, \dots, f_n] \in \mathbb{K}^n$ eindeutig bestimmt. Dann gehören zu $f + g$ und $t \cdot f$ die Tupel $[f_1 + g_1, \dots, f_n + g_n] = [f_1, \dots, f_n] + [g_1, \dots, g_n]$ bzw. $[t \cdot f_1, \dots, t \cdot f_n] = t \cdot [f_1, \dots, f_n]$, mit der Addition bzw. Skalarmultiplikation auf \mathbb{K}^n , die uns bereits geläufig ist. Satz 5.4.4 stellt jetzt sicher, daß es sich bei \mathbb{K}^n um einen Vektorraum gemäß Definition 5.4.1 handelt.

Die Menge \mathcal{P} der Polynome auf \mathbb{C} ist eine Teilmenge von $\mathcal{F}(\mathbb{C}, \mathbb{C})$, dem Vektorraum aller Funktionen von \mathbb{C} nach \mathbb{C} , gemäß Satz 5.4.4. Da für Polynome p und q auch $t \cdot p + s \cdot q$ ein Polynom ist, haben wir die Teilraumeigenschaft 5.71 von Satz 5.4.3 nachgewiesen. Damit ist \mathcal{P} ein Vektorraum über \mathbb{C} .

5.5 Vektorräume mit Norm und Skalarprodukt

Im Folgenden werden wir die Eigenschaften eines Skalarprodukts und den Zusammenhang mit einer Norm untersuchen. Dabei werden die Überlegungen an keiner Stelle einfacher, wenn wir

uns nur auf die Vektorräume \mathbb{R}^n oder \mathbb{C}^n beziehen. Einzig nötig ist ein Vektorraum V über $\mathbb{K} = \mathbb{R}$, oder $\mathbb{K} = \mathbb{C}$, für den wir zusätzlich Eigenschaften annehmen werden.

5.5.1 Definition (Norm) Eine Norm auf einem Vektorraum (V, \mathbb{K}) ist eine Abbildung

$$\|\cdot\|: V \rightarrow \mathbb{R}$$

mit folgenden Eigenschaften:

- i) $\|\mathbf{x}\| \geq 0$, $\|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}$, (Definitheit)
- ii) $\|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|$, (Homogenität)
- iii) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, (Dreiecksungleichung)

für alle $\mathbf{x}, \mathbf{y} \in V$, $\lambda \in \mathbb{K}$.

5.5.2 Definition (Skalarprodukt) Ein Skalarprodukt auf dem Vektorraum (V, \mathbb{K}) über $\mathbb{K} = \mathbb{R}$, oder $\mathbb{K} = \mathbb{C}$ ist eine Abbildung

$$\langle \cdot | \cdot \rangle: V \times V \rightarrow \mathbb{K}$$

mit folgenden Eigenschaften:

- i) $\langle \mathbf{x} | \mathbf{x} \rangle \geq 0$, $\langle \mathbf{x} | \mathbf{x} \rangle = 0 \iff \mathbf{x} = \mathbf{0}$, (Definitheit)
- ii) $\langle \mathbf{x} | \mathbf{y} \rangle = \overline{\langle \mathbf{y} | \mathbf{x} \rangle}$, (Anti)symmetrie)
- iii) $\langle \mathbf{z} | \lambda\mathbf{x} + \gamma\mathbf{y} \rangle = \lambda\langle \mathbf{z} | \mathbf{x} \rangle + \gamma\langle \mathbf{z} | \mathbf{y} \rangle$, (Sesquilinearität)

für alle $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$, $\lambda, \gamma \in \mathbb{K}$.

Eine solche Abbildung $\langle \cdot | \cdot \rangle$ nennt man auch *sesquilinear*, oder *Sesquilinearform*, wenn es sich bei \mathbb{K} um \mathbb{C} handelt, denn sie ist nur in der zweiten Komponente linear, in der ersten aber *antilinear*, wie man sofort aus ii) und iii) schließt:

$$\begin{aligned} \langle \lambda\mathbf{x} + \gamma\mathbf{y} | \mathbf{z} \rangle &= \overline{\langle \mathbf{z} | \lambda\mathbf{x} + \gamma\mathbf{y} \rangle} = \overline{\lambda \langle \mathbf{z} | \mathbf{x} \rangle + \gamma \langle \mathbf{z} | \mathbf{y} \rangle} = \bar{\lambda} \overline{\langle \mathbf{z} | \mathbf{x} \rangle} + \bar{\gamma} \overline{\langle \mathbf{z} | \mathbf{y} \rangle} \\ &= \bar{\lambda} \langle \mathbf{x} | \mathbf{z} \rangle + \bar{\gamma} \langle \mathbf{y} | \mathbf{z} \rangle. \end{aligned}$$

Für $\mathbb{K} = \mathbb{R}$ ist ii) einfach die Symmetrie, aus der mit iii) die Linearität in beiden Komponenten von $\langle \cdot | \cdot \rangle$, die *Bilinearität* folgt.

5.5.3 Satz (CAUCHY-SCHWARZ-Ungleichung) Für alle $\mathbf{x}, \mathbf{y} \in V$ gilt

$$|\langle \mathbf{x} | \mathbf{y} \rangle| \leq \langle \mathbf{x} | \mathbf{x} \rangle^{1/2} \langle \mathbf{y} | \mathbf{y} \rangle^{1/2}. \quad (5.74)$$

Gleichheit gilt genau dann, wenn \mathbf{x} und \mathbf{y} linear abhängig sind.

Beweis. Wir gehen von $\mathbb{K} = \mathbb{C}$ aus, da die Beweisführung leicht auf die einfachere Situation $\mathbb{K} = \mathbb{R}$ übertragen werden kann. Für den Fall $\mathbf{x} = \mathbf{0}$, oder $\mathbf{y} = \mathbf{0}$ ist die Behauptung trivial. Daher können wir für unsere Überlegungen von $\mathbf{y} \neq \mathbf{0}$ ausgehen. Aus i) folgt $\langle \mathbf{x} - \lambda\mathbf{y} | \mathbf{x} - \lambda\mathbf{y} \rangle \geq 0$

0 für alle $\mathbf{x}, \mathbf{y} \in V$ und $\lambda \in \mathbb{C}$. Wir haben also $0 \leq \langle \mathbf{x} | \mathbf{x} \rangle - \bar{\lambda} \langle \mathbf{y} | \mathbf{x} \rangle - \lambda \langle \mathbf{x} | \mathbf{y} \rangle + |\lambda|^2 \langle \mathbf{y} | \mathbf{y} \rangle$. Wir wählen $\lambda := \langle \mathbf{y} | \mathbf{x} \rangle / \langle \mathbf{y} | \mathbf{y} \rangle$ und erhalten damit

$$0 \leq \langle \mathbf{x} | \mathbf{x} \rangle - 2 \frac{|\langle \mathbf{x} | \mathbf{y} \rangle|^2}{\langle \mathbf{y} | \mathbf{y} \rangle} + \frac{|\langle \mathbf{x} | \mathbf{y} \rangle|^2}{\langle \mathbf{y} | \mathbf{y} \rangle} = \langle \mathbf{x} | \mathbf{x} \rangle - \frac{|\langle \mathbf{x} | \mathbf{y} \rangle|^2}{\langle \mathbf{y} | \mathbf{y} \rangle}.$$

Multiplikation mit $\langle \mathbf{y} | \mathbf{y} \rangle$ und Umstellen nach $|\langle \mathbf{x} | \mathbf{y} \rangle|^2$ ergibt die quadrierte Fassung der CAUCHY-SCHWARZSchen Ungleichung. Die Monotonie der Wurzel (vergl. die Lösung zu Übung 5.2.5) zeigt die behauptete Ungleichung.

Es ist klar, daß in der Ungleichung das Gleichheitszeichen gilt, wenn \mathbf{x} und \mathbf{y} linear abhängig sind (man überlege sich das). Wir müssen auch die Umkehrung zeigen. D. h. wir gehen von $|\langle \mathbf{x} | \mathbf{y} \rangle| = \langle \mathbf{x} | \mathbf{x} \rangle^{1/2} \langle \mathbf{y} | \mathbf{y} \rangle^{1/2}$ aus und müssen zeigen, daß $\gamma \mathbf{x} + \delta \mathbf{y} = \mathbf{0}$ für geeignete $\gamma, \delta \in \mathbb{C}$, $(\gamma, \delta) \neq (0, 0)$ gilt. Ist $\mathbf{y} = \mathbf{0}$, so können wir $\gamma = 0$ und $\delta = 1$ wählen (entsprechend, falls $\mathbf{x} = \mathbf{0}$ sein sollte). Wir dürfen also wieder von $\mathbf{x} \neq \mathbf{0}$ und $\mathbf{y} \neq \mathbf{0}$ ausgehen. Aus der EULER-Darstellung $\langle \mathbf{x} | \mathbf{y} \rangle = |\langle \mathbf{x} | \mathbf{y} \rangle| e^{i\varphi}$ ergibt sich mit $\gamma := 1$ und $\delta := -\frac{\langle \mathbf{x} | \mathbf{x} \rangle^{1/2}}{\langle \mathbf{y} | \mathbf{y} \rangle^{1/2}} e^{-i\varphi}$

$$\begin{aligned} \langle \mathbf{x} + \delta \mathbf{y} | \mathbf{x} + \delta \mathbf{y} \rangle &= \langle \mathbf{x} | \mathbf{x} \rangle + 2 \operatorname{Re} \delta \langle \mathbf{x} | \mathbf{y} \rangle + |\delta|^2 \langle \mathbf{y} | \mathbf{y} \rangle \\ &= \langle \mathbf{x} | \mathbf{x} \rangle - 2 \operatorname{Re} \frac{\langle \mathbf{x} | \mathbf{x} \rangle^{1/2}}{\langle \mathbf{y} | \mathbf{y} \rangle^{1/2}} e^{-i\varphi} \langle \mathbf{x} | \mathbf{y} \rangle + \langle \mathbf{x} | \mathbf{x} \rangle \\ &= 2 \langle \mathbf{x} | \mathbf{x} \rangle - 2 \frac{\langle \mathbf{x} | \mathbf{x} \rangle^{1/2}}{\langle \mathbf{y} | \mathbf{y} \rangle^{1/2}} |\langle \mathbf{x} | \mathbf{y} \rangle| = 0. \end{aligned}$$

Aus der Definitheit des Skalarprodukts folgt $\mathbf{x} + \delta \mathbf{y} = \mathbf{0}$, d. h. \mathbf{x} und \mathbf{y} sind linear abhängig. \square

5.5.4 Korollar Durch $\|\mathbf{x}\| := \sqrt{\langle \mathbf{x} | \mathbf{x} \rangle}$ wird eine Norm auf V definiert.

Beweis. Die Eigenschaften i) und ii) sind leicht einzusehen. Wir müssen die Dreiecksungleichung nachweisen:

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 &= \langle \mathbf{x} + \mathbf{y} | \mathbf{x} + \mathbf{y} \rangle = \|\mathbf{x}\|^2 + \langle \mathbf{x} | \mathbf{y} \rangle + \langle \mathbf{y} | \mathbf{x} \rangle + \|\mathbf{y}\|^2 \\ &= \|\mathbf{x}\|^2 + 2 \operatorname{Re} \langle \mathbf{x} | \mathbf{y} \rangle + \|\mathbf{y}\|^2 \\ &\leq \|\mathbf{x}\|^2 + 2 |\operatorname{Re} \langle \mathbf{x} | \mathbf{y} \rangle| + \|\mathbf{y}\|^2 \\ &\leq \|\mathbf{x}\|^2 + 2 |\langle \mathbf{x} | \mathbf{y} \rangle| + \|\mathbf{y}\|^2 \\ &\leq \|\mathbf{x}\|^2 + 2 \|\mathbf{x}\| \|\mathbf{y}\| + \|\mathbf{y}\|^2 = (\|\mathbf{x}\| + \|\mathbf{y}\|)^2. \end{aligned}$$

Die Monotonie der Wurzel liefert schließlich den Beweis der Behauptung. \square

Mit diesem Ergebnis lautet die CAUCHY-SCHWARZ-Ungleichung jetzt

$$|\langle \mathbf{x} | \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|. \quad (5.75)$$

Vektorräume mit so reichhaltiger Struktur haben eigene Namen erhalten:

5.5.5 Definition Ein Vektorraum V über \mathbb{K} , der mit einer Norm $\|\cdot\|$ versehen ist, heißt normierter Raum, den wir, wenn wir ausführlich sein wollen, durch $(V, \|\cdot\|)$ kennzeichnen. Hat er ein Skalarprodukt $\langle \cdot | \cdot \rangle$, so wird er als Skalarprodukt-Raum bezeichnet, mit der ausführlichen Notation $(V, \langle \cdot | \cdot \rangle)$.

Haben diese Räume eine weitere Eigenschaft, eine Vollständigkeit gegenüber Grenzwertbildung, so werden sie als *Banachraum* bzw. *Hilbertraum* bezeichnet. Diese zusätzliche Eigenschaft ist für die endlichdimensionalen Räume, die wir im Rahmen der linearen Algebra überwiegend behandeln, immer erfüllt. Ein Hilbertraum ist immer auch ein Banachraum, da das Skalarprodukt eine kanonische Norm erzeugt. Die Umkehrung gilt im Allgemeinen nicht.

5.5.6 Das Skalarprodukt auf \mathbb{C}^n Als Anwendung behandeln wir noch einmal das kanonische Skalarprodukt 5.70 auf \mathbb{C}^n . Für $\mathbf{x} = [x_1, \dots, x_n]^t \in \mathbb{C}^n$ und $\mathbf{y} = [y_1, \dots, y_n]^t \in \mathbb{C}^n$ ist es folgendermaßen definiert:

$$\langle \mathbf{x} | \mathbf{y} \rangle := \sum_{k=1}^n \overline{x_k} y_k.$$

Wir prüfen die Bedingungen von 5.5.2 nach. $\langle \mathbf{x} | \mathbf{x} \rangle \geq 0$ ist klar.

$$\begin{aligned} \langle \mathbf{x} | \mathbf{x} \rangle &= \sum_{k=1}^n |x_k|^2 = 0 \Leftrightarrow |x_k| = 0 \text{ für } k = 1, \dots, n \Leftrightarrow \mathbf{x} = [0, \dots, 0]^t = \mathbf{0}. \\ \langle \mathbf{x} | \mathbf{y} \rangle &= \sum_{k=1}^n \overline{x_k} y_k = \sum_{k=1}^n \overline{x_k} \overline{y_k} = \overline{\sum_{k=1}^n y_k x_k} = \overline{\langle \mathbf{y} | \mathbf{x} \rangle}. \\ \langle \mathbf{x} | t\mathbf{y} + s\mathbf{z} \rangle &= \sum_{k=1}^n \overline{x_k} (ty_k + sz_k) = t \sum_{k=1}^n \overline{x_k} y_k + s \sum_{k=1}^n \overline{x_k} z_k = t \langle \mathbf{x} | \mathbf{y} \rangle + s \langle \mathbf{x} | \mathbf{z} \rangle. \end{aligned}$$

Unsere allgemeinen Überlegungen zum Skalarprodukt gelten jetzt für diese spezielle Version. Wir wissen also, daß durch $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x} | \mathbf{x} \rangle}$ eine Norm auf \mathbb{C}^n gegeben ist, für die die CAUCHY-SCHWARZ-Ungleichung und die Dreiecksungleichung gilt. Schreiben wir uns das einmal explizit auf:

$$\begin{aligned} \left| \sum_{k=1}^n \overline{x_k} y_k \right| &\leq \sqrt{\sum_{k=1}^n |x_k|^2} \sqrt{\sum_{k=1}^n |y_k|^2}, \\ \sqrt{\sum_{k=1}^n |x_k + y_k|^2} &\leq \sqrt{\sum_{k=1}^n |x_k|^2} + \sqrt{\sum_{k=1}^n |y_k|^2}. \end{aligned}$$

Man sieht, daß diese konkrete Realisierung durch die Koordinaten, die man durchaus als weniger abstrakt als unsere Überlegungen aus den letzten Abschnitten empfinden könnte, den Beweis dieser Ungleichungen nicht erleichtert hätte (wer das anders sieht, kann ja einmal versuchen, die Beweise in dieser Koordinatenversion zu führen).

Wir werden immer wieder auf kreis- oder kugelförmige Umgebungen eines Punktes \mathbf{x}_0 stoßen, für die wir eine einheitliche Notation benötigen:

5.5.7 Definition Auf einem normierten Raum $(V, \|\cdot\|)$ bezeichnen wir die Mengen

$$U_r(x_0) := \{ x \in V \mid \|x - x_0\| < r \} \quad (5.76)$$

$$\bar{U}_r(x_0) := \{ x \in V \mid \|x - x_0\| \leq r \} \quad (5.77)$$

als offene bzw. abgeschlossene Kugeln mit Mittelpunkt x_0 und Radius r .

In der Analysis werden wir diesen Umgebungen auch begegnen, nur wird dort traditionell das Symbol *Epsilon* ε für den Radius verwendet und von der ε -Umgebung des Punktes x_0 gesprochen ($\varepsilon > 0$ kennzeichnet in der Mathematik üblicherweise eine kleine Größe). Normalerweise meint man damit immer die offene Kugel um x_0 mit Radius $\varepsilon > 0$.

Für $(\mathbb{R}^n, \|\cdot\|)$ mit der euklidischen Norm $\|[x_1, \dots, x_n]^t\| = \sqrt{\sum_{k=1}^n x_k^2}$, handelt es sich bei $U_r(x_0)$ für $n = 1$ um das offene Intervall $(x_0 - r, x_0 + r)$, für $n = 2$ um die offene Kreisscheibe mit Radius r und Mittelpunkt $x_0 \in \mathbb{R}^2$ und endlich für $n = 3$ tatsächlich um eine Kugel im landläufigen Sinne, allerdings ohne die Punkte der Kugelschale. Durch die Kugelschale $K_r(x_0) := \{ x \in \mathbb{R}^n \mid \|x - x_0\| = r \}$ unterscheidet sich $\bar{U}_r(x_0)$ von $U_r(x_0)$ (vergl. (5.1.16)).

Versehen wir \mathbb{R}^n mit anderen Normen, dann werden die Mengen $U_r(x_0)$ bzw. $\bar{U}_r(x_0)$ mangels besserer Bezeichnung gemäß der Definition weiterhin als Kugeln bezeichnet, auch wenn sie gar nicht mehr unserer Vorstellung von Kugeln entsprechen.

5.5.8 A Eine wichtige Klasse von Normen sind die sogenannten L^p -Normen $\|\cdot\|_p$. Für $1 \leq p \leq \infty$ ist ihre Definition unter iii) bzw. iv) für den Vektorraum \mathbb{R}^2 wiedergegeben. Die L^2 -Norm ist die uns schon bekannte euklidische Norm. Skizzieren Sie die abgeschlossenen Einheitskugeln $\bar{U}_1(\mathbf{0})$ für die L^1 -, die L^2 -, die L^3 - und die L^∞ -Norm. Zeigen Sie, daß es sich bei $\|\cdot\|_1$ und $\|\cdot\|_\infty$ tatsächlich um Normen handelt.

i) $\bar{U}_1(\mathbf{0}) = \{ [x_1, x_2]^t \in \mathbb{R}^2 \mid |x_1| + |x_2| \leq 1 \}$ gehört zur L^1 -Norm:

$$\|[x_1, x_2]^t\|_1 := |x_1| + |x_2|.$$

ii) $\bar{U}_1(\mathbf{0}) = \{ [x_1, x_2]^t \in \mathbb{R}^2 \mid |x_1|^2 + |x_2|^2 \leq 1 \}$ gehört zur L^2 -, oder euklidischen Norm:

$$\|[x_1, x_2]^t\|_2 := \sqrt{|x_1|^2 + |x_2|^2}.$$

iii) $\bar{U}_1(\mathbf{0}) = \{ [x_1, x_2]^t \in \mathbb{R}^2 \mid |x_1|^p + |x_2|^p \leq 1 \}$ gehört zur L^p -Norm:

$$\|[x_1, x_2]^t\|_p := \sqrt[p]{|x_1|^p + |x_2|^p}, p \geq 1.$$

iv) $\bar{U}_1(\mathbf{0}) = \{ [x_1, x_2]^t \in \mathbb{R}^2 \mid \max\{|x_1|, |x_2|\} \leq 1 \}$ gehört zur L^∞ -Norm:

$$\|[x_1, x_2]^t\|_\infty := \max\{|x_1|, |x_2|\}.$$

Die Bezeichnung $\|\cdot\|_\infty$ kommt nicht von ungefähr, denn es gilt $\lim_{p \rightarrow \infty} \|x\|_p = \|x\|_\infty$.

5.5.9 A Zeigen Sie: Für eine Norm gilt die sog. *umgekehrte Dreiecksungleichung*

$$\|x - y\| \geq |\|x\| - \|y\||. \quad (5.78)$$

5.5.10 A (Satz von RIESZ) Eine *Linearform* ℓ auf einem Vektorraum V über dem Körper \mathbb{K} ist eine lineare Abbildung von V nach \mathbb{K} . Zeigen Sie: Jede Linearform ℓ auf $V = \mathbb{C}^n$ ist von der Form

$$\ell(\mathbf{x}) = \langle \mathbf{y}_\ell | \mathbf{x} \rangle, \quad (5.79)$$

mit einem durch ℓ eindeutig bestimmten Vektor $\mathbf{y}_\ell \in \mathbb{C}^n$. *Hinweis:* Entwickeln Sie \mathbf{x} in der kanonischen Basis von \mathbb{C}^n .

5.5.11 A (Polarisationsgleichung) Ein Skalarprodukt definiert über $\|\mathbf{x}\| := \sqrt{\langle \mathbf{x} | \mathbf{x} \rangle}$ die kanonisch zugeordnete Norm. Umgekehrt kann man das Skalarprodukt aus dieser Norm auch wieder zurückgewinnen. Zeigen Sie dazu die sogenannte *Polarisationsgleichung*

$$\langle \mathbf{x} | \mathbf{y} \rangle = \frac{1}{4} [\|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2] \quad (5.80)$$

$$\langle \mathbf{x} | \mathbf{y} \rangle = \frac{1}{4} [\|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2 - i(\|\mathbf{x} + i\mathbf{y}\|^2 - \|\mathbf{x} - i\mathbf{y}\|^2)] \quad (5.81)$$

Dabei gilt (5.80) für Vektorräume über \mathbb{R} und (5.81) für Vektorräume über \mathbb{C} .

5.5.12 A Zeigen Sie: Für ein Skalarprodukt $\langle \cdot | \cdot \rangle$ auf V gilt:

$$\langle \mathbf{x} | \mathbf{y} \rangle = 0 \text{ f. a. } \mathbf{y} \in V \Leftrightarrow \mathbf{x} = \mathbf{0}.$$

Schließen Sie daraus: Aus $\langle \mathbf{x} | \mathbf{y} \rangle = \langle \mathbf{z} | \mathbf{y} \rangle$ für alle $\mathbf{y} \in V$ folgt $\mathbf{x} = \mathbf{z}$.

5.5.13 A Folgern Sie aus $\|\|\mathbf{a}\|^2 \mathbf{b} - \langle \mathbf{a} | \mathbf{b} \rangle \mathbf{a}\|^2 \geq 0$ die CAUCHY-SCHWARZ-Ungleichung.

5.5.14 Die Hölder- und Minkowski-Ungleichung* Es seien $\mathbf{x} := [x_1, \dots, x_n]^t$ und $\mathbf{y} := [y_1, \dots, y_n]^t$ Vektoren aus \mathbb{C}^n (oder \mathbb{R}^n), sowie $1 < p < \infty$. Der Nachweis, daß

$$\|\mathbf{x}\|_p := \sqrt[p]{\sum_{k=1}^n |x_k|^p} = \left[\sum_{k=1}^n |x_k|^p \right]^{\frac{1}{p}} \quad (5.82)$$

eine Norm definiert, steht und fällt mit dem Beweis der Dreiecksungleichung. Das war schon für $p = 2$ nicht ganz einfach, wo die CAUCHY-SCHWARZsche Ungleichung $|\langle \mathbf{x} | \mathbf{y} \rangle| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$ das entscheidende Werkzeug war. Diese Rolle übernimmt im allgemeinen Fall die HÖLDER-Ungleichung

$$\sum_{k=1}^n |x_k y_k| \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q. \quad (5.83)$$

Dabei ist $q > 1$ die Lösung der Gleichung $\frac{1}{p} + \frac{1}{q} = 1$. Für $p = 2$ ergibt sich daraus die CAUCHY-SCHWARZsche Ungleichung:

$$|\langle \mathbf{x} | \mathbf{y} \rangle| = \left| \sum_{k=1}^n \overline{x_k} y_k \right| \leq \sum_{k=1}^n |x_k y_k| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

Ausgangspunkt für (5.83) ist die YOUNGSche Ungleichung in der Form

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \quad (5.84)$$

für $a, b > 0$ und $\frac{1}{p} + \frac{1}{q} = 1$, die wir an anderer Stelle beweisen werden (Satz 11.9.16). Wir definieren $a_i := \frac{|x_i|}{\|x\|_p}$, $b_i := \frac{|y_i|}{\|y\|_q}$ und haben dann nach (5.84)

$$\begin{aligned} \frac{1}{\|x\|_p \|y\|_q} \sum_{i=1}^n |x_i y_i| &= \sum_{i=1}^n a_i b_i \leq \frac{1}{p} \sum_{i=1}^n a_i^p + \frac{1}{q} \sum_{i=1}^n b_i^q = \frac{1}{p} \sum_{i=1}^n \frac{|x_i|^p}{\|x\|_p^p} + \frac{1}{q} \sum_{i=1}^n \frac{|y_i|^q}{\|y\|_q^q} \\ &= \frac{1}{p} + \frac{1}{q} = 1. \end{aligned}$$

Daraus folgt unmittelbar (5.83). Mit dieser Ungleichung lässt sich jetzt die sogenannte MINKOWSKI-Ungleichung beweisen, die einfach die Dreiecksungleichung für $\|\cdot\|_p$ darstellt. Unter Verwendung von $q(p-1) = p$ und $\frac{1}{q} = \frac{p-1}{p}$ schätzen wir folgendermaßen ab:

$$\begin{aligned} \|x + y\|_p^p &= \sum_{i=1}^n |x_i + y_i|^{p-1} \leq \sum_{i=1}^n |x_i| |x_i + y_i|^{p-1} + \sum_{i=1}^n |y_i| |x_i + y_i|^{p-1} \\ &\stackrel{5.83}{\leq} \|x\|_p \left[\sum_{i=1}^n |x_i + y_i|^{(p-1)q} \right]^{\frac{1}{q}} + \|y\|_p \left[\sum_{i=1}^n |x_i + y_i|^{(p-1)q} \right]^{\frac{1}{q}} \\ &= (\|x\|_p + \|y\|_p) \left[\sum_{i=1}^n |x_i + y_i|^p \right]^{\frac{1}{p}(p-1)} = (\|x\|_p + \|y\|_p) \cdot \|x + y\|_p^{p-1}. \end{aligned}$$

Daraus erhalten wir sofort die MINKOWSKI-Ungleichung

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p \quad (5.85)$$

Den Vektorraum $(\mathbb{C}^n, \|\cdot\|_p)$ bezeichnet man als L^p -Raum.

6 Matrizen

In einer Gießerei werden Legierungen aus Kupfer (Cu), Zinn (Sn) und Zink (Zn) benötigt. Als Restposten alter Aufträge liegen noch Legierungen mit 80% Cu, 15% Sn, 5% Zn, sowie 85% Cu, 12% Sn, 3% Zn und 75% Cu, 10% Sn, 15% Zn vor. Kann daraus durch Zusammenschmelzen geeigneter Anteile dieser Legierungen der neue Auftrag mit 81% Cu, 12% Sn und 7% Zn bedient werden?

Der erste Schritt zu einer Lösung dieser Aufgabe besteht darin, die einzelnen Legierungen übersichtlich zusammenzufassen. Wir bedienen uns dabei der Vektorrechnung, indem wir jeder Legierung einen Vektor zuordnen, dessen Einträge die prozentualen Anteile an Cu, Sn und Zn sind. Zur ersten Legierung gehört demnach der Vektor $[80, 15, 5]$, zur zweiten $[85, 12, 3]$ und zur dritten $[75, 10, 15]$. Der Vektor des Auftrags ist $[81, 12, 7]$.

Wir versuchen aus x Anteilen der ersten Legierung, y Anteilen der zweiten und z Anteilen der dritten die gewünschte Legierung zu mischen:

$$x \begin{bmatrix} 80 \\ 15 \\ 5 \end{bmatrix} + y \begin{bmatrix} 85 \\ 12 \\ 3 \end{bmatrix} + z \begin{bmatrix} 75 \\ 10 \\ 15 \end{bmatrix} = \begin{bmatrix} 80x + 85y + 75z \\ 15x + 12y + 10z \\ 5x + 3y + 15z \end{bmatrix} = \begin{bmatrix} 81 \\ 12 \\ 7 \end{bmatrix}.$$

Daraus erhalten wir das lineare Gleichungssystem

$$\begin{aligned} 80x + 85y + 75z &= 81 \\ 15x + 12y + 10z &= 12 \\ 5x + 3y + 15z &= 7, \end{aligned}$$

für das wir ein systematisches Lösungsverfahren benötigen.

6.1 Lineare Gleichungssysteme und das Gauss-Verfahren

Ein *lineares Gleichungssystem (LGS)* besteht im Allgemeinen aus m Gleichungen der Form

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = b_i, \quad i = 1, \dots, m$$

für die n Unbekannten $x_1, \dots, x_n \in \mathbb{K}$. Dabei sind für $i = 1, \dots, n$ und $j = 1, \dots, m$ die Zahlen $a_{ij} \in \mathbb{K}$ – die *Koeffizienten* des Gleichungssystems – fest gewählt. Ausgeschrieben:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ a_{31}x_1 + a_{32}x_2 + \cdots + a_{3n}x_n &= b_3 \\ \vdots &\quad \vdots &\quad \vdots &\quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m. \end{aligned} \tag{6.1}$$

Sind die Zahlen b_1, \dots, b_m auf der rechten Seite alle gleich Null, so heißt das Gleichungssystem *homogen*, andernfalls *inhomogen*.

Um zu dem Lösungsverfahren für (6.1) zu gelangen, das als GAUSS-Verfahren bekannt ist, machen wir uns klar, welche Operationen eine Gleichung in eine äquivalente überführt: Zu einer Gleichung kann auf beiden Seiten dieselbe Zahl addiert werden. Außerdem kann man beide Seiten mit ein und derselben Zahl multiplizieren, solange es sich dabei nicht um Null handelt. Wenn wir diese beiden Möglichkeiten geschickt kombinieren, können wir (6.1) soweit vereinfachen, daß die Existenz einer Lösung leicht zu erkennen und diese gegebenenfalls auch einfach zu bestimmen ist. Das Verfahren verläuft nach folgendem Schema. Wir dürfen o. B. d. A. davon ausgehen, daß $a_{11} \neq 0$ ist, denn andernfalls können wir uns die Gleichungen so vertauscht denken, daß das erste Element der ersten Gleichung nicht verschwindet (sollten alle Elemente a_{1j} der ersten Spalte gleich Null sein, dann wenden wir unser Verfahren auf die erste Spalte an, die wenigstens einen nicht verschwindenden Koeffizienten a_{ij} aufweist). Sollte $a_{21} = 0$ sein, dann ist für den Moment an der zweiten Gleichung nichts zu ändern. Andernfalls dürfen wir die erste Gleichung mit a_{21} und die zweite mit a_{11} multiplizieren:

$$\begin{aligned} a_{21}a_{11}x_1 + a_{21}a_{12}x_2 + \cdots + a_{21}a_{1n}x_n &= a_{21}b_1 \\ a_{11}a_{21}x_1 + a_{11}a_{22}x_2 + \cdots + a_{11}a_{2n}x_n &= a_{11}b_2. \end{aligned}$$

Jetzt können wir auf beiden Seiten der zweiten Gleichung die Zahl $a_{21}b_1$ abziehen. Auf der rechten erhalten wir dabei $a_{11}b_2 - a_{21}b_1$. Der Trick ist nun, für die linke Seite die Zahl $a_{21}b_1$ durch den Ausdruck $a_{21}a_{11}x_1 + a_{21}a_{12}x_2 + \cdots + a_{21}a_{1n}x_n$ zu ersetzen, von dem die erste Gleichung ja behauptet, daß er mit $a_{21}b_1$ übereinstimmt. Wenn das Gleichungssystem lösbar ist, kann das zu keinem Widerspruch führen, da die erste Gleichung dann durch geeignete Zahlen x_1, \dots, x_n erfüllt werden kann. Wie man sieht ist der Sinn dieses Verfahrens, die Unbekannte x_1 aus der zweiten Gleichung zu eliminieren:

$$\begin{aligned} a_{21}a_{11}x_1 + a_{21}a_{12}x_2 + \cdots &= a_{21}b_1 \\ (a_{11}a_{22} - a_{21}a_{12})x_2 + \cdots &= a_{11}b_2 - a_{21}b_1. \end{aligned}$$

Wenn wir die erste Gleichung durch Division mit a_{21} wieder in ihre Ausgangsform zurück verwandeln, ergibt sich

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots &= b_1 \\ (a_{11}a_{22} - a_{21}a_{12})x_2 + \cdots &= a_{11}b_2 - a_{21}b_1 \\ a_{31}x_1 + a_{32}x_2 + \cdots &= b_3 \\ \vdots &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots &= b_m. \end{aligned}$$

Diesen Vorgang werden wir künftig meist dadurch abkürzen, daß wir einfach das a_{21} -fache der ersten Gleichung von der modifizierten zweiten abziehen, ohne dafür die erste Gleichung tatsächlich zu verändern. Nun wenden wir dieses Verfahren auf die dritte, die vierte etc. an bis

wir bei der letzten angelangt sind. Wir erhalten dabei ein neues Gleichungssystem der Form

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + & \cdots & + & a_{1n}x_n = b_1 \\ \tilde{a}_{22}x_2 & + & \tilde{a}_{23}x_3 & + & \cdots & + & \tilde{a}_{2n}x_n & = & \tilde{b}_1 \\ \tilde{a}_{32}x_2 & + & \tilde{a}_{33}x_3 & + & \cdots & + & \tilde{a}_{3n}x_n & = & \tilde{b}_3 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ \tilde{a}_{m2}x_2 & + & \tilde{a}_{m3}x_3 & + & \cdots & + & \tilde{a}_{mn}x_n & = & \tilde{b}_m \end{array}$$

mit neuen Koeffizienten \tilde{a}_{ij} , die in der ersten Spalte, bis auf den ersten, alle verschwinden. Jetzt wenden wir dieses Verfahren auf die zweite Spalte an. Dabei arbeiten wir nun mit der zweiten Gleichung. Wir können o. B. d. A. wieder $\tilde{a}_{22} \neq 0$ annehmen, denn andernfalls vertauschen wir die Gleichungen 2, ..., m, bis der Eintrag \tilde{a}_{22} ungleich Null ist. Sollten alle Koeffizienten $\tilde{a}_{22}, \dots, \tilde{a}_{m2}$ verschwinden, so machen wir mit der dritten Spalte weiter, vertauschen gegebenenfalls die Gleichungen bis $\tilde{a}_{23} \neq 0$ gilt, usw. Wie oben beschrieben bringen wir die Koeffizienten \tilde{a}_{32} bis \tilde{a}_{m2} , aber auch a_{12} zum Verschwinden (bzw. \tilde{a}_{33} bis \tilde{a}_{m3} und a_{13}). Das Gleichungssystem hat mittlerweile die Form

$$\begin{array}{ccccccccc} a_{11}x_1 & & + & \hat{a}_{13}x_3 & + & \cdots & + & \hat{a}_{1n}x_n & = \hat{b}_1 \\ \tilde{a}_{22}x_2 & + & \tilde{a}_{23}x_3 & + & \cdots & + & \tilde{a}_{2n}x_n & = & \tilde{b}_1 \\ \hat{a}_{33}x_3 & + & \cdots & + & \hat{a}_{3n}x_n & = & \hat{b}_3 \\ \vdots & & & & \vdots & & \vdots \\ \hat{a}_{m3}x_3 & + & \cdots & + & \hat{a}_{mn}x_n & = & \hat{b}_m \end{array}$$

bzw., falls die zweite Spalte bereits bei der Bearbeitung der ersten mit verschwunden ist:

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & & + & \hat{a}_{14}x_4 & + & \cdots & + & \hat{a}_{1n}x_n = \hat{b}_1 \\ & & \tilde{a}_{23}x_3 & + & \tilde{a}_{24}x_4 & + & \cdots & + & \tilde{a}_{2n}x_n = \tilde{b}_1 \\ & & \hat{a}_{34}x_4 & + & \cdots & + & \hat{a}_{3n}x_n & = & \hat{b}_3 \\ & & \vdots & & & & \vdots & & \vdots \\ & & \hat{a}_{m4}x_4 & + & \cdots & + & \hat{a}_{mn}x_n & = & \hat{b}_m \end{array}$$

\hat{a}_{ij} und \hat{b}_i sind die Koeffizienten, die sich bei diesem zweiten Durchgang ergeben haben. Auf diese Weise fahren wir fort, bis die letzte Spalte abgearbeitet ist. Dabei kann es durchaus passieren, daß bei einem Arbeitsschritt eine ganze Gleichung verschwindet. Das bedeutet, daß von den m Ausgangsgleichungen wenigstens eine überflüssig ist und daß von den m Unbekannten, die durch die m Gleichungen festgelegt werden sollen tatsächlich höchstens $m - 1$ zu bestimmen sind. Pro verschwundener Gleichung erhöht sich die Anzahl der Unbekannten, die nicht festgelegt werden können um eins. Daß solche Situationen auftreten ist zu erwarten. Man muß sich ja nur vorstellen, daß man zwei Gleichungen für drei Unbekannte hat, sagen wir

$$\begin{aligned} 2x_1 + 3x_2 - 4x_3 &= -3 \\ 5x_1 - 2x_2 + 3x_3 &= 10. \end{aligned}$$

Sicher werden wir nichts Neues erfahren, wenn wir das Doppelte der ersten von der zweiten abziehen und das Ergebnis als dritte Gleichung verwenden. Obwohl es sich dann um drei Glei-

chungen handelt, liefert eine von ihnen keine weiteren Informationen. Man kann sich vorstellen, daß solche Beispiele mit wachsender Anzahl von Gleichungen beliebig verwickelt werden können. Das GAUSS-Verfahren entdeckt sie jedoch alle.

Es ist auch möglich, daß bei einem Arbeitsschritt nur die linke Seite, nicht aber die rechte verschwindet. In diesem Fall ist das Gleichungssystem nicht lösbar, denn unter der Annahme, daß es lösbar ist (und nur unter dieser Prämisse können wir das Verfahren rechtfertigen) sind wir zu einem Widerspruch in Form einer unlösbaren Gleichung gelangt. Das bedeutet, daß die Annahme der Lösbarkeit nicht wahr sein kann.

Nach Beendigung des Verfahrens ist die ursprünglich rechteckige Gestalt des Gleichungssystems in eine Stufenform verwandelt worden, aus der die Lösungen, falls sie existieren, leicht bestimmt werden können. Ein typisches Ergebnis könnte etwa folgendermaßen aussehen:

$$\begin{array}{lll} a_{11}x_1 + a_{12}x_2 & + \bar{a}_{15}x_5 + \bar{a}_{16}x_6 & = \bar{b}_1 \\ \bar{a}_{23}x_3 & + \bar{a}_{25}x_5 + \bar{a}_{26}x_6 & = \bar{b}_2 \\ \bar{a}_{34}x_4 + \bar{a}_{35}x_5 + \bar{a}_{36}x_6 & & = \bar{b}_3 \\ & \bar{a}_{47}x_7 & = \bar{b}_4, \end{array}$$

oder mit konkreten Zahlen

$$\begin{array}{lll} x_1 - 2x_2 & + 7x_5 - 5x_6 & = 1 \\ x_3 & - x_5 - 8x_6 & = 4 \\ x_4 + 9x_5 + 7x_6 & & = 6 \\ & & x_7 = 2. \end{array}$$

Hier wären die gesuchten Größen x_1, x_3, x_4 und x_7 für jede Wahl von x_2, x_5 und x_6 eindeutig zu berechnen:

$$x_1 = 1 + 2x_2 - 7x_5 + 5x_6, \quad x_3 = 4 + x_5 + 8x_6, \quad x_4 = 6 - 9x_5 - 7x_6, \quad x_7 = 2.$$

Anders gesagt: Wir können das System nach x_1, x_3, x_4 und x_7 auflösen, während x_2, x_5 und x_6 nicht weiter bestimmbar sind. Jede Wahl von x_2, x_5 und x_6 führt nach unserer Lösungsformel zu einer Lösung. Es gibt daher unendlich viele.

Und wie sehen diese Lösungen aus?

Wenn wir vereinbaren, die Lösungen als Elemente $[x_1, x_2, x_3, x_4, x_5, x_6, x_7]$ aus den Vektorraum \mathbb{K}^7 aufzufassen (in diesem Beispiel), dann erhalten wir alle Lösungen als Elemente einer Menge L, nämlich der Menge aller $[x_1, x_2, x_3, x_4, x_5, x_6, x_7] \in \mathbb{K}^7$, für die x_1, x_3, x_4 und x_7 wie oben als Funktionen der beliebig wählbaren x_2, x_5 und x_6 bestimmt sind:

$$L = \left\{ \begin{array}{c|c} \begin{bmatrix} 1 + 2x_2 - 7x_5 + 5x_6 \\ x_2 \\ 4 + x_5 + 8x_6 \\ 6 - 9x_5 - 7x_6 \\ x_5 \\ x_6 \\ 2 \end{bmatrix} & x_2, x_5, x_6 \in \mathbb{K} \end{array} \right\}.$$

Um die Lösungsstruktur zu erkennen, ergänzen wir die Koeffizienten, die Eins oder Null sind:

$$L = \left\{ \begin{array}{c|c} \begin{bmatrix} 1 + 2x_2 - 7x_5 + 5x_6 \\ 0 + 1x_2 + 0x_5 + 0x_6 \\ 4 + 0x_2 + 1x_5 + 8x_6 \\ 6 + 0x_2 - 9x_5 - 7x_6 \\ 0 + 0x_2 + 1x_5 + 0x_6 \\ 0 + 0x_2 + 0x_5 + 1x_6 \\ 2 + 0x_2 + 0x_5 + 0x_6 \end{bmatrix} & x_2, x_5, x_6 \in \mathbb{K} \end{array} \right\}.$$

Berücksichtigen wir nun noch die Rechenregeln für Vektoren in \mathbb{K}^7 , so lässt sich das systematisch und übersichtlich folgendermaßen darstellen:

$$L = \left\{ \begin{array}{c|c} \begin{bmatrix} 1 \\ 0 \\ 4 \\ 6 \\ 0 \\ 0 \\ 2 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + x_5 \begin{bmatrix} -7 \\ 0 \\ 1 \\ -9 \\ 1 \\ 0 \end{bmatrix} + x_6 \begin{bmatrix} 5 \\ 0 \\ 8 \\ -7 \\ 0 \\ 1 \end{bmatrix} & x_2, x_5, x_6 \in \mathbb{K} \end{array} \right\}.$$

Durch Einsetzen bestätigt man, daß $[1, 0, 4, 6, 0, 0, 2]$ eine Lösung des inhomogenen Gleichungssystems ist. Die Vektoren $[2, 1, 0, 0, 0, 0, 0]$, $[-7, 0, 1, -9, 1, 0, 0]$ und $[5, 0, 8, -7, 0, 1, 0]$ sind Lösungen des homogenen Systems

$$\begin{aligned} x_1 - 2x_2 &+ 7x_5 - 5x_6 = 0 \\ x_3 &- x_5 - 8x_6 = 0 \\ x_4 + 9x_5 + 7x_6 &= 0 \\ &x_7 = 0. \end{aligned}$$

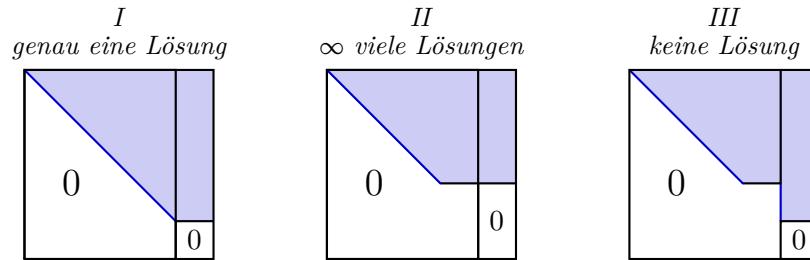
Diese Zweiteilung ist hier kein Zufall, sondern spiegelt die wesentliche Struktur der Lösungen linearer Gleichungssysteme wieder. Zum Schluß wollen wir uns noch klar machen, wie denn die eindeutige Lösbarkeit zu erkennen ist. Bei dem hier vorgestellten vollständigen GAUSS-Verfahren, bei dem nicht nur die Elemente unter der Diagonalen, sondern auch die oberhalb zum Verschwinden gebracht werden, ist das denkbar einfach: Das Verfahren muß bei einer voll besetzten Diagonale enden. Die gesuchte Lösung steht dann auf der rechten Seite. Das könnte z. B. folgendermaßen aussehen

$$\begin{aligned} x_1 &= 1 \\ x_2 &= 4 \\ x_3 &= 2 \\ x_4 &= 3, \end{aligned}$$

wenn wir ursprünglich von vier Gleichungen ausgegangen wären.

Es gibt drei typische Zustände für das GAUSS-Verfahren, die (eventuell nach Vertauschen von geeigneten Spalten) immer zu erreichen sind. An ihnen ist die *Anzahl* möglicher Lösungen sofort zu erkennen, auch wenn das für die Lösungen selbst i. A. noch nicht gilt. Es gibt durchaus Situationen, in denen diese Information bereits völlig ausreichend ist (vergl. 6.4)

Diese Zustände sind in der folgenden Skizze schematisch wiedergegeben. Die stark gezeichneten Linien meinen dabei Einträge $\neq 0$ während die unterlegten Flächen beliebige Einträge symbolisieren.



Ein Beispiel für die Situation *genau eine Lösung* ist etwa

$$\begin{aligned} x_1 - 2x_2 &= -2x_4 = 5 \\ x_2 + x_3 + 2x_4 &= 4 \\ x_3 + 3x_4 &= 2 \\ x_4 &= 3. \end{aligned}$$

Wenn man nur an der Anzahl der Lösungen interessiert ist, dann ist man mit dieser sog. *oberen Dreiecksform* fertig. Die letzte Gleichung legt nämlich x_4 als 3 fest. Das in die vorletzte Gleichung eingesetzt, ermöglicht es x_3 als $2 - 3 \cdot 3 = -7$ zu bestimmen. Auf diese Weise könnte man fortfahren und auch noch x_2 und x_1 ausrechnen. Wenn man die konkrete Lösung gar nicht wissen muß, begnügt man sich damit festzustellen, daß dieses Verfahren immer durchzuführen ist und immer zu genau einem Ergebnis führt.

Die Situation ∞ viele Lösungen könnte folgendermaßen aussehen:

$$\begin{aligned} x_1 + 2x_3 - x_4 &+ 7x_5 - 5x_6 - 2x_2 = 2 \\ x_3 + 2x_4 &- x_5 - 8x_6 = 3 \\ x_4 &+ 9x_5 + 7x_6 = 1 \\ x_7 &= 4. \end{aligned}$$

Hier ist x_7 durch die Zahl 4 bestimmt. x_4 lässt sich in Abhängigkeit von x_5 und x_6 berechnen: $x_4 = 1 - 9x_5 - 7x_6$. Mit dieser Lösung könnte man aus der zweiten Gleichung auf x_3 und damit endlich auf x_1 schließen. Auf diese Weise haben wir uns noch einmal klar gemacht, daß das Gleichungssystem auf jeden Fall lösbar ist. Da die Variablen x_2 , x_5 und x_6 aber nicht weiter bestimmt werden können, liefert jede ihrer Belegungen mit konkreten Zahlen eine Lösung. Das zeigt deutlich, daß es sich um unendlich viele Lösungen handelt. Wieder ist es ausreichend,

dieses Ergebnis an der oberen Dreiecksform abzulesen, wenn man sich nur über die Anzahl der Lösungen informieren will.

Schließlich die Situation *keine Lösung*. Sie ist leicht daran zu erkennen, daß die linke Seite einer Gleichung bei einer Umformung des GAUSS-Verfahrens komplett verschwindet, während die rechte Seite von Null verschieden bleibt.

6.2 Die Matrix zum LGS

Wir führen den Begriff *Matrix* und eine Multiplikation zwischen Matrix und Vektor ein, um die Lösungsstruktur linearer Gleichungssysteme besser herauszuarbeiten. Diese neuen Begriffe sind dabei zunächst nur bequeme Abkürzungen, die uns unsere Aufgabe erleichtern. Aber wie so oft in der Mathematik werden diese Dinge zusehends ein Eigenleben entwickeln, das weit über die ursprüngliche Aufgabenstellung hinausreicht.

Für das LGS (6.1) definieren wir die $m \times n$ -Matrix A der Koeffizienten $a_{11}, a_{12}, \dots, a_{mn}$ als das rechteckige Zahlenschema

$$A := \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}, \quad (6.2)$$

mit m Zeilen und n Spalten. Einen Vektor x schreiben wir ab jetzt als Spaltenvektor, oder anders gesagt, als eine $n \times 1$ -Matrix:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Die sogenannte *Transponierte* A^t einer $m \times n$ -Matrix A entsteht, indem man die Zeilen zu Spalten macht, oder indem man die Einträge außerhalb der Diagonalen an dieser spiegelt. Man erhält die $n \times m$ -Matrix

$$A^t := \begin{bmatrix} a_{11} & a_{21} & a_{31} & \cdots & a_{m1} \\ a_{12} & a_{22} & a_{32} & \cdots & a_{m2} \\ a_{13} & a_{23} & a_{33} & \cdots & a_{m3} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & a_{3n} & \cdots & a_{mn} \end{bmatrix}. \quad (6.3)$$

Insbesondere ist also $x = [x_1, x_2, \dots, x_n]^t$.

Eine $m \times n$ -Matrix wird nach folgender Vorschrift mit einem Vektor $\mathbf{x} \in \mathbb{K}^n$ multipliziert

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ a_{31} & a_{32} & \cdots & a_{3n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} := \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ a_{31}x_1 + a_{32}x_2 + \cdots + a_{3n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix}. \quad (6.4)$$

Der Ergebnisvektor ist aus \mathbb{K}^m . Man sieht, daß die Multiplikation gerade so konstruiert ist, daß die Einträge dieses Vektors die linke Seite des LGS (6.1) wiedergeben. Führen wir noch $\mathbf{b} := [b_1, b_2, b_3, \dots, b_m]^t \in \mathbb{K}^m$ ein, so läßt sich (6.1) in die übersichtliche Form

$$A\mathbf{x} = \mathbf{b} \quad (6.5)$$

bringen (wie üblich schreiben wir meist $A\mathbf{x}$ statt $A \cdot \mathbf{x}$). Ist die rechte Seite \mathbf{b} nicht der Nullvektor, so nennen wir das LGS *inhomogen*, andernfalls *homogen*. Eine Lösung \mathbf{x} des inhomogenen Systems heißt *inhomogene*, eine des homogenen *homogene Lösung*. Die folgende Regel läßt sich einfach nachrechnen:

$$A(s\mathbf{x} + t\mathbf{y}) = sA\mathbf{x} + tA\mathbf{y} \quad (6.6)$$

für alle $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$ und $s, t \in \mathbb{K}$, denn sie läßt sich auf das Distributivgesetz $a_{ij}(sx_j + ty_j) = sa_{ij}x_j + ta_{ij}y_j$ gewöhnlicher Zahlen $s, t, x_j, y_j, a_{ij} \in \mathbb{K}$ zurückführen. Diese Gleichung zeigt, daß die Multiplikation mit der Bildung von Linearkombinationen verträglich ist, in dem Sinne, daß das Produkt mit einer Linearkombination die Linearkombination der Produkte ist. Wir sprechen künftig in diesem Zusammenhang von der *Linearität* der Multiplikation von Matrix und Vektor.

Jetzt läßt sich die Lösungsstruktur eines inhomogenen LGSs leicht verstehen. Ist \mathbf{x}_1 eine inhomogene Lösung des LGSs und \mathbf{x}_0 eine homogene – es gilt also $A\mathbf{x}_1 = \mathbf{b}$ und $A\mathbf{x}_0 = \mathbf{0}$ – so ist $\mathbf{x}_1 + t\mathbf{x}_0$ wieder eine inhomogene Lösung:

$$A(\mathbf{x}_1 + t\mathbf{x}_0) = A\mathbf{x}_1 + tA\mathbf{x}_0 = \mathbf{b} + t \cdot \mathbf{0} = \mathbf{b}.$$

Man kann zu einer inhomogenen Lösung beliebige homogene addieren, ohne die Eigenschaft inhomogene Lösung zu sein zu stören. Um es ganz deutlich zu sagen:

Eine beliebige Lösung \mathbf{x} eines inhomogenen LGSs läßt sich immer aus einer inhomogenen Lösung \mathbf{x}_1 durch Addition einer geeigneten homogenen Lösung \mathbf{x}_0 gewinnen: $\mathbf{x} = \mathbf{x}_0 + \mathbf{x}_1$.

Ist nämlich \mathbf{x} eine beliebige inhomogene Lösung, dann gilt $A(\mathbf{x} - \mathbf{x}_1) = A\mathbf{x} - A\mathbf{x}_1 = \mathbf{b} - \mathbf{b} = \mathbf{0}$. Also ist $\mathbf{x}_0 := \mathbf{x} - \mathbf{x}_1$ eine homogene Lösung und $\mathbf{x} = \mathbf{x}_1 + \mathbf{x} - \mathbf{x}_1 = \mathbf{x}_1 + \mathbf{x}_0$.

Diese einfache Lösungsstruktur begegnet uns bei allen Arten linearer Gleichungen, etwa bei linearen Rekurrenzgleichungen 9, oder bei linearen Differentialgleichungen.

6.3 Das Schema zum Gauss-Verfahren

6.3.1 Beispiel

Wir lösen das inhomogene LGS

$$\begin{array}{lll} \text{I} & 2x_1 + 5x_2 + 7x_3 + 3x_4 + 3x_5 = & 0 \\ \text{II} & 3x_1 + 6x_2 + 6x_3 + 3x_4 + 3x_5 = & 6 \\ \text{III} & 2x_1 + 4x_2 + 4x_3 + 2x_4 + 3x_5 = & 3 \\ \text{IV} & x_1 + 3x_2 + 5x_3 + 2x_4 + 3x_5 = & -3. \end{array}$$

Dabei stehen uns die folgenden *Zeilenumformungen* zur Verfügung (vergl. 6.1):

1. Multiplikation einer Zeile mit einer Zahl $a \neq 0$.
2. Addiere / subtrahiere eine Zeile zu / von einer anderen.
3. Vertausche zwei Zeilen / Spalten.

Hier könnte das bedeuten: Teile Gleichung II durch 3 und vertausche II mit I. Ziehe das Doppelte der Gleichung I von II ab. Wir werden das beim *GAUSS-Verfahren*, das wir an diesem Beispiel vorführen wollen, folgendermaßen notieren: $\text{II} : 3$, $\text{I} \leftrightarrow \text{II}$, $\text{II} - 2 \cdot \text{I}$, etc.

Bei allen Rechenoperationen sind letztlich nur die Koeffizienten der Variablen x_1, x_2, \dots, x_5 betroffen. Wenn wir etwa das Doppelte der vierten Gleichung von der dritten abziehen wollen, $\text{III} - 2 \cdot \text{IV}$, so erhalten wir

$$(2 - 2) \cdot x_1 + (4 - 6) \cdot x_2 + (4 - 10) \cdot x_3 + (2 - 4) \cdot x_4 + (3 - 6) \cdot x_5 = 3 + 6.$$

Das *GAUSS-Verfahren* erfordert viele dieser Rechenschritte. Um den Schreibaufwand dabei in Grenzen zu halten, lässt man die Variablen x_1, \dots, x_5 einfach weg und notiert das Ergebnis einer Umformung nur als Zahlenreihe

$$0 \quad -2 \quad -6 \quad -2 \quad -3 \quad 9.$$

Dabei steht an der ersten Position der Koeffizient von x_1 , an der zweiten der von x_2 usw. und an der letzten die rechte Seite der Gleichung. Daher besteht der erste Schritt beim *GAUSS-Verfahren* darin, das Gleichungssystem in ein Zahlenschema zu verwandeln, in dem nur noch die Koeffizienten auftauchen:

	x_1	x_2	x_3	x_4	x_5	
I	2	5	7	3	3	0
II	3	6	6	3	3	6
III	2	4	4	2	3	3
IV	1	3	5	2	3	-3
I	1	2	2	1	1	2
II	2	5	7	3	3	0
III	2	4	4	2	3	3
IV	1	3	5	2	3	-3

Nachdem Zeile I mit Zeile II vertauscht wurde, ist die neu entstandene Zeile I die Arbeitszeile. Mit ihr werden die anderen verändert. Dabei wird versucht, durch geschickte Addition oder Subtraktion geeigneter Vielfacher des ersten Elements die restlichen Spaltelemente zum Verschwinden zu bringen.

	x_1	x_2	x_3	x_4	x_5	
I	1	2	2	1	1	2
II	0	1	3	1	1	-4
III	0	0	0	0	1	-1
IV	0	1	3	1	2	-5
						I - 2 · II
I	1	0	-4	-1	-1	10
II	0	1	3	1	1	-4
III	0	0	0	0	1	-1
IV	0	0	0	0	1	-1
I	1	0	-4	-1	0	9
II	0	1	3	1	0	-3
III	0	0	0	0	1	-1

Anschließend ist Zeile II die Arbeitszeile, mit der die nicht verschwindenden Spalteneinträge 2 und 1 der zweiten Spalte beseitigt werden. Im vorletzten Schritt ist Zeile IV überflüssig geworden, da sie mit Zeile III übereinstimmt. Das Verfahren ist beendet, wenn jede weitere Zeilenumformung bereits beseitigte Einträge wieder zum Erscheinen bringen würde. Die Lösung des LGS erhalten wir, wenn wir das System in die Gleichungen zurück übersetzen, die es repräsentiert:

$$\begin{array}{lllll} \text{I} & x_1 & -4x_3 - x_4 & = & 9 \\ \text{II} & & x_2 + 3x_3 + x_4 & = & -3 \\ \text{III} & & & & x_5 = -1. \end{array}$$

x_1, x_2 lassen sich in Abhängigkeit von x_3 und x_4 angeben. $x_5 = -1$ ist konstant. Damit können wir den Lösungsvektor $\mathbf{x} = [x_1, x_2, x_3, x_4, x_5]^t$ folgendermaßen aufschreiben:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 9 + 4x_3 + 1x_4 \\ -3 - 3x_3 - 1x_4 \\ 0 + 1x_3 + 0x_4 \\ 0 + 0x_3 + 1x_4 \\ -1 + 0x_3 + 0x_4 \end{bmatrix} = \begin{bmatrix} 9 \\ -3 \\ 0 \\ 0 \\ -1 \end{bmatrix} + x_3 \begin{bmatrix} 4 \\ -3 \\ 1 \\ 0 \\ 0 \end{bmatrix} + x_4 \begin{bmatrix} 1 \\ -1 \\ 0 \\ 1 \\ 0 \end{bmatrix}.$$

Das bedeutet, die Lösungsmenge L ist

$$L = \left\{ \begin{bmatrix} 9 \\ -3 \\ 0 \\ 0 \\ -1 \end{bmatrix} + t \begin{bmatrix} 4 \\ -3 \\ 1 \\ 0 \\ 0 \end{bmatrix} + s \begin{bmatrix} 1 \\ -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \mid t, s \in \mathbb{R} \right\}.$$

6.3.2 A Finden Sie die Lösung für das Beispiel der Gießerei aus der Einführung

$$\begin{aligned} 80x + 85y + 75z &= 81 \\ 15x + 12y + 10z &= 12 \\ 5x + 3y + 15z &= 7. \end{aligned}$$

6.3.3 A Finden Sie alle Lösungen des Gleichungssystems

$$\begin{aligned} 2x + 3y - 16z + 14w &= 16 \\ 7x + 7y - 34z + 20w &= -1 \\ -3x - y + 2z + 8w &= 33 \\ 31x + 22y - 94z + 14w &= -151. \end{aligned}$$

6.3.4 A Ein magisches Quadrat

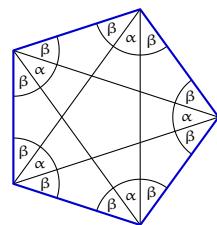
$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

hat gleiche Spalten-, Zeilen- und Diagonalensumme s und nur Einträge aus \mathbb{N} . Das führt auf die acht linearen Gleichungen

$$\begin{aligned} a + b + c &= s \\ d + e + f &= s \\ g + h + i &= s \\ a + d + g &= s \\ b + e + h &= s \\ c + f + i &= s \\ a + e + i &= s \\ c + e + g &= s. \end{aligned}$$

Finden sie ein magisches Quadrat mit lauter verschiedenen Einträgen.

6.3.5 A Stellen Sie für die beiden Winkel α und β ein lineares Gleichungssystem auf und lösen Sie es.



6.3.6 Die Grenzen des GAUSS-Verfahrens Das GAUSS-Verfahren besteht im Wesentlichen in der geschickten Anwendung *äquivalenter Zeilenumformungen*: Multiplikation einer Zeile mit einer Zahl $c \neq 0$ und Addition des Vielfachen einer anderen Zeile. Sie heißen äquivalent, weil sich durch sie die Lösungsmenge des Systems nicht ändert. Das ist eine elementare Überlegung.

Weil sie aber so grundlegend ist, soll sie doch kurz vorgeführt werden. Wir gehen von einem Gleichungssystem

$$\begin{array}{cccccc} \vdots & & \vdots & & \vdots & \vdots \\ a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n & = & b_i \\ \vdots & & \vdots & & \vdots & \vdots \\ a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n & = & b_k \\ \vdots & & \vdots & & \vdots & \vdots \end{array}$$

aus und addieren das c -Fache der i -ten Gleichung zur k -ten:

$$\begin{array}{cccccc} \vdots & & \vdots & & \vdots & \vdots \\ a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n & = & b_i \\ \vdots & & \vdots & & \vdots & \vdots \\ (a_{k1} + ca_{i1})x_1 + (a_{k2} + ca_{i2})x_2 + \cdots + (a_{kn} + ca_{in})x_n & = & b_k + cb_i \\ \vdots & & \vdots & & \vdots & \vdots \end{array}$$

Ist $\mathbf{x} = [x_1, x_2, \dots, x_n]^t$ eine Lösung des ersten Gleichungssystems, dann löst \mathbf{x} auch das zweite. Dort hat sich ja nur die k -te Gleichung verändert, alle anderen behielten ihre ursprüngliche Form. Sie bleiben also weiterhin gültig. Es ist demnach nur zu zeigen, daß die neue k -te Gleichung erfüllt wird. Das ist simpel, denn dafür müssen wir ihre Terme nur geeignet umsortieren:

$$\begin{aligned} & (a_{k1} + ca_{i1})x_1 + (a_{k2} + ca_{i2})x_2 + \cdots + (a_{kn} + ca_{in})x_n \\ &= a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n + c(a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n) \\ &= b_k + cb_i. \end{aligned}$$

Nun sei \mathbf{x} eine Lösung des zweiten Systems. Jetzt müssen wir zeigen, daß die k -te Gleichung des ersten Systems erfüllt wird, denn die anderen sind ja weiterhin in beiden Systemen identisch. Insbesondere gilt die i -te Gleichung, so daß wir $0 = c(a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n) - cb_i$ in der folgenden Rechnung verwenden dürfen:

$$\begin{aligned} & a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n \\ &= a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n + c(a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n) - cb_i \\ &= (a_{k1} + ca_{i1})x_1 + (a_{k2} + ca_{i2})x_2 + \cdots + (a_{kn} + ca_{in})x_n - cb_i \\ &= b_k + cb_i + cb_i = b_k. \end{aligned}$$

Damit haben wir gezeigt, daß \mathbf{x} genau dann eine Lösung des ersten Systems ist, wenn \mathbf{x} eine Lösung des zweiten darstellt. Dasselbe gilt natürlich auch, wenn die Zeilenumformung nur in der Multiplikation mit einer Zahl $c \neq 0$ besteht.

Streng genommen dürfte das GAUSS-Verfahren in jedem Schritt nur in einer solchen äquivalenten Zeilenumformung bestehen. Wir haben aber mit einer Zeile i und einem Arbeitselement $a_{ir} \neq 0$ immer gleich alle Einträge a_{kr} oberhalb ($k < i$) und alle unterhalb ($k > i$) in einem Arbeitsgang eliminiert. Das ist jedoch gerechtfertigt, denn diesen Vorgang kann man sich leicht in

eine mehrfache Anwendung elementarer Zeilenumformungen aufgelöst denken, wie wir sie oben beschrieben haben. Dieses Vorgehen ist natürlich vernünftig, da das GAUSS-Verfahren sonst in endlose Schreibarbeit ausufern würde. Man kann sogar von diesem starren Verfahren abweichen und ‚Schwächen‘ eines Gleichungssystems ausnutzen, die darin bestehen können, daß durch geschickte Kombination von Gleichungen ein beträchtlicher Teil der Koeffizienten mit einem Mal zum Verschwinden gebracht werden können. Bei dem Gleichungssystem

	x_1	x_2	x_3	x_4	x_5	
I	3	1	0	0	0	1
II	1	1	6	1	1	1
III	4	4	24	2	3	3
IV	-1	-1	0	0	1	-3

bietet es sich natürlich an, nicht sofort mit dem Arbeitselement 3 der ersten Zeile zu arbeiten, sondern zunächst die günstige Lage der Koeffizienten in der zweiten, dritten und vierten Zeile auszunutzen:

I	3	1	0	0	0	1	
II	1	1	6	1	1	1	$3 II - I$
III	0	0	0	-2	-1	-1	$III \leftrightarrow IV$
IV	0	0	6	1	2	-2	
I	3	1	0	0	0	1	
II	0	2	18	3	3	2	
III	0	0	6	1	2	-2	
IV	0	0	0	-2	-1	-1	

Das ist solange gerechtfertigt, wie wir bei solchen Abkürzungen nur äquivalente Zeilenumformungen verwenden. Wenn man aber sehr geniale, weil sehr komplexe Möglichkeiten der Vereinfachung entdeckt, kann es leicht passieren, daß man den Überblick verliert und unbemerkt Zeilenumformungen durchführt, die nicht mehr äquivalent sind. Wie so etwas aussehen könnte, zeigt folgendes einfache Beispiel:

6.3.7 Beispiel

I	3	1	0	0	0	1	$I + II$
II	1	1	6	1	1	1	$II + I$
III	1	2	2	1	3	3	$III - IV$
IV	1	-1	0	0	1	-3	$IV - III$
I	4	2	6	1	1	2	
II	4	2	6	1	1	2	$II - I$
III	0	3	2	1	2	6	
IV	0	-3	-2	-1	-2	-6	$IV + III$
I	4	2	6	1	1	2	
III	0	3	2	1	2	6	

Wie durch ein Wunder sind wir auf einmal zwei Gleichungen losgeworden. Da wir bei diesem ‚Trick‘ aber gar nicht auf eine spezielle Form der Zeileneinträge angewiesen sind, sollten wir mißtrauisch werden und unser Verfahren noch einmal überdenken. Dabei ist es gar nicht so offensichtlich, wo die beiden Gleichungen eigentlich geblieben sind. Schließlich haben wir ja nur Gleichungen addiert und subtrahiert. Eine Lösung des Ausgangssystems ist daher auch eine des entstandenen Systems. Da dieses aber nur noch aus zwei Gleichungen besteht, kann es *mehr* Lösungen als das Ausgangssystem haben. Unsere Umformungen sind daher wohl nicht mehr äquivalent. Das erkennt man immer daran, daß es einem nicht gelingt, das Verfahren auf die äquivalenten Zeilenumformungen zurückzuführen: Zur ersten Gleichung haben wir die zweite addiert, was eine äquivalente Umformung wäre, würden wir nicht gleichzeitig die erste zur zweiten addieren. Das läßt sich nicht mehr in ein Nacheinander äquivalenter Zeilenumformungen auflösen. Dadurch, daß wir in einem ersten Schritt die zweite Gleichung zur ersten addieren, haben wir die erste verändert, so daß wir die zweite Zeilenumformung II + I mangels unveränderter ersten Gleichung gar nicht mehr ausführen können.

Dieses Beispiel ist zugegebenermaßen sehr durchsichtig und der beschriebene Fehler daher leicht zu vermeiden. Das ist aber nicht immer so. Die Situation kann durchaus verwickelter sein und dann zu nicht äquivalenten Zeilenumformungen führen, die nicht mehr so leicht zu entdecken sind. Wir rechnen ein Beispiel auf zwei Weisen vollständig durch. Einmal, unter (zu) geschickter Ausnutzung von Abkürzungen und einmal nach dem Standardverfahren, wie wir es in Abschnitt 6.1 eingeführt haben.

6.3.8 Beispiel Wir lösen das homogene Gleichungssystem

$$\begin{aligned} 2x_1 + x_2 + 3x_3 + x_4 &= 0 \\ x_1 + 2x_2 + x_3 + 2x_4 + 3x_5 &= 0 \\ -x_1 + x_2 + 3x_4 + x_5 &= 0. \end{aligned}$$

Wie üblich machen wir das mit dem schematischen GAUSS-Verfahren:

I	2	1	3	1	0	I + 2 III	I	2	1	3	1	0	I + 2 III
II	1	2	1	2	3	2 II - I	II	1	2	1	2	3	II + III
III	-1	1	0	3	1	III + II	III	-1	1	0	3	1	
I	0	3	3	7	2	I + II - 2 III	I	0	3	3	7	2	I - II
II	0	3	-1	3	6		II	0	3	1	5	4	
III	0	3	1	5	4	III - II	III	-1	1	0	3	1	III ↔ I
I	0	0	0	0	0		I	-1	1	0	3	1	
II	0	3	-1	3	6	2 II + III	II	0	3	1	5	4	II - III/2
III	0	0	2	2	-2	III/2	III	0	0	2	2	-2	III/2
I	0	0	0	0	0		I	-1	1	0	3	1	-3 I + II
II	0	6	0	8	10	II/2	II	0	3	0	4	5	
III	0	0	1	1	-1		III	0	0	1	1	-1	
I	0	0	0	0	0		I	3	0	0	-5	2	
II	0	3	0	4	5		II	0	3	0	4	5	
III	0	0	1	1	-1		III	0	0	1	1	-1	

Das linke System führt auf die Lösungen

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1x_1 + 0x_4 + 0x_5 \\ 0x_1 - \frac{4}{3}x_4 - \frac{5}{3}x_5 \\ 0x_1 - 1x_4 + 1x_5 \\ 0x_1 + 1x_4 + 1x_5 \\ 0x_1 + 0x_4 + 1x_5 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \frac{x_4}{3} \begin{bmatrix} 0 \\ -4 \\ -3 \\ 3 \\ 0 \end{bmatrix} + \frac{x_5}{3} \begin{bmatrix} 0 \\ -5 \\ 3 \\ 0 \\ 3 \end{bmatrix},$$

mit beliebigen x_1, x_4 und x_5 , während das rechte

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} \frac{5}{3}x_4 - \frac{2}{3}x_5 \\ -\frac{4}{3}x_4 - \frac{5}{3}x_5 \\ -1x_4 + 1x_5 \\ 1x_4 + 0x_5 \\ 0x_4 + 1x_5 \end{bmatrix} = \frac{x_4}{3} \begin{bmatrix} 5 \\ -4 \\ -3 \\ 3 \\ 0 \end{bmatrix} + \frac{x_5}{3} \begin{bmatrix} -2 \\ -5 \\ 3 \\ 0 \\ 3 \end{bmatrix},$$

mit beliebigen x_4 und x_5 ergibt. Keiner der drei Basisvektoren $\mathbf{a} := [1, 0, 0, 0, 0]^t$, $\mathbf{b} := [0, -4, -3, 3, 0]^t$ und $\mathbf{c} := [0, -5, 3, 0, 3]^t$ des linken Schemas ist eine Lösung des Ausgangssystems, wie man durch Einsetzen leicht erkennt. Dagegen sind die Basisvektoren $\mathbf{d} := [5, -4, -3, 3, 0]^t$ und $\mathbf{e} := [-2, -5, 3, 0, 3]^t$ des rechten Schemas sehr wohl Lösungen der Aufgabe. Wie hängen die beiden Lösungsmengen zusammen? Die vermeintliche Lösungsmenge $\{ r\mathbf{a} + s\mathbf{b} + t\mathbf{c} \mid r, s, t \in \mathbb{R} \}$, die das Verfahren auf der linken Seite erzeugt, umfaßt die tatsächliche Lösungsmenge $\mathcal{L} := \{ u\mathbf{d} + v\mathbf{e} \mid u, v \in \mathbb{R} \}$, die das traditionelle GAUSS-Verfahren hervorbringt, denn es gilt $\mathbf{d} = 5\mathbf{a} + \mathbf{b}$ und $\mathbf{e} = -2\mathbf{a} + \mathbf{c}$. Sie hat aber mehr Elemente, denn \mathbf{a} ist offensichtlich keine Lösung des Ausgangssystems und kann daher auch nicht in \mathcal{L} liegen. Der Grund für dieses Ergebnis ist nicht in den Umformungen I+2 III und 2 II-I der ersten beiden Gleichungen zu suchen, denn das kann in äquivalente Zeilenumformungen aufgelöst werden: Zuerst 2 II-I und dann I+2 III. Das Problem macht in der dritten Gleichung der Schritt III+II, sollte man annehmen. Die Kombination mit 2 II-I läßt sich aber wieder durch äquivalente Umformungen erreichen: Zuerst III+II und dann 2 II-I. Dasselbe gilt für die erste und die dritte Zeilenumformung: Zuerst I+2 III und dann III+II. Was das Problem erzeugt ist demnach die Kombination aller drei Umformungen in einem Arbeitsschritt. Man kann sich vorstellen, daß man bei mehr Gleichungen noch viel undurchsichtiger Konstellationen produzieren kann (ein neuer Gesichtspunkt dazu ist im Beispiel 6.4.8 zu finden). Sie alle haben aber ein Erkennungsmerkmal, nämlich, daß nicht klar ist, welche Zeile die eigentliche Arbeitszeile ist, mit der andere Zeilen verändert werden. Und das ist auch schon die Strategie, mit der solche Situationen einfach zu vermeiden sind: Die Arbeitszeile darf im gleichen Arbeitsschritt nie selbst verändert werden (allenfalls darf sie mit einer geeigneten Zahl $a \neq 0$ multipliziert werden, um dafür nicht das Gleichungssystem noch einmal aufschreiben zu müssen).

6.4 Lineare Unabhängigkeit

6.4.1 Definition Eine Teilmenge $\mathcal{B} \subset V$ des Vektorraums V über \mathbb{K} heißt linear unabhängig (mitunter durch l. u. abgekürzt), falls jede Nullkombination aus \mathcal{B} , d. h. jede Gleichung der Form

$$\lambda_1 \mathbf{b}_1 + \lambda_2 \mathbf{b}_2 + \cdots + \lambda_n \mathbf{b}_n = \mathbf{0}, \quad (6.7)$$

mit Vektoren $\mathbf{b}_1, \dots, \mathbf{b}_n$ von \mathcal{B} und Koeffizienten $\lambda_1, \dots, \lambda_n$ aus \mathbb{K} nur durch $\lambda_1 = \lambda_2 = \dots = \lambda_n = 0$ zu erfüllen ist. Eine solche Nullkombination bezeichnen wir als trivial. Damit können wir die lineare Unabhängigkeit von \mathcal{B} kurz und prägnant folgendermaßen ausdrücken: \mathcal{B} ist genau dann linear unabhängig, wenn jede Nullkombination aus \mathcal{B} trivial ist. \mathcal{B} heißt linear abhängig (l. a.), falls \mathcal{B} nicht linear unabhängig ist.

Enthält $\mathcal{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ nur endlich viele Elemente, dann lässt sich die lineare Unabhängigkeit durch Lösen einer einzigen linearen (Vektor-)Gleichung überprüfen, nämlich $\lambda_1 \mathbf{b}_1 + \cdots + \lambda_n \mathbf{b}_n = \mathbf{0}$. Diese deckt alle möglichen Nullkombinationen ab. Für $V = \mathbb{C}^n$ ist sie zu einem homogenen linearen Gleichungssystem aus n Gleichungen äquivalent, das aus den n Vektorkomponenten entsteht. Es wird mit dem GAUSS-Verfahren gelöst. Betrachten wir das folgende Beispiel.

6.4.2 Beispiel

$$\mathcal{B} := \left\{ \begin{bmatrix} 1 \\ 3 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ 2 \\ 2 \\ 7 \end{bmatrix}, \begin{bmatrix} 0 \\ 4 \\ 1 \\ 9 \end{bmatrix} \right\}.$$

Die Nullkombination

$$\lambda_1 \begin{bmatrix} 1 \\ 3 \\ 2 \\ 3 \end{bmatrix} + \lambda_2 \begin{bmatrix} -1 \\ 2 \\ 2 \\ 7 \end{bmatrix} + \lambda_3 \begin{bmatrix} 0 \\ 4 \\ 1 \\ 9 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

führt auf das homogene lineare Gleichungssystem

$$\begin{array}{rclcl} \lambda_1 & - & \lambda_2 & = & 0 \\ 3\lambda_1 & + & 2\lambda_2 & + & 4\lambda_3 = 0 \\ 2\lambda_1 & + & 2\lambda_2 & + & \lambda_3 = 0 \\ 3\lambda_1 & + & 7\lambda_2 & + & 9\lambda_3 = 0, \end{array}$$

das wir wie üblich behandeln – allerdings nur soweit, bis wir die Anzahl der Lösungen erkennen können, denn die Lösungen selbst sind normalerweise uninteressant. Sind die Vektoren nämlich linear unabhängig, dann ist die Lösung der Nullvektor $[\lambda_1, \lambda_2, \lambda_3]^t = \mathbf{0}$. Da $\mathbf{0}$ immer eine Lösung darstellt, sind die Vektoren genau dann linear unabhängig, wenn es nur eine Lösung

(nämlich $\mathbf{0}$) gibt. Das ist zweifelsfrei an einer oberen Dreiecksform und nicht verschwindenden Diagonaleinträgen über die gesamte Breite des Gleichungssystems abzulesen (vergl. Seite 104):

	λ_1	λ_2	λ_3	
I	1	-1	0	
II	3	2	4	II - 3I
III	2	3	1	III - 2I
IV	3	7	9	IV - 3I
I	1	-1	0	
II	0	5	4	
III	0	5	1	III - II
IV	0	10	9	IV - 2II
I	1	-1	0	
II	0	5	4	
III	0	0	-3	
IV	0	0	1	

Zeile III als Gleichung zurück übersetzt lautet einfach $-3\lambda_3 = 0$, also $\lambda_3 = 0$. Das in die Gleichung für Zeile II eingesetzt ergibt $5\lambda_2 + 4\lambda_3 = 5\lambda_2 = 0$, d. h. $\lambda_2 = 0$. Diese Ergebnisse in die erste Gleichung $\lambda_1 - \lambda_2 = 0$ eingesetzt ergibt schließlich auch $\lambda_1 = 0$. Damit ist nachgerechnet, daß eine Nullkombination der Vektoren $[1, 3, 2, 3]^t$, $[-1, 2, 3, 7]$ und $[0, 4, 1, 9]^t$ trivial sein muß. Sie sind damit linear unabhängig.

Diese Rechnung haben wir hier noch einmal zur Demonstration vorgeführt, um daran zu erinnern, wieso eine vollbesetzte Diagonale zur linearen Unabhängigkeit der beteiligten Vektoren führt. In einer konkreten Rechnung ist das aber nicht mehr nötig, da die Überlegung jedes mal nach genau diesem Schema verläuft und nie mit Überraschungen aufwarten wird.

6.4.3 Beispiel

$$\mathcal{C} := \left\{ \begin{bmatrix} 1 \\ -1 \\ 6 \\ 4 \\ 2 \end{bmatrix}, \begin{bmatrix} -11 \\ -2 \\ 2 \\ 19 \\ 7 \end{bmatrix}, \begin{bmatrix} -3 \\ 2 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 10 \\ 6 \\ 21 \\ -9 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ 5 \\ -4 \\ 0 \end{bmatrix} \right\}.$$

Eine Nullkombination dieser Vektoren führt, wie im Beispiel 6.4.2, auf das Schema

	λ_1	λ_2	λ_3	λ_4	λ_5	
I	1	-3	10	2	-11	
II	-1	2	6	3	-2	II + I
III	6	0	21	5	2	III - 6I
IV	4	1	-9	-4	19	IV - 4I
V	2	1	1	0	7	V - 2I

	λ_1	λ_2	λ_3	λ_4	λ_5	
I	1	-3	10	2	-11	
II	0	-1	16	5	-13	
III	0	18	-39	-7	68	III + 18 II
IV	0	13	-49	-12	63	IV + 13 II
V	0	7	-19	-4	29	V + 7 II
I	1	-3	10	2	-11	
II	0	-1	16	5	-13	
III	0	0	249	83	-166	III/83
IV	0	0	159	53	-106	IV/53
V	0	0	93	31	-62	V/31
I	1	-3	10	2	-11	
II	0	-1	16	5	-13	
III	0	0	3	1	-2	
IV	0	0	3	1	-2	IV - III
V	0	0	3	1	-2	V - III
I	1	-3	10	2	-11	
II	0	-1	16	5	-13	
III	0	0	3	1	-2	
IV	0	0	0	0	0	
V	0	0	0	0	0	

Die dritte Zeile bedeutet $3\lambda_3 + \lambda_4 - 2\lambda_5 = 0$. Sie ist z. B. für $\lambda_4 = -3, \lambda_5 = 0$ und $\lambda_3 = 1$ zu erfüllen. Eingesetzt in die Gleichung $-\lambda_2 + 16\lambda_3 + 5\lambda_4 - 13\lambda_5 = 0$ für Zeile II erhalten wir $\lambda_2 = 1$. Das in die Gleichung $\lambda_1 - 3\lambda_2 + 10\lambda_3 + 2\lambda_4 - 11\lambda_5 = 0$ für die erste Zeile eingesetzt liefert schließlich $\lambda_1 = -1$. Mit

$$-\begin{bmatrix} 1 \\ -1 \\ 6 \\ 4 \\ 2 \end{bmatrix} + \begin{bmatrix} -3 \\ 2 \\ 0 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 10 \\ 6 \\ 21 \\ -9 \\ 1 \end{bmatrix} - 3 \begin{bmatrix} 2 \\ 3 \\ 5 \\ -4 \\ 1 \end{bmatrix} + 0 \begin{bmatrix} -11 \\ -2 \\ 2 \\ 19 \\ 7 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

haben wir daher eine nicht triviale Nullkombination der Vektoren von \mathcal{C} gefunden. Es gibt sogar noch weitere, aber für die lineare Abhängigkeit von Vektoren ist es nur wichtig *eine* zu finden. Es ist sogar ausreichend zweifelsfrei nachzuweisen, daß es eine geben muß. Dafür ist es normalerweise gar nicht nötig, sie tatsächlich auszurechnen. Hier genügt es, wenn man das GAUSS-Verfahren bis zur oberen Dreiecksform durchführt und aus dessen stumpfer Form auf die Existenz nicht trivialer Nullkombinationen zu schließen (vergl. Situation II auf Seite 104).

6.4.4 Beispiel Enthält $\mathcal{B} := \{ \mathbf{b}_1, \dots, \mathbf{b}_n, \dots \}$ abzählbar unendlich viele Elemente, dann muß man die Gleichungen $\lambda_1 \mathbf{b}_1 + \dots + \lambda_n \mathbf{b}_n = \mathbf{0}$ für alle $n \in \mathbb{N}$ lösen. Sollte allerdings nur eine dieser Gleichungen eine nicht triviale Lösung aufweisen, so ist \mathcal{B} linear abhängig. Wir wählen als Beispiel im Vektorraum \mathcal{P} aller reellen (oder komplexen) Polynome die Menge \mathcal{B} aller *Monome*, also aller Polynome p_k , die durch $p_k(x) := x^k$ definiert sind: $\mathcal{B} := \{ p_k \mid k \in \mathbb{N}_0 \}$.

Wir nehmen an, es gäbe eine nicht triviale Nullkombination $\lambda_0 p_0 + \lambda_1 p_1 + \cdots + \lambda_n p_n = \sum_{k=0}^n \lambda_k p_k = 0$. Durch $p := \sum_{k=0}^n \lambda_k p_k$ ist insbesondere ein Polynom $p \in \mathcal{P}$ definiert: $p(x) = \sum_{k=0}^n \lambda_k p_k(x) = \sum_{k=0}^n \lambda_k x^k$. Laut unserer Annahme hätte es die merkwürdige Eigenschaft $p(x) = 0$ für alle $x \in \mathbb{R}$, d.h. dieses Polynom ist das sogenannte *Nullpolynom*. Wir müssen zeigen, daß alle Koeffizienten λ_k verschwinden. Das folgt aus einem etwas allgemeineren Ergebnis, nämlich der Eindeutigkeit der Koeffizienten eines Polynoms (Näheres im Abschnitt 11.8 über Polynome, Satz 11.8.2). Da das Nullpolynom mit den Koeffizienten $\lambda_k = 0$ darstellbar ist, gibt es gar keine anderen. Das heißt, es muß $\lambda_k = 0$ für $k = 0, \dots, n$ gelten. Damit haben wir nachgewiesen, daß \mathcal{B} linear unabhängig ist.

Die folgende Übung zeigt, daß es auch einen anderen Weg gibt die lineare Unabhängigkeit von $\{p_k \mid k \in \mathbb{N}_0\}$ zu überprüfen, wenn man dafür auf die elementare Ableitungsregel $p'_k(x) = kx^{k-1}$ für Polynome zurückgreift.

6.4.5 A Zeigen Sie, daß $\mathcal{B} := \{p_k \mid k \in \mathbb{N}_0\}$ linear unabhängig ist. Gehen Sie dabei folgendermaßen vor:

Starten Sie mit einer Nullkombination $0 = \lambda_0 p_0 + \lambda_1 p_1 + \cdots + \lambda_n p_n$. Durch wiederholtes Ableiten und Auswerten an der Stelle $x = 0$ erhalten Sie daraus ein lineares Gleichungssystem für die λ_i , das bereits in Diagonalform vorliegt.

6.4.6 Beispiel. Eine linear unabhängige Menge kann sogar mehr als nur abzählbar unendlich viele Elemente enthalten. Bezeichnen wir für jedes $0 \neq t \in \mathbb{R}$ mit e_t die Exponentialfunktionen $e_t(x) := e^{tx}$, so ist die Menge $\mathcal{E} := \{e_t \mid t \in \mathbb{R} \setminus \{0\}\}$ linear unabhängig. Offensichtlich enthält \mathcal{E} so viele Elemente wie $\mathbb{R} \setminus \{0\}$ (und \mathbb{R} läßt sich nicht mehr abzählen, wie wir aus den Aufgaben wissen → ...). Um das einzusehen, müssen wir zeigen, daß jede Nullkombination $\lambda_1 e_{t_1} + \lambda_2 e_{t_2} + \cdots + \lambda_n e_{t_n} = 0$ trivial ist. Dabei sind die Faktoren t_i paarweise verschieden. Machen wir eine kleine Anleihe an unser Schulwissen, nämlich $e'_t(x) = te^{tx}$ (vergl. (11.40)). Damit läßt sich die Nullkombination beliebig oft ableiten. Wir erhalten $t_1 \lambda_1 e_{t_1} + t_2 \lambda_2 e_{t_2} + \cdots + t_n \lambda_n e_{t_n} = 0$, dann $t_1^2 \lambda_1 e_{t_1} + t_2^2 \lambda_2 e_{t_2} + \cdots + t_n^2 \lambda_n e_{t_n} = 0$ und nach der k -ten Ableitung $t_1^k \lambda_1 e_{t_1} + t_2^k \lambda_2 e_{t_2} + \cdots + t_n^k \lambda_n e_{t_n} = 0$. An einer Stelle $x \in \mathbb{R}$ ausgewertet bedeutet das

$$t_1^k \lambda_1 e^{t_1 x} + t_2^k \lambda_2 e^{t_2 x} + \cdots + t_n^k \lambda_n e^{t_n x} = 0.$$

Wählen wir dabei die Stelle $x = 0$, so erhalten wir wegen $e^0 = 1$ das folgende System linearer Gleichungen für die Unbekannten λ_i :

$$\begin{aligned} t_1 \lambda_1 + t_2 \lambda_2 + \cdots + t_n \lambda_n &= 0 \\ t_1^2 \lambda_1 + t_2^2 \lambda_2 + \cdots + t_n^2 \lambda_n &= 0 \\ &\vdots \\ t_1^n \lambda_1 + t_2^n \lambda_2 + \cdots + t_n^n \lambda_n &= 0. \end{aligned}$$

Die Lösung erfordert eine geschickte Anwendung des GAUSS-Verfahrens. Wir führen das aber nicht für allgemeines n vor. Die Idee ist bereits für $n = 4$ zu erkennen und kann dann leicht

verallgemeinert werden:

	λ_1	λ_2	λ_3	λ_4	
I	t_1	t_2	t_3	t_4	
II	t_1^2	t_2^2	t_3^2	t_4^2	II - t_1 I
III	t_1^3	t_2^3	t_3^3	t_4^3	III - t_1 II
IV	t_1^4	t_2^4	t_3^4	t_4^4	IV - t_1 III
I	t_1	t_2	t_3	t_4	
II	0	$t_2(t_2 - t_1)$	$t_3(t_3 - t_1)$	$t_4(t_4 - t_1)$	
III	0	$t_2^2(t_2 - t_1)$	$t_3^2(t_3 - t_1)$	$t_4^2(t_4 - t_1)$	III - t_2 II
IV	0	$t_2^3(t_2 - t_1)$	$t_3^3(t_3 - t_1)$	$t_4^3(t_4 - t_1)$	IV - t_2 III
I	t_1	t_2	t_3	t_4	
II	0	$t_2(t_2 - t_1)$	$t_3(t_3 - t_1)$	$t_4(t_4 - t_1)$	
III	0	0	$t_3(t_3 - t_2)(t_3 - t_1)$	$t_4(t_4 - t_2)(t_4 - t_1)$	
IV	0	0	$t_3^2(t_3 - t_2)(t_3 - t_1)$	$t_4^2(t_4 - t_2)(t_4 - t_1)$	IV - t_3 III
I	t_1	t_2	t_3	t_4	
II	0	$t_2(t_2 - t_1)$	$t_3(t_3 - t_1)$	$t_4(t_4 - t_1)$	
III	0	0	$t_3(t_3 - t_2)(t_3 - t_1)$	$t_4(t_4 - t_2)(t_4 - t_1)$	
IV	0	0	0	$t_4(t_4 - t_3)(t_4 - t_2)(t_4 - t_1)$	

Das ist eine obere Dreiecksform mit nicht verschwindender Diagonalen. Daher hat unser Gleichungssystem die einzige mögliche Lösung $\lambda_1 = \dots = \lambda_n = 0$. Das bedeutet, daß alle Nullkombinationen in \mathcal{E} trivial sind. \mathcal{E} ist also, wie behauptet, linear unabhängig.

Wir wollen noch die Verbindung unserer Definition der linearen Unabhängigkeit mit einer gängigen Vorstellung aufzeigen. Linear unabhängige Vektoren sollten die Eigenschaft haben, daß sich keiner der beteiligten durch die anderen linear kombinieren läßt.

6.4.7 Lemma Eine Teilmenge \mathcal{B} eines Vektorraumes V ist genau dann linear unabhängig, wenn kein Vektor aus \mathcal{B} eine Linearkombination der übrigen Vektoren aus \mathcal{B} ist.

Beweis. Ist \mathcal{B} linear unabhängig und $\mathbf{b} \in \mathcal{B}$ eine Linearkombination von anderen Vektoren $\mathbf{b}_1, \dots, \mathbf{b}_n \in \mathcal{B}$, also $\mathbf{b} = \lambda_1 \mathbf{b}_1 + \dots + \lambda_n \mathbf{b}_n$, dann würde $\lambda_1 \mathbf{b}_1 + \dots + \lambda_n \mathbf{b}_n - 1 \cdot \mathbf{b} = \mathbf{0}$ eine nicht triviale Nullkombination mit Vektoren aus \mathcal{B} ergeben, im Widerspruch zur linearen Unabhängigkeit von \mathcal{B} . Daher kann kein Element von \mathcal{B} eine Linearkombination der übrigen Elemente aus \mathcal{B} sein.

Für die Umkehrung gehen wir davon aus, daß kein Element aus \mathcal{B} eine Linearkombination der übrigen Elemente von \mathcal{B} sein kann und nehmen an, $\lambda_1 \mathbf{b}_1 + \lambda_2 \mathbf{b}_2 + \dots + \lambda_n \mathbf{b}_n = \mathbf{0}$ wäre eine nicht triviale Nullkombination von Vektoren $\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n \in \mathcal{B}$. Das bedeutet, daß wenigstens einer der Koeffizienten $\lambda_1, \lambda_2, \dots, \lambda_n$ von Null verschieden sein müßte. Wir können o. B. d. A. $\lambda_1 \neq 0$ annehmen. Dann ließe sich die Nullkombination nach \mathbf{b}_1 auflösen: $\mathbf{b}_1 = -\frac{\lambda_2}{\lambda_1} \mathbf{b}_2 - \dots - \frac{\lambda_n}{\lambda_1} \mathbf{b}_n$. \mathbf{b}_1 wäre eine Linearkombination der übrigen Vektoren aus \mathcal{B} , im Widerspruch zu unserer Annahme. \square

6.4.8 Beispiel Mit den Ergebnissen unserer Untersuchungen zur linearen Unabhängigkeit können wir jetzt noch etwas besser verstehen, was der eigentliche Grund dafür war, daß die Modifikation des GAUSS-Verfahrens auf Seite 112 falsche Ergebnisse lieferte. Dazu fassen wir die drei Zeilen des Gleichungssystems als Zeilenvektoren z_1, z_2 und z_3 auf. Im ersten Arbeitsschritt haben wir aus $\mathcal{Z}_1 := \{z_1, z_2, z_3\}$ die Menge $\mathcal{Z}_2 := \{z_1 + 2z_3, 2z_2 - z_1, z_3 + z_2\}$ mit den neuen Zeilenvektoren $\tilde{z}_1 := z_1 + 2z_3, \tilde{z}_2 := 2z_2 - z_1$ und $\tilde{z}_3 := z_3 + z_2$ des zweiten Gleichungssystems gemacht. Egal ob \mathcal{Z}_1 linear unabhängig ist oder nicht, die Menge \mathcal{Z}_2 ist sicher linear abhängig, denn $1 \cdot (z_1 + 2z_3) + 1 \cdot (2z_2 - z_1) - 2 \cdot (z_3 + z_2)$ ist eine nicht triviale Nullkombination ihrer Elemente. Das bedeutet, wenn man die zweite Zeile \tilde{z}_2 zur ersten \tilde{z}_1 addiert und das Doppelte der Dritten \tilde{z}_3 von ihr abzieht, kann man sie zum Verschwinden bringen, unabhängig davon, was ursprünglich in z_1, z_2 und z_3 stand. Wäre also der Übergang von \mathcal{Z}_1 zu \mathcal{Z}_2 durch äquivalente Zeilenumformungen zustande gekommen, dann hätten wir eine universelle Methode entdeckt, wie wir in jedem Gleichungssystem eine Zeile zum Verschwinden bringen können. Wiederholte Anwendung unseres Verfahrens würde schließlich dafür sorgen, daß wir jedes Gleichungssystem auf zwei Gleichungen reduzieren könnten. Das ist offensichtlich unsinnig, denn dann wäre kein Gleichungssystem mit mehr als zwei Unbekannten eindeutig lösbar.

6.4.9 A

$$\text{i) } \mathcal{B} := \left\{ \begin{bmatrix} 2 \\ 1 \\ 3 \\ 0 \\ 12 \end{bmatrix}, \begin{bmatrix} -10 \\ 3 \\ 2 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 10 \\ 11 \\ 4 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 2 \\ -2 \\ 5 \end{bmatrix} \right\}, \quad \mathcal{C} := \left\{ \begin{bmatrix} 1 \\ 14 \\ 19 \\ -60 \\ 0 \end{bmatrix}, \begin{bmatrix} 9 \\ -3 \\ -12 \\ 63 \\ -87 \end{bmatrix}, \begin{bmatrix} 11 \\ 5 \\ 18 \\ -45 \\ -19 \end{bmatrix}, \begin{bmatrix} 3 \\ 2 \\ 6 \\ -16 \\ -4 \end{bmatrix} \right\}.$$

Zeigen Sie, daß \mathcal{B} linear unabhängig und \mathcal{C} linear abhängig ist.

*ii) $\mathcal{S} := \{s_t \mid t \in \mathbb{R}_+\}$ mit $s_t(x) := \sin(tx)$ ist linear unabhängig.

6.4.10 A Zeigen Sie, daß der Inkreismittelpunkt m eines Dreiecks mit den Ecken a, b und c durch

$$m = \frac{\|b - c\|}{U} \cdot a + \frac{\|c - a\|}{U} \cdot b + \frac{\|a - b\|}{U} \cdot c \quad (6.8)$$

gegeben ist. Dabei bezeichnet $U := \|a - b\| + \|b - c\| + \|c - a\|$ den Umfang des Dreiecks.

Schneiden Sie dafür die Winkelhalbierenden durch a und c . Als Richtungsvektoren der Geraden eignen sich $\frac{b-a}{\|b-a\|} + \frac{c-a}{\|c-a\|}$ bzw. $\frac{b-c}{\|b-c\|} - \frac{c-a}{\|c-a\|}$.

Zur Lösung der Gleichung verwenden Sie, daß die Menge $\{b - a, b - c\}$ linear unabhängig ist.

Berechnen Sie m für $a := [0, 0]^t, b := [14, 0]^t$ und $c := [9, 12]^t$. Bestimmen Sie auch den Radius des Kreises und den Flächeninhalt des Dreiecks (Vergl. auch Aufgabe 5.2.7).

6.5 Basis und Dimension

6.5.1 Definition Für einen Vektorraum V über \mathbb{K} sei eine Menge $\mathcal{B} \subseteq V$ gegeben. Die lineare Hülle $lh(\mathcal{B})$ von \mathcal{B} ist die Menge aller Linearkombinationen, die sich mit Elementen aus \mathcal{B} bilden lassen. \mathcal{B} heißt erzeugend, falls $lh(\mathcal{B}) = V$ gilt. \mathcal{B} heißt Basis von V , falls \mathcal{B} erzeugend und linear unabhängig ist. V heißt endlich erzeugt, falls \mathcal{B} nur endlich viele Elemente enthält.

Von einer Basis \mathcal{B} wird man erwarten, daß sich jedes Element aus V durch die Basisvektoren aufbauen läßt. Das ist gewährleistet, weil \mathcal{B} erzeugend ist. Die lineare Unabhängigkeit sorgt dafür, daß \mathcal{B} nicht unnötig viele Elemente enthält (vergl. Lemma 6.4.7) und daß die Darstellung durch die Basisvektoren eindeutig ist.

6.5.2 Lemma Sei \mathcal{B} eine Basis für den Vektorraum V . Dann läßt sich jeder Vektor aus V eindeutig als Linearkombination von Elementen aus \mathcal{B} darstellen.

Beweis. Sei $x \in V$. Da \mathcal{B} erzeugend ist, sind alle Elemente von V Linearkombinationen von Vektoren aus \mathcal{B} : $x = \sum_{k=1}^n x_k \mathbf{b}_k$, $\mathbf{b}_k \in \mathcal{B}$, $k = 1, \dots, n$. Läßt sich x auch noch auf eine andere Weise durch Vektoren von \mathcal{B} kombinieren, etwa $x = \sum_{k=1}^m y_k \tilde{\mathbf{b}}_k$, $\tilde{\mathbf{b}}_k \in \mathcal{B}$, $k = 1, \dots, m$, dann wäre durch $0 = \sum_{k=1}^n x_k \mathbf{b}_k - \sum_{k=1}^m y_k \tilde{\mathbf{b}}_k$ eine nicht triviale Nullkombination von Vektoren aus \mathcal{B} gegeben. Da \mathcal{B} linear unabhängig ist, erhalten wir einen Widerspruch. \square

6.5.3 Satz (Basis-Austausch-Satz) V sei durch eine Basis $\mathcal{B} := \{ \mathbf{b}_1, \dots, \mathbf{b}_n \}$ endlich erzeugt. In der Entwicklung eines Vektors $x \in V$ sei der Basisvektor \mathbf{b}_j vorhanden. Ersetzt man \mathbf{b}_j in \mathcal{B} durch x , so ist die entstandene Menge $\mathcal{B}' := \{ \mathbf{b}_k \mid k = 1, \dots, n, k \neq j \} \cup \{ x \}$ ebenfalls eine Basis für V .

Beweis. Laut Voraussetzung gilt $x = \sum_{k \neq j}^n x_k \mathbf{b}_k + x_j \mathbf{b}_j$, $x_j \neq 0$. Wir können o. B. d. A. von $x_j = 1$ ausgehen, indem wir x durch x/x_j ersetzen. Zunächst zeigen wir, daß \mathcal{B}' linear unabhängig ist. Aus

$$\mathbf{0} = \sum_{k \neq j}^n \lambda_k \mathbf{b}_k + \lambda_j x = \sum_{k \neq j}^n \lambda_k \mathbf{b}_k + \lambda_j \sum_{k \neq j}^n x_k \mathbf{b}_k + \lambda_j \mathbf{b}_j = \sum_{k \neq j}^n (\lambda_k + \lambda_j x_k) \mathbf{b}_k + \lambda_j \mathbf{b}_j$$

folgt, wegen der linearen Unabhängigkeit von \mathcal{B} , zuerst $\lambda_j = 0$ und dann $0 = \lambda_k + \lambda_j x_k = \lambda_k$ für $k \neq j$. Daher ist jede Nullkombination aus \mathcal{B}' trivial und \mathcal{B}' linear unabhängig.

\mathcal{B}' ist erzeugend: Jedes $y \in V$ hat eine Darstellung $y = \sum_{k \neq j}^n y_k \mathbf{b}_k + y_j \mathbf{b}_j$. Mit $\mathbf{b}_j = x - \sum_{k \neq j}^n x_k \mathbf{b}_k$ erhalten wir daraus $y = \sum_{k \neq j}^n (y_k - y_j x_k) \mathbf{b}_k + y_j x \in lh(\mathcal{B}')$. Das zeigt $V = lh(\mathcal{B}')$. \square

6.5.4 Korollar Für eine Basis $\mathcal{B} = \{ \mathbf{b}_1, \dots, \mathbf{b}_n \}$ von V kann eine linear unabhängige Menge $\mathcal{C} = \{ \mathbf{c}_1, \dots, \mathbf{c}_m \}$ nicht mehr Elemente als \mathcal{B} enthalten. Hat sie gleich viele Elemente, so ist sie ebenfalls eine Basis. Insbesondere hat jede Basis von V dieselbe Anzahl von Basisvektoren.

Beweis. Wir nehmen $m > n$ an. Der Vektor c_1 läßt sich in der Basis \mathcal{B} darstellen. O. B. d. A. sei der Basisvektor b_1 in dieser Entwicklung vertreten (andernfalls nummerieren wir \mathcal{B} geeignet um). Nach Satz 6.5.3 ist $\mathcal{B}_1 := \{ c_1, b_2, \dots, b_n \}$ eine Basis von V . In dieser Basis läßt sich c_2 darstellen:

$$c_2 = \alpha_1 c_1 + \sum_{k=2}^n \alpha_k b_k.$$

Da \mathcal{C} linear unabhängig ist, muß für $k \geq 2$ wenigstens ein α_k von Null verschieden sein. Wir können o. B. d. A. $\alpha_2 \neq 0$ annehmen. Damit ist $\mathcal{B}_2 := \{ c_1, c_2, b_3, \dots, b_n \}$ eine Basis für V . Auf diese Weise können wir fortfahren, bis auch der Basisvektor b_n durch c_n ersetzt worden ist. Dann ist $\mathcal{B}_n := \{ c_1, c_2, \dots, c_n \}$ eine Basis von V , in der die überzähligen Vektoren c_{n+1}, \dots, c_m darstellbar sein müßten. Das ist ein Widerspruch zur linearen Unabhängigkeit von \mathcal{C} (vergl. Lemma 6.4.7). Daher muß $n = m$ gelten. Dann ist \mathcal{C} selbst eine Basis von V . Jetzt seien \mathcal{B} und \mathcal{C} zwei Basen. Dann kann die linear unabhängige Menge \mathcal{C} nach den bisherigen Ergebnissen nicht mehr Elemente als \mathcal{B} enthalten – weniger aber auch nicht, da die Argumentation mit vertauschten Rollen von \mathcal{B} und \mathcal{C} wiederholt werden kann. \square

6.5.5 Satz Jeder endlich erzeugte Vektorraum hat eine endliche Basis.

Beweis. Der Vektorraum V (o. B. d. A. $\neq \{\mathbf{0}\}$) wird durch die Vektoren einer Menge $\mathcal{C} := \{c_1, \dots, c_m\}$ linear erzeugt. Ist \mathcal{C} linear unabhängig, so handelt es sich bereits um eine Basis. Andernfalls gibt es nach Lemma 6.4.7 einen Vektor aus \mathcal{C} , sagen wir c_m , der durch die anderen linear kombiniert werden kann. Dann sollte auch die Menge $\mathcal{C}_1 := \{c_1, \dots, c_{m-1}\}$ noch erzeugend sein. Denn jeder Vektor $x \in V$ lässt sich als Linearkombination der Vektoren c_1, \dots, c_{m-1} und c_m gewinnen. Ersetzen wir darin c_m durch seine Linearkombination aus den Vektoren c_1, \dots, c_{m-1} , so sind im Ergebnis auch nur diese $m - 1$ Vektoren aus \mathcal{C}_1 an der Linearkombination von x beteiligt.

Nun, da \mathcal{C}_1 ebenfalls erzeugend ist, können wir die Argumentation wiederholen: Entweder ist \mathcal{C}_1 bereits linear unabhängig und damit eine Basis, oder wir gehen (o. B. d. A.) zu der erzeugenden Menge $\mathcal{C}_2 := \{c_1, \dots, c_{m-2}\}$ über. Dieser Vorgang muß spätestens nach $m - 1$ Schritten stoppen und eine linear unabhängige Menge liefern. Laut Konstruktion ist sie erzeugend und daher eine Basis von V . \square

6.5.6 Definition Ein endlich erzeugter Vektorraum V heißt **endlichdimensional**. Die Dimension $\dim V$ von V ist die Anzahl der Basisvektoren einer Basis. V bezeichnen wir als **unendlichdimensional**, falls er keine Basis mit endlich vielen Elementen hat.

$\dim V$ ist die größte Anzahl an Elementen, die eine linear unabhängige Teilmenge von V haben kann.

6.5.7 Korollar Hat ein Vektorraum linear unabhängige Mengen mit beliebig, oder gar unendlich vielen Elementen, so ist er unendlich dimensional.

Beweis. Gäbe es eine endliche Basis, sagen wir mit n Elementen, so dürfte keine linear unabhängige Menge mehr als n Elemente enthalten, im Widerspruch zur Annahme. \square

Wenn nichts anderes gesagt wird, gehen wir im Folgenden immer von endlichdimensionalen Vektorräumen aus.

6.5.8 Satz (Basisergänzungssatz) In einem endlichdimensionalen Vektorraum lässt sich jede linear unabhängige Menge zu einer Basis ergänzen.

Beweis. $\mathcal{C} := \{c_1, \dots, c_m\}$ sei eine linear unabhängige Menge in einem Vektorraum V der Dimension n . Wir können von $m < n$ ausgehen, da andernfalls \mathcal{C} bereits eine Basis wäre (Korollar 6.5.4). Dann ist $V_0 := \text{Ih } \mathcal{C}$ ein echter Teilraum von V (weil andernfalls \mathcal{C} eine Basis von V mit weniger Elementen als $\dim V$ wäre). Es gibt daher ein $x_1 \notin V_0$. Dann ist $\mathcal{C}_1 := \mathcal{C} \cup \{x_1\}$ ebenfalls linear unabhängig. Andernfalls gäbe es nämlich eine nicht triviale Nullkombination $\mathbf{0} = \sum_{k=1}^m \lambda_k c_k + \lambda_0 x_1$. Dabei müßte $\lambda_0 \neq 0$ (o. B. d. A. = 1) sein, da \mathcal{C} ja linear unabhängig ist. Wir könnten dann nach x_1 auflösen und erhalten x_1 als Linearkombination der Vektoren aus \mathcal{C} , was zum Widerspruch $x_1 \in V_1$ führen würde. \mathcal{C}_1 ist also tatsächlich linear unabhängig. Enthält diese Menge bereits n Elemente, so handelt es sich um eine Basis.

Andernfalls können wir das Verfahren wiederholen und erhalten linear unabhängige Mengen $\mathcal{C}_2 := \mathcal{C}_1 \cup \{\mathbf{x}_2\} = \mathcal{C} \cup \{\mathbf{x}_1, \mathbf{x}_2\}$, usw. Nach $n - m$ Schritten haben wir eine linear unabhängige Menge $\mathcal{C}_{n-m} := \mathcal{C} \cup \{\mathbf{x}_1, \dots, \mathbf{x}_{n-m}\}$ mit n Elementen, also eine Basis, die die linear unabhängigen Vektoren aus \mathcal{C} enthält. \square

Bemerkung: Die linear unabhängige Menge, von der in dem Satz die Rede ist, kann durchaus auch nur ein Element enthalten.

6.5.9 Beispiel Die Vektorräume \mathbb{R}^n und \mathbb{C}^n haben, wie erwartet, die Dimension n . Wir müssen dafür nur eine Basis mit n Elementen finden. Ohne groß darüber nachzudenken verwenden wir eine solche Basis bereits, wenn wir einen Vektor $\mathbf{x} \in \mathbb{R}^n$ oder $\mathbf{x} \in \mathbb{C}^n$ durch seinen Koordinatenvektor $[x_1, x_2, x_3, \dots, x_n]^t$ angeben. Es gilt nämlich

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \dots + x_n \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} = \sum_{i=1}^n x_i \mathbf{e}_i.$$

Dabei ist \mathbf{e}_i , $i = 1, 2, \dots, n$, der Spaltenvektor, der an der Stelle i den Eintrag 1 und an allen anderen Stellen den Eintrag 0 hat.

$$\mathcal{E} := \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \right\} = \{ \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \dots, \mathbf{e}_{n-1}, \mathbf{e}_n \}. \quad (6.9)$$

Diese Vektoren sind die Richtungsvektoren der Koordinatenachsen. Die Zahl x_i misst den Anteil, den die Richtung \mathbf{e}_i am Vektor \mathbf{x} hat. Offensichtlich sind diese Vektoren erzeugend und linear unabhängig, also eine Basis. Wir bezeichnen sie mit \mathcal{E} und nennen sie die *kanonische Basis* von \mathbb{R}^n bzw. \mathbb{C}^n :

6.5.10 Beispiel Für \mathbb{C}^n definiert jede Menge aus n linear unabhängigen Vektoren eine Basis. Beispielsweise ist

$$\mathcal{B} := \left\{ \begin{bmatrix} 1 \\ 4 \\ i \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \begin{bmatrix} i \\ 1 \\ 3 \end{bmatrix} \right\},$$

eine Basis für den \mathbb{C}^3 . Überzeugen wir uns davon, indem wir das homogene Gleichungssystem

lösen, das zu einer Nullkombination der drei Vektoren aus \mathcal{B} gehört:

I	1	1	i	
II	4	2	1	II - 4 I
III	i	0	3	III - i I
I	1	1	i	
II	0	-2	1 - 4i	
III	0	-i	4	2 III - i II
I	1	1	i	
II	0	-2	1 - 4i	
III	0	0	4 - i	

Jeder Vektor x aus \mathbb{C}^3 muß sich jetzt auf eindeutige Weise als Linearkombination der Basisvektoren b_1, b_2, b_3 von \mathcal{B} schreiben lassen: $x = \sum_{k=1}^3 x_k b_k$. Man nennt das den *Vektor x nach der Basis \mathcal{B} entwickeln*. Um das konkret durchzuführen, müssen die *Entwicklungscoeffizienten* x_k gefunden werden. Dabei handelt es sich um ein lineares Gleichungssystem, das mit dem GAUSS-Verfahren gelöst werden kann. Als Beispiel wählen wir $x := [3, 1, i]^t$:

$$x_1 \begin{bmatrix} 1 \\ 4 \\ i \end{bmatrix} + x_2 \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} i \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ i \end{bmatrix} \Leftrightarrow \begin{array}{l} x_1 + x_2 + ix_3 = 3 \\ 4x_1 + 2x_2 + x_3 = 1 \\ ix_1 + 3x_3 = i \end{array}$$

	x_1	x_2	x_3	
I	1	1	i	3
II	4	2	1	1
III	i	0	3	i
I	1	1	i	3
II	0	-2	1 - 4i	-11
III	0	-i	4	-2i
I	2	0	1 - 2i	-5
II	0	-2	1 - 4i	-11
III	0	0	4 - i	7i
I	$2(4 - i)$	0	0	$(4 - i)I - (1 - 2i)III$
II	0	$-2(4 - i)$	0	$(4 - i)II - (1 - 4i)III$
III	0	0	$4 - i$	$(4 + i)III$
I	17	0	0	$(4 + i)I/2$
II	0	17	0	$-(4 + i)II/2$
III	0	0	17	$(4 + i)III$

Die Lösung ist $x_1 = -\frac{1}{17}(67 + 21i)$, $x_2 = \frac{1}{17}(146 + 28i)$ und $x_3 = \frac{1}{17}(-7 + 28i)$. Die Entwicklung von x nach der Basis \mathcal{B} haben wir damit berechnet:

$$\begin{bmatrix} 3 \\ 1 \\ i \end{bmatrix} = -\frac{1}{17}(67 + 21i) \begin{bmatrix} 1 \\ 4 \\ i \end{bmatrix} + \frac{1}{17}(146 + 28i) \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} + \frac{1}{17}(-7 + 28i) \begin{bmatrix} i \\ 1 \\ 3 \end{bmatrix}.$$

6.5.11 Komplexes LGS mit reellem GAUSS-Verfahren Das letzte Beispiel zeigt, daß sich am GAUSS-Verfahren für komplexe Zahlen nichts ändert – außer, daß es recht mühsam werden kann. Es gibt aber eine Methode, ein komplexes Gleichungssystem $Cz = w$ in ein reelles zu übersetzen, allerdings zu dem Preis, daß aus einem komplexen $m \times n$ -System ein reelles $2m \times 2n$ -System wird. Dafür zerlegt man die Matrix C in ihren Real- und Imaginärteil A bzw. B (reelle $m \times n$ -Matrizen) und genauso den gesuchten Vektor $z \in \mathbb{C}^n$ sowie die Inhomogenität $w \in \mathbb{C}^m$:

$$C = A + iB, \quad z = x + iy, \quad w = u + iv, \quad x, y \in \mathbb{R}^n, \quad u, v \in \mathbb{R}^m.$$

Dann gilt $Cz = (A + iB)(x + iy) = Ax - By + i(Bx + Ay) = u + iv$, d. h., wir erhalten ein gekoppeltes, reelles Gleichungssystem: $Ax - By = u$ und $Bx + Ay = v$. Das bauen wir zu einem reellen $2m \times 2n$ -System um

$$\begin{bmatrix} A & -B \\ B & A \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix}. \quad (6.10)$$

Dieses (viermal so große) reelle Gleichungssystem lösen wir mit dem reellen GAUSS-Verfahren.

6.5.12 Beispiel

Ein komplexes 3×3 -Gleichungssystem:

$$C = \begin{bmatrix} 1+i & -i & 2 \\ i & 1+2i & 2+4i \\ 3 & 1+2i & 4+i \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 2 \\ 3 & 1 & 4 \end{bmatrix} + i \begin{bmatrix} 1 & -1 & 0 \\ 1 & 2 & 4 \\ 0 & 2 & 1 \end{bmatrix} = A + iB,$$

$$w = \begin{bmatrix} 16+3i \\ -8+34i \\ 15+18i \end{bmatrix} = \begin{bmatrix} 16 \\ -8 \\ 15 \end{bmatrix} + i \begin{bmatrix} 3 \\ 34 \\ 18 \end{bmatrix} = u + iv.$$

Das führt auf das reelle 6×7 -Schema für das GAUSS-Verfahren:

	x_1	x_2	x_3	y_1	y_2	y_3	
I	1	0	2	-1	1	0	16
II	0	1	2	-1	-2	-4	-8
III	3	1	4	0	-2	-1	15
IV	1	-1	0	1	0	2	3
V	1	2	4	0	1	2	34
VI	0	2	1	3	1	4	18
I	1	0	2	-1	1	0	16
II	0	1	2	-1	-2	-4	-8
III	0	1	-2	3	-5	-1	-33
IV	0	-1	-2	2	-1	2	-13
V	0	2	2	1	0	2	18
VI	0	2	1	3	1	4	18
I	1	0	2	-1	1	0	16
II	0	1	2	-1	-2	-4	-8
III	0	0	-4	4	-3	3	-25
IV	0	0	0	1	-3	-2	-21
V	0	0	-2	3	4	10	34
VI	0	0	-3	5	5	12	34
I	2	0	0	2	-1	3	7
II	0	2	0	2	-7	-5	-41
III	0	0	-4	4	-3	3	-25
IV	0	0	0	1	-3	-2	-21
V	0	0	0	2	11	17	93
VI	0	0	0	8	29	39	211

	x_1	x_2	x_3	y_1	y_2	y_3		
I	2	0	0	0	5	7	49	$(17I - 5V)/2$
II	0	2	0	0	-1	-1	1	$(17II + V)/2$
III	0	0	-4	0	9	11	59	$(17III - 9V)/2$
IV	0	0	0	1	-3	-2	-21	$17IV + 3V$
V	0	0	0	0	17	21	135	
VI	0	0	0	0	53	55	379	$-(17VI - 53V)/178$
I	17	0	0	0	0	7	79	$(I - 7VI)/17$
II	0	17	0	0	0	2	76	$(II - 2VI)/17$
III	0	0	-34	0	0	-1	-106	$-(III + VI)/34$
IV	0	0	0	17	0	29	48	$(IV - 29VI)/17$
V	0	0	0	0	17	21	135	$(V - 21VI)/17$
VI	0	0	0	0	0	1	4	
I	1	0	0	0	0	0	3	
II	0	1	0	0	0	0	4	
III	0	0	1	0	0	0	3	
IV	0	0	0	1	0	0	-4	
V	0	0	0	0	1	0	3	
VI	0	0	0	0	0	1	4	

Daher ist $z = \begin{bmatrix} 3 \\ 4 \\ 3 \end{bmatrix} + i \begin{bmatrix} -4 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 - 4i \\ 4 + 3i \\ 3 + 4i \end{bmatrix}$.

Kontrolle:

$$\begin{aligned} Cz &= \begin{bmatrix} 1+i & -i & 2 \\ i & 1+2i & 2+4i \\ 3 & 1+2i & 4+i \end{bmatrix} \begin{bmatrix} 3-4i \\ 4+3i \\ 3+4i \end{bmatrix} \\ &= \begin{bmatrix} (1+i)(3-4i) - i(4+3i) + 2(3+4i) \\ i(3-4i) + (1+2i)(4+3i) + (2+4i)(3+4i) \\ 3(3-4i) + (1+2i)(4+3i) + (4+i)(3+4i) \end{bmatrix} \\ &= \begin{bmatrix} 7-i+3-4i+6+8i \\ 4+3i-2+11i-10+20i \\ 9-12i-2+11i+8+19i \end{bmatrix} = \begin{bmatrix} 16+3i \\ -8+34i \\ 15+18i \end{bmatrix} = w. \end{aligned}$$

Ändern wir das Beispiel so ab, daß C einen nichttrivialen Kern hat:

$$C = \begin{bmatrix} 1+i & -i & 2 \\ i & 1+2i & 2+4i \end{bmatrix}, w = \begin{bmatrix} 16+3i \\ -8+34i \end{bmatrix}.$$

	x_1	x_2	x_3	y_1	y_2	y_3		
I	1	0	2	-1	1	0	16	
II	0	1	2	-1	-2	-4	-8	
III	1	-1	0	1	0	2	3	III - I
IV	1	2	4	0	1	2	34	IV - I
I	1	0	2	-1	1	0	16	
II	0	1	2	-1	-2	-4	-8	
III	0	-1	-2	2	-1	2	-13	III + II
IV	0	2	2	1	0	2	18	IV - 2II
I	1	0	2	-1	1	0	16	I + IV
II	0	1	2	-1	-2	-4	-8	II + IV
III	0	0	0	1	-3	-2	-21	III ↔ IV
IV	0	0	-2	3	4	10	34	

	x_1	x_2	x_3	y_1	y_2	y_3		
I	1	0	0	2	5	10	50	I - 2 IV
II	0	1	0	2	2	6	26	II - 2 IV
III	0	0	-2	3	4	10	34	III - 3 IV
IV	0	0	0	1	-3	-2	-21	
I	1	0	0	0	11	14	92	
II	0	1	0	0	8	10	68	
III	0	0	-2	0	13	16	97	
IV	0	0	0	1	-3	-2	-21	

Das führt auf

$$\begin{aligned} x_1 &= 92 & -11 & y_2 & -14 & y_3 \\ x_2 &= 68 & -8 & y_2 & -10 & y_3 \\ x_3 &= -48.5 & +6.5 & y_2 & +8 & y_3 \quad \text{also} \quad z = \begin{bmatrix} 92 - 21i \\ 68 \\ -48.5 \end{bmatrix} + y_2 \begin{bmatrix} 3i - 11 \\ i - 8 \\ 6.5 \end{bmatrix} + y_3 \begin{bmatrix} 2i - 14 \\ -10 \\ i + 8 \end{bmatrix}, \\ y_1 &= -21 & +3 & y_2 & +2 & y_3 \\ y_2 &= 0 & +1 & y_2 & +0 & y_3 \\ y_3 &= 0 & +0 & y_2 & +1 & y_3 \quad y_2, y_3 \in \mathbb{R}. \end{aligned}$$

Das ist zwar ein Ergebnis, aber die allgemeine Lösungsform $z = z_1 + t \cdot z_0$, $t \in \mathbb{C}$, lässt sich nicht ganz leicht erkennen. Die spezielle inhomogene Lösung $z_1 = [92 - 21i, 68, -48.5]^t$ ist natürlich gut sichtbar, aber die homogene Lösung z_0 nicht, wenn man nicht schon weiß, daß bei diesem einfachen Beispiel der Kern von C eindimensional ist. Daher müssen $[3i - 11, i - 8, 6.5]^t$ und $[2i - 14, -10, i + 8]$ linear abhängig sein (der Streckungsfaktor ist $\frac{2}{13}(8+i)$), so daß man z. B. den ersten Vektor als z_0 wählen kann. Bei größeren Systemen mit größeren Kernen wird die Analyse der homogenen Lösungen allerdings weiteren Rechenaufwand erfordern. Es wäre daher wünschenswert, das Verfahren so abzuändern, daß die komplexen Lösungen ohne weitere Nachbearbeitung herauskommen. Dafür sollte man das GAUSS-Verfahren derart steuern, daß die Ausgangsform in ihrer Struktur erhalten wird, daß das Ergebnis also wieder die Gestalt von (6.10) aufweist:

	x_1	x_2	x_3	y_1	y_2	y_3		
I	1	0	2	-1	1	0	16	I + III
II	0	1	2	-1	-2	-4	-8	II + III
III	0	0	0	1	-3	-2	-21	
IV	0	0	-2	3	4	10	34	IV - 3 III
I	1	0	2	0	-2	-2	-5	13 I + 2 IV
II	0	1	2	0	-5	-6	-29	13 II + 5 IV
III	0	0	0	1	-3	-2	-21	13 III + 3 IV
IV	0	0	-2	0	13	16	97	
I	13	0	22	0	0	6	129	
II	0	13	16	0	0	2	108	
III	0	0	-6	13	0	22	18	
IV	0	0	-2	0	13	16	97	

Das läßt sich jetzt zwanglos in eine komplexe Lösung übersetzen:

$$\begin{aligned} z &= \frac{1}{13} \begin{bmatrix} 129 + 18i \\ 108 + 97i \\ 0 \end{bmatrix} + \frac{x_3}{13} \begin{bmatrix} -22 + 6i \\ -16 + 2i \\ 13 \end{bmatrix} + \frac{y_3}{13} \begin{bmatrix} -6 - 22i \\ -2 - 16i \\ 13i \end{bmatrix} \\ &= \frac{1}{13} \begin{bmatrix} 129 + 18i \\ 108 + 97i \\ 0 \end{bmatrix} + \frac{x_3 + iy_3}{13} \begin{bmatrix} -22 + 6i \\ -16 + 2i \\ 13 \end{bmatrix}. \end{aligned}$$

6.5.13 Komplexe Inverse mit reellem GAUSS-Verfahren Die Berechnung einer inversen Matrix läßt sich natürlich auch mit dem reellen GAUSS-Verfahren durchführen (das ist ein Vorgriff auf das Thema *Inverse Matrix* (siehe 6.9), das an dieser Stelle der Vollständigkeit halber

mitbehandelt wird): $C = A + iB$ habe die Inverse $C^{-1} = D + iE$. Das ist äquivalent zu $AD - BE + i(AE + BD) = \mathbb{1}$, oder $AD - BE = \mathbb{1}$ und $AE + BD = 0$, also zu

$$\begin{bmatrix} A & -B \\ B & A \end{bmatrix} \begin{bmatrix} D \\ E \end{bmatrix} = \begin{bmatrix} \mathbb{1} \\ 0 \end{bmatrix}.$$

Diese Matrixgleichung lösen wir mit dem erweiterten GAUSS-Verfahren. Nehmen wir zur Demonstration das einfache Beispiel $C := \begin{bmatrix} 2-3i & 4+i \\ 4-i & i \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 4 & 0 \end{bmatrix} + i \begin{bmatrix} -3 & 1 \\ -1 & 1 \end{bmatrix}$:

I	2	4	3	-1	1	0	
II	4	0	1	-1	0	1	
III	-3	1	2	4	0	0	$II - 2I$
IV	-1	1	4	0	0	0	$2III + 3I$
							$2IV + I$
I	2	4	3	-1	1	0	
II	0	-8	-5	1	-2	1	$2I + II$
III	0	14	13	5	3	0	$4III + 7II$
IV	0	6	11	-1	1	0	$4IV + 3II$
I	4	0	1	-1	0	1	$(17I - III)/2$
II	0	-8	-5	1	-2	1	$(17II + 5III)/4$
III	0	0	17	27	-2	7	
IV	0	0	29	-1	-2	3	$(17IV - 29III)/8$
I	34	0	0	-22	1	5	$(100I - 22IV)/34$
II	0	-34	0	38	-11	13	$-(100II + 38IV)/34$
III	0	0	17	27	-2	7	
IV	0	0	0	-100	3	-19	$(100III + 27IV)/17$
I	100	0	0	0	1	27	
II	0	100	0	0	29	-17	
III	0	0	100	0	-7	11	
IV	0	0	0	100	-3	19	

Die zugehörige Matrixgleichung hat jetzt die Form

$$\begin{bmatrix} \mathbb{1} & 0 \\ 0 & \mathbb{1} \end{bmatrix} \begin{bmatrix} D \\ E \end{bmatrix} = \begin{bmatrix} D \\ E \end{bmatrix} = \frac{1}{100} \begin{bmatrix} 1 & 27 \\ 29 & -17 \\ -7 & 11 \\ -3 & 19 \end{bmatrix}.$$

Also ist $C^{-1} = \frac{1}{100} \begin{bmatrix} 1 & 27 \\ 29 & -17 \end{bmatrix} + \frac{i}{100} \begin{bmatrix} -7 & 11 \\ -3 & 19 \end{bmatrix} = \frac{1}{100} \begin{bmatrix} 1-7i & 27+11i \\ 29-3i & -17+19i \end{bmatrix}$.

6.5.14 Orthonormalbasis Beispiel 6.5.10 zeigt, daß es mühsam sein kann, die Entwicklungskoeffizienten eines Vektors $x \in V$ für eine Basis $\mathcal{B} := \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ von V zu bestimmen. Normalerweise läuft das auf ein inhomogenes Gleichungssystem hinaus, das mit dem GAUSS-Verfahren gelöst werden muß. Es gibt aber besonders gute Basen, für die das Auffinden der Koeffizienten sehr einfach wird. Wenn uns ein Skalarprodukt auf V zur Verfügung steht, wenn V also ein Hilbertraum ist, lassen sich Basen \mathcal{B} mit folgenden Eigenschaften bilden.

6.5.15 Definition Eine Basis $\mathcal{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ für einen Hilbertraum V über \mathbb{K} heißt orthonormal, falls sie folgende Eigenschaften aufweist:

- i) Die Basisvektoren sind paarweise orthogonal zueinander: $\mathbf{b}_i \perp \mathbf{b}_j$, $i \neq j = 1, \dots, n$.
- ii) Alle Basisvektoren sind auf die Länge 1 normiert: $\|\mathbf{b}_i\| = 1$, $i = 1, \dots, n$.

\mathcal{B} heißt Orthonormalbasis (ONB).

Die Orthonormalität einer Basis $\mathcal{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ lässt sich mit dem KRONECKER-Symbol δ_{ik} , das durch

$$\delta_{ik} := \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases} \quad (6.11)$$

definiert ist, folgendermaßen wiedergeben: Für alle $i, k = 1, \dots, n$ gilt

$$\langle \mathbf{b}_i | \mathbf{b}_k \rangle = \delta_{ik}. \quad (6.12)$$

Die Entwicklung eines Vektors $x = x_1 \mathbf{b}_1 + x_2 \mathbf{b}_2 + \dots + x_n \mathbf{b}_n$ nach einer ONB $\mathcal{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ reduziert sich auf die Berechnung von Skalarprodukten:

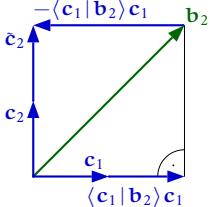
$$\begin{aligned} \langle \mathbf{b}_1 | x \rangle &= \langle \mathbf{b}_1 | x_1 \mathbf{b}_1 + x_2 \mathbf{b}_2 + \dots + x_n \mathbf{b}_n \rangle \\ &= x_1 \langle \mathbf{b}_1 | \mathbf{b}_1 \rangle + x_2 \langle \mathbf{b}_1 | \mathbf{b}_2 \rangle + \dots + x_n \langle \mathbf{b}_1 | \mathbf{b}_n \rangle = \sum_{k=1}^n x_k \cdot \delta_{1k} = x_1. \end{aligned}$$

Genauso folgen $x_2 = \langle \mathbf{b}_2 | x \rangle, \dots, x_n = \langle \mathbf{b}_n | x \rangle$.

6.5.16 Das GRAM-SCHMIDT-Verfahren Das GRAM-SCHMIDT-Verfahren, erzeugt aus einer beliebigen Basis von $(V, \langle \cdot | \cdot \rangle)$ eine Orthonormalbasis.

Wir stellen die Idee an einer Basis $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ vor. Aus \mathcal{B} soll eine ONB $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$ entstehen. Wir starten mit $\mathbf{c}_1 := \frac{1}{\|\mathbf{b}_1\|} \mathbf{b}_1$. Anschließend konstruieren wir aus \mathbf{c}_1 und \mathbf{b}_2 einen Vektor \mathbf{c}_2 , der orthogonal zu \mathbf{c}_1 ist. Die Idee dazu lässt sich aus der Skizze leicht ablesen:

$$\tilde{\mathbf{c}}_2 := \mathbf{b}_2 - \langle \mathbf{c}_1 | \mathbf{b}_2 \rangle \mathbf{c}_1.$$



Dann gilt nämlich $\langle \mathbf{c}_1 | \tilde{\mathbf{c}}_2 \rangle = \langle \mathbf{c}_1 | \mathbf{b}_2 \rangle - \langle \mathbf{c}_1 | \mathbf{b}_2 \rangle \langle \mathbf{c}_1 | \mathbf{c}_1 \rangle = 0$. Den zweiten Basisvektor \mathbf{c}_2 erhalten wir durch Normierung von $\tilde{\mathbf{c}}_2$: $\mathbf{c}_2 := \frac{1}{\|\tilde{\mathbf{c}}_2\|} \tilde{\mathbf{c}}_2$. Damit sind die ersten beiden Basisvektoren bereits konstruiert. Sehen wir, ob wir das Verfahren richtig erweitern können:

$$\tilde{\mathbf{c}}_3 := \mathbf{b}_3 - \langle \mathbf{c}_1 | \mathbf{b}_3 \rangle \mathbf{c}_1 - \langle \mathbf{c}_2 | \mathbf{b}_3 \rangle \mathbf{c}_2.$$

Wir erhalten $\langle \mathbf{c}_1 | \tilde{\mathbf{c}}_3 \rangle = \langle \mathbf{c}_1 | \mathbf{b}_3 \rangle - \langle \mathbf{c}_1 | \mathbf{b}_3 \rangle \langle \mathbf{c}_1 | \mathbf{c}_1 \rangle - \langle \mathbf{c}_2 | \mathbf{b}_3 \rangle \langle \mathbf{c}_1 | \mathbf{c}_2 \rangle = 0$ und genauso $\langle \mathbf{c}_2 | \tilde{\mathbf{c}}_3 \rangle = \langle \mathbf{c}_2 | \mathbf{b}_3 \rangle - \langle \mathbf{c}_1 | \mathbf{b}_3 \rangle \langle \mathbf{c}_2 | \mathbf{c}_1 \rangle - \langle \mathbf{c}_2 | \mathbf{b}_3 \rangle \langle \mathbf{c}_2 | \mathbf{c}_2 \rangle = 0$. Damit ist $\mathbf{c}_1 \perp \tilde{\mathbf{c}}_3$ und $\mathbf{c}_2 \perp \tilde{\mathbf{c}}_3$ gewährleistet. Also ist $\mathbf{c}_3 := \frac{1}{\|\tilde{\mathbf{c}}_3\|} \tilde{\mathbf{c}}_3$ der letzte Vektor der ONB \mathcal{C} .

An diesem Beispiel ist klar zu erkennen, daß es sich um ein rekursives Verfahren handelt. Von $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ ausgehend werden nacheinander die Vektoren $\mathbf{c}_1, \mathbf{c}_2$ usw. berechnet. Ist \mathbf{c}_k für ein $k < n$ bereits konstruiert, dann wird zunächst

$$\tilde{\mathbf{c}}_{k+1} := \mathbf{b}_{k+1} - \sum_{i=1}^k \langle \mathbf{c}_i | \mathbf{b}_{k+1} \rangle \mathbf{c}_i \quad (6.13)$$

gebildet. Dieser Vektor ist zu allen $\mathbf{c}_1, \dots, \mathbf{c}_k$ orthogonal, denn für $1 \leq j \leq k$ gilt

$$\langle \mathbf{c}_j | \tilde{\mathbf{c}}_{k+1} \rangle := \langle \mathbf{c}_j | \mathbf{b}_{k+1} \rangle - \sum_{i=1}^k \langle \mathbf{c}_i | \mathbf{b}_{k+1} \rangle \langle \mathbf{c}_j | \mathbf{c}_i \rangle = \langle \mathbf{c}_j | \mathbf{b}_{k+1} \rangle - \sum_{i=1}^k \langle \mathbf{c}_i | \mathbf{b}_{k+1} \rangle \delta_{ji} = 0.$$

Anschließend wird dieser Vektor normiert: $\mathbf{c}_k := \frac{1}{\|\tilde{\mathbf{c}}_k\|} \tilde{\mathbf{c}}_k$. Auf diese Weise entsteht die ONB $\mathcal{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ aus $\mathcal{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$.

6.5.17 Beispiel

Wir gehen von der Basis

$$\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\} := \left\{ \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix}, \begin{bmatrix} i \\ i \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \\ i \end{bmatrix} \right\}$$

aus. Dann ist $\mathbf{c}_1 = \frac{1}{5} [3, 0, 4]^t$,

$$\tilde{\mathbf{c}}_2 = \begin{bmatrix} i \\ i \\ -1 \end{bmatrix} - \frac{1}{25}(3i-4) \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix} = \frac{1}{25} \begin{bmatrix} 25i-9i+12 \\ 25i \\ -25-12i+16 \end{bmatrix} = \frac{1}{25} \begin{bmatrix} 12+16i \\ 25i \\ -9-12i \end{bmatrix}$$

und $\mathbf{c}_2 = \frac{1}{25\sqrt{2}} [12+16i, 25i, -9-12i]^t$. Bevor wir $\tilde{\mathbf{c}}_3$ bestimmen, berechnen wir die beteiligten Skalarprodukte.

$$\langle \mathbf{c}_1 | \mathbf{b}_3 \rangle = \frac{1}{5}(3+4i), \quad \langle \mathbf{c}_2 | \mathbf{b}_3 \rangle = \frac{1}{25\sqrt{2}}(12-16i-75i-9i-12) = -2\sqrt{2}i.$$

Damit erhalten wir

$$\begin{aligned} \tilde{\mathbf{c}}_3 &= \begin{bmatrix} 1 \\ 3 \\ i \end{bmatrix} - \frac{3+4i}{25} \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix} + \frac{2i}{25} \begin{bmatrix} 12+16i \\ 25i \\ -9-12i \end{bmatrix} = \frac{1}{25} \begin{bmatrix} 25-9-12i+24i-32 \\ 75-50 \\ 25i-12-16i-18i+24 \end{bmatrix} \\ &= \frac{1}{25} \begin{bmatrix} -16+12i \\ 25 \\ 12-9i \end{bmatrix}. \end{aligned}$$

Die ONB ist damit

$$\mathcal{C} = \{ \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3 \} = \left\{ \frac{1}{5} \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix}, \frac{\sqrt{2}}{50} \begin{bmatrix} 12 + 16i \\ 25i \\ -9 - 12i \end{bmatrix}, \frac{\sqrt{2}}{50} \begin{bmatrix} -16 + 12i \\ 25 \\ 12 - 9i \end{bmatrix} \right\}.$$

Jetzt stellen wir noch den Vektor $\mathbf{x} := [1, 2, 3]^t$ in dieser Basis dar:

$$\langle \mathbf{c}_1 | \mathbf{x} \rangle = 3, \langle \mathbf{c}_2 | \mathbf{x} \rangle = -\frac{3\sqrt{2}}{10} (1 + 2i) \text{ und } \langle \mathbf{c}_3 | \mathbf{x} \rangle = \frac{\sqrt{2}}{10} (14 + 3i) \text{ ergibt}$$

$$\mathbf{x} = 3 \cdot \mathbf{c}_1 - \frac{3\sqrt{2}}{10} (1 + 2i) \cdot \mathbf{c}_2 + \frac{\sqrt{2}}{10} (14 + 3i) \cdot \mathbf{c}_3.$$

6.6 Matrizen als lineare Abbildungen

Im Zusammenhang mit linearen Gleichungssystemen sind Matrizen eine bequeme Methode das Wesentliche des Systems, nämlich die Koeffizienten a_{ij} , zusammenzufassen. Mit der maßgeschneiderten Multiplikation zwischen $m \times n$ -Matrix A und Spaltenvektor $x \in \mathbb{K}^n$ ergibt sich die prägnante Form $Ax = b$ für das Gleichungssystem. In diesem Zusammenhang hat A eine passive, sozusagen eine ordnende Aufgabe, die wir verwenden, um das GAUSS-Verfahren übersichtlich zu gestalten, oder um die Lösungsstruktur herauszuarbeiten.

Wir können der Matrix A in $Ax = b$ aber auch eine aktive Rolle zuweisen, indem wir diese Gleichung als Abbildung lesen: Der Vektor $x \in \mathbb{K}^n$ wird durch A auf den Vektor $b \in \mathbb{K}^m$ abgebildet. Die Multiplikation zwischen A und x bestimmt dabei die konkrete Abbildungsvorschrift. Der Einfachheit halber bezeichnen wir diese Abbildung wieder mit demselben Symbol A (wir werden aber Situationen kennenlernen, in denen wir zwischen solchen Abbildungen und ihren Realisierungen durch Matrizen unterscheiden müssen). A hat dabei eine entscheidende Eigenschaft, nämlich die *Linearität*

$$A(sx + ty) = sAx + tAy \quad \text{für alle } s, t \in \mathbb{K}, x, y \in \mathbb{K}^n,$$

die sie gegenüber einer beliebigen Abbildung zwischen \mathbb{K}^n und \mathbb{K}^m auszeichnet. Wir sprechen in diesem Zusammenhang von einer *linearen Abbildung*. Gelegentlich heben wir den Abbildungscharakter von A durch die Schreibweise

$$A : \mathbb{K}^n \rightarrow \mathbb{K}^m; x \mapsto Ax$$

hervor. Der Linearität ist es zu verdanken, daß solche Abbildungen durch die Kenntnis weniger Daten bereits eindeutig festgelegt sind. Um zu verstehen, was genau damit gemeint ist, entwickeln wir x in der kanonischen Basis \mathcal{E} : $x = \sum_{i=1}^n x_i e_i$. Wenden wir darauf die Abbildung A an, so folgt aus deren Linearität

$$Ax = A\left(\sum_{i=1}^n x_i e_i\right) = \sum_{i=1}^n x_i Ae_i.$$

Das zeigt, daß das Bild Ax von x als Linearkombination $\sum_{i=1}^n x_i Ae_i$ der Bilder Ae_i der kanonischen Basisvektoren e_i entsteht. Daher ist die Wirkung der linearen Abbildung A wirklich bereits durch wenige Daten festgelegt, nämlich durch die n Bilder Ae_1, \dots, Ae_n der kanonischen Basisvektoren e_1, \dots, e_n . Diese Bilder müßten im Zahlenschema der Matrix wiederzufinden sein:

$$Ae_1 = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ \vdots \\ a_{m1} \end{bmatrix} =: a_1,$$

$$\begin{aligned} A\mathbf{e}_2 &= \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \\ \vdots \\ a_{m2} \end{bmatrix} =: \mathbf{a}_2, \\ \\ A\mathbf{e}_n &= \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} a_{1n} \\ a_{2n} \\ a_{3n} \\ \vdots \\ a_{mn} \end{bmatrix} =: \mathbf{a}_n. \end{aligned}$$

Und tatsächlich sind es einfach die Spaltenvektoren \mathbf{a}_i der Matrix. Um auf diesen Sachverhalt hinzuweisen, vereinbaren wir die folgende Schreibweise für eine Matrix:

$$A =: [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n]. \quad (6.14)$$

D.h.:

In den Spalten einer Matrix stehen die Bilder der kanonischen Basisvektoren.

Daraus ist sofort ersichtlich, daß es sich bei Ax einfach um eine Linearkombination der Spaltenvektoren \mathbf{a}_i handelt, mit den Koordinaten x_i als Linearfaktoren:

$$Ax = x_1 \mathbf{a}_1 + x_2 \mathbf{a}_2 + \dots + x_n \mathbf{a}_n = \sum_{i=1}^n x_i \mathbf{a}_i. \quad (6.15)$$

Dieses Ergebnis läßt sich natürlich auch direkt aus der Definition (6.4) für die Multiplikation einer Matrix A mit einem Vektor x gewinnen:

$$\begin{aligned} Ax &= \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n \\ a_{31}x_1 + a_{32}x_2 + \dots + a_{3n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n \end{bmatrix} = \begin{bmatrix} x_1 a_{11} \\ x_1 a_{21} \\ x_1 a_{31} \\ \vdots \\ x_1 a_{m1} \end{bmatrix} + \begin{bmatrix} x_2 a_{12} \\ x_2 a_{22} \\ x_2 a_{32} \\ \vdots \\ x_2 a_{m2} \end{bmatrix} + \dots + \begin{bmatrix} x_n a_{1n} \\ x_n a_{2n} \\ x_n a_{3n} \\ \vdots \\ x_n a_{mn} \end{bmatrix} \\ &= x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \\ \vdots \\ a_{m2} \end{bmatrix} + \dots + x_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ a_{3n} \\ \vdots \\ a_{mn} \end{bmatrix} = x_1 \mathbf{a}_1 + x_2 \mathbf{a}_2 + \dots + x_n \mathbf{a}_n. \end{aligned}$$

Können wir also aus der Abbildungsvorschrift erkennen, was aus den Basisvektoren wird, so kennen wir die zugehörige Matrix. Als Beispiel dafür nehmen wir eine Drehung D_α im \mathbb{R}^2 um einen festen Winkel α . Aus der Skizze entnehmen wir die Bilder der Basisvektoren:

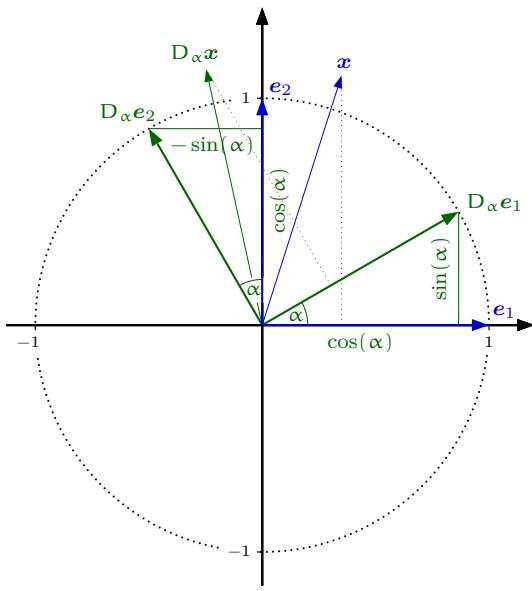


Abb. 6.1 Eine ebene Drehung

$$\mathbf{e}_1 \mapsto \begin{bmatrix} \cos(\alpha) \\ \sin(\alpha) \end{bmatrix}, \quad \mathbf{e}_2 \mapsto \begin{bmatrix} -\sin(\alpha) \\ \cos(\alpha) \end{bmatrix}.$$

Daher ist

$$D(\alpha) = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}. \quad (6.16)$$

Als weiteres Beispiel für diese Regel verschaffen wir uns die Matrix, die zur Hintereinanderausführung zweier linearer Abbildungen $A : \mathbb{K}^n \rightarrow \mathbb{K}^m$ und $B : \mathbb{K}^m \rightarrow \mathbb{K}^k$ gehört – diese Hintereinanderausführung ergibt eine lineare Abbildung von \mathbb{K}^n nach \mathbb{K}^k (wie man sich schnell überzeugt), die wir $B \cdot A$, meist aber einfach BA schreiben. Diese Matrix bezeichnen wir als *Matrixprodukt* von B mit A . Die Spaltenvektoren von BA erhalten wir durch Auswertung dieser Abbildung auf den kanonischen Basisvektoren \mathbf{e}_i :

$$BA\mathbf{e}_i = B(A\mathbf{e}_i) = B(\mathbf{a}_i) = B\mathbf{a}_i.$$

Die Darstellung gemäß (6.14) für BA durch die Spaltenvektoren lautet daher

$$BA = [B\mathbf{a}_1, B\mathbf{a}_2, \dots, B\mathbf{a}_n]. \quad (6.17)$$

Das ist eine konkrete Rechenvorschrift für die Matrixmultiplikation. Wir haben sie auf die Multiplikation Matrix · Vektor (6.4) zurückgeführt: Die Spaltenvektoren von BA entstehen einfach durch Multiplikation der Spaltenvektoren \mathbf{a}_i von A mit B . Zum Beispiel für

$$A := \begin{bmatrix} -1 & 3 & 4 & 1 \\ 2 & 3 & 3 & -1 \end{bmatrix} \quad \text{und} \quad B := \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 0 & 2 \end{bmatrix}$$

erhalten wir

$$BA = \left[\begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right] = \begin{bmatrix} 3 & 9 & 10 & -1 \\ 0 & 9 & 11 & 1 \\ 4 & 6 & 6 & -2 \end{bmatrix}.$$

Ein interessanteres Beispiel ist die Hintereinanderausführung zweier Drehungen $D(\beta)$ und $D(\alpha)$. Sie muß die Drehung $D(\alpha + \beta)$ ergeben:

$$\begin{aligned} D(\alpha)D(\beta) &= \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix} \begin{bmatrix} \cos(\beta) & -\sin(\beta) \\ \sin(\beta) & \cos(\beta) \end{bmatrix} \\ &= \begin{bmatrix} \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta) & -\cos(\alpha)\sin(\beta) - \sin(\alpha)\cos(\beta) \\ \sin(\alpha)\cos(\beta) + \cos(\alpha)\sin(\beta) & -\sin(\alpha)\sin(\beta) + \cos(\alpha)\cos(\beta) \end{bmatrix} \\ &= \begin{bmatrix} \cos(\alpha + \beta) & -\sin(\alpha + \beta) \\ \sin(\alpha + \beta) & \cos(\alpha + \beta) \end{bmatrix} = D(\alpha + \beta). \end{aligned}$$

Vergleichen wir die Einträge, so finden wir die sogenannten *Additionssätze* für die trigonometrischen Funktionen sin und cos (vergl. auch Seite 79 und 272)

$$\sin(\alpha + \beta) = \sin(\alpha)\cos(\beta) + \sin(\beta)\cos(\alpha), \quad (6.18)$$

$$\cos(\alpha + \beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta). \quad (6.19)$$

Der Vollständigkeit halber seien noch zwei Abbildungen erwähnt, für die die zugehörigen Matrizen leicht anzugeben sind: Die erste ist die Abbildung $\mathbb{O}: \mathbb{K}^n \rightarrow \mathbb{K}^m$, die jedem Vektor $x \in \mathbb{K}^n$ den Vektor $\mathbf{0} \in \mathbb{K}^m$ zuordnet: $\mathbb{O}(x) := \mathbf{0}$. Das ist die sogenannte *Nullabbildung*. Die zweite ist die *identische Abbildung* $\text{id}: \mathbb{K}^n \rightarrow \mathbb{K}^n$, $\text{id}(x) := x$. Die zugehörigen Matrizen werden als *Nullmatrix* \mathbb{O}_n bzw. *Einheitsmatrix* $\mathbb{1}_n$ bezeichnet (wenn die Dimension n klar ist, meist einfach \mathbb{O} bzw. $\mathbb{1}$, wobei sich \mathbb{O} nicht durchgesetzt hat, weil es bequemer ist, dafür 0 zu schreiben und darauf zu vertrauen, daß aus dem Kontext klar wird, was gemeint ist). Sie sind offensichtlich durch

$$\mathbb{O}_n = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad \mathbb{1}_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \quad (6.20)$$

gegeben. Der Eintrag von $\mathbb{1}_n$ an der Position (i, k) ist δ_{ik} (vergl. (6.11)).

6.7 Direkte Zerlegung eines Vektorraums

6.7.1 Definition Für zwei Teilräume T_1 und T_2 eines Vektorraums V heißt

$$T_1 + T_2 := \{x_1 + x_2 \mid x_1 \in T_1, x_2 \in T_2\} \quad (6.21)$$

Summe der Teilräume T_1 und T_2 . Gilt $T_1 \cap T_2 = \{\mathbf{0}\}$, so heißt die Summe direkt. In diesem Fall schreibt man $T_1 \oplus T_2$.

6.7.2 Lemma $T_1 \oplus T_2$ ist ein Teilraum von V . Für einen Vektor $x \in T_1 \oplus T_2$ ist die Zerlegung $x = x_1 + x_2$ in Vektoren $x_1 \in T_1$ und $x_2 \in T_2$ eindeutig. Es gilt

$$\dim T_1 \oplus T_2 = \dim T_1 + \dim T_2. \quad (6.22)$$

Beweis. Jeder Vektor $x \in T_1 \oplus T_2$ hat per Definition eine Zerlegung $x = x_1 + x_2$, mit Vektoren $x_1 \in T_1$ und $x_2 \in T_2$. Gäbe es eine weitere Zerlegung $x = \tilde{x}_1 + \tilde{x}_2$ dieser Art, so müßte $x_1 - \tilde{x}_1 = x_2 - \tilde{x}_2$ gelten. Die linke Seite dieser Gleichung ist ein Vektor aus T_1 , die rechte einer aus T_2 . Dieser Vektor würde in $T_1 \cap T_2 = \{\mathbf{0}\}$ liegen, was $x_1 = \tilde{x}_1$ und $\tilde{x}_2 = x_2$ zur Folge hätte. Also gibt es keine von $x_1 + x_2$ verschiedene Zerlegung von x .

Für $x = x_1 + x_2 \in T_1 \oplus T_2$, $y = y_1 + y_2 \in T_1 \oplus T_2$ und $t, s \in \mathbb{K}$ ist $tx + sy = tx_1 + sy_1 + tx_2 + sy_2$ wieder in $T_1 \oplus T_2$. Da es sich bei T_1 und T_2 nämlich um Teilräume handelt, ist $tx_1 + sy_1 \in T_1$ und $tx_2 + sy_2 \in T_2$.

Seien $B_1 := \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ und $B_2 := \{\mathbf{c}_1, \dots, \mathbf{c}_m\}$ Basen von T_1 bzw. T_2 . B_1 und B_2 sind offensichtlich disjunkt, denn ein gemeinsames Element müßte in $T_1 \cap T_2$ liegen und daher $\mathbf{0}$ sein.

$B_1 \cup B_2$ ist linear unabhängig: Aus $\mathbf{0} = \sum_{k=1}^n \lambda_k \mathbf{b}_k + \sum_{k=1}^m \gamma_k \mathbf{c}_k$ folgt $\mathbf{0} = \sum_{k=1}^n \lambda_k \mathbf{b}_k$ und $\mathbf{0} = \sum_{k=1}^m \gamma_k \mathbf{c}_k$, denn die Zerlegung auch von $\mathbf{0}$ in $\mathbf{0} \in T_1$ und $\mathbf{0} \in T_2$ ist eindeutig. Die Basiseigenschaft von B_1 und B_2 ergibt $\lambda_1 = \dots = \lambda_n = 0$ und $\gamma_1 = \dots = \gamma_m = 0$.

$B_1 \cup B_2$ ist offensichtlich erzeugend und daher sogar eine Basis für $T_1 \oplus T_2$. Deshalb gilt $\dim T_1 \oplus T_2 = n + m = \dim T_1 + \dim T_2$. \square

6.7.3 Satz Für einen Teilraum T von V gibt es einen weiteren Teilraum S mit der Eigenschaft $V = T \oplus S$. Insbesondere gilt dann $\dim V = \dim T + \dim S$.

Beweis. Dieses Lemma ist im Wesentlichen ein Ergebnis des Basisergänzungssatzes 6.5.8. Der Teilraum T ist insbesondere ein Vektorraum und hat nach 6.5.5 eine Basis $\mathcal{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_m\}$ mit einem m , nicht größer als $n := \dim V$. Ist $m = n$, so handelt es sich um den trivialen Fall $T = V$, für den $S := \{\mathbf{0}\}$ die formalen Kriterien des Lemmas erfüllt. Wir können also $m < n$ annehmen. Nach Satz 6.5.8 läßt sich \mathcal{B} zu einer Basis $\mathcal{B} \cup \{\mathbf{c}_1, \dots, \mathbf{c}_{n-m}\}$ von V ergänzen. Wir wählen $S := \text{lh}\{\mathbf{c}_1, \dots, \mathbf{c}_{n-m}\}$. Dann gilt $T \cap S = \{\mathbf{0}\}$, denn ein Vektor x im Schnitt dieser Teilräume besäße Darstellungen $x = \sum_{k=1}^m \lambda_k \mathbf{b}_k = \sum_{k=1}^{n-m} \gamma_k \mathbf{c}_k$. Ihre Differenz ergäbe eine nicht triviale Nullkombination der Basisvektoren, würden nicht alle Koeffizienten λ_k und γ_k einfach Null sein. Daher ist $x = \mathbf{0}$. Bei $T + S$ handelt es sich damit sogar um die direkte Summe $T \oplus S$ von T und S . Die Vereinigung der Basen von T und S ist laut Konstruktion eine Basis für V , was sofort zu $V = T \oplus S$ und $\dim V = \dim T + \dim S$ führt. \square

Wie der Beweis zeigt, ist der Teilraum S durch T normalerweise nicht eindeutig festgelegt. Es gibt viele Möglichkeiten, eine Basis von T zu einer Basis von V zu ergänzen. Verfügen wir allerdings über ein Skalarprodukt auf V , dann kann S durch die Zusatzforderung, daß alle Vektoren aus S auf T senkrecht stehen sollen, eindeutig festgelegt werden.

6.7.4 Definition Sei U eine nicht leere Teilmenge des Vektorraums V . Dann heißt $U^\perp := \{x \in V \mid y \perp x \text{ f.a. } y \in U\}$ der zu U orthogonale Teilraum. Falls U nur einen Vektor x enthält, schreiben wir x^\perp statt $\{x\}^\perp$.

Es ist eine gute Übung nachzurechnen, daß es sich bei U^\perp tatsächlich um einen Teilraum von V handelt und daß $\{\mathbf{0}\}^\perp = V$, $V^\perp = \{\mathbf{0}\}$ gilt.

6.7.5 Satz Für eine nicht leere Teilmenge $U \subseteq V$ ist $U^{\perp\perp}$ die lineare Hülle von U , also der kleinste Teilraum von V , der U umfaßt: $U^{\perp\perp} = \text{lh}(U)$. Wenn U sogar ein Teilraum von V ist, gilt also $U^{\perp\perp} = U$ und

$$V = U \oplus U^\perp. \quad (6.23)$$

Beweis. Ein Vektor $x \in \text{lh}(U)$ läßt sich als Linearkombination $x = \sum_{i=1}^p \lambda_i x_i$ von Vektoren $x_i \in U$ schreiben. Für jedes $y \in U^\perp$ gilt daher $\langle x | y \rangle = \sum_{i=1}^p \lambda_i \langle x_i | y \rangle = 0$. Das bedeutet $x \in U^{\perp\perp}$. Damit haben wir $\text{lh}(U) \subseteq U^{\perp\perp}$. Falls $\text{lh}(U) = V$ gelten sollte, ist nichts weiter zu zeigen. Wir können also von $\text{lh}(U) \subset V$ ausgehen. Für die umgekehrte Inklusion nehmen wir zunächst an, daß U nicht nur eine Teilmenge, sondern sogar ein (echter) Teilraum von V ist und zeigen erst einmal (6.23). Als endlich erzeugter Vektorraum hat U eine Basis $\mathcal{B} := \{b_1, \dots, b_k\}$, die mit dem GRAM-SCHMIDTSchen Orthonormalisierungsverfahren 6.5.16 zu einer ONB für U gemacht werden kann. Wir können also davon ausgehen, daß \mathcal{B} eine ONB ist. Nach dem Basisergänzungssatz 6.5.8 kann \mathcal{B} zu einer Basis $\{b_1, \dots, b_k, c_1, \dots, c_\ell\}$ von V ergänzt werden. Darauf wenden wir wieder das GRAM-SCHMIDTSche Orthonormalisierungsverfahren an. Dabei ändern sich die Vektoren b_i nicht. Deshalb können wir annehmen, daß die Vektoren c_j untereinander und auch zu allen b_i orthogonal sind. Das bedeutet, daß die Teilräume U und $W := \text{lh}(\{c_1, \dots, c_\ell\})$ orthogonal zueinander sind und zusammen eine direkte Zerlegung von V ergeben: $V = U \oplus W$. Offensichtlich gilt $W \subseteq U^\perp$. Jeder Vektor $x \in U^\perp$ läßt sich eindeutig in einen Anteil $x_1 \in U$ und einen Anteil $x_2 \in W$ zerlegen: $x = x_1 + x_2$. Das ergibt $\langle x | x_1 \rangle = 0 = \langle x_1 | x_1 \rangle + \langle x_2 | x_1 \rangle = \langle x_1 | x_1 \rangle$, also $x_1 = \mathbf{0}$ und damit $x \in W$. Daher folgt $U^\perp \subseteq W$ und daraus $U^\perp = W$. Damit ist (6.23) gezeigt. Für eine orthogonale direkte Zerlegung $V = U \oplus W$ ist also $W = U^\perp$. Wenden wir das auf die orthogonale Zerlegung $V = U^\perp \oplus U$ an, so erhalten wir $U = U^{\perp\perp}$.

Kommen wir auf den allgemeinen Fall zurück. Ein Vektor, der orthogonal zu U ist, ist auch orthogonal zu allen Linearkombinationen von Vektoren aus U , also orthogonal zu $\text{lh}(U)$. Das bedeutet $U^\perp \subseteq \text{lh}(U)^\perp$. Die umgekehrte Inklusion ist offensichtlich, so daß $U^\perp = \text{lh}(U)^\perp$ folgt. Jetzt ergibt sich $U^{\perp\perp} = \text{lh}(U)^{\perp\perp} = \text{lh}(U)$ aus den oben erhaltenen Ergebnissen, angewandt auf den Teilraum $\text{lh}(U)$. \square

6.7.6 Korollar Sei $\mathbf{0} \neq x \in V$. Dann gilt $\dim x^\perp = \dim V - 1$.

Beweis. Das folgt sofort aus $V = \text{lh}(x) \oplus x^\perp$, aus Lemma 6.7.2 und $\dim \text{lh}(x) = 1$. \square

6.8 Die Dimensionsformel

6.8.1 Definition Für eine lineare Abbildung $A : V \rightarrow W$ wird $\ker A := \{x \in V \mid Ax = \mathbf{0}\} \subseteq V$ als Kern von A bezeichnet. Mit $\text{im } A := \{y \in W \mid \exists_{x \in V} y = Ax\}$ wird, wie üblich, das Bild von A bezeichnet (vergl. 2.7).

Man überlege sich, daß $\ker A$ und $\text{im } A$ Teilräume von V bzw. W sind.

6.8.2 Lemma Eine lineare Abbildung $A : V \rightarrow W$ ist genau dann injektiv, wenn $\ker A = \{\mathbf{0}\}$ gilt. Für $A := [a_1, \dots, a_n]$ ist das genau dann der Fall, wenn die Menge $\{a_1, \dots, a_n\}$ der Spaltenvektoren von A linear unabhängig ist.

Beweis. Für eine lineare Abbildung A gilt immer $A\mathbf{0} = \mathbf{0}$ ($A\mathbf{0} = A\mathbf{0} = 0A\mathbf{0} = \mathbf{0}$). Ist A injektiv und $x \in \ker A$, so folgt $Ax = \mathbf{0} = A\mathbf{0}$, also $x = \mathbf{0}$. Daher muß $\ker A = \{\mathbf{0}\}$ gelten.

Ist $\ker A = \{\mathbf{0}\}$ und $Ax = Ay$, so folgt aus der Linearität $Ax - Ay = A(x - y) = \mathbf{0}$, also $x - y \in \ker A$ und daher $x = y$. Damit ist A eine injektive Abbildung.

Für einen Vektor $x = [x_1, \dots, x_n]^t \in V$ gilt $Ax = x_1 a_1 + \dots + x_n a_n$. $Ax = \mathbf{0}$ für $x \neq \mathbf{0}$ ist daher äquivalent dazu, daß es eine nicht triviale Nullkombination $x_1 a_1 + \dots + x_n a_n = \mathbf{0}$ der Spaltenvektoren von A gibt. Das ist äquivalent zur linearen Abhängigkeit von $\{a_1, \dots, a_n\}$. Damit haben wir gezeigt: A ist genau dann nicht injektiv, wenn $\{a_1, \dots, a_n\}$ linear abhängig ist. Damit ist alles gezeigt. \square

Injektive Abbildungen erhalten die lineare Unabhängigkeit von Mengen:

6.8.3 Proposition Für eine injektive lineare Abbildung $A : V \rightarrow W$ und eine linear unabhängige Menge $B = \{b_1, \dots, b_m\} \subset V$ ist $A(B) := \{Ab_1, \dots, Ab_m\} \subset W$ wieder linear unabhängig. Ist A sogar bijektiv, so gilt $\dim V = \dim W$ und $A(B)$ ist eine Basis für W .

Beweis. Aus $\sum_{k=1}^m \lambda_k Ab_k = \mathbf{0}$ folgt wegen der Linearität von A : $A \sum_{k=1}^m \lambda_k b_k = \mathbf{0}$, also $\sum_{k=1}^m \lambda_k b_k \in \ker A = \{\mathbf{0}\}$. Da B linear unabhängig ist, geht das nur für $\lambda_1 = \dots = \lambda_m = 0$. Daher enthält $A(B)$ nur triviale Nullkombinationen und ist somit ebenfalls linear unabhängig.

Für eine bijektive Abbildung A und eine Basis $B = \{b_1, \dots, b_n\}$ von V ist $A(B)$ lineare unabhängig. Für die Basiseigenschaft fehlt uns noch, daß $A(B)$ für W erzeugend ist:

Jedes $y \in W$ ist das Bild genau eines Vektors $x \in V$, mit einer Basisdarstellung $x = \sum_{k=1}^n x_k b_k$. Daher haben wir $y = Ax = \sum_{k=1}^n x_k Ab_k \in \text{lh } A(B)$. Das zeigt, daß $A(B)$ erzeugend ist.

$A(B)$ hat wie B genau n Elemente, so daß $\dim V = n = \dim W$ gelten muß. \square

6.8.4 Satz (Dimensionsformel) Für eine lineare Abbildung $A : V \rightarrow W$ zwischen den Vektorräumen V und W gilt

$$\dim V = \dim \ker A + \dim \text{im } A. \quad (6.24)$$

Beweis. Die Beweisidee besteht darin, den Vektorraum V gemäß Lemma 6.7.3 in eine direkte Summe aus $\ker A$ und einem Teilraum V_1 aufzuspalten: $V = \ker A \oplus V_1$. $\ker A$ ist der Anteil von V , der auf $\mathbf{0}$ abgebildet wird. Dieser Teilraum trägt also zur Dimension des Bildes im A nichts bei. Dann muß V_1 der Teilraum sein, der den von $\mathbf{0}$ verschiedenen Anteil des Bildes erzeugt. Tatsächlich bildet A den Teilraum V_1 bijektiv auf $\text{im } A$ ab, so daß nach Proposition 6.8.3 $\dim V_1 = \dim \text{im } A$ gilt. Zusammen mit $\dim V = \dim \ker A + \dim V_1$ nach Lemma 6.7.2, ergibt das die Dimensionsformel.

Wir müssen somit nur noch zeigen, daß die Einschränkung $A|_{V_1} : V_1 \rightarrow \text{im } A$ bijektiv ist.

Zunächst untersuchen wir den Kern von $A|_{V_1}$.

$x \in \ker A|_{V_1}$ bedeutet $x \in V_1$ und $A|_{V_1}x = Ax = \mathbf{0}$. Das heißt $x \in V_1 \cap \ker A = \{\mathbf{0}\}$, also $x = \mathbf{0}$. Folglich ist $\ker A|_{V_1} = \{\mathbf{0}\}$ und daher $A|_{V_1}$ injektiv (Lemma 6.8.2).

Natürlich gilt $\text{im } A|_{V_1} \subseteq \text{im } A$, denn $A|_{V_1}$ ist ja die Einschränkung von A . Wir zeigen die umgekehrte Inklusion.

Zu jedem $y \in \text{im } A$ gibt es ein $x \in V$ mit $y = Ax$. x läßt sich gemäß Lemma 6.7.2 eindeutig in seinen Anteil $x_0 \in \ker A$ und $x_1 \in V_1$ zerlegen: $x = x_0 + x_1$. Dann gilt $y = Ax_0 + Ax_1 = Ax_1 = A|_{V_1}x_1 \in \text{im } A|_{V_1}$. Da y beliebig war, haben wir $\text{im } A \subseteq \text{im } A|_{V_1}$, insgesamt also $\text{im } A|_{V_1} = \text{im } A$ gezeigt. $A|_{V_1}$ ist daher surjektiv und somit bijektiv. \square

Eine der wichtigsten Folgerungen der Dimensionsformel ist, daß eine lineare Abbildung von V in denselben Vektorraum V schon dann bijektiv, also (auf ganz V) invertierbar ist, wenn sie injektiv, oder wenn sie surjektiv ist. Die jeweils zur Bijektivität fehlende Eigenschaft ist dann automatisch erfüllt. Für eine Abbildung $V \rightarrow W$ dagegen benötigt man normalerweise beide Eigenschaften.

6.8.5 Korollar *Eine lineare Abbildung $A : V \rightarrow V$ ist genau dann invertierbar, wenn sie injektiv, oder wenn sie surjektiv ist.*

Beweis. Ist A etwa injektiv, so gilt $\dim \ker A = 0$, denn $\ker A = \{\mathbf{0}\}$ nach Lemma 6.8.2. Nach der Dimensionsformel muß daher $\dim V = \dim \text{im } A$ gelten, d. h. $\text{im } A$ ist ein Teilraum von V mit derselben Dimension wie V . Das geht nur für $\text{im } A = V$, also wenn A surjektiv und damit bijektiv ist.

Ist A surjektiv, so gilt $\text{im } A = V$ und daher $\dim V = \dim \text{im } A$. Laut Dimensionsformel folgt daraus $\dim \ker A = 0$, was nur für den 0-dimensionalen Teilraum $\{\mathbf{0}\} = \ker A$ möglich ist. Nach Lemma 6.8.2 ist A injektiv und daher auch bijektiv.

Die Umkehrung ist trivial. \square

6.8.6 A Zeigen Sie: Für eine lineare Abbildung $A : V \rightarrow W$ handelt es sich bei $\ker A$ und $\text{im } A$ um einen Teilraum von V bzw. von W .

6.8.7 A Für eine bijektive lineare Abbildung $A : V \rightarrow W$ existiert die inverse Abbildung $A^{-1} : W \rightarrow V$. Zeigen Sie, daß diese ebenfalls linear ist.

6.9 Die inverse Matrix

6.9.1 Definition Die Menge der quadratischen $n \times n$ -Matrizen mit Einträgen aus einem Körper \mathbb{K} bezeichnen wir durch $M_n(\mathbb{K})$, oder, wenn der Körper aus dem Kontext heraus klar ist, auch einfach durch M_n .

6.9.2 Satz Für eine Matrix $A \in M_n(\mathbb{K})$, ($\mathbb{K} = \mathbb{R}$ oder $\mathbb{K} = \mathbb{C}$) folgt aus der Existenz einer Matrix $B \in M_n(\mathbb{K})$, die eine der beiden Bedingungen

$$BA = \mathbb{1} \quad \text{oder} \quad AB = \mathbb{1} \quad (6.25)$$

erfüllt, daß sie invertierbar ist und daß $A^{-1} = B$ gilt.

Beweis. Aus $BA = \mathbb{1}$ folgt für ein Element $x \in \ker A$: $x = BAx = \mathbf{0}$. Das heißt, $\ker A = \{\mathbf{0}\}$. Daher ist A injektiv und nach Korollar 6.8.5 bijektiv, also invertierbar. Die Inverse ist bereits durch B gegeben, denn aus $BA = \mathbb{1}$ folgt durch Multiplikation mit A^{-1} von rechts: $B = BAA^{-1} = A^{-1}$.

Gilt $AB = \mathbb{1}$, dann liegt jedes $y \in \mathbb{K}^n$ im Bild von A , denn $y = ABy \in \text{im } A$. Daher ist A surjektiv und nach Korollar 6.8.5 invertierbar. Wie oben zeigt man $B = A^{-1}$. \square

6.9.3 Korollar Für zwei Matrizen $A, B \in M_n(\mathbb{K})$ ist das Produkt AB genau dann invertierbar, wenn A und B invertierbar sind. In diesem Fall gilt $(AB)^{-1} = B^{-1}A^{-1}$.

Aus Satz 6.9.2 leiten wir uns ein Verfahren her, mit dem man gleichzeitig erfährt, ob A^{-1} existiert und wie A^{-1} gegebenenfalls aussieht. Dafür müssen wir nur die Gleichung $AB = \mathbb{1}$ nach B auflösen. Allerdings wissen wir im Augenblick noch nicht, wie das gehen könnte. Bis-her haben wir nur lineare Vektorgleichungen $Ab = e$ mit Hilfe des GAUSS-Verfahrens nach einem gesuchten Vektor b aufgelöst, nicht eine Matrixgleichung $AB = \mathbb{1}$ nach einer Matrix B . Tatsächlich läßt sich letzteres aber auf das GAUSS-Verfahren zurückführen. Denn die Matrix $B = [b_1, \dots, b_n]$ ist bekannt, sobald wir ihre Spaltenvektoren b_i kennen. Schreiben wir die Gleichung $AB = \mathbb{1}$ ausführlich auf und verwenden dabei $\mathbb{1} = [e_1, \dots, e_n]$, mit den kanoni-schen Basisvektoren e_i :

$$AB = A[b_1, b_2, \dots, b_n] = [Ab_1, Ab_2, \dots, Ab_n] = [e_1, e_2, \dots, e_n] = \mathbb{1}.$$

Die Matrixgleichung $AB = \mathbb{1}$ ist also äquivalent zu den n Vektorgleichungen

$$Ab_1 = e_1, \quad Ab_2 = e_2, \quad \dots \quad Ab_n = e_n.$$

Falls B existiert, müßten diese Gleichungen eindeutig nach den Vektoren b_i auflösbar sein. Damit haben wir schon einen möglichen Lösungsweg gefunden: Wir müssen jede der n Gleichen-gungen $Ab_i = e_i$ mit dem GAUSS-Verfahren nach b_i auflösen. Das hört sich allerdings nach viel Arbeit an. Überlegen wir uns doch einmal, was im Einzelnen dafür zu tun wäre. Zunächst würden wir die Gleichung $Ab_1 = e_1$ in das zugehörige GAUSS-Schema verwandeln. Bei einer

inhomogenen Gleichung wie dieser, besteht das Schema aus der Matrix A , erweitert um die Inhomogenität e_1 :

$$\left| \begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & 1 \\ a_{21} & a_{22} & \cdots & a_{2n} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & 0 \end{array} \right|$$

Jetzt verwenden wir das GAUSS-Verfahren. Da b_1 eindeutig ist, muß es bis zur vollbesetzten Diagonalform durchführbar sein. Wir können sogar annehmen, daß die Diagonale nur die 1 als Einträge aufweist. Aus der rechten Seite e_1 wird dabei der gesuchte Vektor $b_1 = [b_{11}, b_{21}, \dots, b_{n1}]^t$:

$$\left| \begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & 1 \\ a_{21} & a_{22} & \cdots & a_{2n} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & 0 \end{array} \right| \xrightarrow[\text{Verf.}]{\text{GAUSS}} \left| \begin{array}{cccc|c} 1 & 0 & \cdots & 0 & b_{11} \\ 0 & 1 & \cdots & 0 & b_{21} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & b_{n1} \end{array} \right|$$

Dasselbe machen wir zur Bestimmung von b_2 aus $Ab_2 = e_2$, von b_3 usw., bis wir schließlich b_n aus $Ab_n = e_n$ bestimmt haben:

$$\begin{array}{c} \left| \begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & 1 \\ a_{21} & a_{22} & \cdots & a_{2n} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & 0 \end{array} \right| \quad \left| \begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & 0 \\ a_{21} & a_{22} & \cdots & a_{2n} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & 0 \end{array} \right| \quad \left| \begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & 0 \\ a_{21} & a_{22} & \cdots & a_{2n} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & 1 \end{array} \right| \\ \downarrow \qquad \downarrow \qquad \downarrow \\ \left| \begin{array}{cccc|c} 1 & 0 & \cdots & 0 & b_{11} \\ 0 & 1 & \cdots & 0 & b_{21} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & b_{n1} \end{array} \right| \quad \left| \begin{array}{cccc|c} 1 & 0 & \cdots & 0 & b_{12} \\ 0 & 1 & \cdots & 0 & b_{22} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & b_{n2} \end{array} \right| \quad \cdots \quad \left| \begin{array}{cccc|c} 1 & 0 & \cdots & 0 & b_{1n} \\ 0 & 1 & \cdots & 0 & b_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & b_{nn} \end{array} \right| \end{array}$$

Für jeden der gesuchten Vektoren b_1 bis b_n ist ein GAUSS-Verfahren erforderlich. Die entscheidende Beobachtung ist aber, daß es sich dabei jedes mal um *dasselbe* Verfahren handelt, d. h., daß dabei jeweils dieselbe Abfolge von Zeilenumformungen durchgeführt werden. Die einzelnen Rechenschritte hängen nämlich nur von den Einträgen von A , nicht aber von den rechten Seiten e_1, \dots, e_n ab. Daher können wir die n GAUSS-Verfahren zu einem einzigen *erweiterten GAUSS-Verfahren* bündeln, indem wir die rechten Seiten e_1, \dots, e_n alle auf einmal berücksichtigen:

$$\left| \begin{array}{cccc|cc} a_{11} & a_{12} & \cdots & a_{1n} & 1 & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & a_{2n} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & 0 & 0 & \cdots & 1 \end{array} \right| \xrightarrow[\text{Verf.}]{\text{GAUSS}} \left| \begin{array}{cccc|cc} 1 & 0 & \cdots & 0 & b_{11} & b_{12} & \cdots & b_{1n} \\ 0 & 1 & \cdots & 0 & b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & b_{n1} & b_{n2} & \cdots & b_{nn} \end{array} \right|$$

In der praktischen Umsetzung geht man wie folgt vor. Anstatt die quadratische Matrix A daraufhin zu untersuchen, ob sie injektiv ist, was nach Lemma 6.8.2 die Bestimmung des Kerns von

A bedeutet würde, oder, was auf dasselbe hinausläuft, die Spaltenvektoren von A auf lineare Unabhängigkeit zu prüfen, versucht man einfach das erweiterte GAUSS-Verfahren durchzuführen. Gelingt das, dann wissen wir erstens, daß A invertierbar ist und wir kennen zweitens bereits A^{-1} . Gelingt es nicht, dann ist A nicht invertierbar. Für diese Erkenntnis haben wir dann nur etwa so viel Arbeit aufgewandt, wie nötig gewesen wäre, um zu zeigen, daß $Ax = \mathbf{0}$ eine Lösung $x \neq \mathbf{0}$ hat, oder daß die Spaltenvektoren von A linear abhängig sind. Dafür hätten wir nämlich dasselbe GAUSS-Verfahren durchführen müssen.

6.9.4 Beispiel Wir versuchen, die inverse Matrix der folgenden Matrix zu bestimmen:

$$A := \begin{bmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \\ 3 & 4 & 1 & 2 \\ 2 & 3 & 4 & 1 \end{bmatrix}.$$

I	1	2	3	4	1	0	0	0	
II	4	1	2	3	0	1	0	0	II - 4 I
III	3	4	1	2	0	0	1	0	III - 3 I
IV	2	3	4	1	0	0	0	1	IV - 2 I
I	1	2	3	4	1	0	0	0	I . 7
II	0	-7	-10	-13	-4	1	0	0	II . (-1)
III	0	-2	-8	-10	-3	0	1	0	III . (-7)
IV	0	-1	-2	-7	-2	0	0	1	IV . (-7)
I	7	14	21	28	7	0	0	0	I - 2 II
II	0	7	10	13	4	-1	0	0	
III	0	14	56	70	21	0	-7	0	III - 2 II
IV	0	7	14	49	14	0	0	-7	IV - II
I	7	0	1	2	-1	2	0	0	36 I - III
II	0	7	10	13	4	-1	0	0	18 II - 5 III
III	0	0	36	44	13	2	-7	0	
IV	0	0	4	36	10	1	0	-7	9 IV - III
I	252	0	0	28	-49	70	7	0	(10 I - IV)/7
II	0	126	0	14	7	-28	35	0	(20 II - IV)/7
III	0	0	36	44	13	2	-7	0	(70 III - 11 IV)/7
IV	0	0	0	280	77	7	7	-63	9 IV/7
I	360	0	0	0	-81	99	9	9	
II	0	360	0	0	9	-81	99	9	
III	0	0	360	0	9	9	-81	99	
IV	0	0	0	360	99	9	9	-81	

Das zeigt bereits die Existenz der Matrix A^{-1} und daß sie durch

$$A^{-1} = \frac{1}{360} \begin{bmatrix} -81 & 99 & 9 & 9 \\ 9 & -81 & 99 & 9 \\ 9 & 9 & -81 & 99 \\ 99 & 9 & 9 & -81 \end{bmatrix} = \frac{1}{40} \begin{bmatrix} -9 & 11 & 1 & 1 \\ 1 & -9 & 11 & 1 \\ 1 & 1 & -9 & 11 \\ 11 & 1 & 1 & -9 \end{bmatrix}$$

gegeben ist. Es gibt eigentlich nie einen Grund, das Verfahren mit einer fehlerhaften inversen Matrix zu beenden. Wir können uns immer davon überzeugen, daß wir die richtige gefunden haben:

$$\begin{aligned} & \begin{bmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \\ 3 & 4 & 1 & 2 \\ 2 & 3 & 4 & 1 \end{bmatrix} \cdot \frac{1}{40} \begin{bmatrix} -9 & 11 & 1 & 1 \\ 1 & -9 & 11 & 1 \\ 1 & 1 & -9 & 11 \\ 11 & 1 & 1 & -9 \end{bmatrix} \\ &= \frac{1}{40} \begin{bmatrix} -9 + 5 + 44 & 11 - 18 + 7 & 5 + 22 - 27 & 3 + 33 - 36 \\ -36 + 3 + 33 & 44 - 9 + 5 & 7 + 11 - 18 & 5 + 22 - 27 \\ -27 + 5 + 22 & 33 - 36 + 3 & 5 + 44 - 9 & 7 + 11 - 18 \\ -18 + 7 + 11 & 22 - 27 + 5 & 3 + 33 - 36 & 5 + 44 - 9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \mathbf{1}. \end{aligned}$$

Laut Satz 6.9.2 haben wir tatsächlich die Inverse von \mathbf{A} berechnet.

Versuchen wir es mit der Matrix

$$\mathbf{B} := \begin{bmatrix} 2 & 1 & 2 & 3 \\ -6 & -5 & 4 & 11 \\ 3 & 3 & 1 & -1 \\ -1 & 2 & 3 & 2 \end{bmatrix}.$$

I	2	1	2	3	1	0	0	0	
II	-6	-5	4	11	0	1	0	0	II + 3 I
III	3	3	1	-1	0	0	1	0	2 III - 3 I
IV	-1	2	3	2	0	0	0	1	2 IV + I
I	2	1	2	3	1	0	0	0	2 I + II
II	0	-2	10	20	3	1	0	0	
III	0	3	-4	-11	-3	0	2	0	2 III + 3 II
IV	0	5	8	7	1	0	0	2	2 IV + 5 II
I	4	0	14	26	5	1	0	0	
II	0	-2	10	20	3	1	0	0	
III	0	0	22	38	3	3	4	0	
IV	0	0	66	114	17	5	0	4	IV - 3 III
I	4	0	14	26	5	1	0	0	
II	0	-2	10	20	3	1	0	0	
III	0	0	22	38	3	3	4	0	
IV	0	0	0	0	8	-4	-12	4	

Die letzte Zeile zeigt, daß \mathbf{B}^{-1} nicht existiert.

6.9.5 A Prüfen Sie, ob die folgenden Matrizen invertierbar sind und bestimmen Sie gegebenenfalls die Inversen.

$$\text{i) } \mathbf{A} := \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix} \quad \text{ii) } \mathbf{B} := \begin{bmatrix} 1 & 0 & 3 & 4 \\ 5 & 6 & 0 & 8 \\ 9 & 10 & 11 & 0 \\ 0 & 14 & 15 & 16 \end{bmatrix}$$

$$\begin{array}{ll}
 \text{iii)} & C(x) := \begin{bmatrix} 1 & x & x^2 & x^3 \\ x & 1 & x & x^2 \\ x^2 & x & 1 & x \\ x^3 & x^2 & x & 1 \end{bmatrix} \\
 \text{iv)} & D := \begin{bmatrix} a & b \\ c & d \end{bmatrix} \\
 \text{v)} & E := \begin{bmatrix} 2 & -2 & 1 & 0 \\ 2i & i & -2i & 0 \\ i & 2i & 2i & i \\ 0 & 0 & 2-i & 0 \end{bmatrix} \\
 \text{vi)} & F := \begin{bmatrix} 1 & 0 & 3 & 2 & 2 \\ 0 & 3 & 3 & 1 & 1 \\ 2 & 1 & 3 & 1 & 0 \\ 0 & 1 & 1 & 2 & 2 \\ 4 & 0 & 0 & 4 & 2 \end{bmatrix}
 \end{array}$$

6.10 Die adjungierte Matrix

6.10.1 Definition Für eine $m \times n$ -Matrix A mit den Einträgen $a_{ij} \in \mathbb{K}$ ($\mathbb{K} = \mathbb{R}$, oder $\mathbb{K} = \mathbb{C}$) an der Stelle (i, j) ist ihre adjungierte Matrix A^* die $n \times m$ -Matrix, mit den Einträgen $\overline{a_{ji}}$ an der Stelle (i, j) . A^* entsteht also aus A , indem man alle Einträge durch ihre konjugiert komplexen Werte ersetzt (natürlich nur für $\mathbb{K} = \mathbb{C}$) und diese Matrix anschließend transponiert:

$$A^* = \overline{A^t} \quad (6.26)$$

A^* wird auch die Adjungierte von A genannt.

Ausführlich aufgeschrieben bedeutet das

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & \color{blue}{a_{1j}} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & \color{blue}{a_{2j}} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ \color{green}{a_{i1}} & \color{green}{a_{i2}} & \color{green}{a_{i3}} & \cdots & \color{red}{a_{ij}} & \cdots & \color{green}{a_{in}} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & \color{blue}{a_{mj}} & \cdots & a_{mn} \end{bmatrix}^* = \begin{bmatrix} \overline{a_{11}} & \overline{a_{21}} & \cdots & \overline{a_{i1}} & \cdots & \overline{a_{m1}} \\ \overline{a_{12}} & \overline{a_{22}} & \cdots & \overline{a_{i2}} & \cdots & \overline{a_{m2}} \\ \overline{a_{13}} & \overline{a_{23}} & \cdots & \overline{a_{i3}} & \cdots & \overline{a_{m3}} \\ \vdots & \vdots & & \vdots & & \vdots \\ \overline{a_{1j}} & \overline{a_{2j}} & \cdots & \color{red}{\overline{a_{ij}}} & \cdots & \overline{a_{mj}} \\ \vdots & \vdots & & \vdots & & \vdots \\ \overline{a_{1n}} & \overline{a_{2n}} & \cdots & \overline{a_{in}} & \cdots & \overline{a_{mn}} \end{bmatrix}.$$

In der knapperen Formulierung mit Hilfe der Spaltenvektoren \mathbf{a}_i :

$$A^* = [\mathbf{a}_1, \dots, \mathbf{a}_n]^* = \begin{bmatrix} \mathbf{a}_1^* \\ \vdots \\ \mathbf{a}_n^* \end{bmatrix},$$

mit den Zeilenvektoren $\mathbf{a}_j^* := [\overline{a_{1j}}, \overline{a_{2j}}, \dots, \overline{a_{mj}}]$.

Einen solchen Zeilenvektor sehen wir als eine $1 \times m$ -Matrix an.

Für $\mathbf{a} := [a_1, \dots, a_m]^t \in \mathbb{K}^m$ und $\mathbf{b} := [b_1, \dots, b_m]^t \in \mathbb{K}^m$ gilt

$$\langle \mathbf{a} | \mathbf{b} \rangle = \sum_{k=1}^m \overline{a_j} b_j = [\overline{a_1}, \overline{a_2}, \dots, \overline{a_m}] \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} = \mathbf{a}^* \mathbf{b}. \quad (6.27)$$

Das läßt sich verallgemeinern: Für eine $m \times n$ -Matrix A und eine $m \times k$ -Matrix B gilt

$$\begin{aligned} A^*B &= \begin{bmatrix} a_1^* \\ a_2^* \\ \vdots \\ a_n^* \end{bmatrix} [b_1, b_2, \dots, b_k] = \begin{bmatrix} a_1^*b_1 & a_1^*b_2 & \cdots & a_1^*b_k \\ a_2^*b_1 & a_2^*b_2 & \cdots & a_2^*b_k \\ \vdots & \vdots & & \vdots \\ a_n^*b_1 & a_n^*b_2 & \cdots & a_n^*b_k \end{bmatrix} \\ &= \begin{bmatrix} \langle a_1 | b_1 \rangle & \langle a_1 | b_2 \rangle & \cdots & \langle a_1 | b_k \rangle \\ \langle a_2 | b_1 \rangle & \langle a_2 | b_2 \rangle & \cdots & \langle a_2 | b_k \rangle \\ \vdots & \vdots & & \vdots \\ \langle a_n | b_1 \rangle & \langle a_n | b_2 \rangle & \cdots & \langle a_n | b_k \rangle \end{bmatrix}. \end{aligned} \quad (6.28)$$

6.10.2 Satz Die zu einer $m \times n$ -Matrix A adjungierte Matrix A^* erfüllt die Gleichung

$$\langle x | Ay \rangle = \langle A^*x | y \rangle \quad (6.29)$$

f. a. $x \in \mathbb{K}^m$, $y \in \mathbb{K}^n$. A^* wird durch diese Beziehung eindeutig festgelegt.

Für eine $n \times k$ -Matrix B gilt

$$(AB)^* = B^*A^*. \quad (6.30)$$

Beweis. Für diese Gleichung gibt es mehrere Beweismöglichkeiten. Die kürzeste, möglicherweise aber nicht die anschaulichste, ist die folgende. Wir verwenden $x := [x_1, \dots, x_m]^t$, $y := [y_1, \dots, y_n]^t$ und $(Ay)_i = \sum_{j=1}^n a_{ij}y_j$, sowie $(A^*x)_j = \sum_{i=1}^m \overline{a_{ij}}x_i (= \sum_{i=1}^m (A^*)_{ji}x_i)$:

$$\begin{aligned} \langle x | Ay \rangle &= \sum_{i=1}^m \overline{x_i}(Ay)_i = \sum_{i=1}^m \overline{x_i} \sum_{j=1}^n a_{ij}y_j = \sum_{j=1}^n \sum_{i=1}^m \overline{x_i} a_{ij} y_j \\ &= \sum_{j=1}^n \overline{\sum_{i=1}^m a_{ij}x_i} y_j = \sum_{j=1}^n \overline{(A^*x)_j} y_j = \langle A^*x | y \rangle. \end{aligned}$$

Man beachte, daß das erste Skalarprodukt auf \mathbb{K}^m und das letzte auf \mathbb{K}^n wirkt.

Zu Gleichung (6.30). Sei $z := [z_1, \dots, z_k]^t \in \mathbb{K}^k$. Wir wenden (6.29) nacheinander auf A und dann auf B und ein Mal auf AB als Ganzes an:

$$\langle x | ABz \rangle = \langle A^*x | Bz \rangle = \langle B^*A^*x | z \rangle = \langle (AB)^*x | z \rangle.$$

Das gilt für alle $x \in \mathbb{K}^m$ und $z \in \mathbb{K}^k$. Aus Übung 5.5.12 wissen wir, daß daraus $(AB)^*x = B^*A^*x$ für alle $x \in \mathbb{K}^m$ folgt. Das ist äquivalent zu (6.30). Übrigens ergibt sich so auch die Eindeutigkeit von A^* : Aus $\langle A'x | y \rangle = \langle x | Ay \rangle$ für alle $x \in \mathbb{K}^m$, $y \in \mathbb{K}^n$ folgt $\langle A'x | y \rangle = \langle A^*x | y \rangle$ und daraus $A'x = A^*x$ für alle $x \in \mathbb{K}^m$, also $A' = A^*$.

Eine anschaulichere Beweismöglichkeit für Gleichung (6.29) besteht darin, zuerst (6.30) und zwar zunächst für ein Matrixprodukt der Form C^*B zu zeigen (dafür muß C eine $n \times m$ -Matrix sein):

$$(C^*B)^* \stackrel{(6.28)}{=} \begin{bmatrix} \langle c_1 | b_1 \rangle & \cdots & \langle c_1 | b_k \rangle \\ \langle c_2 | b_1 \rangle & \cdots & \langle c_2 | b_k \rangle \\ \vdots & & \vdots \\ \langle c_n | b_1 \rangle & \cdots & \langle c_n | b_k \rangle \end{bmatrix}^* = \begin{bmatrix} \overline{\langle c_1 | b_1 \rangle} & \overline{\langle c_2 | b_1 \rangle} & \cdots & \overline{\langle c_n | b_1 \rangle} \\ \vdots & \vdots & & \vdots \\ \overline{\langle c_1 | b_k \rangle} & \overline{\langle c_2 | b_k \rangle} & \cdots & \overline{\langle c_n | b_k \rangle} \end{bmatrix}$$

$$= \begin{bmatrix} \langle \mathbf{b}_1 | \mathbf{c}_1 \rangle & \langle \mathbf{b}_1 | \mathbf{c}_2 \rangle & \cdots & \langle \mathbf{b}_1 | \mathbf{c}_n \rangle \\ \vdots & \vdots & & \vdots \\ \langle \mathbf{b}_k | \mathbf{c}_1 \rangle & \langle \mathbf{b}_k | \mathbf{c}_2 \rangle & \cdots & \langle \mathbf{b}_k | \mathbf{c}_n \rangle \end{bmatrix} = \mathbf{B}^* \mathbf{C}.$$

Jetzt ist (6.29) nur noch eine Frage des richtigen Einsetzens ($\mathbf{C} := \mathbf{A}^*$):

$$(\mathbf{AB})^* = ((\mathbf{A}^*)^* \mathbf{B})^* = \mathbf{B}^* \mathbf{A}^*.$$

Durch Anwendung dieser Rechenregel auf den Spezialfall $(\mathbf{A}^* \mathbf{x})^*$ erhalten wir (6.29)

$$\langle \mathbf{x} | \mathbf{A} \mathbf{y} \rangle = \mathbf{x}^* (\mathbf{A} \mathbf{y}) = (\mathbf{x}^* \mathbf{A}) \mathbf{y} = (\mathbf{A}^* \mathbf{x})^* \mathbf{y} = \langle \mathbf{A}^* \mathbf{x} | \mathbf{y} \rangle. \quad \square$$

6.10.3 Korollar Eine Matrix $\mathbf{U} \in M_n(\mathbb{K})$ vermittelt genau dann eine längentreue lineare Abbildung, also eine Abbildung, die für alle $\mathbf{x} \in \mathbb{K}^n$ die Gleichung $\|\mathbf{Ux}\| = \|\mathbf{x}\|$ erfüllt, wenn sie winkeltreu ist, d. h., wenn sie $\langle \mathbf{Ux} | \mathbf{Uy} \rangle = \langle \mathbf{x} | \mathbf{y} \rangle$ für alle $\mathbf{x}, \mathbf{y} \in \mathbb{K}^n$ erfüllt. Das wiederum ist äquivalent zu $\mathbf{U}^* = \mathbf{U}^{-1}$.

Beweis. Wir zeigen hier nur den Fall $\mathbb{K} = \mathbb{C}$. Für $\mathbb{K} = \mathbb{R}$ verlaufen die Überlegungen analog. Die Längentreue ergibt mit der Polarisationsgleichung (5.81)

$$\begin{aligned} \langle \mathbf{Ux} | \mathbf{Uy} \rangle &= \frac{1}{4} [\|\mathbf{Ux} + \mathbf{Uy}\|^2 - \|\mathbf{Ux} - \mathbf{Uy}\|^2 - i(\|\mathbf{Ux} + i\mathbf{Uy}\|^2 - \|\mathbf{Ux} - i\mathbf{Uy}\|^2)] \\ &= \frac{1}{4} [\|\mathbf{U(x+y)}\|^2 - \|\mathbf{U(x-y)}\|^2 - i(\|\mathbf{U(x+iy)}\|^2 - \|\mathbf{U(x-iy)}\|^2)] \\ &= \frac{1}{4} [\|\mathbf{x+y}\|^2 - \|\mathbf{x-y}\|^2 - i(\|\mathbf{x+iy}\|^2 - \|\mathbf{x-iy}\|^2)] = \langle \mathbf{x} | \mathbf{y} \rangle \end{aligned}$$

für alle $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$. Daher ist \mathbf{U} winkeltreu. Die Umkehrung folgt natürlich aus $\|\mathbf{Ux}\|^2 = \langle \mathbf{Ux} | \mathbf{Ux} \rangle = \langle \mathbf{x} | \mathbf{x} \rangle = \|\mathbf{x}\|^2$. Mit unseren Überlegungen zur Adjungierten können wir aus der Winkeltreue

$$\langle \mathbf{x} | \mathbf{y} \rangle = \langle \mathbf{Ux} | \mathbf{Uy} \rangle = \langle \mathbf{U}^* \mathbf{Ux} | \mathbf{y} \rangle$$

die Gleichung $\mathbf{U}^* \mathbf{U} = \mathbf{1}$ folgern. Aus Satz 6.9.2 folgt, daß \mathbf{U} invertierbar und daß \mathbf{U}^{-1} durch \mathbf{U}^* gegeben ist. Die Umkehrung sieht man genauso einfach: $\langle \mathbf{x} | \mathbf{y} \rangle = \langle \mathbf{U}^{-1} \mathbf{Ux} | \mathbf{y} \rangle = \langle \mathbf{U}^* \mathbf{Ux} | \mathbf{y} \rangle = \langle \mathbf{Ux} | \mathbf{Uy} \rangle$. \square

6.10.4 Definition Eine lineare Abbildung $\mathbf{U} \in M_n(\mathbb{K})$ mit der Eigenschaft $\mathbf{U}^* \mathbf{U} = \mathbf{1}$ heißt unitär. Im Fall $\mathbb{K} = \mathbb{R}$ wird sie auch oft orthogonal genannt.

6.10.5 A Bestimmen Sie die adjungierte Matrix von

$$\mathbf{A} := \begin{bmatrix} 2i & (2-4i)e^{i\varphi} & \frac{2i}{1+2i} \\ 0 & 42i & -3e^{-4i} \\ \frac{-2i}{1-2i} & 5+2i & 9 \end{bmatrix}$$

und die Inverse von

$$\mathbf{U} := \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}.$$

6.10.6 Die duale Basis Um einen Vektor $\mathbf{x} \in V$ aus einem Vektorraum über \mathbb{K} nach einer Basis $\mathcal{B} := \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ zu entwickeln, müssen die Koeffizienten x_1, x_2, \dots, x_n in

$$\mathbf{x} = \sum_{k=1}^n x_k \mathbf{b}_k$$

bestimmt werden. Die Lösung fassen wir übersichtlich in einem Koeffizientenvektor $\mathbf{x}_{\mathcal{B}} := [x_1, \dots, x_n]_{\mathcal{B}}^t \in \mathbb{K}^n$ zusammen. Der Index \mathcal{B} bei $[x_1, \dots, x_n]_{\mathcal{B}}^t$ soll deutlich machen, auf welche Basis sich die Koeffizienten x_1, \dots, x_n beziehen. Das wird besonders dann wichtig, wenn mehrere Basen im Spiel sind, wie bei Basiswechseln (vergl. Abschnitt 6.11), da einem Vektor $[x_1, \dots, x_n]^t$ sonst nicht anzusehen ist, zu welcher Basis er gehört. Wir vereinbaren den Index generell wegzulassen, wenn es sich bei \mathcal{B} um die kanonische Basis (6.9) \mathcal{E} in \mathbb{K}^n handelt. Auch dann, wenn in einem Problemkreis nur eine einzige Basis relevant ist, werden wir den Index normalerweise ebenfalls einsparen.

Wir haben im Abschnitt 6.5.14 gesehen, wie leicht es ist, in einem Hilbertraum V einen Vektor \mathbf{x} nach einer ONB $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ zu entwickeln. Die Koeffizienten x_j sind nämlich einfach durch die Skalarprodukte $\langle \mathbf{b}_j | \mathbf{x} \rangle$ gegeben:

$$\langle \mathbf{b}_j | \mathbf{x} \rangle = \left\langle \mathbf{b}_j \mid \sum_{k=1}^n x_k \mathbf{b}_k \right\rangle = \sum_{k=1}^n x_k \langle \mathbf{b}_j | \mathbf{b}_k \rangle = \sum_{k=1}^n \delta_{jk} x_k = x_j.$$

Wünschenswert wäre es, wenn uns eine solche Rechnung auch für eine beliebige Basis eines Hilbertraums möglich wäre. Hier kommt die sogenannte *duale Basis* $\mathcal{B}' := \{\mathbf{b}^1, \dots, \mathbf{b}^n\}$ von \mathcal{B} ins Spiel. Ihre Vektoren haben die maßgeschneiderte Eigenschaft

$$\langle \mathbf{b}^j | \mathbf{b}_k \rangle = \delta_{jk}. \quad (6.31)$$

Jetzt erhalten wir die Entwicklungskoeffizienten x_j von \mathbf{x} einfach durch $\langle \mathbf{b}^j | \mathbf{x} \rangle$:

$$\langle \mathbf{b}^j | \mathbf{x} \rangle = \left\langle \mathbf{b}^j \mid \sum_{k=1}^n x_k \mathbf{b}_k \right\rangle = \sum_{k=1}^n x_k \langle \mathbf{b}^j | \mathbf{b}_k \rangle = \sum_{k=1}^n \delta_{jk} x_k = x_j.$$

Es stellt sich nur noch die Frage, wie man die duale Basis aus \mathcal{B} bestimmt. Eigentlich wissen wir das schon, wir müssen die Aufgabe nur von der richtigen Warte aus betrachten. Dabei hilft uns (6.28). In Verbindung mit (6.31) haben wir nämlich für die Matrizen $B := [\mathbf{b}_1, \dots, \mathbf{b}_n]$ und $B' := [\mathbf{b}^1, \dots, \mathbf{b}^n]$

$$B'^* B = \begin{bmatrix} \langle \mathbf{b}^1 | \mathbf{b}_1 \rangle & \langle \mathbf{b}^1 | \mathbf{b}_2 \rangle & \cdots & \langle \mathbf{b}^1 | \mathbf{b}_n \rangle \\ \langle \mathbf{b}^2 | \mathbf{b}_1 \rangle & \langle \mathbf{b}^2 | \mathbf{b}_2 \rangle & \cdots & \langle \mathbf{b}^2 | \mathbf{b}_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{b}^n | \mathbf{b}_1 \rangle & \langle \mathbf{b}^n | \mathbf{b}_2 \rangle & \cdots & \langle \mathbf{b}^n | \mathbf{b}_n \rangle \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbb{1}.$$

Das heißt: $B'^* = B^{-1}$, oder $B' = (B^{-1})^* = (B^*)^{-1}$. Die Inverse von B^* hat die duale Basis als Spaltenvektoren. Es ist klar, daß wir im Allgemeinen nicht billiger davon kommen können, als die Inverse der Matrix B zu berechnen. Denn bei Lichte besehen ist die duale Basis nur eine andere Lösung für die Aufgabe, für jedes $\mathbf{x} \in V$ die Gleichung $B\mathbf{x}_{\mathcal{B}} = \mathbf{x}$ nach $\mathbf{x}_{\mathcal{B}}$ aufzulösen. Und das läuft nun mal auf die Berechnung von B^{-1} hinaus.

6.10.7 Beispiel Wir wollen die duale Basis von

$$\mathcal{B} := \left\{ \begin{bmatrix} 2 \\ 2i \\ i \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ i \\ 2i \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ -2i \\ 2i \\ 2-i \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ i \\ 0 \end{bmatrix} \right\}$$

finden. In Übung 6.9.5 haben wir die Inverse der zugehörigen Matrix B bestimmt:

$$B^{-1} = \frac{1}{30} \begin{bmatrix} 5 & -10i & 0 & 3(2+i) \\ -10 & -10i & 0 & 6(2+i) \\ 0 & 0 & 0 & 6(2+i) \\ 15 & 30i & -30i & -27(2+i) \end{bmatrix}.$$

Die duale Basis ist demnach

$$\mathcal{B}' = \left\{ \frac{1}{30} \begin{bmatrix} 5 \\ 10i \\ 0 \\ 3(2-i) \end{bmatrix}, \frac{1}{15} \begin{bmatrix} -5 \\ 5i \\ 0 \\ 3(2-i) \end{bmatrix}, \frac{1}{5} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 2-i \end{bmatrix}, \frac{1}{10} \begin{bmatrix} 5 \\ -10i \\ 10i \\ -9(2-i) \end{bmatrix} \right\}.$$

6.10.8 Beispiel Wir bestimmen die duale Basis $\mathcal{B}' := \{\mathbf{b}^1, \mathbf{b}^2\}$ einer beliebigen Basis $\mathcal{B} := \{\mathbf{b}_1, \mathbf{b}_2\}$ des \mathbb{R}^2 . Hier ist die Situation besonders übersichtlich. Das liegt daran, daß es (bis auf Streckung) nur einen Vektor gibt, der auf einem gegebenen senkrecht steht. Diesen erhalten wir einfach durch eine Drehung um $\frac{\pi}{2}$, also durch Anwendung der Matrix

$$J := \begin{bmatrix} \cos(\frac{\pi}{2}) & -\sin(\frac{\pi}{2}) \\ \sin(\frac{\pi}{2}) & \cos(\frac{\pi}{2}) \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}. \quad (6.32)$$

Wie üblich sei $\mathbf{b}_1 = [b_{11}, b_{21}]^t$, $\mathbf{b}_2 = [b_{12}, b_{22}]^t$ und $B := [\mathbf{b}_1, \mathbf{b}_2]$.

$\mathbf{b}^1 \perp \mathbf{b}_2$ bedeutet $\mathbf{b}^1 \sim J\mathbf{b}_2 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} b_{12} \\ b_{22} \end{bmatrix} = \begin{bmatrix} -b_{22} \\ b_{12} \end{bmatrix}$. Für die Normierung berechnen wir $\langle \mathbf{b}^1 | \mathbf{b}_1 \rangle \sim \left\langle \begin{bmatrix} -b_{22} \\ b_{12} \end{bmatrix} \mid \begin{bmatrix} b_{11} \\ b_{21} \end{bmatrix} \right\rangle = -(b_{11}b_{22} - b_{12}b_{21}) = -\det(B)$. Dadurch ist \mathbf{b}^1 bestimmt: $\mathbf{b}^1 = \frac{-1}{\det(B)} J\mathbf{b}_2$. Genauso verfahren wir mit $\mathbf{b}^2 \sim J\mathbf{b}_1$ und erhalten $\mathbf{b}^2 = \frac{1}{\det(B)} J\mathbf{b}_1$. Zusammengefaßt:

$$\mathcal{B}' = \left\{ -\frac{1}{\det(B)} J\mathbf{b}_2, \frac{1}{\det(B)} J\mathbf{b}_1 \right\}, \quad B' = \frac{1}{\det(B)} \begin{bmatrix} b_{22} & -b_{21} \\ -b_{12} & b_{11} \end{bmatrix}. \quad (6.33)$$

Man rechnet leicht $B' = \frac{1}{\det(B)} JBJ^*$ nach und kann sich durch

$$B^* B' = \frac{1}{\det(B)} \begin{bmatrix} \mathbf{b}_1^* \\ \mathbf{b}_2^* \end{bmatrix} [J\mathbf{b}_1, J\mathbf{b}_2] J^* = \begin{bmatrix} \mathbf{b}_1^* \\ \mathbf{b}_2^* \end{bmatrix} [\mathbf{b}^2, -\mathbf{b}^1] J^* = \begin{bmatrix} \langle \mathbf{b}_1 | \mathbf{b}^2 \rangle & -\langle \mathbf{b}_1 | \mathbf{b}^1 \rangle \\ \langle \mathbf{b}_2 | \mathbf{b}^2 \rangle & -\langle \mathbf{b}_2 | \mathbf{b}^1 \rangle \end{bmatrix} J^* = JJ^* = \mathbb{1}$$

noch einmal davon überzeugen, daß B' die Inverse von B^* ist. Als kleines Nebenprodukt unserer Rechnung gewinnen wir eine (nicht sonderlich wichtige) Formel für die Inverse von B:

$$B^{-1} = \frac{1}{\det(B)} JB^* J^*. \quad (6.34)$$

6.10.9 Beispiel (Umkreis)

Drei Punkte \mathbf{a} , \mathbf{b} und \mathbf{c} liegen auf einem eindeutig bestimmten Kreis κ . Dabei vereinbaren wir, eine Gerade als einen Kreis mit unendlichem Radius anzusehen, denn nur so lässt sich die Aussage in ihrer Allgemeinheit aufrecht erhalten. Unsere Aufgabe besteht darin, den Mittelpunkt \mathbf{m} und den Radius r von κ aus \mathbf{a} , \mathbf{b} und \mathbf{c} zu berechnen.

$\mathcal{B} := \{\mathbf{b}_1, \mathbf{b}_2\}$, mit $\mathbf{b}_1 := \mathbf{c} - \mathbf{b}$ und $\mathbf{b}_2 := \mathbf{a} - \mathbf{b}$ ist eine dem Problem angepaßte Basis. $\mathcal{B}' := \{\mathbf{b}^1, \mathbf{b}^2\} = \left\{-\frac{1}{A} J\mathbf{b}_2, \frac{1}{A} J\mathbf{b}_1\right\}$ ist die zugehörige duale Basis (vergl. (6.33)) mit $A := \det(\mathbf{c} - \mathbf{b}, \mathbf{a} - \mathbf{b})$.

\mathbf{m} ist dadurch bestimmt, daß der Abstand von \mathbf{a} , \mathbf{b} und \mathbf{c} jeweils der gesuchte Radius r des Umkreises ist:

$$\|\mathbf{m} - \mathbf{a}\|^2 = \|\mathbf{m}\|^2 + \|\mathbf{a}\|^2 - 2\langle \mathbf{m} | \mathbf{a} \rangle = r^2, \quad \|\mathbf{m} - \mathbf{b}\|^2 = \|\mathbf{m}\|^2 + \|\mathbf{b}\|^2 - 2\langle \mathbf{m} | \mathbf{b} \rangle = r^2,$$

$$\|\mathbf{m} - \mathbf{c}\|^2 = \|\mathbf{m}\|^2 + \|\mathbf{c}\|^2 - 2\langle \mathbf{m} | \mathbf{c} \rangle = r^2.$$

Nach Subtraktion der ersten beiden bzw. der letzten beiden Gleichungen erhalten wir:

$$\|\mathbf{a}\|^2 - \|\mathbf{b}\|^2 = 2\langle \mathbf{m} | \mathbf{a} - \mathbf{b} \rangle = 2\langle \mathbf{m} | \mathbf{b}_2 \rangle, \quad \|\mathbf{c}\|^2 - \|\mathbf{b}\|^2 = 2\langle \mathbf{m} | \mathbf{c} - \mathbf{b} \rangle = 2\langle \mathbf{m} | \mathbf{b}_1 \rangle.$$

Damit gewinnen wir die Entwicklung $\mathbf{m} = \lambda_1 \mathbf{b}^1 + \lambda_2 \mathbf{b}^2$ von \mathbf{m} in der dualen Basis:

$$\langle \mathbf{m} | \mathbf{b}_1 \rangle = \lambda_1 \langle \mathbf{b}^1 | \mathbf{b}_1 \rangle + \lambda_2 \langle \mathbf{b}^2 | \mathbf{b}_1 \rangle = \lambda_1 = \frac{1}{2} (\|\mathbf{c}\|^2 - \|\mathbf{b}\|^2),$$

$$\langle \mathbf{m} | \mathbf{b}_2 \rangle = \lambda_1 \langle \mathbf{b}^1 | \mathbf{b}_2 \rangle + \lambda_2 \langle \mathbf{b}^2 | \mathbf{b}_2 \rangle = \lambda_2 = \frac{1}{2} (\|\mathbf{a}\|^2 - \|\mathbf{b}\|^2).$$

$$\begin{aligned} \mathbf{m} &= \frac{J}{2A} \left(-(\|\mathbf{c}\|^2 - \|\mathbf{b}\|^2)(\mathbf{a} - \mathbf{b}) + (\|\mathbf{a}\|^2 - \|\mathbf{b}\|^2)(\mathbf{c} - \mathbf{b}) \right) \\ &= J \frac{\mathbf{a}(\|\mathbf{b}\|^2 - \|\mathbf{c}\|^2) + \mathbf{b}(\|\mathbf{c}\|^2 - \|\mathbf{a}\|^2) + \mathbf{c}(\|\mathbf{a}\|^2 - \|\mathbf{b}\|^2)}{2 \det(\mathbf{c} - \mathbf{b}, \mathbf{a} - \mathbf{b})} \\ &= J \frac{\mathbf{a}(\|\mathbf{b}\|^2 - \|\mathbf{c}\|^2) + \mathbf{b}(\|\mathbf{c}\|^2 - \|\mathbf{a}\|^2) + \mathbf{c}(\|\mathbf{a}\|^2 - \|\mathbf{b}\|^2)}{2(\det(\mathbf{a}, \mathbf{b}) + \det(\mathbf{b}, \mathbf{c}) + \det(\mathbf{c}, \mathbf{a}))} \end{aligned} \quad (6.35)$$

Diese Formel zeigt sehr deutlich die vollkommene Symmetrie in den Punkten \mathbf{a} , \mathbf{b} und \mathbf{c} , die zu erwarten war, da keiner dieser Punkte vor den anderen ausgezeichnet ist.

Der Radius: Aus der Skizze ist $r = \frac{\|\mathbf{a} - \mathbf{b}\|}{2 \sin(\gamma)} = \frac{\|\mathbf{a} - \mathbf{b}\|}{2\sqrt{1 - \cos^2(\gamma)}}$ abzulesen. Nach dem Satz vom Falskreisbogen (siehe Seite 14-L) ist γ auch bei \mathbf{c} zu finden: $\cos(\gamma) = \frac{\langle \mathbf{a} - \mathbf{c} | \mathbf{b} - \mathbf{c} \rangle}{\|\mathbf{a} - \mathbf{c}\| \|\mathbf{b} - \mathbf{c}\|}$. Daher ist

$$r = \frac{\|\mathbf{a} - \mathbf{b}\|}{2\sqrt{1 - \cos^2(\gamma)}} = \frac{\|\mathbf{a} - \mathbf{b}\| \|\mathbf{a} - \mathbf{c}\| \|\mathbf{b} - \mathbf{c}\|}{2\sqrt{\|\mathbf{a} - \mathbf{c}\|^2 \|\mathbf{b} - \mathbf{c}\|^2 - \langle \mathbf{a} - \mathbf{c} | \mathbf{b} - \mathbf{c} \rangle^2}}$$

$$\begin{aligned}
&= \frac{\|\mathbf{a} - \mathbf{b}\| \|\mathbf{b} - \mathbf{c}\| \|\mathbf{c} - \mathbf{a}\|}{2\sqrt{\det(\mathbf{a} - \mathbf{c}, \mathbf{b} - \mathbf{c})^2}} = \frac{\|\mathbf{a} - \mathbf{b}\| \|\mathbf{b} - \mathbf{c}\| \|\mathbf{c} - \mathbf{a}\|}{2|\det(\mathbf{a} - \mathbf{c}, \mathbf{b} - \mathbf{c})|}, \\
r &= \frac{\|\mathbf{a} - \mathbf{b}\| \|\mathbf{b} - \mathbf{c}\| \|\mathbf{c} - \mathbf{a}\|}{2|\det(\mathbf{a}, \mathbf{b}) + \det(\mathbf{b}, \mathbf{c}) + \det(\mathbf{c}, \mathbf{a})|} = \frac{\|\mathbf{a} - \mathbf{b}\| \|\mathbf{b} - \mathbf{c}\| \|\mathbf{c} - \mathbf{a}\|}{2|\det(\mathbf{a} - \mathbf{b}, \mathbf{c} - \mathbf{b})|}. \tag{6.36}
\end{aligned}$$

Dabei haben wir verwendet, daß für Vektoren \mathbf{x} und \mathbf{y} , die den Winkel α einschließen,

$$\|\mathbf{x}\|^2 \|\mathbf{y}\|^2 - \langle \mathbf{x} | \mathbf{y} \rangle^2 = \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 (1 - \cos^2(\alpha)) = \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 \sin^2(\alpha) = \det(\mathbf{x}, \mathbf{y})^2$$

gilt. Denn $\|\mathbf{x}\| \|\mathbf{y}\| \sin(\alpha)$ ist der Flächeninhalt des von \mathbf{x} und \mathbf{y} aufgespannten Parallelogramms.

Für $\mathbf{a} := [1, 1]^t$, $\mathbf{b} := [6, -1]^t$ und $\mathbf{c} := [7, 6]^t$ ist $A = 37$, $\|\mathbf{a}\|^2 = 2$, $\|\mathbf{b}\|^2 = 37$ und $\|\mathbf{c}\|^2 = 85$. Damit erhalten wir

$$\mathbf{m} = \frac{1}{74} \left(-48 \begin{bmatrix} -1 \\ 1 \end{bmatrix} + 83 \begin{bmatrix} 1 \\ 6 \end{bmatrix} - 35 \begin{bmatrix} -6 \\ 7 \end{bmatrix} \right) = \frac{1}{74} \begin{bmatrix} 341 \\ 205 \end{bmatrix}, \quad r = \frac{\sqrt{29 \cdot 50 \cdot 61}}{74} = \frac{5\sqrt{3538}}{74}.$$

Eine kleine Zusatzüberlegung läßt uns auch eine Formel für die Höhen h_a , h_b , h_c durch \mathbf{a} , \mathbf{b} und \mathbf{c} finden: Der Flächeninhalt des Parallelogramms, zu dem man das Dreieck ergänzen kann, ist $A = \det(\mathbf{b} - \mathbf{a}, \mathbf{c} - \mathbf{a}) = \det(\mathbf{c} - \mathbf{b}, \mathbf{a} - \mathbf{b}) = \det(\mathbf{a} - \mathbf{c}, \mathbf{b} - \mathbf{c}) = h_a \|\mathbf{b} - \mathbf{c}\| = h_b \|\mathbf{c} - \mathbf{a}\| = h_c \|\mathbf{a} - \mathbf{b}\|$. Das ergibt

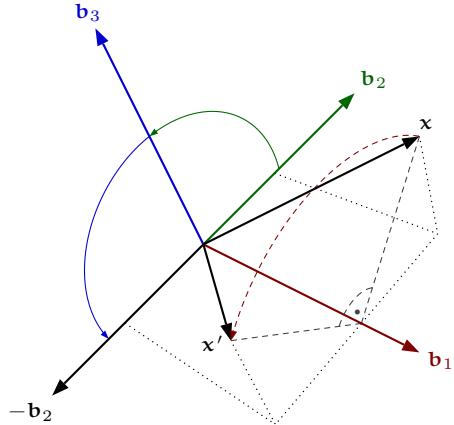
$$h_a = \frac{\det(\mathbf{b} - \mathbf{a}, \mathbf{c} - \mathbf{a})}{\|\mathbf{b} - \mathbf{c}\|}, \quad h_b = \frac{\det(\mathbf{c} - \mathbf{b}, \mathbf{a} - \mathbf{b})}{\|\mathbf{c} - \mathbf{a}\|}, \quad h_c = \frac{\det(\mathbf{a} - \mathbf{c}, \mathbf{b} - \mathbf{c})}{\|\mathbf{a} - \mathbf{b}\|}. \tag{6.37}$$

In unserem Beispiel: $h_a = \frac{37}{\sqrt{50}}$, $h_b = \frac{37}{\sqrt{61}}$ und $h_c = \frac{37}{\sqrt{29}}$.

6.11 Koordinatentransformation

Es gibt immer wieder Aufgaben, deren Lösung in der gegebenen Basisdarstellung schwer, in einer anderen, geeigneteren Basis aber leicht zu finden ist. Wünschenswert wäre dann ein Verfahren, mit dem die beteiligten Vektoren in dieser neuen Basis dargestellt werden. Dann löst man das Problem und transformiert das Ergebnis wieder zurück in die Ausgangsbasis.

Um ein konkretes Beispiel zu haben, stellen wir uns folgende Aufgabe: Wir wollen den Vektor $x := [1, 3, 1]^t$ um 90° um die Drehachse $n := [1, 2, 2]^t$ drehen. Da eine Drehung eine lineare Abbildung ist, könnten wir versuchen, die Drehmatrix zu finden. Wie bei jeder linearen Abbildung brauchen wir dafür nur die Bilder der kanonischen Basisvektoren anzugeben. Das ist zwar richtig, aber es gibt keine Garantie dafür, daß das immer leicht zu bewerkstelligen ist. Und in diesem Fall ist es tatsächlich nicht einfach. Eine dem Problem angepasste Basis $\mathcal{B} := \{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ läßt sich folgendermaßen gewinnen: Wir wählen $\mathbf{b}_1 \sim n$. \mathbf{b}_2 und \mathbf{b}_3 sind eine Ergänzung von \mathbf{b}_1 zu einer ONB von \mathbb{R}^3 . In dieser neuen Basis handelt es sich jetzt um die Aufgabe, eine 90° -Drehung um die \mathbf{b}_1 -Achse auszuführen. Dabei wird \mathbf{b}_1 in \mathbf{b}_1 , \mathbf{b}_2 in $\mathbf{b}'_2 := \mathbf{b}_3$ und \mathbf{b}_3 in $\mathbf{b}'_3 := -\mathbf{b}_2$ abgebildet. Die Darstellung von x in dieser Basis ist leicht. Gemäß Abschnitt 6.5.14 haben wir nur die Skalarprodukte $\langle \mathbf{b}_1 | x \rangle$, $\langle \mathbf{b}_2 | x \rangle$ und $\langle \mathbf{b}_3 | x \rangle$ auszurechnen, um



$$x = \langle \mathbf{b}_1 | x \rangle \mathbf{b}_1 + \langle \mathbf{b}_2 | x \rangle \mathbf{b}_2 + \langle \mathbf{b}_3 | x \rangle \mathbf{b}_3$$

und daher

$$x' = \langle \mathbf{b}_1 | x \rangle \mathbf{b}'_1 + \langle \mathbf{b}_2 | x \rangle \mathbf{b}'_2 + \langle \mathbf{b}_3 | x \rangle \mathbf{b}'_3 = \langle \mathbf{b}_1 | x \rangle \mathbf{b}_1 + \langle \mathbf{b}_2 | x \rangle \mathbf{b}_3 - \langle \mathbf{b}_3 | x \rangle \mathbf{b}_2$$

zu erhalten. In unserem Beispiel wählen wir

$$\mathbf{b}_1 := \frac{1}{3} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \quad \mathbf{b}_2 := \frac{1}{3} \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}, \quad \mathbf{b}_3 := \frac{1}{3} \begin{bmatrix} -2 \\ 2 \\ -1 \end{bmatrix}.$$

Dann ist $\tilde{x}_1 := \langle x | \mathbf{b}_1 \rangle = 3$, $\tilde{x}_2 := \langle x | \mathbf{b}_2 \rangle = 1$ und $\tilde{x}_3 := \langle x | \mathbf{b}_3 \rangle = 1$. Das heißt $x = 3\mathbf{b}_1 + \mathbf{b}_2 + \mathbf{b}_3$ und somit

$$x' = 3\mathbf{b}_1 + \mathbf{b}_3 - \mathbf{b}_2 = \frac{1}{3} \begin{bmatrix} 3 \\ 6 \\ 6 \end{bmatrix} + \frac{1}{3} \begin{bmatrix} -2 \\ 2 \\ -1 \end{bmatrix} + \frac{1}{3} \begin{bmatrix} -2 \\ 2 \\ -1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} -1 \\ 7 \\ 7 \end{bmatrix}.$$

Übrigens darf das Ergebnis x' der Drehung natürlich nicht von der Realisierung der Basisvektoren \mathbf{b}_1 , \mathbf{b}_2 und \mathbf{b}_3 abhängen. Die Wahl von \mathbf{b}_2 ist keineswegs eindeutig. Es muß nur ein Vektor

gesucht werden, der senkrecht auf \mathbf{b}_1 steht. Davon gibt es viele, etwa $\mathbf{b}_2 := \frac{1}{\sqrt{5}} [2, 0, -1]^t$. $\mathbf{b}_3 := \frac{1}{3\sqrt{5}} [-2, 5, -4]$ ist dann über das Kreuzprodukt festgelegt. Es ist eine gute Übung, die Rechnungen mit diesen Vektoren zu wiederholen.

Jetzt wollen wir unser Beispiel ausbauen und beliebige Drehwinkel α zulassen. \mathbf{b}_1 geht dabei weiterhin in \mathbf{b}_1 über, aber \mathbf{b}_2 in $\mathbf{b}'_2 := \cos(\alpha)\mathbf{b}_2 + \sin(\alpha)\mathbf{b}_3$ und \mathbf{b}_3 in $\mathbf{b}'_3 := -\sin(\alpha)\mathbf{b}_2 + \cos(\alpha)\mathbf{b}_3$ (vergl. Abb. 6.1). Die Basisdarstellung $\mathbf{x} = \tilde{x}_1\mathbf{b}_1 + \tilde{x}_2\mathbf{b}_2 + \tilde{x}_3\mathbf{b}_3$ hat sich nicht geändert. Daher wird

$$\begin{aligned}\mathbf{x}' &= \tilde{x}_1\mathbf{b}'_1 + \tilde{x}_2\mathbf{b}'_2 + \tilde{x}_3\mathbf{b}'_3 \\ &= \tilde{x}_1\mathbf{b}_1 + \tilde{x}_2(\cos(\alpha)\mathbf{b}_2 + \sin(\alpha)\mathbf{b}_3) + \tilde{x}_3(-\sin(\alpha)\mathbf{b}_2 + \cos(\alpha)\mathbf{b}_3) \\ &= \tilde{x}_1\mathbf{b}_1 + (\cos(\alpha)\tilde{x}_2 - \sin(\alpha)\tilde{x}_3)\mathbf{b}_2 + (\sin(\alpha)\tilde{x}_2 + \cos(\alpha)\tilde{x}_3)\mathbf{b}_3 \\ &= \tilde{x}'_1\mathbf{b}_1 + \tilde{x}'_2\mathbf{b}_2 + \tilde{x}'_3\mathbf{b}_3.\end{aligned}$$

Man sieht, daß die eigentliche Berechnung der Koordinaten \tilde{x}'_i von \mathbf{x}' bezüglich der Basis \mathcal{B} nur in den Koordinaten \tilde{x}_i von \mathbf{x} bezüglich dieser Basis stattfinden. Die Basisvektoren \mathbf{b}_i selbst nehmen daran gar nicht teil. Sie haben nur eine ordnende Rolle, indem sie sozusagen ihre Koeffizienten aufsammeln. Außerdem braucht man sie natürlich, um das Ergebnis wieder in der Ausgangsbasis angeben zu können, oder einfacher gesagt, um \mathbf{x}' letztendlich auszurechnen. Da uns \mathcal{B} ja bekannt ist, handelt es sich dabei nur noch um simples Einsetzen. Das gestellte Problem wurde im Wesentlichen mit dem Auffinden der Bildkoordinaten \tilde{x}'_i bezüglich \mathcal{B} gelöst. Es sollte daher möglich sein, ein Verfahren anzugeben, das allein mit den Koordinaten auskommt. Wir werden dabei mit Spaltenvektoren arbeiten, deren Einträge sich auf verschiedene Basen beziehen. Damit man dabei den Bezug nicht verliert, vereinbaren wir ein für alle Mal folgende *Regelung*:

Wenn nur eine Basis in Betracht kommt, meistens ist es die kanonische Basis $\mathcal{E} = \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$, dann schreiben wir den Spaltenvektor, der die Koordinaten von \mathbf{x} bzgl. dieser Basis enthält, *ohne* Index

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Das ist die überwiegend anzutreffende Situation. Außerdem identifizieren wir meistens den Vektor \mathbf{x} mit seiner Koordinatenspalte. Wenn dagegen mehrere Basen nebeneinander Verwendung finden, machen wir durch einen Index am Spaltenvektor deutlich, auf welche Basis sich die Koordinaten beziehen. Für eine Basis $\mathcal{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ etwa fügen wir in der Kurznotation \mathbf{x} und in der ausführlichen Koordinatenschreibweise einen Index $_{\mathcal{B}}$ an:

$$\mathbf{x}_{\mathcal{B}} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_{\mathcal{B}} = [x_1, x_2, \dots, x_n]_{\mathcal{B}}^t. \quad (6.38)$$

$\mathbf{x}_{\mathcal{B}}$ ist der Koordinatenvektor bzgl. der Basis \mathcal{B} für den Vektor $\mathbf{x} = x_1 \mathbf{b}_1 + x_2 \mathbf{b}_2 + \dots + x_n \mathbf{b}_n$. In unserem Beispiel sieht das folgendermaßen aus:

$$\mathbf{x} = \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix}, \quad \mathbf{x}_{\mathcal{B}} = \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}_{\mathcal{B}}, \quad \mathbf{x}'_{\mathcal{B}} = \begin{bmatrix} 3 \\ -1 \\ 1 \end{bmatrix}_{\mathcal{B}}, \quad \mathbf{x}' = \frac{1}{3} \begin{bmatrix} -1 \\ 7 \\ 7 \end{bmatrix}.$$

$\mathbf{x}_{\mathcal{B}}$ ist der *Koordinatenvektor von \mathbf{x} bzgl. der Basis \mathcal{B}* .

Kehren wir zu unserem erweiterten Beispiel zurück und arbeiten konsequent mit den Koordinatenvektoren: $\mathbf{x}_{\mathcal{B}} = [\tilde{x}_1, \tilde{x}_2, \tilde{x}_3]^t_{\mathcal{B}}$,

$$\mathbf{x}'_{\mathcal{B}} = \begin{bmatrix} \tilde{x}'_1 \\ \tilde{x}'_2 \\ \tilde{x}'_3 \end{bmatrix}_{\mathcal{B}} = \begin{bmatrix} \tilde{x}_1 \\ \cos(\alpha)\tilde{x}_2 - \sin(\alpha)\tilde{x}_3 \\ \sin(\alpha)\tilde{x}_2 + \cos(\alpha)\tilde{x}_3 \end{bmatrix}_{\mathcal{B}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{bmatrix}_{\mathcal{B}} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix}_{\mathcal{B}} =: D(\alpha)_{\mathcal{B}} \mathbf{x}_{\mathcal{B}}.$$

Die Matrix $D(\alpha)_{\mathcal{B}}$ ist im Wesentlichen eine ebene Drehung, wie in (6.16). Die 2- und die 3-Achse werden um den Winkel α gedreht, während die 1-Achse unverändert bleibt. Die günstige Wahl der Basisvektoren und das Rechnen mit den Koordinatenvektoren hat das deutlich hervortreten lassen.

Drehen wir den Vektor $\mathbf{x} = [1, 3, 1]^t$ um 45° . Zunächst in der Basis \mathcal{B} :

$$\mathbf{x}'_{\mathcal{B}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}_{\mathcal{B}} \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}_{\mathcal{B}} = \begin{bmatrix} 3 \\ 0 \\ \sqrt{2} \end{bmatrix}_{\mathcal{B}}.$$

Daraus erhalten wir den Ergebnisvektor \mathbf{x}' durch

$$\mathbf{x}' = 3 \mathbf{b}_1 + \sqrt{2} \mathbf{b}_3 = \frac{1}{3} \begin{bmatrix} 3 \\ 6 \\ 6 \end{bmatrix} + \frac{1}{3} \begin{bmatrix} -2\sqrt{2} \\ 2\sqrt{2} \\ -\sqrt{2} \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 3 - 2\sqrt{2} \\ 6 + 2\sqrt{2} \\ 6 - \sqrt{2} \end{bmatrix}.$$

Übrigens sollte die Länge von \mathbf{x}' mit der Länge $\sqrt{11}$ von \mathbf{x} übereinstimmen, denn ein Drehung ändert sie nicht: $\|\mathbf{x}'\|^2 = \frac{1}{9} (9 + 8 - 12\sqrt{2} + 36 + 8 + 24\sqrt{2} + 36 + 2 - 12\sqrt{2}) = \frac{99}{9} = 11$.

Die Rücktransformation des Vektors $\mathbf{x}'_{\mathcal{B}}$ aus der Basis \mathcal{B} in die Ausgangsbasis gestaltet sich sehr einfach nach folgendem Schema:

$$\mathbf{x}'_{\mathcal{B}} = \begin{bmatrix} \tilde{x}'_1 \\ \tilde{x}'_2 \\ \tilde{x}'_3 \end{bmatrix}_{\mathcal{B}} \Rightarrow \mathbf{x}' = \tilde{x}'_1 \mathbf{b}_1 + \tilde{x}'_2 \mathbf{b}_2 + \tilde{x}'_3 \mathbf{b}_3 = [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3] \cdot \begin{bmatrix} \tilde{x}'_1 \\ \tilde{x}'_2 \\ \tilde{x}'_3 \end{bmatrix}_{\mathcal{B}} = B \mathbf{x}'_{\mathcal{B}},$$

mit der Transformationsmatrix $B := [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3]$, die einfach die Basisvektoren als Spaltenvektoren enthält. Ist die Basis \mathcal{B} also bekannt, so lässt sich B sofort angeben. Wie oben zu sehen ist, transformiert sie die Koordinaten aus der neuen Basis \mathcal{B} in die Ausgangsbasis. B ist seiner Natur nach eine Rücktransformation, denn meistens will man die Koordinaten eines Vektors

bzgl. der neuen Basis aus den bekannten der Ausgangsbasis berechnen, bei der es sich überwiegend um die kanonische handelt. In der Gleichung $\mathbf{x} = \mathbf{B}\mathbf{x}_{\mathcal{B}}$ ist also normalerweise der Vektor $\mathbf{x}_{\mathcal{B}}$ auf der rechten Seite gesucht, während \mathbf{x} bekannt ist. Das bedeutet, daß wir \mathbf{B} auf die andere Seite bringen müssen – was kein Problem darstellt, wenn \mathbf{B} invertierbar ist: $\mathbf{x}_{\mathcal{B}} = \mathbf{B}^{-1}\mathbf{x}$. Und natürlich ist \mathbf{B} invertierbar, denn seine Spaltenvektoren bilden eine Basis und sind daher linear unabhängig (vegl. Lemma 6.8.2). Weil es sich bei \mathcal{B} sogar um eine ONB handelt, ist die Inverse nach Korollar 6.10.3 einfach durch \mathbf{B}^* gegeben. Damit können wir jetzt das Verfahren systematisch darstellen:

- i) Transformiere den Vektor \mathbf{x} in die neue Basis \mathcal{B} , d. h., berechne den Koordinatenvektor $\mathbf{x}_{\mathcal{B}}$ von \mathbf{x} bzgl. \mathcal{B} : $\mathbf{x}_{\mathcal{B}} = \mathbf{B}^{-1}\mathbf{x}$.
- ii) Auf $\mathbf{x}_{\mathcal{B}}$ wird die lineare Abbildung A angewandt. In der Basis \mathcal{B} ist sie durch die Matrix $A_{\mathcal{B}}$ gegeben: $\mathbf{x}'_{\mathcal{B}} = A_{\mathcal{B}}\mathbf{x}_{\mathcal{B}}$.
- iii) Durch die Matrix \mathbf{B} wird das Ergebnis zurück in die Ausgangsbasis transformiert: $\mathbf{x}' = \mathbf{B}\mathbf{x}'_{\mathcal{B}}$.

Dieser Vorgang kann kompakt und übersichtlich durch $\mathbf{x}' = \mathbf{B}A_{\mathcal{B}}\mathbf{B}^{-1}\mathbf{x}$ zusammengefaßt werden. Wenn wir uns erinnern, daß die Abbildung $A: \mathbf{x} \mapsto \mathbf{x}'$ die Drehung um den Vektor \mathbf{n} darstellt, (deren Matrix uns partout nicht einfallen wollte), dann zeigt obige Gleichung, daß wir A inzwischen kennen:

$$A = \mathbf{B}A_{\mathcal{B}}\mathbf{B}^{-1}. \quad (6.39)$$

Sie ist auch leicht zu interpretieren: Mit \mathbf{B}^{-1} wechselt man in die Basis \mathcal{B} . Dort wird mit der Matrix $A_{\mathcal{B}}$ abgebildet und das Ergebnis anschließend mit \mathbf{B} zurück in die Ausgangsbasis transformiert. In dieser Interpretation ist der Bezug zu unserem konkreten Beispiel in den Hintergrund getreten. Wir haben den allgemeinen Mechanismus einer Koordinatentransformation gefunden: *Man sucht eine geeignete Basis \mathcal{B} (die durchaus nicht immer eine ONB sein muß), in der man die Matrix $A_{\mathcal{B}}$ einer Abbildung leicht angeben kann. Durch (6.39) berechnet man die Matrix A der Abbildung im Ausgangssystem.*

Das läßt sich übersichtlich durch ein sogenanntes *kommutierendes Diagramm* veranschaulichen:

$$\begin{array}{ccc} \mathbb{R}^n & \xrightarrow{A} & \mathbb{R}^n \\ B^{-1} \downarrow & & \uparrow B \\ \mathbb{R}^n & \xrightarrow{A_{\mathcal{B}}} & \mathbb{R}^n \end{array}$$

Statt den direkten Weg $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$ zu gehen, beschreitet man den Umweg $B^{-1}A_{\mathcal{B}}B$ über eine Basisdarstellung des \mathbb{R}^n , in der die Abbildungsmatrix $A_{\mathcal{B}}$ bekannt ist. Das läßt sich leicht auf allgemeine Vektorräume und auch auf lineare Abbildungen zwischen verschiedenen Räumen verallgemeinern:

$$\begin{array}{ccc} V & \xrightarrow{A} & W \\ B \downarrow & & \uparrow C \\ X & \xrightarrow{\tilde{A}} & Y \end{array}$$

Durch eine invertierbare lineare Abbildung B wird der Vektorraum V auf den Vektorraum X abgebildet. Von dort führt die bekannte lineare Abbildung \tilde{A} in einen passenden Vektorraum

\mathcal{Y} , der mit einer invertierbaren Abbildung C in den Zielraum \mathcal{W} transformiert wird. Der allgemeinste Vorgang, den wir als Koordinatentransformation deuten wollen, hat daher die Form

$$\mathcal{A} = C \tilde{\mathcal{A}} B .$$

Für unser Beispiel bedeutet das

$$\begin{aligned} A &= \frac{1}{9} \begin{bmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ 2 & -2 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{bmatrix} \begin{bmatrix} 1 & 2 & 2 \\ 2 & 1 & -2 \\ -2 & 2 & -1 \end{bmatrix} \\ &= \frac{1}{9} \begin{bmatrix} 1 & 2 & -2 \\ 2 & 1 & 2 \\ 2 & -2 & -1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 2 \\ 2 \cos(\alpha) + 2 \sin(\alpha) & \cos(\alpha) - 2 \sin(\alpha) & -2 \cos(\alpha) + \sin(\alpha) \\ -2 \cos(\alpha) + 2 \sin(\alpha) & 2 \cos(\alpha) + \sin(\alpha) & -\cos(\alpha) - 2 \sin(\alpha) \end{bmatrix} \\ &= \frac{1}{9} \begin{bmatrix} 1 + 8 \cos(\alpha) & 2 - 2 \cos(\alpha) - 6 \sin(\alpha) & 2 - 2 \cos(\alpha) + 6 \sin(\alpha) \\ 2 - 2 \cos(\alpha) + 6 \sin(\alpha) & 4 + 5 \cos(\alpha) & 4 - 4 \cos(\alpha) - 3 \sin(\alpha) \\ 2 - 2 \cos(\alpha) - 6 \sin(\alpha) & 4 - 4 \cos(\alpha) + 3 \sin(\alpha) & 4 + 5 \cos(\alpha) \end{bmatrix}. \end{aligned}$$

Für $\alpha = \frac{\pi}{2}$ ergibt sich wieder

$$Ax = \frac{1}{9} \begin{bmatrix} 1 & -4 & 8 \\ 8 & 4 & 1 \\ -4 & 7 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} -1 \\ 7 \\ 7 \end{bmatrix}.$$

Natürlich lässt sich (6.39) auch nach A_B umstellen. Wir haben dann die Transformationsformel einer Abbildung A in die Basis B :

$$A_B = B^{-1}AB . \quad (6.40)$$

6.12 Determinanten

Wir haben im Abschnitt 5.1.13 das Spatprodukt $\det(\mathbf{x}, \mathbf{y}, \mathbf{z})$ dreier Vektoren \mathbf{x} , \mathbf{y} und \mathbf{z} im \mathbb{R}^3 als Volumenfunktion kennengelernt. Es bestimmt, bis auf ein mögliches Vorzeichen, das Volumen des Spats, der von den drei Vektoren aufgespannt wird. Wir haben das Produkt mit dem systematischen Namen Determinante bezeichnet, der für die Volumenfunktion in höheren Dimensionen verwendet wird, mit der wir uns jetzt beschäftigen werden. Um die richtige Definition zu finden, rekapitulieren wir noch einmal die Eigenschaften, die die Determinante im \mathbb{R}^3 charakterisiert (vergl. Satz 5.1.14):

Die Determinante ist eine Abbildung von $\mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3$ nach \mathbb{R} , mit den Eigenschaften:

- i) $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \mapsto \det(\mathbf{x}, \mathbf{y}, \mathbf{z})$ ist in jeder Komponente linear. D. h., für $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^3$ und $t_1, t_2 \in \mathbb{R}$ gilt

$$\det(t_1 \mathbf{x}_1 + t_2 \mathbf{x}_2, \mathbf{y}, \mathbf{z}) = t_1 \det(\mathbf{x}_1, \mathbf{y}, \mathbf{z}) + t_2 \det(\mathbf{x}_2, \mathbf{y}, \mathbf{z}).$$

Entsprechendes gilt für die zweite und die dritte Komponente.

- ii) Die Vertauschung zweier Komponenten ergibt einen Vorzeichenwechsel der Determinante. Also etwa $\det(\mathbf{x}, \mathbf{y}, \mathbf{z}) = -\det(\mathbf{y}, \mathbf{x}, \mathbf{z}) = \det(\mathbf{y}, \mathbf{z}, \mathbf{x}) \dots$
- iii) Für die kanonische Basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ des \mathbb{R}^3 gilt $\det(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = 1$.

Genau genommen würde es ausreichen, unter i) die Linearität in der ersten Komponente zu fordern. Zusammen mit ii) lässt sie sich dann auf alle anderen ausdehnen.

Eine Abbildung mit der Eigenschaft i) heißt *multilinear*, mit der Eigenschaft ii) *alternierend* und *normiert*, wenn iii) erfüllt ist. Kurz gesagt ist eine Determinante daher eine *normierte, alternierende Multilinearform*. In dieser Formulierung ist der zugrunde liegende Raum \mathbb{R}^3 schon nicht mehr erkennbar. Deshalb werden wir ein solches Objekt im \mathbb{R}^n oder \mathbb{C}^n als Determinante bezeichnen und seine Eigenschaften in diesem allgemeinen Rahmen untersuchen. Dabei stellt sich zunächst die Frage nach der Realisierbarkeit, denn in diesen Räumen haben wir keine Hilfsmittel, wie etwa das Kreuzprodukt zu unserer Verfügung. Erfreulicherweise wird die Analyse der Eigenschaften einer Determinante in eine konkrete Rechenvorschrift münden und so die Frage nach der Existenz zur Zufriedenheit beantworten.

6.12.1 Definition Eine alternierende Multilinearform λ auf \mathbb{K}^n ($\mathbb{K} = \mathbb{R}$, oder $\mathbb{K} = \mathbb{C}$) ist eine Abbildung

$$\mathbb{K}^n \times \mathbb{K}^n \times \cdots \times \mathbb{K}^n \ni (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) \mapsto \lambda(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) \in \mathbb{K}$$

mit folgenden Eigenschaften:

- i) Sie ist in jeder Komponente linear, d. h., für jedes $i = 1, \dots, n$ ist

$$\mathbb{K}^n \ni \mathbf{a}_i \mapsto \lambda(\mathbf{a}_1, \dots, \mathbf{a}_i, \dots, \mathbf{a}_n)$$

eine lineare Abbildung von \mathbb{K}^n nach \mathbb{K} .

- ii) Sie ist alternierend, d. h., für zwei beliebige Positionen $i \neq j$ gilt

$$\lambda(\mathbf{a}_1, \dots, \mathbf{a}_i, \dots, \mathbf{a}_j, \dots, \mathbf{a}_n) = -\lambda(\mathbf{a}_1, \dots, \mathbf{a}_j, \dots, \mathbf{a}_i, \dots, \mathbf{a}_n).$$

Ist λ darüber hinaus normiert, d. h. gilt für die kanonische Basis $\{e_1, \dots, e_n\}$ von \mathbb{K}^n $\lambda(e_1, \dots, e_n) = 1$, so nennen wir diese Form Determinante \det auf \mathbb{K}^n . Wir verwenden den Begriff Determinante auch für Matrizen $A \in M_n(\mathbb{K})$, indem wir für eine solche Matrix $A = [a_1, \dots, a_n]$ mit den Spaltenvektoren $a_i \in \mathbb{K}^n$ die Determinante durch $\det(A) := \det(a_1, \dots, a_n)$ definieren.

In dieser Definition steckt auch schon eine Behauptung, weil wir von *der* Determinante auf \mathbb{K}^n sprechen. Wir werden gleich sehen, daß es tatsächlich nur eine Determinante auf \mathbb{K}^n gibt. Wir ziehen zunächst einige Folgerungen aus der Definition:

6.12.2 Satz Alternierende Multilinearformen $\lambda \neq 0$ auf \mathbb{K}^n haben folgende Eigenschaften:

- i) Ist einer der Vektoren a_1, \dots, a_n der Nullvektor, oder sind zwei der Vektoren gleich, so gilt $\lambda(a_1, \dots, a_n) = 0$.
- ii) $\lambda(a_1, \dots, t a_j + a_i, \dots, a_j, \dots, a_n) = \lambda(a_1, \dots, a_i, \dots, a_j, \dots, a_n)$, $t \in \mathbb{K}$,
d. h., man kann beliebige Vielfache eines Vektors zu einem anderen Vektor addieren, ohne daß sich der Wert der Form ändert.
- iii) Ist $\lambda(a_1, \dots, a_n) \neq 0$, so ist die Menge $\{a_1, \dots, a_n\}$ linear unabhängig.
- iv) Zwei alternierende Multilinearformen λ und λ' auf \mathbb{K}^n unterscheiden sich nur durch einen gemeinsamen Faktor: $\lambda' = c \cdot \lambda$, $c \in \mathbb{K}$.
- v) Es gibt nur eine normierte alternierende Multilinearform auf \mathbb{K} , nämlich die Determinante \det .
Insbesondere unterscheiden sich alle anderen alternierenden Multilinearformen von \det jeweils nur um einen festen Vorfaktor.

Beweis. Zu i): Wir können o. B. d. A. von $a_1 = \mathbf{0}$ ausgehen. Die Linearität in der ersten Komponente zeigt dann $\det(\mathbf{0}, a_2, \dots, a_n) = \det(0 \cdot \mathbf{0}, a_2, \dots, a_n) = 0 \cdot \det(\mathbf{0}, a_2, \dots, a_n) = 0$. Gilt $a_i = a_j$ für $i \neq j$, dann ergibt die Vertauschung dieser Vektoren in der Determinante keine Veränderung, muß aber nach 6.12.1 ii) einen Vorzeichenwechsel verursachen. Das heißt: $\lambda(a_1, \dots, a_i, \dots, a_i, \dots, a_n) = -\lambda(a_1, \dots, a_i, \dots, a_i, \dots, a_n) = 0$.

Zu ii): Das folgt aus i) und der Multilinearität:

$$\lambda(\dots, t a_j + a_i, \dots, a_j, \dots) = t \lambda(\dots, a_j, \dots, a_j, \dots) + \lambda(\dots, a_i, \dots, a_j, \dots).$$

Nach i) verschwindet $t \lambda(\dots, a_j, \dots, a_j, \dots)$.

Diese Eigenschaft erinnert an die Umformungen beim GAUSS-Verfahren. Tatsächlich ist sie der Grund dafür, daß große Determinanten mit einer geringfügigen Variation des GAUSS-Verfahrens berechnet werden können.

Zu iii): Wir zeigen: $\{a_1, \dots, a_n\}$ ist linear abhängig $\Rightarrow \lambda(a_1, \dots, a_n) = 0$.

Nach Lemma 6.4.7 ist $\{a_1, \dots, a_n\}$ genau dann linear abhängig, wenn es einen Vektor aus dieser Menge gibt, o. B. d. A. dürfen wir von a_1 ausgehen, der eine Linearkombination der übrigen ist: $a_1 = t_2 a_2 + t_3 a_3 + \dots + t_n a_n$. Die Linearität in der ersten Komponente und i) ergeben dann

$$\begin{aligned} \lambda(a_1, \dots, a_n) &= \lambda(t_2 a_2 + t_3 a_3 + \dots + t_n a_n, a_2, a_3, \dots, a_n) \\ &= t_2 \lambda(a_2, a_2, a_3, \dots, a_n) + t_3 \lambda(a_3, a_2, a_3, \dots, a_n) + \dots \end{aligned}$$

$$+ t_n \lambda(a_1, a_2, a_3, \dots, a_n) = 0.$$

Daher gilt: $\lambda(a_1, \dots, a_n) \neq 0 \Rightarrow \{a_1, \dots, a_n\}$ ist linear unabhängig (vergl. 1.9).

Zu iv): Im Laufe der Untersuchungen zu diesem Punkt werden wir eine konkrete Rechenvorschrift zur Berechnung von $\lambda(a_1, \dots, a_n)$ erarbeiten.

Zunächst untersuchen wir die Multilinearität von λ . Dazu entwickeln wir alle Vektoren a_i in der kanonischen Basis $\{e_1, \dots, e_n\}$:

$$a_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ni} \end{bmatrix} = a_{1i} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + a_{2i} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \cdots + a_{ni} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} = \sum_{k=1}^n a_{ki} e_k.$$

Dann verwenden wir nacheinander die Linearität in jeder Komponente von λ :

$$\begin{aligned} \lambda(a_1, a_2, \dots, a_n) &= \lambda\left(\sum_{k_1=1}^n a_{k_1 1} e_{k_1}, a_2, \dots, a_n\right) \\ &= \sum_{k_1=1}^n a_{k_1 1} \cdot \lambda\left(e_{k_1}, \sum_{k_2=1}^n a_{k_2 2} e_{k_2}, \dots, a_n\right) \\ &= \sum_{k_1=1}^n \sum_{k_2=1}^n a_{k_1 1} a_{k_2 2} \cdot \lambda\left(e_{k_1}, e_{k_2}, \dots, a_n\right) \\ &= \sum_{k_1=1}^n \sum_{k_2=1}^n \cdots \sum_{k_n=1}^n a_{k_1 1} a_{k_2 2} \cdots a_{k_n n} \cdot \lambda(e_{k_1}, e_{k_2}, \dots, e_{k_n}). \end{aligned}$$

Dieser Ausdruck stellt bereits eine Rechenvorschrift dar, wenn wir die Ausdrücke $\lambda(e_{k_1}, e_{k_2}, \dots, e_{k_n})$ berechnen können. Alles andere hängt nur noch von den Koordinaten der Vektoren a_i ab. Aus i) wissen wir, daß λ den Wert Null liefert, wenn zwei Vektoren e_{k_i} und e_{k_j} gleich sind. Das bedeutet, daß in der Vielfachsumme alle Summanden wegfallen, für die wenigstens zwei der Indizes k_1, k_2, \dots, k_n gleich sind. Oder anders ausgedrückt: Die Summe erstreckt sich nur über Indextupel $[k_1, k_2, \dots, k_n]$, in denen keine zwei Einträge k_i und k_j gleich sind. Diese Tupel stellen daher *Permutationen* π von $[1, 2, \dots, n]$ dar. Wir haben ihre Eigenschaften in Abschnitt 2.5.10 eingehend untersucht und verwenden jetzt die dort vereinbarten Notationen und die erzielten Ergebnisse – allerdings benötigen wir letztlich nur eine einzige Tatsache (s. u.). Die Vielfachsumme erstreckt sich also tatsächlich über alle Elemente π aus der Menge S_n der Permutationen der Zahlen $1, \dots, n$. k_1, k_2, \dots sind dann die Bilder von $1, 2, \dots$ der Abbildung π : $k_i = \pi(i)$. Wir erhalten

$$\lambda(a_1, a_2, \dots, a_n) = \sum_{\pi \in S_n} a_{\pi(1)1} a_{\pi(2)2} \cdots a_{\pi(n)n} \cdot \lambda(e_{\pi(1)}, e_{\pi(2)}, \dots, e_{\pi(n)}).$$

Der Ausdruck $\lambda(e_{\pi(1)}, e_{\pi(2)}, \dots, e_{\pi(n)})$ unterscheidet sich von $\lambda(e_1, e_2, \dots, e_n)$ nur um ein Vorzeichen. Durch eine Folge von Transpositionen können die Vektoren

$\mathbf{e}_{\pi(1)}, \mathbf{e}_{\pi(2)}, \dots, \mathbf{e}_{\pi(n)}$ nämlich in die gewohnte Reihenfolge $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ gebracht werden. Deren Abfolge ist zwar nicht eindeutig, aber zwei verschiedene unterscheiden sich immer um eine gerade Anzahl an Transpositionen (mehr wird aus dem Abschnitt 2.5.10 eigentlich nicht gebraucht). Jede der k benötigten Transpositionen erzeugt nach 6.12.1 ii) ein Vorzeichenwechsel. Damit ist $(-1)^k$ der Faktor, der $\lambda(\mathbf{e}_{\pi(1)}, \mathbf{e}_{\pi(2)}, \dots, \mathbf{e}_{\pi(n)})$ von $\lambda(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ unterscheidet. Diesen Faktor bezeichnet man als Vorzeichen $\text{sgn}(\pi)$ von π (Definition 2.5.12).

Als Ergebnis unserer Untersuchung haben wir

$$\lambda(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) = \lambda(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n) \cdot \sum_{\pi \in S_n} \text{sgn}(\pi) \cdot a_{\pi(1)1} a_{\pi(2)2} \cdots a_{\pi(n)n}. \quad (6.41)$$

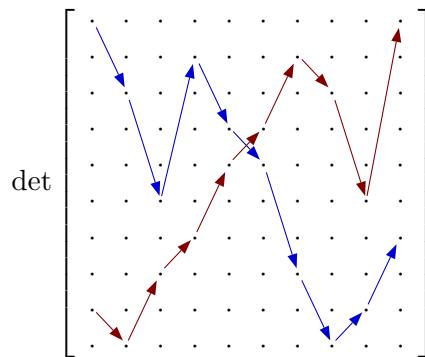
Die Summe ist eine konkrete Rechenvorschrift, die nur von den Koeffizienten der Vektoren \mathbf{a}_i abhängt, aber nicht mehr von λ . Zwei alternierende Multilinearformen $\lambda \neq 0$ und $\lambda' \neq 0$ unterscheiden sich also nur um einen gemeinsamen Faktor voneinander (nämlich um $\lambda(\mathbf{e}_1, \dots, \mathbf{e}_n)/\lambda'(\mathbf{e}_1, \dots, \mathbf{e}_n)$, wenn λ' nicht die triviale Form 0 ist, was von $\lambda'(\mathbf{e}_1, \dots, \mathbf{e}_n) \neq 0$ bereits garantiert wird).

Zu v): Gleichung 6.41 zeigt auch, daß es nur eine alternierende Multilinearform geben kann, die normiert ist. Diese ist nämlich durch die Summe in 6.41 bestimmt, in die nur die Koordinatendarstellungen der Vektoren \mathbf{a}_i eingehen. Der Vorfaktor $\lambda(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$ ist wegen der Normierung 1. Gemäß Definition 6.12.1 gilt

$$\det(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) = \sum_{\pi \in S_n} \text{sgn}(\pi) \cdot a_{\pi(1)1} a_{\pi(2)2} \cdots a_{\pi(n)n}. \quad (6.42)$$

Das ist die *Summendarstellung der Determinante*. □

Diese Darstellung ist für konkrete Rechnungen nicht sonderlich geeignet. Aber sie wird uns dabei helfen bessere Methoden zu finden. Obwohl sie so unhandlich ist, hat sie doch eine sehr anschauliche Deutung, die den ersten Schritt hin zu bequemeren Berechnungsmöglichkeiten darstellt. Es ist das sogenannte *Pfadbild* der Determinante. Dafür sehen wir für den Moment mal von den konkreten Einträgen der Matrix $A = [\mathbf{a}_1, \dots, \mathbf{a}_n]$ ab und stellen sie uns als ein quadratisches Punkteschema vor:



Jeder Summand in (6.42) entspricht dann einem Pfad durch das Punktgitter, der jede Zeile und jede Spalte jeweils genau einmal trifft. Diese Eigenschaft einer Determinante bezeichnen

wir als *Pfadregel*.. Zwei Beispiele sind eingezeichnet, nämlich der Pfad, der zum Summanden $a_{11}a_{32}a_{63}a_{24}a_{45}a_{56}a_{87}a_{108}a_{99}a_{710}$ und der, der zu $-a_{91}a_{102}a_{83}a_{74}a_{55}a_{46}a_{27}a_{38}a_{69}a_{110}$ gehört.

Dieses Bild ist stark genug, um sofort einige praktische Folgerungen ziehen zu können, wenn die Matrix eine spezielle Form hat. Ist A etwa eine obere (oder untere) Dreiecksmatrix

$$\det \begin{bmatrix} a_{11} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & a_{22} & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & a_{33} & \cdot & \cdot & \cdot & \cdot \\ & & & a_{44} & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot & \cdot \\ & & & & & \ddots & \cdot \\ 0 & & & & & & a_{nn} \end{bmatrix} = a_{11}a_{22}a_{33}a_{44} \cdots a_{nn},$$

dann ist die Determinante einfach das Produkt der Diagonalelemente, denn der einzige Pfad, der nicht wenigsten einmal in die Dreieckshälfte, die nur aus Nullen besteht, ausweichen muß, ist der entlang der Diagonalen. Die Permutation, die zu diesem Pfad gehört, ist natürlich id, für die $\text{sgn}(\text{id}) = 1$ gilt.

Diese Situation läßt sich noch verallgemeinern. Dafür benötigen wir aber eine wichtige Folgerung aus dem Eindeutigkeitssatz 6.12.2 iv).

6.12.3 Der Determinanten-Produktsatz

6.12.4 Satz Für zwei Matrizen $A, B \in M_n(\mathbb{K})$ gilt

$$\det(AB) = \det(A)\det(B). \quad (6.43)$$

Das bedeutet insbesondere $\det(AB) = \det(BA)$, obwohl im Allgemeinen $AB \neq BA$ gilt.

Beweis. Die Matrix B sei durch ihre Spaltenvektoren gegeben: $B := [b_1, \dots, b_n]$. Wir definieren die Abbildung $(b_1, \dots, b_n) \mapsto \lambda(b_1, \dots, b_n) := \det(AB) = \det(A[b_1, \dots, b_n]) = \det([Ab_1, \dots, Ab_n]) = \det(AB_1, \dots, AB_n)$. Sie ist multilinear, denn A ist eine lineare Abbildung:

$$\begin{aligned} \lambda(b_1 + t \cdot c_1, \dots, b_n) &= \det(AB_1 + t \cdot Ac_1, \dots, Ab_n) \\ &= \det(AB_1, \dots, Ab_n) + t \cdot \det(Ac_1, \dots, Ab_n) \\ &= \lambda(b_1, \dots, b_n) + t \cdot \lambda(c_1, \dots, b_n). \end{aligned}$$

Das zeigt die Linearität in der ersten Komponente. λ ist auch alternierend, denn das erbt die Abbildung von \det . Daher ist λ eine alternierende Multilinearform. Sie unterscheidet sich von jeder anderen, insbesondere von \det , nur jeweils um eine Konstante. Das heißt, es gilt $\det(AB) = \lambda(b_1, \dots, b_n) = c \cdot \det(b_1, \dots, b_n) = c \cdot \det(B)$, mit einer Konstante c , die wir gleich bestimmen werden. Die Gleichung gilt für jede Matrix $B \in M_n(\mathbb{K})$ und insbesondere auch für $B = 1$: $\det(A) = \det(A1) = c \cdot \det(1) = c$. Damit haben wir $\det(AB) = \det(A)\det(B)$ gezeigt. \square

Der Determinanten-Produktsatz hat viele Anwendungen, von denen wir einige kennenlernen werden. Etwa für Matrizen der Form

$$\left[\begin{array}{cccc|cccc} a_{11} & a_{12} & \cdots & a_{1k} & c_{11} & c_{12} & \cdots & c_{1n} \\ a_{21} & a_{22} & \cdots & a_{2k} & c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} & c_{k1} & c_{k2} & \cdots & c_{kn} \\ \hline 0 & 0 & \cdots & 0 & b_{11} & b_{12} & \cdots & b_{1n} \\ 0 & 0 & \cdots & 0 & b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & b_{n1} & b_{n2} & \cdots & b_{nn} \end{array} \right] = \left[\begin{array}{c|c} A & C \\ \hline 0 & B \end{array} \right].$$

Wir haben die $k \times k$ -Matrix in der linken oberen Ecke mit A , die anschließende $k \times n$ -Matrix mit C , die $n \times n$ -Matrix darunter mit B und die links anschließende Nullmatrix mit 0 bezeichnet. Die Matrix

$$\left[\begin{array}{cc} A & C \\ 0 & B \end{array} \right]$$

ist eine *Blockmatrix*, da ihre Einträge selber Matrizen passender Abmessungen sind. Die Pfade, die einen Beitrag zur Determinante dieser Blockmatrix leisten, müssen sich die ersten k Schritte in der Matrix A aufhalten, da sie sonst unweigerlich in die Matrix 0 eintreten müßten. Dabei werden natürlich auch die ersten k Zeilen alle besucht. Daher kann keiner dieser Pfade in die Matrix C gelangen, denn dafür müßte eine der ersten k Zeilen ein weiteres Mal betreten werden, was nach der Pfadregel verboten ist. Die einzige mögliche Fortsetzung findet er also in der Matrix B . Da keiner dieser Pfade Einträge aus C aufweist, können wir C durch 0 ersetzen, ohne den Wert der Determinante zu ändern. Mit dem Determinanten-Produktsatz folgt

$$\begin{aligned} \det \left(\left[\begin{array}{cc} A & C \\ 0 & B \end{array} \right] \right) &= \det \left(\left[\begin{array}{cc} A & 0 \\ 0 & B \end{array} \right] \right) = \det \left(\left[\begin{array}{cc} A & 0 \\ 0 & \mathbb{1}_n \end{array} \right] \left[\begin{array}{cc} \mathbb{1}_k & 0 \\ 0 & B \end{array} \right] \right) \\ &= \det \left(\left[\begin{array}{cc} A & 0 \\ 0 & \mathbb{1}_n \end{array} \right] \right) \det \left(\left[\begin{array}{cc} \mathbb{1}_k & 0 \\ 0 & B \end{array} \right] \right) = \det(A) \det(B). \end{aligned}$$

Die letzte Gleichheit bedarf noch einer Begründung. Wir zeigen nur, daß

$$\det \left(\left[\begin{array}{cc} \mathbb{1}_k & 0 \\ 0 & B \end{array} \right] \right) = \det(B)$$

gilt. Dafür definieren wir die alternierende Multilinearform λ auf \mathbb{K}^n durch

$$\lambda(\mathbf{b}_1, \dots, \mathbf{b}_n) := \det \left(\left[\begin{array}{cc} \mathbb{1}_k & 0 \\ 0 & [\mathbf{b}_1, \dots, \mathbf{b}_n] \end{array} \right] \right).$$

Die Eigenschaften einer alternierenden Multilinearform sind für λ schnell überprüft: Sie erbt sie alle von der Determinante auf \mathbb{K}^{k+n} . Wenn wir zeigen können, daß λ normiert ist, handelt es sich um die Determinante auf \mathbb{K}^n . Das ist einfach:

$$\lambda(\mathbf{e}_1, \dots, \mathbf{e}_n) = \det \left(\left[\begin{array}{cc} \mathbb{1}_k & 0 \\ 0 & [\mathbf{e}_1, \dots, \mathbf{e}_n] \end{array} \right] \right) = \det \left(\left[\begin{array}{cc} \mathbb{1}_k & 0 \\ 0 & \mathbb{1}_n \end{array} \right] \right) = 1.$$

Also ist $\lambda(\mathbf{b}_1, \dots, \mathbf{b}_n) = \det(\mathbf{B})$.

Halten wir das Ergebnis fest: Für eine obere (untere) Dreiecks-Blockmatrix gilt

$$\det \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ 0 & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{nn} \end{bmatrix} = \det(A_{11}) \det(A_{22}) \cdots \det(A_{nn}). \quad (6.44)$$

Eigentlich haben wir das nur für Blockmatrizen mit zwei Diagonaleinträgen gezeigt, aber das reicht, denn wir können dieses Ergebnis iterieren:

$$\det \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ 0 & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_{nn} \end{bmatrix} = \det(A_{11}) \cdot \det \begin{bmatrix} A_{22} & \cdots & A_{2n} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & A_{nn} \end{bmatrix}.$$

6.12.5 Beispiel

$$\begin{aligned} \det & \begin{bmatrix} 1 & 30 & 0 & 6 & 19 & 10 & 8 & i & 5 & -3 & 5 \\ 1 & 1 & 10 & 6 & 19 & 0 & 7 & 7i & 55 & -3 & 0 \\ 0 & 0 & 0 & 3 & 6 & 8 & 18 & 2i & 5 & 13 & 1 \\ 0 & 0 & 1 & 4 & 8 & 1 & 0 & 0 & 22 & 0 & 81 \\ 0 & 0 & 3 & 14 & 0 & 0 & 3 & 20 & 17 & 9 & i \\ 0 & 0 & 0 & 0 & 0 & 2 & 1 & 5 & 0 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & 7 & 2 & 4 & 10 & 14 & 9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 2 & 2 & 29 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5 & 1 & 12 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 5 \end{bmatrix} \\ &= \det \begin{bmatrix} 1 & 30 \\ 1 & 1 \end{bmatrix} \cdot \det \begin{bmatrix} 0 & 3 & 6 \\ 1 & 4 & 8 \\ 3 & 14 & 0 \end{bmatrix} \cdot \det \begin{bmatrix} 2 & 1 \\ 7 & 2 \end{bmatrix} \cdot 4 \cdot 5 \cdot \det \begin{bmatrix} 2 & 5 \\ 1 & 5 \end{bmatrix} \\ &= -29 \cdot (3 \cdot 8 \cdot 3 + 6 \cdot 14 - 6 \cdot 4 \cdot 3) \cdot (-3) \cdot 20 \cdot 5 = 730800. \end{aligned}$$

Eine der wichtigsten Folgerungen aus dem Determinanten-Produktsatz betrifft die Charakterisierung der Invertierbarkeit einer Matrix. Falls A^{-1} für eine Matrix A existiert, kann $\det(A)$ nicht Null sein, denn $1 = \det(\mathbb{1}) = \det(AA^{-1}) = \det(A)\det(A^{-1})$. Nebenbei erhalten wir daraus $\det(A^{-1}) = \frac{1}{\det(A)}$.

Erfreulicherweise gilt auch die Umkehrung: Aus $\det(A) \neq 0$ folgt die Existenz von A^{-1} . Denn nach Satz 6.12.2 iii) sind die Spaltenvektoren \mathbf{a}_i von $A = [\mathbf{a}_1, \dots, \mathbf{a}_n]$ linear unabhängig. Daher muß $\ker A = \{\mathbf{0}\}$ gelten, denn ein Vektor $\mathbf{0} \neq \mathbf{x} = [x_1, \dots, x_n] \in \ker A$ würde wegen $A\mathbf{x} = x_1\mathbf{a}_1 + \dots + x_n\mathbf{a}_n = \mathbf{0}$ eine nicht triviale Nullkombination der Spaltenvektoren ergeben, im Widerspruch zu deren linearen Unabhängigkeit. Nach Korollar 6.8.5 ist A invertierbar. Wir haben somit gezeigt:

6.12.6 Satz Für eine Matrix $A \in M_n(\mathbb{K})$ gilt:

$$A^{-1} \text{ existiert} \Leftrightarrow \det(A) \neq 0.$$

Wie verhält sich die Determinante einer Matrix unter äquivalenten Zeilenumformungen, wie sie für das GAUSS-Verfahren benötigt werden? Zur Beantwortung dieser Frage führen wir das GAUSS-Verfahren auf die Multiplikation von A mit geeigneten Matrizen D_{ij} und T_{ij} zurück und verwenden dann den Determinanten-Produktsatz. Zwei einfache Rechnungen zeigen die Idee:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & \lambda \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} + \lambda a_{41} & a_{22} + \lambda a_{42} & a_{23} + \lambda a_{43} & a_{24} + \lambda a_{44} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix},$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{41} & a_{42} & a_{43} & a_{44} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix}.$$

Die Matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & \lambda \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = [\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4 + \lambda \mathbf{e}_2] =: D_{24}(\lambda)$$

bewirkt die Addition des λ -Fachen der vierten Zeile zur zweiten und die Matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} = [\mathbf{e}_1, \mathbf{e}_4, \mathbf{e}_3, \mathbf{e}_2] =: T_{24}$$

die Vertauschung der zweiten mit der vierten Zeile. Die Matrizen

$$D_{ij}(\lambda) := [\mathbf{e}_1, \dots, \mathbf{e}_i, \dots, \mathbf{e}_j + \lambda \mathbf{e}_i, \dots, \mathbf{e}_n] \quad (6.45)$$

$$T_{ij} := [\mathbf{e}_1, \dots, \mathbf{e}_j, \mathbf{e}_{i+1}, \dots, \mathbf{e}_{j-1}, \mathbf{e}_i, \dots, \mathbf{e}_n] \quad (6.46)$$

werden also das λ -Fache der j -ten Zeile zur i -ten Zeile addieren, bzw. die i -te mit der j -ten Zeile vertauschen. Nach der Pfadregel gilt $\det(D_{ij}(\lambda)) = 1$, denn es handelt sich um eine obere ($j > i$), oder um eine untere Dreiecksmatrix ($j < i$), mit einer Diagonalen, die nur aus Einsen besteht. Dagegen ist $\det(T_{ij}) = -1$, denn die Vertauschung der i -ten mit der j -ten Spalte ergibt die Einheitsmatrix. Die Anwendung des Determinanten-Produktsatzes,

$$\begin{aligned} \det(D_{ij}(\lambda)A) &= \det(D_{ij}(\lambda)) \det(A) = \det(A), \\ \det(T_{ij}A) &= \det(T_{ij}) \det(A) = -1 \det(A), \end{aligned}$$

zeigt, daß sich $\det(A)$, was die Zeilenvektoren angeht, genauso verhält, wie für die Spaltenvektoren. Bis auf ein mögliches Vorzeichen ändert sich $\det(A)$ nicht, wenn wir in der Matrix A äquivalenten Zeilenumformungen (repräsentiert durch $D_{ij}(\lambda)$ und T_{ij}) und Vertauschungen von Spaltenvektoren vornehmen, bis A in eine obere bzw. unter Dreiecksmatrix D verwandelt wurde. Bis auf besagtes Vorzeichen stimmt $\det(A)$ mit $\det(D)$ überein.

Jetzt untersuchen wir die Situation für $\det(A^t)$. Führen wir hier, in genau derselben Reihenfolge wie für $\det(A)$, die entsprechenden äquivalenten Spaltenumformungen und Vertauschungen von Zeilen durch, so erhalten wir, bis auf dasselbe Vorzeichen, die Determinante von D^t . Wegen $\det(D) = \det(D^t)$ folgt $\det(A) = \det(A^t)$.

6.12.7 Satz Für jede Matrix $A \in M_n(\mathbb{K})$ gilt $\det(A) = \det(A^t)$.

Wir haben diesen Satz durch unsere Vorüberlegungen bereits bewiesen. Meist findet man für ihn aber einen anderen Beweis, der auf die Summendarstellung 6.42 zurückgreift. Es kann nicht schaden, für einen solchen Satz einen weiteren Beweis zu haben.

Beweis.

$$\begin{aligned}\det(A) &= \sum_{\pi \in S_n} \operatorname{sgn}(\pi) \cdot a_{\pi(1)1} a_{\pi(2)2} \cdots a_{\pi(n)n} \\ &= \sum_{\pi \in S_n} \operatorname{sgn}(\pi) \cdot a_{\pi(1)\pi^{-1}(\pi(1))} a_{\pi(2)\pi^{-1}(\pi(2))} \cdots a_{\pi(n)\pi^{-1}(\pi(n))} \\ &\stackrel{\text{i)}}{=} \sum_{\pi \in S_n} \operatorname{sgn}(\pi^{-1}) \cdot a_{1\pi^{-1}(1)} a_{2\pi^{-1}(2)} \cdots a_{n\pi^{-1}(n)} \\ &\stackrel{\text{ii)}}{=} \sum_{\pi \in S_n} \operatorname{sgn}(\pi) \cdot a_{1\pi(1)} a_{2\pi(2)} \cdots a_{n\pi(n)} \stackrel{\text{iii)}}{=} \det(A^t).\end{aligned}$$

Dazu muß etwas gesagt werden.

Zu i): Im Produkt $a_{\pi(1)\pi^{-1}(\pi(1))} a_{\pi(2)\pi^{-1}(\pi(2))} \cdots a_{\pi(n)\pi^{-1}(\pi(n))}$ durchlaufen die Zahlen $\pi(1), \pi(2), \dots, \pi(n)$ die Zahlen von 1 bis n . Durch Umsortierung läßt es sich also in die Form $a_{1\pi^{-1}(1)} a_{2\pi^{-1}(2)} \cdots a_{n\pi^{-1}(n)}$ bringen. Außerdem gilt $\operatorname{sgn}(\pi) = \operatorname{sgn}(\pi^{-1})$, denn $\operatorname{sgn}(\pi) \operatorname{sgn}(\pi^{-1}) = \operatorname{sgn}(\pi\pi^{-1}) = 1$ (vergl. (2.16)).

Zu ii): Wenn π alle Permutationen aus S_n durchläuft, dann stimmt das auch für π^{-1} . Die Summe unter i) ist also nur die umsortierte Summe ii).

Zu iii): Die Summe, die in ii) erhalten wurde, hat wieder die Summenform der Determinante einer Matrix, nur daß gegenüber 6.42 in den Matrixeinträgen die Reihenfolge der Indizes vertauscht ist. Es handelt sich also um die Einträge der Matrix A^t und bei der Determinante daher um $\det(A^t)$. \square

Die Überlegungen, die uns zu Satz 6.12.7 geführt haben, zeigen schon, daß wir zur Berechnung von $\det(A)$ die Matrix A mit Hilfe des GAUSS-Verfahrens und eventuell durch Vertauschen geeigneter Spaltenvektoren in eine obere oder untere Dreiecksmatrix D transformieren können, deren Determinante als Produkt der Diagonalelemente leicht gebildet werden kann. Führt man dabei Buch über die Vorzeichenwechsel, die durch Zeilen- bzw. Spaltenvertauschungen hervorgerufen werden, dann ist $\det(D)$, multipliziert mit dem erhaltenen Vorzeichen gerade

$\det(A)$. Praktischerweise gestattet man es sich dabei auch, wenn es sich anbietet, Spalten oder Zeilen mit geeigneten Faktoren zu multiplizieren, um die Rechnungen zu vereinfachen. Wenn man auch über diese Faktoren Buch führt, kann man sie zuletzt aus dem Ergebnis wieder herausrechnen. Wie das aussehen kann zeigen wir an einem Beispiel.

6.12.8 Beispiel

Wir wollen die Determinante folgender Matrix berechnen.

$$A := \begin{bmatrix} 2 & 5 & 0 & 0 & 2 & 1 \\ 1 & 2 & 6 & 0 & 5 & 2 \\ 0 & 5 & 3 & 5 & 4 & 3 \\ 8 & 5 & 3 & 0 & 1 & 2 \\ 3 & 1 & 1 & 0 & 0 & 3 \\ 0 & 2 & 0 & 1 & 1 & 0 \end{bmatrix}.$$

I	2	5	0	0	2	1			I	1	2	0	0	1	0	III - 5 VI		I	1	2	0	0	1	0		
II	1	2	6	0	5	2			II	0	1	1	3	0	3			II	0	1	1	3	0	3		
III	0	5	3	5	4	3			III	0	0	8	15	-1	18			III	0	0	0	25	-11	26		
IV	8	5	3	0	1	2			IV	0	0	6	8	0	5			IV	0	0	0	4	-2	-3		
V	3	1	1	0	0	3			V	0	0	4	-5	5	-4			V	0	0	4	-5	5	-4		
VI	0	2	0	1	1	0			VI	0	0	3	2	1	4			VI	0	0	3	2	1	4		
I	2	5	0	0	2	1	I ↔ VI		I	1	2	0	0	1	0			I	1	2	0	0	1	0		
II	1	2	6	0	5	2			II	0	1	1	3	0	3			II	0	1	1	3	0	3		
III	0	-5	3	0	-1	3			III	0	0	0	25	-11	26			III	0	0	0	25	-11	26		
IV	8	5	3	0	1	2			IV	0	0	0	4	-2	-3			IV	0	0	0	4	-2	-3		
V	3	1	1	0	0	3			V	0	0	4	-5	5	-4			V	0	0	0	-23	11	-28		
VI	0	2	0	1	1	0			VI	0	0	3	2	1	4			VI	0	0	3	2	1	4		
I	0	2	0	1	1	0		·(-1)	I	1	2	0	0	1	0			I	1	2	0	0	1	0		
II	1	2	6	0	5	2			II	0	1	1	3	0	3			II	0	1	1	3	0	3		
III	0	-5	3	0	-1	3			III	0	0	0	25	-11	26			III	0	0	0	25	-11	26		
IV	8	5	3	0	1	2			IV	0	0	0	4	-2	-3			IV	0	0	0	4	-2	-3		
V	3	1	1	0	0	3			V	0	0	0	-23	11	-28			V	0	0	0	-23	11	-28		
VI	2	5	0	0	2	1			VI	0	0	3	2	1	4			VI	0	0	3	2	1	4		
I	1	2	0	0	1	0		·(-1) ²	I	1	2	0	0	1	0			I	1	2	0	0	1	0		
II	0	2	6	1	5	2	II ↔ V		II	0	1	1	3	0	3			II	0	1	1	3	0	3		
III	0	-5	3	0	-1	3			III	0	0	0	25	-11	26			III	0	0	0	25	-11	26		
IV	0	5	3	8	1	2	IV + III		IV	0	0	0	4	-2	-3			IV	0	0	0	4	-2	-3		
V	0	1	1	3	0	3			V	0	0	0	-23	11	-28			V	0	0	0	-23	11	-28		
VI	0	5	0	2	2	1	VI + III		VI	0	0	0	25	-11	26			VI	0	0	0	25	-11	26		
I	1	2	0	0	1	0		·(-1)	I	1	2	0	0	1	0			I	1	2	0	0	1	0		
II	0	1	1	3	0	3	III + 5 II		II	0	1	1	3	0	3			II	0	1	1	3	0	3		
III	0	-5	3	0	-1	3			III	0	0	3	2	1	4			III	0	0	3	2	1	4		
IV	0	0	6	8	0	5			IV	0	0	0	4	-2	-3			IV	0	0	0	4	-2	-3		
V	0	2	6	1	5	2	V - 2 II		V	0	0	0	2	0	-2			V	0	0	0	2	0	-2		
VI	0	0	3	2	1	4			VI	0	0	0	25	-11	26			VI	0	0	0	25	-11	26		

An dieser Stelle brechen wir das GAUSS-Verfahren ab. Wie man sieht, haben wir eine weitere Spalte dem Schema angefügt, in der wir über Korrekturfaktoren Buch führen, die durch Tauschungen oder Multiplikation einzelner Zeilen mit geeigneten Zahlen entstanden sind. Da unsere Matrix weitgehend dreieckig geworden ist, erledigen wir den Rest mit Hilfe von 6.44 und erhalten

$$\det(A) = \frac{3}{3} \cdot \det \begin{bmatrix} 4 & -2 & -3 \\ 2 & 0 & -2 \\ 25 & -11 & 26 \end{bmatrix} = \det \begin{bmatrix} 4 & -2 & -3 \\ 2 & 0 & -2 \\ 1 & 1 & 44 \end{bmatrix} = -\det \begin{bmatrix} -2 & 4 & -3 \\ 0 & 2 & -2 \\ 1 & 1 & 44 \end{bmatrix}$$

$$= -\frac{1}{2} \det \begin{bmatrix} -2 & 4 & -3 \\ 0 & 2 & -2 \\ 0 & 6 & 85 \end{bmatrix} = -\frac{1}{2} \det \begin{bmatrix} -2 & 4 & -3 \\ 0 & 2 & -2 \\ 0 & 0 & 91 \end{bmatrix} = 182,$$

oder kürzer, unter Verwendung der Merkregel 5.32:

$$\det(A) = \det \begin{bmatrix} 4 & -2 & -3 \\ 2 & 0 & -2 \\ 25 & -11 & 26 \end{bmatrix} = 4 \cdot 25 + 6 \cdot 11 - 8 \cdot 11 + 4 \cdot 26 = 182.$$

6.12.9 Der LAPLACESCHE ENTWICKLUNGSSATZ Neben dem GAUSS-Verfahren gibt es eine rekursivee Methode, um Determinanten zu berechnen. Dabei wird die Determinante einer $n \times n$ -Matrix auf eine Summe über n Determinanten von $(n-1) \times (n-1)$ -Matrizen zurückgeführt. Der eigentliche Grund für diese Darstellung liegt in der Multilinearität der Determinante, die wir nach Satz 6.12.7 jetzt auch für die Zeilenvektoren zur Verfügung haben. Denken wir uns dazu die Matrix A durch Zeilenvektoren $a_i^t = [a_{i1}, a_{i2}, \dots, a_{in}] = \sum_{j=1}^n a_{ij} e_j^t$ gegeben. Benutzen wir die Linearität der Determinante für die i-te Zeile:

$$\det(A) = \det \begin{bmatrix} a_1^t \\ \vdots \\ a_i^t \\ \vdots \\ a_n^t \end{bmatrix} = \sum_{j=1}^n a_{ij} \det \begin{bmatrix} a_1^t \\ \vdots \\ e_j^t \\ \vdots \\ a_n^t \end{bmatrix}.$$

Das ist eigentlich schon der Entwicklungssatz, nur nicht in einer sonderlich anwenderfreundlichen Form. Wir müssen noch etwas nacharbeiten. Dazu schauen wir uns die Determinanten in der Summe genauer an:

$$\begin{aligned} \det \begin{bmatrix} a_1^t \\ \vdots \\ e_j^t \\ \vdots \\ a_n^t \end{bmatrix} &= \det \begin{bmatrix} a_{11} & \cdots & a_{1j-1} & a_{1j} & a_{1j+1} & \cdots & a_{1n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{i-11} & \cdots & a_{i-1j-1} & a_{i-1j} & a_{i-1j+1} & \cdots & a_{i-1n} \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ a_{i+11} & \cdots & a_{i+1j-1} & a_{i+1j} & a_{i+1j+1} & \cdots & a_{i+1n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{nj-1} & a_{nj} & a_{nj+1} & \cdots & a_{nn} \end{bmatrix} \\ &\stackrel{i)}{=} \det \begin{bmatrix} a_{11} & \cdots & a_{1j-1} & 0 & a_{1j+1} & \cdots & a_{1n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{i-11} & \cdots & a_{i-1j-1} & 0 & a_{i-1j+1} & \cdots & a_{i-1n} \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ a_{i+11} & \cdots & a_{i+1j-1} & 0 & a_{i+1j+1} & \cdots & a_{i+1n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{nj-1} & 0 & a_{nj+1} & \cdots & a_{nn} \end{bmatrix} \end{aligned}$$

$$\begin{aligned}
 & \stackrel{\text{ii)}}{=} \det \left[\begin{array}{ccccccc} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ a_{11} & \cdots & a_{1j-1} & 0 & a_{1j+1} & \cdots & a_{1n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{i-11} & \cdots & a_{i-1j-1} & 0 & a_{i-1j+1} & \cdots & a_{i-1n} \\ a_{i+11} & \cdots & a_{i+1j-1} & 0 & a_{i+1j+1} & \cdots & a_{i+1n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{nj-1} & 0 & a_{nj+1} & \cdots & a_{nn} \end{array} \right] \cdot (-1)^{i-1} \\
 & \stackrel{\text{iii)}}{=} \det \left[\begin{array}{c|cccccc} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & a_{11} & \cdots & a_{1j-1} & a_{1j+1} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & a_{i-11} & \cdots & a_{i-1j-1} & a_{i-1j+1} & \cdots & a_{i-1n} \\ 0 & a_{i+11} & \cdots & a_{i+1j-1} & a_{i+1j+1} & \cdots & a_{i+1n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & a_{n1} & \cdots & a_{nj-1} & a_{nj+1} & \cdots & a_{nn} \end{array} \right] \cdot (-1)^{i+j} \\
 & \stackrel{\text{iv)}}{=} \det \left[\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1j-1} & a_{1j+1} & \cdots & a_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{i-11} & \cdots & a_{i-1j-1} & a_{i-1j+1} & \cdots & a_{i-1n} \\ a_{i+11} & \cdots & a_{i+1j-1} & a_{i+1j+1} & \cdots & a_{i+1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{nj-1} & a_{nj+1} & \cdots & a_{nn} \end{array} \right] \cdot (-1)^{i+j} \\
 & =: (-1)^{i+j} M_{ij}.
 \end{aligned}$$

M_{ij} sind die sogenannten *Minoren*, oder *Unterdeterminanten* von A . Es sind die Determinanten der $(n-1) \times (n-1)$ -Matrizen, die aus A durch Streichen der i -ten Zeile und der j -ten Spalte entstehen.

Überlegen wir uns die einzelnen Schritte, die zu diesem Ergebnis geführt haben.

Zu i): Alle Pfade, die zur Determinante beitragen können, müssen die 1 an der Position (i, j) passieren. In keinem Summanden der Summendarstellung 6.42 kann also ein Faktor vorkommen, der oberhalb oder unterhalb der 1 in der j -ten Spalte steht. Wir ändern an dem Wert der Determinante nichts, wenn wir diese Einträge einfach durch 0 ersetzen.

Zu ii): Durch $i - 1$ Vertauschungen benachbarter Zeilen bringen wir die i -te Zeile e_j^t in die erste Zeile.

Zu iii): Weitere $j - 1$ Spaltenvertauschungen transportieren die j -te Spalte in die erste. Das Vorzeichen ändert sich dabei zu $(-1)^{j-1}(-1)^{i-1} = (-1)^{i+j}$.

Zu iv): Das ist nur noch die Anwendung von (6.44).

Die Entwicklung von $\det(A)$ nach der i -ten Zeile haben wir damit begründet. Man kann die Determinante aber auch nach der j -ten Spalte entwickeln. Es ist jedem nahegelegt, die geringfügigen Abänderungen unserer Überlegungen, die dafür nötig sind, selbst nachzuvollziehen.

6.12.10 Satz (LAPLACE scher Entwicklungssatz) *Die Determinante einer $n \times n$ -Matrix A lässt sich nach einer beliebigen Zeile oder Spalte entwickeln. Die Entwicklung nach der i -ten Zeile bzw. der*

j-ten Spalte ist respektive durch

$$\det(A) = \sum_{j=1}^n (-1)^{i+j} a_{ij} M_{ij}, \quad (6.47)$$

$$\det(A) = \sum_{i=1}^n (-1)^{i+j} a_{ij} M_{ij} \quad (6.48)$$

gegeben. Dabei sind die Unterdeterminanten M_{ij} die Determinanten derjenigen Matrizen, die aus A durch Streichen der i-ten Zeile und der j-ten Spalte entstehen.

6.12.11 Beispiel

Wir entwickeln nach der dritten Zeile:

$$\begin{aligned} \det \begin{bmatrix} 3^+2^-1^+4^- \\ 3^-3^+2^-0^+ \\ 3^+2^-0^+0^- \\ 2^-3^+7^-4^+ \end{bmatrix} &= + 3 \cdot \det \begin{bmatrix} 3 & 2 & 1 & 4 \\ 3 & 3 & 2 & 0 \\ \cancel{3} & \cancel{2} & 0 & 0 \\ 2 & 3 & 7 & 4 \end{bmatrix} - 2 \cdot \det \begin{bmatrix} 3 & 2 & 1 & 4 \\ 3 & 3 & 2 & 0 \\ 3 & 2 & 0 & 0 \\ 2 & 3 & 7 & 4 \end{bmatrix} \\ &= 3 \cdot \det \begin{bmatrix} 2 & 1 & 4 \\ 3 & 2 & 0 \\ 3 & 7 & 4 \end{bmatrix} - 2 \cdot \det \begin{bmatrix} 3 & 1 & 4 \\ 3 & 2 & 0 \\ 2 & 7 & 4 \end{bmatrix} \\ &= 3 \cdot \left(4 \cdot \det \begin{bmatrix} 3 & 2 \\ 3 & 7 \end{bmatrix} + 4 \cdot \det \begin{bmatrix} 2 & 1 \\ 3 & 2 \end{bmatrix} \right) - 2 \cdot \left(4 \cdot \det \begin{bmatrix} 3 & 2 \\ 2 & 7 \end{bmatrix} + 4 \cdot \det \begin{bmatrix} 3 & 1 \\ 3 & 2 \end{bmatrix} \right) \\ &= 3 \cdot (4 \cdot (21 - 6) + 4 \cdot (4 - 3)) - 2 \cdot (4 \cdot (21 - 4) + 4 \cdot (6 - 3)) = 32. \end{aligned}$$

Das Vorzeichen $(-1)^{i+j}$ der Unterdeterminante M_{ij} findet man, in dem man in der Matrix A die linke obere Position in Gedanken mit einem + markiert und dann die restlichen Felder des Quadrats schachbrettartig abwechselnd mit – und + versieht:

$$\begin{bmatrix} + & - & + & - \\ - & + & - & + \\ + & - & + & - \\ - & + & - & + \end{bmatrix}$$

An der Position (i, j) dieser *Vorzeichenmatrix* steht das Vorzeichen von M_{ij} .

Das Beispiel zeigt es schon. Man wird den Entwicklungssatz bei großen Matrizen nur dann anwenden wollen, wenn die Matrix *schwach besetzt* ist, d. h., wenn sie nicht allzuviiele Einträge hat, die von 0 verschieden sind. Ansonsten wird die Berechnung schnell sehr umfangreich: Die Determinante einer $n \times n$ -Matrix ergibt n Unterdeterminanten von $(n-1) \times (n-1)$ -Matrizen, die ihrerseits jeweils $n-1$ Unterdeterminanten von $(n-2) \times (n-2)$ -Matrizen erfordern, usw. Insgesamt sind das $n \cdot (n-1) \cdot (n-2) \cdots 2 = n!$ Summanden. Für eine 10×10 -Matrix könnte das bis auf 3628800 Summanden anwachsen.

6.12.12 Die Adjunkte und die inverse Matrix

Wir wenden den LAPLACESchen Entwicklungssatz auf die Matrix an, die aus A entsteht, wenn wir die i-te Zeile a_i^t durch die k-te Zeile a_k^t

ersetzen, mit einem k , das von i verschieden ist. Wir wissen, daß die Determinante dieser Matrix dann Null ergibt, da ja die i -te und die k -te Zeile gleich sind. Wir entwickeln nach der i -ten Zeile und erhalten gemäß (6.47):

$$0 = \sum_{j=1}^n (-1)^{i+j} a_{kj} M_{ij}.$$

Gleichung (6.47) und dieses Ergebnis läßt sich in einer Formel zusammenfassen:

$$\sum_{j=1}^n (-1)^{i+j} a_{kj} M_{ij} = \det(A) \cdot \delta_{ki}. \quad (6.49)$$

Die rechte Seite gibt die Einträge der mit $\det(A)$ multiplizierten Einheitsmatrix wieder. Erinnern wir uns daran, wie der (k, i) -Eintrag eines Matrixprodukts aussieht, nämlich $(AB)_{ki} = \sum_{j=1}^n a_{kj} b_{ji}$, dann erkennen wir, daß die linke Seite von (6.49), bis auf die Reihenfolge der Indizes im zweiten Faktor, diese Form hat. Wir definieren die sogenannte *Adjunkte* \mathcal{A} von A durch die Einträge

$$\mathcal{A}_{ij} := (-1)^{i+j} M_{ij}. \quad (6.50)$$

Es ist die Matrix, die alle Unterdeterminanten, multipliziert mit dem zugehörigen Vorzeichen aus der Vorzeichenmatrix enthält. Gleichung (6.49) gibt dann $(A\mathcal{A}^t)_{ki}$ wieder. Als Ergebnis haben wir

$$A\mathcal{A}^t = \det(A) \cdot \mathbb{1}. \quad (6.51)$$

Sollte also $\det(A) \neq 0$ sein, dann ist nach Satz 6.9.2 die inverse Matrix von A durch

$$A^{-1} = \frac{1}{\det(A)} \cdot \mathcal{A}^t \quad (6.52)$$

gegeben.

6.12.13 Beispiel Beginnen wir mit $A := \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Dann ist $\mathcal{A}_{11} = d$, $\mathcal{A}_{22} = a$, $\mathcal{A}_{12} = -c$ und $\mathcal{A}_{21} = -b$. Unter der Voraussetzung $\det(A) = ad - bc \neq 0$ ist

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \quad (6.53)$$

Zugegeben, dafür hätte man nicht (6.52) bemühen müssen, durch ein bisschen Nachdenken wäre man auch so darauf gekommen. Anders sieht es für eine 3×3 -Matrix aus. Hier kann es durchaus von Vorteil sein, für die inverse Matrix eine Formel zu haben, um etwa in zeitkritischen Rechenprozessen nicht mit dem GAUSS-Verfahren arbeiten zu müssen.

$$A := \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

hat die Adjunkte

$$\mathcal{A} = \begin{bmatrix} a_{22}a_{33} - a_{23}a_{32} & a_{23}a_{31} - a_{21}a_{33} & a_{21}a_{32} - a_{22}a_{31} \\ a_{13}a_{32} - a_{12}a_{33} & a_{11}a_{33} - a_{13}a_{31} & a_{12}a_{31} - a_{11}a_{32} \\ a_{12}a_{23} - a_{13}a_{22} & a_{13}a_{21} - a_{11}a_{23} & a_{11}a_{22} - a_{12}a_{21} \end{bmatrix}$$

und, falls $\det(\mathcal{A}) \neq 0$ ist, die Inverse

$$\mathcal{A}^{-1} = \frac{1}{\det(\mathcal{A})} \begin{bmatrix} a_{22}a_{33} - a_{23}a_{32} & a_{13}a_{32} - a_{12}a_{33} & a_{12}a_{31} - a_{13}a_{22} \\ a_{23}a_{31} - a_{21}a_{33} & a_{11}a_{33} - a_{13}a_{31} & a_{13}a_{21} - a_{11}a_{23} \\ a_{21}a_{32} - a_{22}a_{31} & a_{12}a_{31} - a_{11}a_{32} & a_{11}a_{22} - a_{12}a_{21} \end{bmatrix}. \quad (6.54)$$

Dann ist

$$\begin{bmatrix} 2 & 3 & -1 \\ 5 & 10 & 3 \\ 0 & 4 & 9 \end{bmatrix}^{-1} = \begin{bmatrix} 78 & -31 & 19 \\ -45 & 18 & -11 \\ 20 & -8 & 5 \end{bmatrix}, \quad \begin{bmatrix} 2 & 3i & -1 \\ 5 & 10 & 3 \\ 2 & 4 & 9 \end{bmatrix}^{-1} = \frac{4+3i}{975} \begin{bmatrix} 78 & -4-27i & 10+9i \\ -39 & 20 & -11 \\ 0 & -8+6i & 20-15i \end{bmatrix}.$$

An Gleichung (6.54) sieht man auch, daß die Inverse einer 4×4 -Matrix wohl die Grenze für eine programmierbare Formel darstellt.

6.12.14 Der Rang einer Matrix Die Überlegungen, die uns zum GAUSS-Verfahren zur Berechnung von Determinanten geführt haben, können wir verwenden, um den Begriff *Rang einer Matrix* zu klären. Darunter wollen wir die maximale Anzahl linear unabhängiger Spaltenvektoren einer beliebigen $m \times n$ -Matrix A verstehen. Vorläufig sprechen wir vorsichtshalber von dem Spaltenrang von A . Denn es gibt auch eine maximale Anzahl linear unabhängiger Zeilenvektoren, und wir wissen vorerst nicht, ob beide übereinstimmen. Da das Bild einer Matrix die lineare Hülle ihrer Spaltenvektoren ist, handelt es sich beim Spaltenrang einer Matrix um die Dimension ihres Bildes. Mit der Dimensionsformel haben wir ein einfaches Mittel, um über die Bestimmung von $\ker A$ mit dem GAUSS-Verfahren auf die Dimension des Bildes schließen zu können. Das lässt vermuten, daß dieses Verfahren den Spaltenrang einer Matrix nicht ändert. Sicherlich ändert er sich nicht bei Vertauschung von Spaltenvektoren, aber der wesentliche Bestandteil des GAUSS-Verfahrens besteht in elementaren *Zeileumformungen*. Hier ist es nicht mehr so offensichtlich, daß dabei der Spaltenrang, der ja die Spaltenvektoren betrifft, erhalten bleibt. Daß das doch der Fall ist, liegt an der Invertierbarkeit der quadratischen $m \times m$ -Matrizen $D_{ij}(\lambda)$ und T_{ij} in (6.45) bzw. (6.46), mit denen wir durch $D_{ij}(\lambda)A$ bzw. $T_{ij}A$ die Addition des λ -Fachen der j -ten Zeile zur i -ten bzw. die Vertauschung der j -ten mit der i -ten Zeile bewirken können. Um das einzusehen, definieren wir für eine endliche Menge $\mathcal{B} := \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ von Vektoren den Begriff *Rang* als die größtmögliche Anzahl von Elementen einer linear unabhängigen Teilmenge $\{\mathbf{b}_{\ell_1}, \dots, \mathbf{b}_{\ell_k}\} \subseteq \mathcal{B}$.

6.12.15 Lemma Es sei $T: V \rightarrow W$ eine invertierbare lineare Abbildung zwischen den Vektorräumen V und W .

- i) Die Mengen $\mathcal{B} := \{\mathbf{b}_1, \dots, \mathbf{b}_n\} \subset V$ und $T(\mathcal{B}) := \{T\mathbf{b}_1, \dots, T\mathbf{b}_n\} \subset W$ haben den gleichen Rang.

- ii) Der Rang von \mathcal{B} ist $\dim \text{lh}(\mathcal{B})$.
- iii) Addiert man zu einem Vektor \mathbf{b}_i von \mathcal{B} eine Linearkombination der übrigen und ersetzt den Vektor durch das Ergebnis, so ändert sich der Rang der Menge dabei nicht.

Beweis. i): Der Rang der Menge $\mathcal{B} := \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ sei k und $\mathcal{C} := \{\mathbf{b}_{\ell_1}, \dots, \mathbf{b}_{\ell_k}\}$ eine linear unabhängige Teilmenge von \mathcal{B} mit k Elementen. Nach Proposition 6.8.3 ist $T(\mathcal{C}) = \{T\mathbf{b}_{\ell_1}, \dots, T\mathbf{b}_{\ell_k}\}$ wieder linear unabhängig. Also ist der Rang von $T(\mathcal{B})$ sicher nicht kleiner als k . Wäre er größer, so gäbe es eine linear unabhängige Teilmenge $\mathcal{D} := \{T\mathbf{b}_{t_1}, \dots, T\mathbf{b}_{t_{k+1}}\}$ von $T(\mathcal{B})$. Nach Proposition 6.8.3 wäre dann aber $T^{-1}(\mathcal{D}) = \{\mathbf{b}_{t_1}, \dots, \mathbf{b}_{t_{k+1}}\}$ eine linear unabhängige Teilmenge von \mathcal{B} mit $k+1$ Elementen, im Widerspruch dazu, daß solche Teilmengen höchstens k Elemente haben können.

ii): Die Teilmenge \mathcal{C} von \mathcal{B} ist eine Basis für $\text{lh}(\mathcal{B})$, denn aus der Maximalität von \mathcal{C} als linear unabhängige Teilmenge von \mathcal{B} folgt, daß alle Vektoren in $\mathcal{B} \setminus \mathcal{C}$ Linearkombinationen von Vektoren aus \mathcal{C} sein müssen. Daher ist $\text{lh}(\mathcal{B}) \subseteq \text{lh}(\mathcal{C}) \subseteq \text{lh}(\mathcal{B})$, also $\text{lh}(\mathcal{B}) = \text{lh}(\mathcal{C})$. \mathcal{C} ist erzeugend und natürlich linear unabhängig, also eine Basis. Insbesondere ist der Rang von \mathcal{B} einfach die Dimension von $\text{lh}(\mathcal{B})$.

iii): Wenn wir (o. B. d. A.) \mathbf{b}_1 durch $\tilde{\mathbf{b}}_1 := \mathbf{b}_1 + \sum_{k=2}^n \lambda_k \mathbf{b}_k$ ersetzen, dann hat die Menge $\tilde{\mathcal{B}} := \{\tilde{\mathbf{b}}_1, \dots, \mathbf{b}_n\} \subset V$ jedenfalls keinen größeren Rang als \mathcal{B} , denn $\text{lh}(\tilde{\mathcal{B}}) \subseteq \text{lh}(\mathcal{B})$. War \mathbf{b}_1 kein Element der Basis \mathcal{C} , dann ist $\text{lh}(\mathcal{B}) = \text{lh}(\tilde{\mathcal{B}})$, denn beide Räume enthalten die Basis \mathcal{C} . Für $\mathbf{b}_1 \in \mathcal{C}$ dagegen ist $\tilde{\mathcal{C}} := \{\tilde{\mathbf{b}}_1, \dots, \mathbf{b}_{\ell_k}\}$ nach dem Basis-Austausch-Satz 6.5.3 wieder eine Basis für $\text{lh}(\mathcal{B})$, so daß auch jetzt $\text{lh}(\mathcal{B}) = \text{lh}(\tilde{\mathcal{B}})$ gilt. In jedem Fall stimmt der Rang von $\tilde{\mathcal{B}}$ und \mathcal{B} überein. \square

Dieses Lemma zeigt, daß sich der Spaltenrang k einer Matrix, also der Rang der Menge ihrer Spaltenvektoren, nicht ändert, wenn wir sie durch das GAUSS-Verfahren in eine Matrix mit maximaler Diagonalenlänge k verwandeln:

$$\begin{bmatrix} a_1 & * & \dots & * \\ a_2 & * & \dots & * \\ \ddots & \vdots & \dots & \vdots \\ a_k & * & \dots & * \\ 0 & 0 & & \end{bmatrix} \quad (6.55)$$

Wie sieht das für den Zeilenrang von A aus? Das GAUSS-Verfahren vertauscht Vektoren, was den Zeilenrang natürlich nicht ändert. Außerdem wird zu einem Zeilenvektor das Vielfache anderer Zeilenvektoren addiert. Auch das ändert den Zeilenrang nicht, wie Lemma 6.12.15 iii) lehrt. Bleibt noch die Auswirkung der Vertauschung von Spaltenvektoren auf den Zeilenrang. Wir gehen von A zu A^t über, um zu erkennen, daß wir diesen Fall schon behandelt haben. Jetzt geht es um die Spaltenvektoren von A^t bei Vertauschung zweier Zeilen. Das wird durch Multiplikation von A^t mit einer passenden Matrix T_{ij} erreicht, die den Spaltenrang nicht ändert. Im Ergebnis bleibt auch der Zeilenrang von A unter dem GAUSS-Verfahren konstant. Das bedeutet, daß wir den Spalten- und den Zeilenrang einer Matrix am Endergebnis (6.55) ablesen

können. Und hier ist klar zu erkennen, daß es genauso viel linear unabhängige Spaltenvektoren wie Zeilenvektoren gibt.

6.12.16 Satz Für eine beliebige $m \times n$ -Matrix A stimmt die maximale Anzahl linear unabhängiger Zeilenvektoren mit der maximalen Anzahl linear unabhängiger Spaltenvektoren überein. Diese Zahl wird Rang der Matrix, ran A, genannt.

6.12.17 A Überlegen Sie sich, daß das Ergebnis (6.54) für die Inverse einer 3×3 -Matrix auch direkt aus ein paar Grundprinzipien der Matrizenrechnung gewonnen werden kann. Dafür beschreiben Sie die Matrix A über ihre Zeilenvektoren $\mathbf{a}_1, \mathbf{a}_2$ und \mathbf{a}_3 :

$$A = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \mathbf{a}_3 \end{bmatrix}.$$

Für die inverse Matrix $A^{-1} = [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3]$ muß dann

$$\begin{bmatrix} \mathbf{a}_1 \mathbf{b}_1 & \mathbf{a}_1 \mathbf{b}_2 & \mathbf{a}_1 \mathbf{b}_3 \\ \mathbf{a}_2 \mathbf{b}_1 & \mathbf{a}_2 \mathbf{b}_2 & \mathbf{a}_2 \mathbf{b}_3 \\ \mathbf{a}_3 \mathbf{b}_1 & \mathbf{a}_3 \mathbf{b}_2 & \mathbf{a}_3 \mathbf{b}_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

gelten. Zeigen Sie, daß A^{-1} folgendermaßen berechnet werden kann:

$$A^{-1} = \frac{1}{\langle \mathbf{a}_1^t \times \mathbf{a}_2^t | \mathbf{a}_3^t \rangle} [\mathbf{a}_2^t \times \mathbf{a}_3^t, \mathbf{a}_3^t \times \mathbf{a}_1^t, \mathbf{a}_1^t \times \mathbf{a}_2^t]. \quad (6.56)$$

7 Eigenwerttheorie

7.1 Spektrum und Eigenvektoren

Eine lineare Abbildung A von einem Vektorraum V nach V ist meistens als quadratisches Zahlschema gegeben, dem man im Allgemeinen nicht direkt ansehen kann, wie sie denn genau wirkt. Die Bilder von Vektoren lassen sich leicht berechnen, doch kommt man auf diese Weise einem Verständnis der Abbildung normalerweise nicht näher: Vektoren werden eben gestreckt und gedreht und manche werden auf Null abgebildet und bilden den Kern der Abbildung. Man könnte also auf die Idee kommen, nach Richtungen zu suchen, die durch die Abbildung nicht verändert werden. Diese müssen durch Vektoren $x \neq \mathbf{0}$ beschrieben werden, die von A nur gestreckt werden, sagen wir um den Faktor λ :

$$Ax = \lambda x. \quad (7.1)$$

Einen solchen Vektor x nennt man *Eigenvektor* der Abbildung A . Der zugehörige Streckungsfaktor λ heißt *Eigenwert* von A . Gleichung (7.1) heißt *Eigenwertgleichung*. Die Menge $\sigma(A)$ aller Eigenwerte von A wird *Spektrum* genannt.

Es stellt sich die Frage, ob es überhaupt Eigenvektoren gibt und wenn ja, wie man sie findet. Einer bestimmten Sorte von Eigenvektoren sind wir schon begegnet, nämlich den Vektoren aus $\ker A$. Sie sind Eigenvektoren zum Eigenwert 0: $Ax = \mathbf{0} = 0 \cdot x$. Denken wir aber an eine ebene Drehung um einen festen Winkel, so können wir uns beim besten Willen keinen Vektor vorstellen, der von $\mathbf{0}$ verschieden ist und nur gestreckt wird. Der Grund ist, daß es auch keinen reellen Vektor mit dieser Eigenschaft gibt. Wir werden aber bald sehen, daß es für komplexe Vektorräume immer wenigstens einen Eigenvektor geben muß (manchmal aber auch nicht mehr).

Das Problem bei der Lösung von (7.1) ist, daß uns *zwei* Dinge fehlen, nämlich der Eigenvektor und der Eigenwert. Wäre uns der Eigenwert bekannt, dann ließe sich (7.1) in das Gleichungssystem $Ax - \lambda x = (A - \lambda \mathbb{1})x = \mathbf{0}$ umwandeln, das mit dem GAUSS-Verfahren zu lösen wäre. Hätten wir dagegen x , dann wäre λ leicht aus dem Vergleich von Ax mit x zu bestimmen. Um zu einem Lösungsverfahren zu kommen, müssen wir das Problem so umformen, daß nur noch der Eigenvektor oder der Eigenwert vorkommt. Dafür stehen uns aus der bisher entwickelten Theorie linearer Abbildungen alle Werkzeuge zur Verfügung:

$$\begin{aligned} & \exists_{\mathbf{0} \neq x} Ax = \lambda x \\ \Leftrightarrow & \exists_{\mathbf{0} \neq x} (A - \lambda \mathbb{1})x = \mathbf{0} \\ \Leftrightarrow & \ker(A - \lambda \mathbb{1}) \neq \{\mathbf{0}\} \\ \Leftrightarrow & A - \lambda \mathbb{1} \text{ ist nicht invertierbar} \end{aligned}$$

$$\Leftrightarrow \det(A - \lambda \mathbb{1}) = 0.$$

Die letzte Bedingung folgt aus Satz 6.12.6. Sie hängt nur noch von den möglichen Eigenwerten λ ab und eignet sich daher dazu, erst einmal alle Eigenwerte zu bestimmen. $\det(A - \lambda \mathbb{1})$ ist ein Polynom der Variablen λ . Es wird *charakteristisches Polynom* von A genannt und mit $\chi(\lambda)$ bezeichnet. Die Eigenwerte von A sind die Nullstellen des charakteristischen Polynoms. Hat man diese bestimmt, so lassen sich die zugehörigen Eigenvektoren x mit Hilfe des GAUSS-Verfahrens aus der homogenen Gleichung $(A - \lambda \mathbb{1})x = \mathbf{0}$ bestimmen. Da nach dem Fundamentalsatz der Algebra 11.8.23 jedes Polynom wenigstens eine Nullstelle haben muß (die komplex sein könnte), gibt es für jede lineare Abbildung wenigstens einen Eigenvektor.

7.1.1 Beispiel

Für $A := \begin{bmatrix} 11 & -2 \\ 3 & 4 \end{bmatrix}$ ist $A - \lambda \mathbb{1} = \begin{bmatrix} 11 & -2 \\ 3 & 4 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 11 - \lambda & -2 \\ 3 & 4 - \lambda \end{bmatrix}$ und $\chi(\lambda) = (11 - \lambda)(4 - \lambda) + 6 = \lambda^2 - 15\lambda + 50 = (\lambda - 10)(\lambda - 5)$. Daher ist $\sigma(A) = \{5, 10\}$. Der Eigenvektor x zum Eigenwert 5 bestimmt sich aus der Gleichung $(A - 5 \mathbb{1})x = \begin{bmatrix} 6 & -2 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

Das zugehörige GAUSS-Verfahren führt auf das Schema $\begin{array}{cc|c} 6 & -2 & 0 \\ 3 & -1 & 0 \end{array}$. Da wir $\lambda = 5$ gerade so bestimmt haben, daß für die Gleichung $(A - 5 \mathbb{1})x = \mathbf{0}$ eine nicht triviale Lösung x existiert, muß während des GAUSS-Verfahrens wenigstens eine der Zeilen verschwinden. In diesem einfachen Fall haben wir aber nur zwei. Wir können also einfach die zweite Zeile streichen und müssen nur noch $6x_1 - 2x_2 = 0$ lösen. Wir sind nur an einer Lösung interessiert. Daher wählen wir z. B. $x_2 = 3$ und finden damit $x_1 = 1$. Der Eigenvektor zum Eigenwert 5 ist $[1, 3]^t$. Genauso findet man den Eigenvektor $[2, 1]^t$ zum Eigenwert 10. Daß die beiden Eigenvektoren linear unabhängig sind ist hier kein Zufall, wie der folgende Satz beweist. \mathbb{R}^2 hat eine Basis aus Eigenvektoren von A . Daß das aber nicht immer so ist, zeigt die Abbildung $B := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Man überzeugt sich leicht von $\sigma(B) = \{0\}$. Ein Eigenvektor ist offensichtlich $x_1 = [1, 0]^t$. Gäbe es noch einen weiteren x_2 , so daß $\{x_1, x_2\}$ eine Basis bildet, dann würde jeder Vektor aus \mathbb{R}^2 auf Null abgebildet. B müßte die Nullabbildung sein, was offensichtlich nicht der Fall ist. Also hat B (bis auf Vielfache) nur den einen Eigenvektor x_1 .

7.1.2 Satz

Die Eigenvektoren zu verschiedenen Eigenwerten sind linear unabhängig.

Beweis. Wir führen den Beweis erst einmal nur für zwei verschiedene Eigenwerte $\lambda_1 \neq \lambda_2$ einer linearen Abbildung A . Die zugehörigen Eigenvektoren seien x_1 und x_2 . Für eine Nullkombination $\alpha_1 x_1 + \alpha_2 x_2 = \mathbf{0}$ folgt $(A - \lambda_1 \mathbb{1})(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1(Ax_1 - \lambda_1 x_1) + \alpha_2(Ax_2 - \lambda_1 x_2) = \alpha_2(\lambda_2 - \lambda_1)x_2 = \mathbf{0}$. Wegen $x_2 \neq \mathbf{0}$ und $\lambda_1 \neq \lambda_2$ muß $\alpha_2 = 0$ gelten. Das bedeutet $\alpha_1 x_1 = \mathbf{0}$. Wegen $x_1 \neq \mathbf{0}$ folgt auch $\alpha_1 = 0$, d. h. die Nullkombination ist trivial. $\{x_1, x_2\}$ ist daher linear unabhängig.

Jetzt der allgemeine Fall: $Ax_i = \lambda_i x_i$ und $\lambda_i \neq \lambda_j$ für $i \neq j$, $i, j \in \{1, \dots, n\}$. $\sum_{i=1}^n \alpha_i x_i = \mathbf{0}$

sei eine Nullkombination der Eigenvektoren. Dann folgt für ein beliebiges $k \in \{1, \dots, n\}$

$$\prod_{j \neq k} (A - \lambda_j \mathbb{1}) \sum_{i=1}^n \alpha_i x_i = \sum_{i=1}^n \alpha_i \prod_{j \neq k} (A - \lambda_j \mathbb{1}) x_i = \alpha_k \prod_{j \neq k} (\lambda_k - \lambda_j) x_k = \mathbf{0}.$$

Das bedeutet $\alpha_k = 0$ für jedes $k \in \{1, \dots, n\}$, d. h. die Nullkombination ist trivial. Somit ist $\{x_1, \dots, x_n\}$ linear unabhängig. \square

7.2 Selbstdjungierte lineare Abbildungen

7.2.1 Definition Sei V ein endlich dimensionaler Vektorraum über \mathbb{K} , versehen mit einem Skalarprodukt $\langle \cdot | \cdot \rangle$ und A eine lineare Abbildung von V nach V mit der Eigenschaft

$$\langle x | A y \rangle = \langle Ax | y \rangle \quad (7.2)$$

für alle $x, y \in V$. Dann heißt A selbstdjungiert.

Ist die lineare Abbildung A durch eine quadratische Matrix auf $V = \mathbb{K}^n$ gegeben, so ist A genau dann selbstdjungiert, wenn

$$A = A^* \quad (7.3)$$

gilt (vergl. Satz 6.10.2). Insbesondere sind also die Diagonalelemente a_{ii} von A reell.

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ \overline{a_{12}} & a_{22} & a_{23} & \cdots & a_{2n} \\ \overline{a_{13}} & \overline{a_{23}} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \overline{a_{1n}} & \overline{a_{2n}} & \overline{a_{3n}} & \cdots & a_{nn} \end{bmatrix}.$$

Für die folgenden Untersuchungen gehen wir generell von einem Vektorraum V mit Skalarprodukt $\langle \cdot | \cdot \rangle$ aus.

7.2.2 Lemma Für eine selbstdjungierte lineare Abbildung ist das Spektrum reell.

Beweis. Es sei $x \neq \mathbf{0}$ ein Eigenvektor zum Eigenwert $\lambda \in \mathbb{C}$. Dann gilt

$$\begin{aligned} \langle x | Ax \rangle &= \langle x | \lambda x \rangle = \lambda \langle x | x \rangle = \\ \langle Ax | x \rangle &= \langle \lambda x | x \rangle = \bar{\lambda} \langle x | x \rangle. \end{aligned}$$

Das bedeutet $(\lambda - \bar{\lambda}) \langle x | x \rangle = 0$. Wegen der positiven Definitheit des Skalarprodukts ist $\langle x | x \rangle \neq 0$, so daß nur $\lambda - \bar{\lambda} = 0$ bleibt. Das heißt aber $\lambda \in \mathbb{R}$. Das Spektrum $\text{sp}(A)$ ist also in \mathbb{R} enthalten. \square

7.2.3 Lemma Für eine selbstdjungierte lineare Abbildung stehen die Eigenvektoren zu verschiedenen Eigenwerten senkrecht aufeinander.

Beweis. Es seien $\lambda \neq \gamma$ zwei Eigenwerte zu den Eigenvektoren \mathbf{x} bzw. \mathbf{y} . Dann gilt

$$\begin{aligned}\langle \mathbf{x} | A\mathbf{y} \rangle &= \langle \mathbf{x} | \gamma\mathbf{y} \rangle = \gamma\langle \mathbf{x} | \mathbf{y} \rangle = \\ \langle A\mathbf{x} | \mathbf{y} \rangle &= \langle \lambda\mathbf{x} | \mathbf{y} \rangle = \lambda\langle \mathbf{x} | \mathbf{y} \rangle.\end{aligned}$$

Daraus folgt $(\lambda - \gamma)\langle \mathbf{x} | \mathbf{y} \rangle = 0$. Da $\lambda - \gamma \neq 0$ gilt, muß $\langle \mathbf{x} | \mathbf{y} \rangle = 0$ erfüllt sein, d. h. \mathbf{x} und \mathbf{y} sind zueinander orthogonal. \square

7.2.4 Theorem Für jede selbstadjungierte lineare Abbildung A auf einem endlichdimensionalen Vektorraum V gibt es eine ONB aus Eigenvektoren von A .

Beweis. Wir führen den Beweis durch Induktion nach der Dimension n von V . Der Fall $n = 1$ ist trivial.

$n \rightarrow n + 1$: Es gilt $\dim V = n + 1$. Ist $A = 0$, so leistet jede ONB von V das Gewünschte. Sei also $A \neq 0$. Dann gibt es wenigstens einen Eigenwert λ von A , denn das charakteristische Polynom von A hat in \mathbb{C} wenigstens eine Nullstelle (Fundamentalsatz der Algebra 11.8.23). Also gibt es einen Eigenvektor \mathbf{x} ($\|\mathbf{x}\| = 1$) zu λ . Wir definieren den Vektorraum $V_1 := \mathbf{x}^\perp$ als den zu \mathbf{x} orthogonalen Teilraum von V . Nach Proposition 6.7.6 hat V_1 die Dimension n . Die Beweisidee besteht nun darin, nachzuweisen, daß A den Vektorraum V_1 in sich abbildet und daß A auf V_1 wieder selbstadjungiert ist. Dann besagt die Induktionsvoraussetzung, daß wir eine ONB $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ von V_1 aus Eigenvektoren von A haben, die zusammen mit \mathbf{x} eine ONB aus Eigenvektoren für ganz V ergibt.

Sei also $\mathbf{y} \in V_1$, dann ist $A\mathbf{y} \in V_1$, also $A\mathbf{y} \perp \mathbf{x}$ nachzuweisen. Das folgt aber sofort aus

$$\langle \mathbf{x} | A\mathbf{y} \rangle = \langle A\mathbf{x} | \mathbf{y} \rangle = \lambda\langle \mathbf{x} | \mathbf{y} \rangle = 0.$$

Die Selbstadjungiertheit von A auf V_1 ist nun klar, da $\langle \mathbf{y} | Az \rangle = \langle A\mathbf{y} | z \rangle$ für alle $\mathbf{y}, z \in V_1$, erst recht also auch für alle $\mathbf{y}, z \in V$ gilt. \square

7.2.5 A Bestimmen Sie die Spektren und die zugehörigen Eigenvektoren der sogenannten PAULI-Matrizen

$$\sigma_1 := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \sigma_2 := \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad \sigma_3 := \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \quad (7.4)$$

7.2.6 Definition Eine selbstadjungierte lineare Abbildung P mit der Eigenschaft

$$P^2 = P \quad (7.5)$$

heißt (orthogonale) Projektion oder Projektor. $[P] := \text{im } P$ bezeichnet ihren Projektionsraum.

Offensichtlich ist mit P auch $\mathbb{1} - P$ eine Projektion, denn $\mathbb{1} - P$ ist selbstadjungiert und es gilt $(\mathbb{1} - P)^2 = \mathbb{1} - 2P + P^2 = \mathbb{1} - P$. Es gibt allgemeinere Projektionen, nämlich lineare Abbildungen P , die zwar die Projektionseigenschaft $P^2 = P$ haben, aber nicht selbstadjungiert sind. Da

wir uns hier nur für orthogonale Projektionen interessieren, sprechen wir einfach von Projektionen und meinen damit orthogonale Projektionen. Wegen

$$(\mathbb{1} - P)P = P - P^2 = \mathbb{O} = P(\mathbb{1} - P)$$

wird $\mathbb{1} - P$ als die *zu P orthogonale Projektion* bezeichnet und mitunter auch P^\perp geschrieben. Die Projektionsräume $[P]$ und $[\mathbb{1} - P]$ sind nämlich orthogonal zueinander: Für ein $x \in [P]$ und ein $y \in [\mathbb{1} - P]$ gilt $x = Px$ und $y = (\mathbb{1} - P)y$ (warum?). Das hat

$$\langle x | y \rangle = \langle Px | (\mathbb{1} - P)y \rangle = \langle x | P(\mathbb{1} - P)y \rangle = 0,$$

also $Px \perp Py$ zur Folge. Das zeigt bereits $[P] \perp [\mathbb{1} - P]$. Tatsächlich gilt sogar noch mehr:

7.2.7 Lemma *Für eine Projektion P gilt $[P]^\perp = [\mathbb{1} - P]$.*

$[P]$ und $[\mathbb{1} - P]$ sind also nicht nur orthogonal zueinander, sondern $[\mathbb{1} - P]$ ist bereits der *ganze* zu $[P]$ orthogonale Teilraum.

Beweis. Da wir $[P] \perp [\mathbb{1} - P]$ bereits wissen, haben wir schon $[\mathbb{1} - P] \subseteq [P]^\perp$. Um die umgekehrte Inklusion zu zeigen, sei $y \in [P]^\perp$. Dann gilt $\langle y | Px \rangle = \langle Py | x \rangle = 0$ für alle $x \in V$. Nach Übung 5.5.12 ist $Py = \mathbf{0}$, also $(\mathbb{1} - P)y = y$. Daher liegt y in $[\mathbb{1} - P]$. Das zeigt $[P]^\perp \subseteq [\mathbb{1} - P]$. \square

Die Aussage dieses Lemmas lässt sich auch so formulieren:

$$[P]^\perp = [P^\perp]. \quad (7.6)$$

Der zum Projektionsraum $[P]$ orthogonale Teilraum $[P]^\perp$ ist der Projektionsraum $[\mathbb{1} - P] = [P^\perp]$ der zu P orthogonalen Projektion P^\perp . Dieser Sachverhalt rechtfertigt zur Genüge die Namensgebung für $\mathbb{1} - P$.

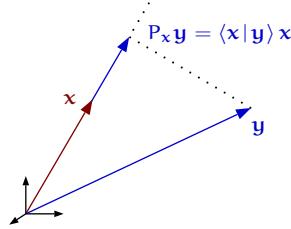
Offensichtlich ist auch $\mathbb{1}$ und \mathbb{O} eine Projektion. Projektionen, die von diesen verschieden sind, heißen *nicht trivial*.

7.2.8 Lemma *Das Spektrum einer nicht trivialen Projektion ist immer die Menge $\{0, 1\}$. Der Eigenraum zum Eigenwert 1 ist der Projektionsraum von P und der zum Eigenwert 0 ist der Projektionsraum von $\mathbb{1} - P$.*

Beweis. Ist $P \neq \mathbb{0}$ und $P \neq \mathbb{1}$, so gibt es immer einen Vektor $z \neq \mathbf{0}$ im Projektionsraum und einen dazu orthogonalen $y \neq \mathbf{0}$ (warum?). Damit folgt $Pz = z$ und $(\mathbb{1} - P)y = y$, also $Py = \mathbf{0}$. Folglich liegen schon 1 und 0 im Spektrum von P : $\{0, 1\} \subseteq \text{sp}(P)$. Bleibt zu zeigen, daß es keine weiteren Elemente haben kann. Aus $Pz = \lambda z$ für ein $z \neq \mathbf{0}$ folgt nämlich $\lambda z = Pz = PPz = \lambda Pz = \lambda^2 z$, d. h. $\lambda^2 = \lambda$. Da λ als Eigenwert einer selbstadjungierten Abbildung reell sein muß, bleiben nur die Möglichkeiten $\lambda = 0$, oder $\lambda = \pm 1$. Aber $Pz = -z$ führt auf den Widerspruch

$$0 < \langle z | z \rangle = \langle Pz | Pz \rangle = \langle z | Pz \rangle = -\langle z | z \rangle < 0.$$

Damit haben wir auch $\text{sp}(P) \subseteq \{0, 1\}$. Der Rest ist klar. \square



Die Dimension des Projektionsraumes $[P]$ nennt man auch *Dimension der Projektion*. Am leichtesten sind eindimensionale Projektionen zu bestimmen. Ihr Projektionsraum wird durch einen normierten Vektor x aufgespannt. Die zugehörige Projektion bezeichnen wir mit P_x . Wenn wir uns daran erinnern, daß das Skalarprodukt $\langle x|y \rangle$ gerade die Länge der Projektion von y auf die Richtung x beschreibt, ist die Wirkung von P_x leicht anzugeben:

$$P_x y = \langle x|y \rangle x. \quad (7.7)$$

Um die Matrix für P_x zu finden, gehen wir von der Koordinatendarstellung $x = [x_1, \dots, x_n]^t \in \mathbb{C}^n$ aus:

$$\begin{aligned} P_x y &= \begin{bmatrix} x_1 \langle x|y \rangle \\ x_2 \langle x|y \rangle \\ \vdots \\ x_n \langle x|y \rangle \end{bmatrix} = \begin{bmatrix} x_1 \bar{x_1} y_1 + x_1 \bar{x_2} y_2 + x_1 \bar{x_3} y_3 + \cdots + x_1 \bar{x_n} y_n \\ x_2 \bar{x_1} y_1 + x_2 \bar{x_2} y_2 + x_2 \bar{x_3} y_3 + \cdots + x_2 \bar{x_n} y_n \\ \vdots \\ x_n \bar{x_1} y_1 + x_n \bar{x_2} y_2 + x_n \bar{x_3} y_3 + \cdots + x_n \bar{x_n} y_n \end{bmatrix} \\ &= \begin{bmatrix} x_1 \bar{x_1} & x_1 \bar{x_2} & x_1 \bar{x_3} & \cdots & x_1 \bar{x_n} \\ x_2 \bar{x_1} & x_2 \bar{x_2} & x_2 \bar{x_3} & \cdots & x_2 \bar{x_n} \\ \vdots & \vdots & \vdots & & \vdots \\ x_n \bar{x_1} & x_n \bar{x_2} & x_n \bar{x_3} & \cdots & x_n \bar{x_n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \end{aligned}$$

Damit ist die Matrix bestimmt:

$$P_x = \begin{bmatrix} x_1 \bar{x_1} & x_1 \bar{x_2} & x_1 \bar{x_3} & \cdots & x_1 \bar{x_n} \\ x_2 \bar{x_1} & x_2 \bar{x_2} & x_2 \bar{x_3} & \cdots & x_2 \bar{x_n} \\ \vdots & \vdots & \vdots & & \vdots \\ x_n \bar{x_1} & x_n \bar{x_2} & x_n \bar{x_3} & \cdots & x_n \bar{x_n} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}^* = [x_i \bar{x_j}]_{i,j=1,\dots,n}. \quad (7.8)$$

7.2.9 A Zeigen Sie, daß durch Gleichung (7.7) tatsächlich eine Projektion definiert wird.

7.2.10 Lemma Zu jedem Teilraum T von V gibt es genau eine Projektion P , die T als Projektionsraum hat: $T = [P]$. Wir nennen P den Projektor auf den Teilraum T .

Die Projektoren auf die trivialen Teilräume $T = \mathbf{0}$ und $T = V$ sind natürlich $P = 0$ bzw. $P = \mathbf{1}$.

Beweis. T ist Teilvektorraum von V und hat daher nach Satz 6.5.5 eine Basis, die nach dem GRAM-SCHMIDTSchen Orthonormalisierungsverfahren 6.5.16 zu einer ONB $\{\mathbf{b}_1, \dots, \mathbf{b}_k\}$ gemacht werden kann. Jetzt läßt sich P durch

$$P \mathbf{x} := \sum_{i=1}^k \langle \mathbf{x} | \mathbf{b}_i \rangle \mathbf{b}_i, \quad (7.9)$$

$\mathbf{x} \in V$, definieren. Wir rechnen die Projektionseigenschaft nach:

$$\begin{aligned} P^2\mathbf{x} &= P(P\mathbf{x}) = \sum_{i=1}^k \langle P\mathbf{x} | \mathbf{b}_i \rangle \mathbf{b}_i = \sum_{i=1}^k \sum_{j=1}^k \langle \langle \mathbf{x} | \mathbf{b}_j \rangle \mathbf{b}_j | \mathbf{b}_i \rangle \mathbf{b}_i = \sum_{i=1}^k \sum_{j=1}^k \langle \mathbf{x} | \mathbf{b}_j \rangle \langle \mathbf{b}_j | \mathbf{b}_i \rangle \mathbf{b}_i \\ &= \sum_{i=1}^k \sum_{j=1}^k \langle \mathbf{x} | \mathbf{b}_j \rangle \delta_{ji} \mathbf{b}_i = \sum_{i=1}^k \langle \mathbf{x} | \mathbf{b}_i \rangle \mathbf{b}_i = P\mathbf{x}. \end{aligned}$$

Das zeigt $P^2 = P$. Bleibt noch $P = P^*$ zu zeigen: Für alle $\mathbf{x}, \mathbf{y} \in V$ gilt

$$\begin{aligned} \langle \mathbf{x} | P\mathbf{y} \rangle &= \left\langle \mathbf{x} \middle| \sum_{i=1}^k \langle \mathbf{y} | \mathbf{b}_i \rangle \mathbf{b}_i \right\rangle = \sum_{i=1}^k \overline{\langle \mathbf{y} | \mathbf{b}_i \rangle} \langle \mathbf{x} | \mathbf{b}_i \rangle = \sum_{i=1}^k \langle \mathbf{x} | \mathbf{b}_i \rangle \langle \mathbf{b}_i | \mathbf{y} \rangle = \left\langle \sum_{i=1}^k \langle \mathbf{x} | \mathbf{b}_i \rangle \mathbf{b}_i \middle| \mathbf{y} \right\rangle \\ &= \langle P\mathbf{x} | \mathbf{y} \rangle. \end{aligned}$$

Für den Projektor P gilt laut Konstruktion $P\mathbf{x} \in T$, d.h., wir haben bereits $[P] \subseteq T$. Jeder Vektor $\mathbf{x} \in T$ hat eine Basisdarstellung $\mathbf{x} = \sum_{i=1}^k \lambda_i \mathbf{b}_i$. Wegen $P\mathbf{b}_j = \sum_{i=1}^k \langle \mathbf{b}_j | \mathbf{b}_i \rangle \mathbf{b}_i = \sum_{i=1}^k \delta_{ji} \mathbf{b}_i = \mathbf{b}_j$ folgt sofort $P\mathbf{x} = \mathbf{x} \in [P]$. Das zeigt die andere Inklusion $T \subseteq [P]$ und damit $T = [P]$. \square

7.2.11 Definition

Eine selbstdrajngierte Abbildung A heißt positiv, wenn für alle $\mathbf{x} \in V$

$$\langle \mathbf{x} | A\mathbf{x} \rangle \geq 0 \quad (7.10)$$

gilt. Dafür schreibt man auch $A \geq 0$.

7.2.12 Satz

Eine selbstdrajngierte Abbildung A ist genau dann positiv, wenn $\sigma(A) \subseteq \mathbb{R}_0^+$ gilt.

Beweis. Zunächst sei $A \geq 0$ und \mathbf{x} ein normierter Eigenvektor zum Eigenwert λ . Dann gilt $0 \leq \langle \mathbf{x} | A\mathbf{x} \rangle = \langle \mathbf{x} | \lambda\mathbf{x} \rangle = \lambda \langle \mathbf{x} | \mathbf{x} \rangle = \lambda$. Das zeigt $\text{sp}(A) \subset \mathbb{R}_0^+$. Gehen wir umgekehrt von dieser Eigenschaft aus, dann sind also alle Eigenwerte $\lambda_1, \dots, \lambda_n$ nicht negativ. $\mathcal{B} := \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ sei eine ONB aus Eigenvektoren. Dann lässt sich jeder Vektor \mathbf{x} nach dieser Basis entwickeln: $\mathbf{x} = \sum_{k=1}^n x_k \mathbf{b}_k$. Es folgt

$$\begin{aligned} \langle \mathbf{x} | A\mathbf{x} \rangle &= \left\langle \sum_{k=1}^n x_k \mathbf{b}_k \middle| \sum_{l=1}^n x_l A\mathbf{b}_l \right\rangle = \left\langle \sum_{k=1}^n x_k \mathbf{b}_k \middle| \sum_{l=1}^n x_l \lambda_l \mathbf{b}_l \right\rangle \\ &= \sum_{k=1}^n \sum_{l=1}^n \overline{x_k} x_l \lambda_l \langle \mathbf{b}_k | \mathbf{b}_l \rangle = \sum_{k=1}^n |x_k|^2 \lambda_k \geq 0. \quad \square \end{aligned}$$

Dieses Ergebnis wird zur Identifikation lokaler Extrema in der mehrdimensionalen Analysis gebraucht. Insbesondere ist jede Projektion eine positive Abbildung, denn ihr Spektrum ist in $\{0, 1\} \subset \mathbb{R}_0^+$ enthalten.

7.3 Funktionalkalkül

7.3.1 A Zeigen Sie:

- i) Für jede lineare Abbildung A ist die Abbildung A^*A positiv.
- ii) A ist genau dann positiv, wenn es eine positive Abbildung W mit der Eigenschaft $A = W^2$ gibt. W wird als *Wurzel* von A bezeichnet.

7.3.2 A (*) Zeigen Sie, daß auf Vektorräumen V über \mathbb{C} eine lineare Abbildung A genau dann selbstadjungiert ist, wenn $\langle x|Ax \rangle$ für alle $x \in V$ reell ist. Gehen Sie dabei folgendermaßen vor:

- i) Zeigen Sie, daß für jede lineare Abbildung A die *sesquilineare Abbildung* $V \times V \ni (x, y) \mapsto \langle x|Ay \rangle$ eine *Polarisationsgleichung* erfüllt:

$$\begin{aligned} \langle x|Ay \rangle &= \frac{1}{4} [\langle x+y|A(x+y) \rangle - \langle x-y|A(x-y) \rangle \\ &\quad - i(\langle x+iy|A(x+iy) \rangle - \langle x-iy|A(x-iy) \rangle)]. \end{aligned}$$

- ii) Verwenden Sie Identitäten der Art $\langle x+iy|A(x+iy) \rangle = \langle y-ix|A(y-ix) \rangle$ etc., um aus i) auf $\langle x|Ay \rangle = \overline{\langle y|Ax \rangle}$ zu schließen.

7.3.3 Theorem Sind A und B selbstadjungierte Abbildungen auf dem Vektorraum V , die miteinander vertauschen ($AB = BA$), so gibt es eine gemeinsame ONB aus Eigenvektoren von A und B für V .

Beweis. $V_{\lambda_1} := \{y \in V \mid Ay = \lambda_1 y\}$ sei der Eigenraum von A zum Eigenwert λ_1 (also die Menge aller Vektoren, die Eigenvektor zum selben Eigenwert λ_1 sind). Ist $\lambda_2 \neq \lambda_1$ ein weiterer Eigenwert von A , so sind die zugehörigen Eigenräume V_{λ_1} und V_{λ_2} nach Lemma 7.2.3 orthogonal zueinander (d. h. jeder Vektor aus V_{λ_1} ist zu jedem Vektor aus V_{λ_2} orthogonal). Die Abbildung B lässt den Teilraum V_{λ_1} invariant, d. h. es gilt $BV_{\lambda_1} \subseteq V_{\lambda_1}$: Für alle $y \in V_{\lambda_1}$ gilt

$$ABy = BAY = B\lambda_1 y = \lambda_1 BY.$$

By ist also ebenfalls ein Eigenvektor von A zum Eigenwert λ_1 und liegt somit wieder in V_{λ_1} . Damit ist B eine selbstadjungierte Abbildung auf dem Vektorraum V_{λ_1} und hat nach Theorem 7.2.4 eine ONB aus Eigenvektoren für V_{λ_1} . Auf V_{λ_1} gibt es demnach eine *gemeinsame* ONB aus Eigenvektoren von A und B . Diese Überlegungen gelten natürlich für alle Eigenräume $V_{\lambda_1}, V_{\lambda_2}, \dots, V_{\lambda_k}$ zu den verschiedenen Eigenwerten $\lambda_1, \lambda_2, \dots, \lambda_k$ von A . Auf jedem dieser paarweise orthogonalen Räumen gibt es eine gemeinsame ONB aus Eigenvektoren von A und B . Zusammengenommen erhalten wir ein Orthonormalsystem aus gemeinsamen Eigenvektoren. Wir müssen uns nur noch überlegen, daß dieses System für V erzeugend ist, um die behauptete gemeinsame ONB nachgewiesen zu haben. Das liegt an Theorem 7.2.4, nach dem es eine ONB aus Eigenvektoren von A für V gibt. Gruppieren wir diese nach den verschiedenen Eigenwerten $\lambda_1, \lambda_2, \dots, \lambda_k$, so erzeugen sie jeweils die Räume $V_{\lambda_1}, V_{\lambda_2}, \dots, V_{\lambda_k}$. Daher ist der Aufspann dieser Räume und damit auch der Aufspann des gemeinsamen Orthonormalsystems erzeugend für V . \square

7.4 Normale Abbildungen

7.4.1 Definition Eine lineare Abbildung T auf einem Vektorraum heißt normal, wenn T mit T^* vertauscht.

Natürlich ist jede selbstadjungierte Abbildung normal, die Umkehrung gilt aber nicht. Die Abbildung

$$U := \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}.$$

z. B. ist normal, denn $UU^* = \mathbb{1} = U^*U$, aber $U \neq U^*$.

7.4.2 Korollar Für jede normale Abbildung auf einem komplexen Vektorraum V gibt es eine ONB aus Eigenvektoren von T .

Beweis. Da T mit T^* vertauscht, gilt das auch für die selbstadjungierten Abbildungen

$$R := \frac{1}{2}(T + T^*) \quad \text{und} \quad I := \frac{1}{2i}(T - T^*).$$

Nach Theorem 7.3.3 gibt es für V eine gemeinsame Eigenvektorbasis $\mathcal{B} := \{ \mathbf{b}_1, \dots, \mathbf{b}_n \}$ für R und I zu den resp. Eigenwerten $\lambda_1, \dots, \lambda_n$ bzw. $\gamma_1, \dots, \gamma_n$. Aus

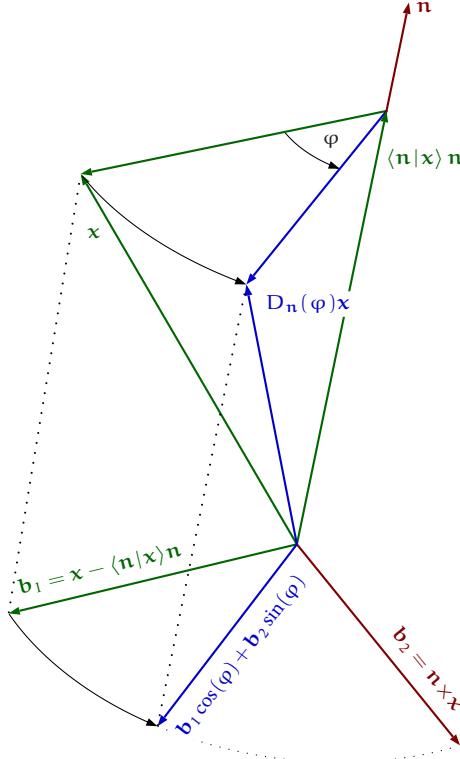
$$T = R + iI$$

folgt $T\mathbf{b}_k = R\mathbf{b}_k + iI\mathbf{b}_k = (\lambda_k + i\gamma_k)\mathbf{b}_k$. \mathcal{B} ist daher auch eine ONB aus Eigenvektoren für T (zu den komplexen Eigenwerten $\lambda_1 + i\gamma_1, \dots, \lambda_n + i\gamma_n$). \square

Die wichtigsten normalen Abbildungen sind neben den selbstadjungierten die *unitären* Abbildungen, auf reellen Vektorräumen auch als *orthogonale* Abbildungen bezeichnet (Definition 6.10.4). Für unitäre Abbildungen U existiert die Inverse und diese ist durch die Adjungierte U^* gegeben (vergl. Korollar 6.10.3). Als normale Abbildung hat U eine ONB aus Eigenvektoren. Ist U durch eine Matrix gegeben, dann stehen in ihren Spalten die Bilder $\mathbf{u}_1, \dots, \mathbf{u}_n$ der kanonischen Basisvektoren $\mathbf{e}_1, \dots, \mathbf{e}_n$. Im Hinblick auf die Winkeltreue von U ist klar, dass $\mathcal{B} := \{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ wieder eine ONB für V bestimmt. Das lässt zwei Interpretationen für die Wirkungsweise von U zu. Zunächst die *passive*, die U als eine Koordinatentransformation von Darstellungen bzgl. der Basis \mathcal{B} in die kanonische Basis ansieht. Oder die *aktive*, für die wir allerdings reelle Vektorräume voraussetzen, in der U eine winkel- und längentreue Abbildung ist, die auf den Basisvektoren \mathbf{e}_k wie $U\mathbf{e}_k = \mathbf{u}_k$ wirkt, die also ein Achsendreibein $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ (im Falle $\dim V = 3$) wieder in ein solches $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ überführt. Wenn letzteres ein Rechtssystem ist ($\det(U) = 1$), handelt es sich um eine *Drehung*. Ist es ein Linkssystem ($\det(U) = -1$), dann ist die Drehung mit einer *Spiegelung* (einer reellen selbstadjungierten linearen Abbildung S mit der Eigenschaft $S^2 = \mathbb{1}$) kombiniert.

7.4.3 A Zeigen Sie mit Hilfe der folgenden Skizze, daß sich eine Drehung um den Winkel φ , mit einer Drehachse, die durch den normierten Richtungsvektor $\mathbf{n} \in \mathbb{R}^3$ gegeben ist, durch die Drehmatrix $D_{\mathbf{n}}(\varphi)$ beschreiben läßt:

$$D_{\mathbf{n}}(\varphi) = P_{\mathbf{n}} + (1 - P_{\mathbf{n}}) \cos(\varphi) + R_{\mathbf{n}} \sin(\varphi). \quad (7.11)$$



Dabei ist $P_{\mathbf{n}}$ die eindimensionale Projektion auf die Richtung \mathbf{n} gemäß (7.8) und $R_{\mathbf{n}}$ die Matrix, die zum Kreuzprodukt mit \mathbf{n} gehört: $R_{\mathbf{n}}\mathbf{x} := \mathbf{n} \times \mathbf{x}$.

i) Zeigen Sie dafür zunächst

$$R_{\mathbf{n}} = \begin{bmatrix} 0 & -n_3 & n_2 \\ n_3 & 0 & -n_1 \\ -n_2 & n_1 & 0 \end{bmatrix}. \quad (7.12)$$

ii) Zeigen Sie $\|\mathbf{b}_1\| = \|\mathbf{b}_2\|$, $\mathbf{b}_1 \perp \mathbf{b}_2$, $\mathbf{b}_1 \perp \mathbf{n}$ und $\mathbf{b}_2 \perp \mathbf{n}$. Folgern Sie daraus, daß $\mathbf{b}_1 \cos(\varphi) + \mathbf{b}_2 \sin(\varphi)$ der um den Winkel φ in Richtung \mathbf{b}_2 gedrehte Vektor \mathbf{b}_1 ist.

iii) Entnehmen Sie der Skizze

$$D_{\mathbf{n}}(\varphi)\mathbf{x} = P_{\mathbf{n}}\mathbf{x} + \mathbf{b}_1 \cos(\varphi) + \mathbf{b}_2 \sin(\varphi)$$

und folgern Sie daraus (7.11).

iv) Bestimmen Sie für die kanonischen Basisvektoren \mathbf{e}_1 , \mathbf{e}_2 und \mathbf{e}_3 die Drehungen $D_{\mathbf{e}_1}(\varphi)$, $D_{\mathbf{e}_2}(\varphi)$ und $D_{\mathbf{e}_3}(\varphi)$.

- v) Bestimmen Sie $D_n(\frac{\pi}{3})$ für $n := \frac{1}{3}[1, 2, 2]^t$. Drehen Sie damit den Vektor $x := [2, 3, 6]^t$. Welche Länge hat $D_n(\frac{\pi}{3})x$?

7.4.4 A Zeigen Sie: Das Spektrum einer unitären Matrix besteht aus Zahlen des komplexen Einheitskreises (daher auch der Name *unitär*).

7.4.5 A Wir haben oben eine reelle unitäre Matrix U mit positiver Determinante als Drehung bezeichnet. Im \mathbb{R}^3 sollte U also eine Drehung mit Drehachse n und Drehwinkel α sein.

Machen Sie sich dafür folgende Punkte klar:

- Das Spektrum von U ist durch $\{1, \lambda, \bar{\lambda}\}$ gegeben, mit einem λ auf dem Einheitskreis von \mathbb{C} . Die zugehörige Eigenvektorbasis sei $\{b_1, b_2, b_3\}$. Dann ist b_1 eine natürliche Wahl für die Richtung der Drehachse. Dafür mache man sich klar, daß $b_1 \in \mathbb{R}^3$ gewählt werden kann (während b_2 und b_3 normalerweise in \mathbb{C}^3 liegen).
- Sei $b_2 = e + if$, mit Vektoren e und f aus \mathbb{R}^3 . Zeigen Sie, daß dann $b_3 = e - if$ wählbar ist. Verwenden Sie $b_2 \perp b_3$ um $e \perp f$ und $\|e\| = \|f\| = \frac{1}{2}\sqrt{2}$ nachzuweisen.
- Zeigen Sie schließlich

$$Ue = \cos(\alpha)e + \sin(\alpha)f, \quad Uf = -\sin(\alpha)e + \cos(\alpha)f.$$

Dabei ist α durch die EULER-Formel $\lambda = e^{-i\alpha} = \cos(\alpha) - i \sin(\alpha)$ definiert (natürlich nur bis auf uninteressante Vielfache von 2π).

- Führen Sie das Verfahren für die beiden unitären Matrizen

$$U := \frac{1}{3} \begin{bmatrix} 1 & 2 & 2 \\ 2 & -2 & 1 \\ 2 & 1 & -2 \end{bmatrix} \quad \text{und} \quad V := \frac{1}{3} \begin{bmatrix} 2 & 1 & 2 \\ 1 & 2 & -2 \\ -2 & 2 & 1 \end{bmatrix}$$

durch.

7.4.6 A

- Zeigen Sie, daß R_n eine *schiefsymmetrische Abbildung* ist, daß also $R_n^* = -R_n$ gilt.
- Zeigen Sie $R_{a \times b} = ba^t - ab^t$ und damit die GRASSMANN-Identität

$$a \times (b \times c) = \langle a | c \rangle b - \langle a | b \rangle c.$$

- Folgern Sie aus ii)

$$(a \times b) \times (c \times d) = \det(a, b, d)c - \det(a, b, c)d.$$

- Folgern Sie aus ii) die LAGRANGE-Identität

$$\langle a \times b | c \times d \rangle = \langle a | c \rangle \langle b | d \rangle - \langle a | d \rangle \langle b | c \rangle.$$

- v) Zeigen Sie $R_a^* R_a = \|a\|^2 \mathbf{1} - aa^t$ und damit $\|a \times b\| = \|a\| \|b\| \sin(\varphi)$. Dabei ist φ der von a und b eingeschlossene Winkel (vergl. auch (5.23)).
- vi) Folgern Sie aus iv) die GRASSMANN-PLÜCKER-Identität für 2×2 -Matrizen

$$\det(a, b) \det(c, d) + \det(b, c) \det(a, d) + \det(c, a) \det(b, d) = 0.$$

Verwenden Sie dabei die Einbettung $\varepsilon : a = [a_1, a_2]^t \mapsto \varepsilon(a) := [a_1, a_2, 0]^t$ des \mathbb{R}^2 in den \mathbb{R}^3 und die lineare Abbildung $J : [x_1, x_2, x_3]^t \mapsto [x_2, -x_1, x_3]$. Zeigen Sie zunächst $\varepsilon(a) \times \varepsilon(b) = \det(a, b) e_3$ und $\langle \varepsilon(a) | J \varepsilon(c) \rangle = \det(a, c)$. Benutzen Sie anschließend die LAGRANGE-Identität für die Vektoren $\varepsilon(a), \varepsilon(b), J\varepsilon(c)$ und $J\varepsilon(d)$.

Alternativ könnten Sie auch folgendermaßen vorgehen:

Zeigen Sie, daß $\lambda(a, b) := \det(a, b) \det(c, d) + \det(b, c) \det(a, d) + \det(c, a) \det(b, d)$ bei festgehaltenen Vektoren c und d eine alternierende Multilinearform des \mathbb{R}^2 definiert, daß also für eine geeignete Konstante k die Gleichung $\lambda(a, b) = k \cdot \det(a, b)$ gilt. Bestimmen Sie k .

8 Hauptachsentransformation

In diesem Kapitel soll die Eigenwerttheorie an dem geometrischen Beispiel der quadratischen Formen und der zugehörigen Quadriken angewandt werden. Im \mathbb{R}^2 handelt es sich dabei, wenn man von Sonderfällen absieht, im Wesentlichen darum, die genaue Lage von verschobenen und gedrehten Ellipsen, Hyperbeln, oder Parabeln zu bestimmen. Die Hauptarbeit wird darin bestehen, die Richtungen der Hauptachsen dieser Figuren zu finden. Sie werden sich als die Eigenvektoren einer selbstadjungierten Matrix herausstellen, die der Quadrik zugeordnet werden kann. Über die zugehörigen Eigenwerte lässt sich auf die Länge dieser Achsen schließen. Die Ideen, die wir hier entwickeln, lassen sich auch auf höhere Dimensionen ohne Änderung anwenden, allerdings erhöht sich dabei der Rechenaufwand erheblich. Wir streben hier jedoch keine systematische Klassifikation quadratischer Formen an, weshalb wir die Situation im \mathbb{R}^3 nur durch eine Übungsaufgabe streifen werden. Zunächst aber stellen wir die Quadriken in ihrer einfachsten Form als klassische Kegelschnitte vor, da erfahrungsgemäß nicht davon ausgegangen werden kann, daß diese aus der Schule bekannt sind.

8.1 Kegelschnitte

8.1.1 Ellipse Eine Ellipse ist der geometrische Ort aller Punkte p , deren Abstände d_1 und d_2 von zwei Punkten f_1 und f_2 , den sogenannten *Brennpunkten*, die konstante Summe $d_1 + d_2 = 2a$ haben.

8.1.2 Satz Eine Ellipse E ist durch

$$E = \left\{ [x, y]^t \in \mathbb{R}^2 \mid \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \right\} \quad (8.1)$$

gegeben. Die Tangente t bzw. die Normale n im Punkt $p := [x_0, y_0]^t \in E$ sind die Geraden

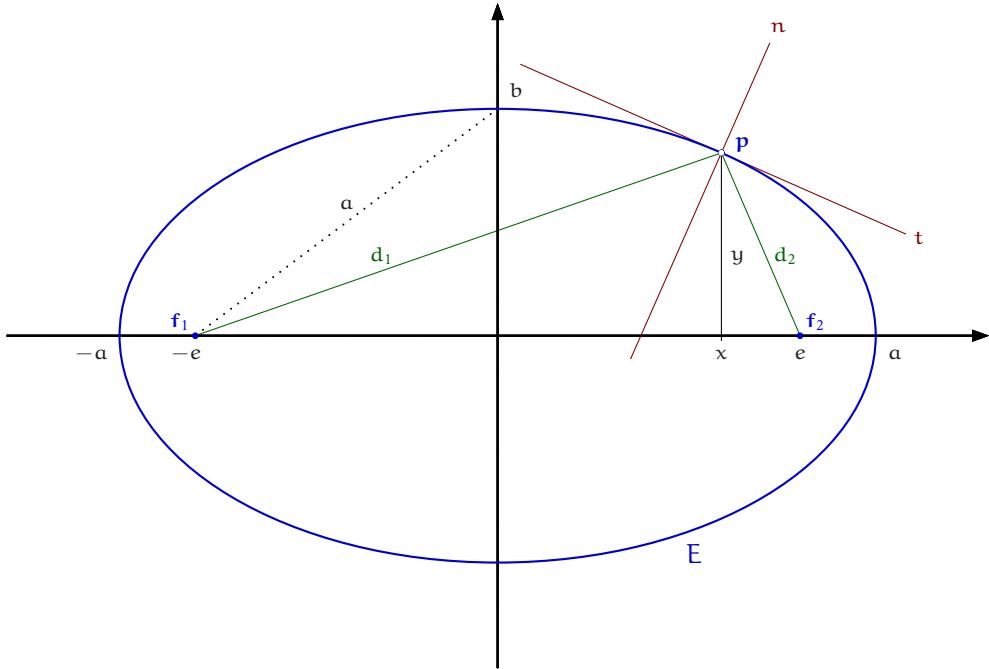
$$\begin{aligned} t &= \left\{ [x, y]^t \in \mathbb{R}^2 \mid \frac{xx_0}{a^2} + \frac{yy_0}{b^2} = 1 \right\} = \left\{ [x, y]^t \in \mathbb{R}^2 \mid \frac{(x - x_0)x_0}{a^2} + \frac{(y - y_0)y_0}{b^2} = 0 \right\}, \\ n &= \left\{ [x, y]^t \in \mathbb{R}^2 \mid \frac{(x - x_0)y_0}{b^2} - \frac{(y - y_0)x_0}{a^2} = 0 \right\}. \end{aligned}$$

n halbiert den Winkel $\angle f_1 p f_2$. Dabei ist $b := \sqrt{a^2 - e^2}$ und $e := \frac{1}{2} \|f_1 - f_2\|$.

Beweis: Aus der Skizze entnimmt man

$$d_{1/2} = \sqrt{(e \pm x)^2 + y^2} = \sqrt{x^2 + y^2 + e^2 \pm 2xe}.$$

Es folgt $4xe = d_1^2 - d_2^2 = d_1^2 - (2a - d_1)^2 = 4ad_1 - 4a^2$, also $d_1 = a + \frac{e}{a}x$. Damit erhalten wir nacheinander



$$\begin{aligned} d_1^2 &= x^2 + y^2 + e^2 + 2ex = \frac{e^2}{a^2}x^2 + 2ex + a^2, \\ x^2\left(1 - \frac{e^2}{a^2}\right) + y^2 &= x^2\frac{b^2}{a^2} + y^2 = a^2 - e^2 = b^2. \end{aligned}$$

Teilen wir diese Gleichung durch b^2 , so erhalten wir die Ellipsengleichung (8.1).

Umgekehrt ergibt sich aus (8.1) auch die definierende Eigenschaft $d_1 + d_2 = 2a$ der Ellipse. Zunächst haben wir $y^2 = b^2 - \frac{b^2}{a^2}x^2$, woraus bereits $|x| \leq a$ zu erkennen ist. Damit erhalten wir

$$d_{1/2}^2 = \left(1 - \frac{b^2}{a^2}\right)x^2 + b^2 + e^2 \pm 2xe = \frac{e^2}{a^2}x^2 + a^2 \pm 2xe = \left(a \pm \frac{e}{a}x\right)^2.$$

Aus $|x| \leq a$ folgt $\frac{e}{a}|x| \leq e \leq a$, so daß $d_{1/2} = a \pm \frac{e}{a}x$ und daher $d_1 + d_2 = 2a$ gilt.

Um zu zeigen, daß t wirklich die Tangente in $p = [x_0, y_0]^t \in E$ ist, nehmen wir an, es gäbe einen weiteren Ellipsenpunkt $p_1 := [x_1, y_1]^t$ auf t (in der Skizze sind die Koordinaten x und y von p durch x_0 und y_0 zu ersetzen, um eine Unterscheidung zu den Punkten $[x, y]^t$ auf t zu erhalten). Dann haben wir

$$\frac{x_0^2}{a^2} + \frac{y_0^2}{b^2} = 1, \quad \frac{x_0 x_1}{a^2} + \frac{y_0 y_1}{b^2} = 1, \quad \frac{x_1^2}{a^2} + \frac{y_1^2}{b^2} = 1.$$

Das hat $\frac{(x_0 - x_1)^2}{a^2} + \frac{(y_0 - y_1)^2}{b^2} = 0$ zur Folge, wie man leicht durch Ausmultiplizieren bestätigt. Da das nur für $x_0 = x_1$ und $y_0 = y_1$ zu erfüllen ist, kann es über p_0 hinaus keinen weiteren

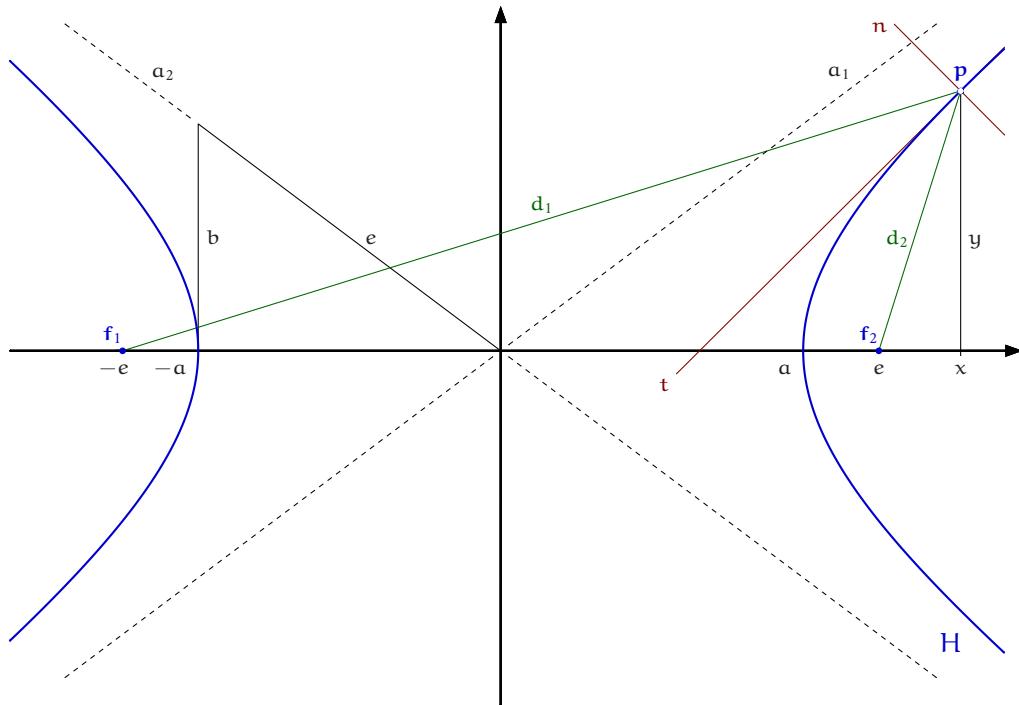
gemeinsamen Punkt von E und t geben. Das ist genau die Eigenschaft, die die Tangente in $p_0 \in E$ auszeichnet.

Die Gerade n schneidet die x -Achse offensichtlich bei $x_n = \frac{e^2}{a^2} x_0$. Sie teilt die Strecke $\|f_1 - f_2\|$ im Verhältnis

$$\frac{e - \frac{e^2}{a^2} x_0}{e + \frac{e^2}{a^2} x_0} = \frac{a - \frac{e}{a} x_0}{a + \frac{e}{a} x_0} = \frac{d_2}{d_1},$$

d. h., im Verhältnis der beim Winkel $\angle f_1 p f_2$ anliegenden Seiten d_2 und d_1 . Daher halbiert n diesen Winkel (vergl. Übung 8.1.6). Daß es sich bei n um die Normale im Punkt p handelt, ist eine einfache Übungsaufgabe. \square

8.1.3 Hyperbel Eine Hyperbel ist der geometrische Ort aller Punkte p , deren Abstände d_1 und d_2 von zwei Punkten f_1 und f_2 , den *Brennpunkten*, die konstante Differenz $|d_1 - d_2| = 2a$ haben.



8.1.4 Satz Eine Hyperbel H ist durch

$$H = \left\{ [x, y]^t \in \mathbb{R}^2 \mid \frac{x^2}{a^2} - \frac{y^2}{b^2} = 1 \right\} \quad (8.2)$$

gegeben. Sie hat die Asymptoten $a_{1/2}(x) := \pm \frac{b}{a} x$. Die Tangente t und die Normale n im Punkt $p := [x_0, y_0]^t \in H$ sind die Geraden

$$t = \left\{ [x, y]^t \in \mathbb{R}^2 \mid \frac{xx_0}{a^2} - \frac{yy_0}{b^2} = 1 \right\} = \left\{ [x, y]^t \in \mathbb{R}^2 \mid \frac{(x - x_0)x_0}{a^2} - \frac{(y - y_0)y_0}{b^2} = 0 \right\},$$

$$\mathbf{n} = \left\{ [x, y]^t \in \mathbb{R}^2 \mid \frac{(x - x_0)y_0}{b^2} + \frac{(y - y_0)x_0}{a^2} = 0 \right\}.$$

t halbiert den Winkel $\angle f_1 p f_2$. Dabei ist $b := \sqrt{e^2 - a^2}$ und $e := \frac{1}{2} \|f_1 - f_2\|$.

Der Beweis verläuft weitgehend wie der für die Ellipse.

Beweis. Aus der Skizze entnimmt man

$$d_{1/2} = \sqrt{(x \pm e)^2 + y^2} = \sqrt{x^2 + y^2 + e^2 \pm 2xe}.$$

Es folgt $4ex = d_1^2 - d_2^2 = d_1^2 - (d_1 - 2a)^2 = 4ad_1 - 4a^2$, also $d_1 = \frac{e}{a}x + a$ und $d_2 = \frac{e}{a}x - a$ ($d_2 \geq \frac{e}{a}a - a = e - a \geq 0$ für $x \geq a$). Für $x \leq -a$ ist $d_1 = -\frac{e}{a}x - a$ und $d_2 = -\frac{e}{a}x + a$.

Damit erhalten wir nacheinander

$$\begin{aligned} d_1^2 &= x^2 + y^2 + e^2 + 2ex = a^2 + 2ex + \frac{e^2}{a^2}x^2, \\ x^2 \left(1 - \frac{e^2}{a^2}\right) + y^2 &= -\frac{b^2}{a^2}x^2 + y^2 = a^2 - e^2 = -b^2. \end{aligned}$$

Division durch $-b^2$ ergibt die Hyperbelgleichung (8.2).

Umgekehrt ergibt sich aus (8.2) auch die definierende Eigenschaft $|d_1 - d_2| = 2a$ der Hyperbel. Zunächst haben wir $y^2 = \frac{b^2}{a^2}x^2 - b^2$, woraus bereits $|x| \geq a$ abzulesen ist. Damit erhalten wir

$$d_{1/2}^2 = \left(1 + \frac{b^2}{a^2}\right)x^2 - b^2 + e^2 \pm 2xe = \frac{e^2}{a^2}x^2 + a^2 \pm 2xe = \left(\frac{e}{a}x \pm a\right)^2.$$

Es folgt $d_{1/2} = \left|\frac{e}{a}x \pm a\right|$ und $d_1 d_2 = \left|\frac{e^2}{a^2}x^2 - a^2\right| = \frac{e^2}{a^2}x^2 - a^2$, denn $x^2 \geq a^2$. Das bedeutet

$$(d_1 - d_2)^2 = d_1^2 + d_2^2 - 2d_1 d_2 = 2\frac{e^2}{a^2}x^2 + 2a^2 - 2\left(\frac{e^2}{a^2}x^2 - a^2\right) = 4a^2.$$

Die Wurzel aus beiden Seiten der Gleichung zeigt $|d_1 - d_2| = 2a$.

Die Tangenteneigenschaft von t : Wir nehmen an, daß $p_1 := [x_1, y_1]^t$ ein gemeinsamer Punkt von t und H ist, verschieden vom Tangentenpunkt $p = [x_0, y_0]^t \in H \cap t$ (in der Skizze sind die Koordinaten x und y von p wieder durch x_0 und y_0 zu ersetzen). Dann haben wir

$$\frac{x_0^2}{a^2} - \frac{y_0^2}{b^2} = 1, \quad \frac{x_0 x_1}{a^2} - \frac{y_0 y_1}{b^2} = 1, \quad \frac{x_1^2}{a^2} - \frac{y_1^2}{b^2} = 1.$$

Daraus erhalten wir

$$\frac{(x_0 - x_1)^2}{a^2} - \frac{(y_0 - y_1)^2}{b^2} = 0 \quad \text{und} \quad \frac{(x_0 - x_1)x_0}{a^2} - \frac{(y_0 - y_1)y_0}{b^2} = 0,$$

was durch Ausmultiplizieren leicht bestätigt werden kann. Das bedeutet

$$\frac{(y_0 - y_1)^2}{(x_0 - x_1)^2} = \frac{b^2}{a^2} \quad \text{und} \quad \frac{(y_0 - y_1)^2 y_0^2}{(x_0 - x_1)^2 x_0^2} = \frac{b^4}{a^4},$$

woraus $y_0^2 = \frac{b^2}{a^2} x_0^2$ folgt. Mit der Hyperbelgleichung $1 = \frac{x_0^2}{a^2} - \frac{y_0^2}{b^2} = \frac{x_0^2}{a^2} - \frac{x_0^2 b^2}{a^2 b^2} = 0$ finden wir einen Widerspruch. Also ist p der einzige gemeinsame Punkt von t und H .

Die Gerade t schneidet die x -Achse offensichtlich bei $x_t = \frac{a^2}{x_0}$. Sie teilt die Strecke $\|f_1 - f_2\|$ im Verhältnis

$$\frac{e - \frac{a^2}{x_0}}{\frac{a^2}{x_0} + e} = \frac{\frac{e}{a} x_0 - a}{\frac{e}{a} x_0 + a} = \frac{d_2}{d_1},$$

d.h., im Verhältnis der beim Winkel $\angle f_1 p f_2$ anliegenden Seiten d_2 und d_1 . Daher halbiert t diesen Winkel (vergl. Übung 8.1.6). Daß es sich bei n um die Normale im Punkt p handelt, ist eine einfache Übungsaufgabe.

Die Asymptoten der Hyperbel sind die Ursprungsgeraden mit den Gleichungen $a_{1/2}(x) = \pm \frac{b}{a} x$. Um das einzusehen, bestimmen wir die Funktionsgleichungen h_{\pm} der Hyperbel, indem wir (8.2) nach y auflösen:

$$h_{\pm}(x) = \pm \frac{b}{a} \sqrt{x^2 - a^2}.$$

h_{\pm} sind die Hyperbeläste oberhalb bzw. unterhalb der x -Achse. Wir zeigen nur den Fall $a_1(x) - h_+(x) \xrightarrow{x \rightarrow \infty} 0$. Für $x > 0$ erhalten wir, wenn wir uns an das dritte Binom erinnern,

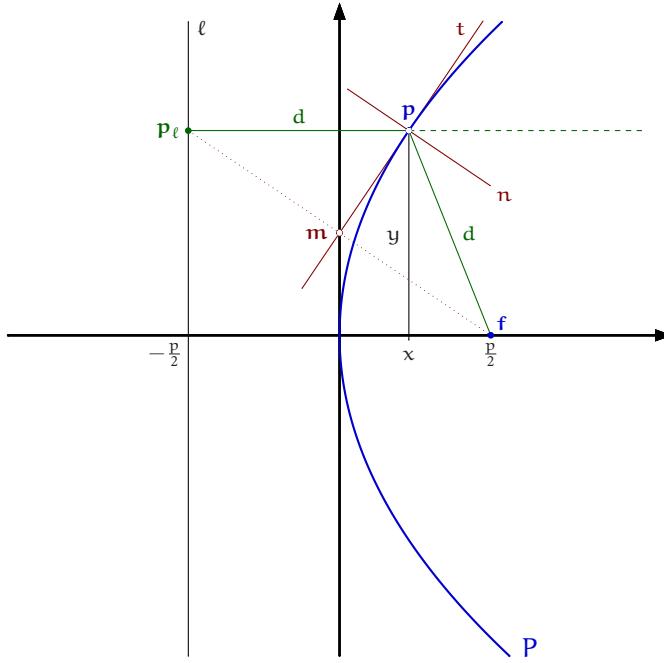
$$\begin{aligned} 0 &\leq a_1(x) - h_+(x) = \frac{b}{a} x - \frac{b}{a} |x| \sqrt{1 - \frac{a^2}{x^2}} = \frac{b}{a} x \left(1 - \sqrt{1 - \frac{a^2}{x^2}}\right) \\ &= \frac{b}{a} x \frac{\left(1 - \sqrt{1 - \frac{a^2}{x^2}}\right)\left(1 + \sqrt{1 - \frac{a^2}{x^2}}\right)}{\left(1 + \sqrt{1 - \frac{a^2}{x^2}}\right)} = \frac{b}{a} x \frac{\frac{a^2}{x^2}}{\left(1 + \sqrt{1 - \frac{a^2}{x^2}}\right)} \\ &\leq \frac{b}{a} x \frac{\frac{a^2}{x^2}}{1} = \frac{ab}{x} \xrightarrow{x \rightarrow \infty} 0. \end{aligned}$$

Für $x \rightarrow -\infty$ müssen wir

$$0 \leq h_-(x) - a_1(x) = -\frac{b}{a} |x| \sqrt{1 - \frac{a^2}{x^2}} - \frac{b}{a} x = \frac{b}{a} |x| \left(1 - \sqrt{1 - \frac{a^2}{x^2}}\right)$$

abschätzen, denn wegen $x < 0$ ist $-x = |x|$. Das verläuft ab jetzt genau so, wie wir es oben vorgeführt haben. \square

8.1.5 Parabel Eine Parabel P ist der geometrische Ort aller Punkte p , für die der Abstand d zu einem festen Punkt f , dem *Brennpunkt*, und zu einer Geraden ℓ , der sogenannten *Leitgeraden*, gleich groß ist.



Mit p bezeichnen wir den Abstand von f zu ℓ . Wir legen den Koordinatenursprung in die Mitte zwischen f und ℓ und die x -Achse senkrecht zu ℓ . Der Skizze entnehmen wir:

$$y^2 + \left(\frac{p}{2} - x\right)^2 = \left(\frac{p}{2} + x\right)^2, \text{ also } y^2 + x^2 - px + \frac{p^2}{4} = x^2 + px + \frac{p^2}{4}, \text{ d.h.:}$$

$$P = \{[x, y]^t \in \mathbb{R}^2 \mid y^2 = 2px\}, \quad (8.3)$$

$$t = \{[x, y]^t \in \mathbb{R}^2 \mid yy_0 = p(x + x_0)\} = \{[x, y]^t \in \mathbb{R}^2 \mid y_0(y - y_0) - p(x - x_0) = 0\},$$

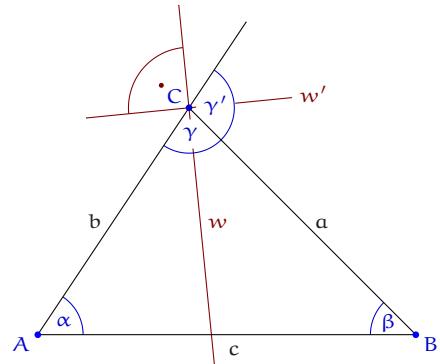
$$n = \{[x, y]^t \in \mathbb{R}^2 \mid p(y - y_0) + y_0(x - x_0) = 0\}.$$

t ist die Tangente an P in einem Parabelpunkt $p = [x_0, y_0]^t$. Sie ist die Winkelhalbierende im Dreieck $f p p_\ell$. Wird $p < 0$ gewählt, so entsteht eine nach links geöffnete Parabel. Wir haben nur noch die Tangenteneigenschaft von t zu zeigen. Dazu sei $p_1 := [x_1, y_1]^t$ ein Punkt auf $t \cap H$, der von $p := [x_0, y_0]^t$ verschieden ist (in der Skizze ist wieder x durch x_0 und y durch y_0 zu ersetzen). Wir haben also $y_0^2 = 2px_0$, $y_1^2 = 2px_1$ und $y_1y_0 = p(x_1 + x_0)$. Daraus folgt $(y_0 - y_1)^2 = y_0^2 - 2px_0 + y_1^2 - 2px_1 + 2px_0 + 2px_1 - 2y_1y_0 = 0$, also $y_0 = y_1$ und dann auch $x_0 = x_1$. Damit kann es nur einen gemeinsamen Punkt von P und t geben.

Der Mittelpunkt $m = \frac{1}{2}([-\frac{p}{2}, y_0]^t + [\frac{p}{2}, 0]^t) = [0, \frac{y_0}{2}]^t$ zwischen p_ℓ und f liegt auf der Tangente, denn er erfüllt die Tangentengleichung: $\frac{y_0}{2}y_0 = px_0$. Da das Dreieck $f p p_\ell$ laut Parabeldefinition gleichschenklig ist, handelt es sich bei t um die Winkelhalbierende in diesem Dreieck. Dessen Nebenwinkel hat die Normale n in p als Winkelhalbierende. Diese Eigenschaft ist der Grund dafür, f als *Brennpunkt* von P zu bezeichnen. Denken wir uns nämlich die Parabelinnenseite aus spiegelndem Material gefertigt, dann wird jeder parallel zur x -Achse einfallende Lichtstrahl in den Punkt f reflektiert.

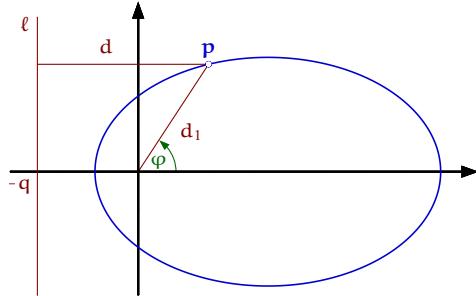
8.1.6 A Zwei elementare Eigenschaften der Winkelhalbierenden in einem Dreieck. Zeigen Sie:

- i) Eine Gerade w durch C ist genau dann die Winkelhalbierende, wenn sie die gegenüberliegende Seite c im Verhältnis $a : b$ der anliegenden Seiten a und b teilt.
- ii) Die zu w orthogonale Gerade w' durch C ist die sogenannte *äußere Winkelhalbierende*. Sie halbiert den *Nebenwinkel* $\gamma' := \pi - \gamma$.



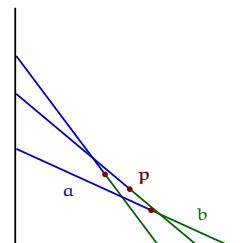
8.1.7 A Es gibt für Kegelschnitte einheitliche Definitionen, von denen wir zwei untersuchen wollen, nämlich die Konstruktion über eine Leitgerade und die Polardarstellung. Zeigen Sie:

- i) Ein Kegelschnitt K ist der geometrische Ort aller Punkte p , für die das Verhältnis $\frac{d_1}{d}$ aus dem Abstand d_1 von einem festen Punkt, dem Brennpunkt, und dem Abstand d von einer Geraden ℓ , der Leitgeraden, eine feste Zahl $\varepsilon > 0$ ist. Für $0 < \varepsilon < 1$ handelt es sich um eine Ellipse, für $\varepsilon = 1$ um eine Parabel und für $\varepsilon > 1$ um eine Hyperbel.
- ii) $K = \left\{ \frac{p}{1 - \varepsilon \cos(\varphi)} e^{i\varphi} \mid 0 \leq \varphi < 2\pi \right\}$ ist die Polardarstellung eines Kegelschnitts. $\varepsilon = 0$ ergibt einen Kreis, $0 < \varepsilon < 1$ eine Ellipse, $\varepsilon = 1$ eine Parabel und $\varepsilon > 1$ eine Hyperbel.



8.1.8 A

Eine Leiter, die an eine Hauswand gelehnt ist, gleitet zu Boden. Welche Bahn beschreibt dabei ein beliebiger, fest gewählter Punkt p auf dieser Leiter?



8.2 Quadratische Formen

Der allgemeinste quadratische Ausdruck im Bereich gewöhnlicher reeller Funktionen ist durch $f(x) = ax^2 + bx + c$ gegeben, mit Konstanten $a \neq 0$, b und c . Die richtige Verallgemeinerung auf den \mathbb{R}^2 , \mathbb{R}^3 oder \mathbb{R}^n ist eine *quadratische Form*:

8.2.1 Definition Eine quadratische Form ist eine Funktion $Q: \mathbb{R}^n \rightarrow \mathbb{R}$ der folgenden Art

$$Q(\mathbf{x}) = \langle \mathbf{x} | A\mathbf{x} \rangle + \langle \mathbf{b} | \mathbf{x} \rangle + c \quad (8.4)$$

mit einer $n \times n$ -Matrix A , einem Vektor $\mathbf{b} \in \mathbb{R}^n$ und einer Zahl $c \in \mathbb{R}$. Die zu Q gehörende Quadrik \mathcal{Q} ist die Menge

$$\mathcal{Q} := \{ \mathbf{x} \in \mathbb{R}^n \mid Q(\mathbf{x}) = 0 \}. \quad (8.5)$$

Die Matrix A der quadratischen Form Q kann o. B. d. A. selbstadjungiert gewählt werden, wie die folgende Rechnung zeigt:

$$\begin{aligned} Q(\mathbf{x}) &= \langle \mathbf{x} | A\mathbf{x} \rangle + \langle \mathbf{b} | \mathbf{x} \rangle + c = \langle A^*\mathbf{x} | \mathbf{x} \rangle + \langle \mathbf{b} | \mathbf{x} \rangle + c \\ &= \langle \mathbf{x} | A^*\mathbf{x} \rangle + \langle \mathbf{b} | \mathbf{x} \rangle + c \\ &= \frac{1}{2} (\langle \mathbf{x} | A\mathbf{x} \rangle + \langle \mathbf{x} | A^*\mathbf{x} \rangle) + \langle \mathbf{b} | \mathbf{x} \rangle + c \\ &= \langle \mathbf{x} | \frac{1}{2}(A + A^*)\mathbf{x} \rangle + \langle \mathbf{b} | \mathbf{x} \rangle + c. \end{aligned}$$

Die quadratische Form ändert sich also nicht, wenn wir A durch die selbstadjungierte Matrix $\frac{1}{2}(A + A^*)$ ersetzen. Wir gehen daher für die folgenden Überlegungen von einer selbstadjungierten Matrix A aus und haben damit zur Untersuchung der Form die Eigenwerttheorie zur Verfügung. Bevor wir aber allgemeine quadratische Formen untersuchen, stellen wir den Zusammenhang mit den Kegelschnitten her. Die Gleichungen (8.1) und (8.2) für die Ellipse bzw. Hyperbel lassen sich leicht als Quadrik einer quadratischen Form identifizieren:

$$1 = \frac{x^2}{a^2} \pm \frac{y^2}{b^2} = \left\langle \begin{bmatrix} x \\ y \end{bmatrix} \mid \begin{bmatrix} \frac{1}{a^2} & 0 \\ 0 & \pm \frac{1}{b^2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right\rangle.$$

Die zugehörige Matrix

$$\tilde{A} := \begin{bmatrix} \frac{1}{a^2} & 0 \\ 0 & \pm \frac{1}{b^2} \end{bmatrix}$$

liegt offensichtlich bereits in Diagonalform vor. Aus den Eigenwerten $\frac{1}{a^2}$ und $\frac{1}{b^2}$ bzw. $-\frac{1}{b^2}$ lassen sich die Längen der Hauptachsen leicht ablesen. \mathbf{b} und c sind Null. Eine um \mathbf{m}_1 in x -Richtung und um \mathbf{m}_2 in y -Richtung verschobene Ellipse bzw. Hyperbel hat die Gleichung

$$1 = \frac{(x - m_1)^2}{a^2} \pm \frac{(y - m_2)^2}{b^2} = \langle \mathbf{x} - \mathbf{m} | \tilde{A}(\mathbf{x} - \mathbf{m}) \rangle$$

$$= \langle \mathbf{x} | \tilde{A} \mathbf{x} \rangle - 2 \langle \tilde{A} \mathbf{m} | \mathbf{x} \rangle + \langle \mathbf{m} | \tilde{A} \mathbf{m} \rangle, \quad \mathbf{m} := \begin{bmatrix} m_1 \\ m_2 \end{bmatrix}.$$

Offensichtlich hat also der Vektor \mathbf{b} in einer allgemeinen quadratischen Form etwas mit der Verschiebung der Quadrik zu tun, auch wenn aus ihm der Verschiebungsvektor meistens nicht direkt ablesbar ist. Hier ist $\mathbf{b} = -2\tilde{A}\mathbf{m}$. Wenn wir vor der Verschiebung $T(\mathbf{x}) := \mathbf{x} + \mathbf{m}$ noch eine Koordinatentransformation $B = [\mathbf{b}_1, \mathbf{b}_2]$ durchführen (mit der Orthonormalbasis $\mathcal{B} := \{\mathbf{b}_1, \mathbf{b}_2\}$), die eine Drehung der Quadrik bewirkt, dann haben wir die allgemeinste Form einer Ellipsen- bzw. Hyperbelgleichung gefunden:

$$1 = \langle B^*(\mathbf{x} - \mathbf{m}) | \tilde{A} B^*(\mathbf{x} - \mathbf{m}) \rangle = \langle \mathbf{x} - \mathbf{m} | \tilde{A} B^*(\mathbf{x} - \mathbf{m}) \rangle.$$

Die Matrix $A := B \tilde{A} B^*$ wird nun i. Allg. nicht mehr diagonal sein (wohl aber selbstadjungiert – nachrechnen).

Auch wenn sich auf diese Weise nicht alle Quadriken gewinnen lassen, denn von Sonderfällen wie Doppelkegel, oder Doppelpunkten abgesehen, bilden die Parabeln eine Ausnahme, lässt sich aus der vorgestellten Konstruktion doch eine Methode erkennen, die auch zur Untersuchung allgemeiner Quadriken verwendet werden kann. Wir haben die Konstruktion nur schrittweise rückgängig zu machen.

Im ersten Schritt werden die Eigenwerte $\lambda_1, \dots, \lambda_n$ und die zugehörige ONB $\mathcal{B} := \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ aus Eigenvektoren der Matrix A in (8.4) bestimmt. Die Koordinatentransformation $\mathbf{x} = B\mathbf{x}_{\mathcal{B}}$ von der Koordinatendarstellung bzgl. \mathcal{B} in die kanonische Basis \mathcal{E} (gemäß unserer Vereinbarung auf Seite 152, daß sich eine Koordinatendarstellung \mathbf{x} ohne Index immer auf die kanonische Basis $\mathcal{E} = \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ bezieht, führt die Gleichung $Q(\mathbf{x}) = 0$ aus (8.4) in

$$\begin{aligned} 0 &= \langle B\mathbf{x}_{\mathcal{B}} | A B\mathbf{x}_{\mathcal{B}} \rangle + \langle \mathbf{b} | B\mathbf{x}_{\mathcal{B}} \rangle + c \\ &= \langle \mathbf{x}_{\mathcal{B}} | B^* A B\mathbf{x}_{\mathcal{B}} \rangle + \langle B^* \mathbf{b} | \mathbf{x}_{\mathcal{B}} \rangle + c \\ &= \langle \mathbf{x}_{\mathcal{B}} | A_{\mathcal{B}} \mathbf{x}_{\mathcal{B}} \rangle + \langle \mathbf{b}_{\mathcal{B}} | \mathbf{x}_{\mathcal{B}} \rangle + c \end{aligned}$$

über. Dabei ist $A_{\mathcal{B}} := B^* A B = \text{diag}[\lambda_1, \dots, \lambda_n]$ eine Diagonalmatrix, deren Einträge die Eigenwerte $\lambda_1, \dots, \lambda_n$ von A sind. $\mathbf{b}_{\mathcal{B}} := B^* \mathbf{b}$ ist die Koordinatendarstellung von \mathbf{b} bzgl. \mathcal{B} .

Im zweiten Schritt wird obige Gleichung in Koordinatenform ausgeschrieben.

$$0 = \lambda_1 \tilde{x}_1^2 + \dots + \lambda_n \tilde{x}_n^2 + \tilde{b}_1 \tilde{x}_1 + \dots + \tilde{b}_n \tilde{x}_n + c.$$

Für die Eigenwerte $\lambda_i \neq 0$ kann $\lambda_i \tilde{x}_i^2 + \tilde{b}_i \tilde{x}_i$ durch quadratische Ergänzung in ein Binom verwandelt werden.

$$\lambda_i \tilde{x}_i^2 + \tilde{b}_i \tilde{x}_i = \lambda_i \left(\tilde{x}_i^2 + 2 \frac{\tilde{b}_i}{2\lambda_i} \tilde{x}_i + \frac{\tilde{b}_i^2}{4\lambda_i^2} \right) - \frac{\tilde{b}_i^2}{4\lambda_i} = \lambda_i (\tilde{x}_i - \tilde{m}_i)^2 - \lambda_i \tilde{m}_i^2,$$

mit $\tilde{m}_i := \frac{\tilde{b}_i}{2\lambda_i}$. Sind etwa im \mathbb{R}^2 alle Eigenwerte von Null verschieden, so erhält man die transformierte Gleichung der Quadrik in der Form einer verschobenen Ellipse oder Hyperbel:

$$\tilde{c} = \frac{(\tilde{x}_1 - \tilde{m}_1)^2}{1/\lambda_1} + \frac{(\tilde{x}_2 - \tilde{m}_2)^2}{1/\lambda_2}. \quad (8.6)$$

Dabei ist $\tilde{c} := \frac{\tilde{b}_1^2}{4\lambda_1} + \frac{\tilde{b}_2^2}{4\lambda_2} - c$. Jetzt erkennt man eine Ellipse daran, daß entweder alle Eigenwerte positiv oder alle negativ sind (sind sie z. B. alle positiv, dann muß $\tilde{c} > 0$ gelten, damit die Menge \mathcal{Q} nicht leer ist), während bei einer Hyperbel die Eigenwerte unterschiedliche Vorzeichen haben. In jedem Fall ergeben sich die Längen der Hauptachsen als $\sqrt{\frac{|\tilde{c}|}{|\lambda_i|}}$. Der Verschiebungsvektor der Quadrik hat in der Basis \mathcal{B} offensichtlich die Koordinaten \tilde{m}_1 und \tilde{m}_2 . Mit B wird er in die kanonische Basis zurück transformiert:

$$\mathbf{m} = B \begin{bmatrix} \tilde{m}_1 \\ \tilde{m}_2 \end{bmatrix} .$$

Ist einer der Eigenwerte Null, sagen wir λ_1 , dann läßt sich für \tilde{x}_1 keine quadratische Ergänzung durchführen. Statt (8.6) erhalten wir

$$0 = (\tilde{x}_2 - \tilde{m}_2)^2 + \frac{\tilde{b}_1}{\lambda_2} \tilde{x}_1 + \frac{c}{\lambda_2} - \tilde{m}_2^2.$$

Ist $\tilde{b}_1 \neq 0$, dann läßt sich das mit einem geeigneten \tilde{m}_1 in die Gestalt

$$(\tilde{x}_2 - \tilde{m}_2)^2 = -\frac{\tilde{b}_1}{\lambda_2} (\tilde{x}_1 - \tilde{m}_1)$$

bringen. Das ist eine verschobene Parabel. Man überlege sich, welche Möglichkeiten sich noch für $b_1 = 0$ ergeben.

Die gerade vorgestellten Ideen lassen sich auch auf höhere Dimensionen, aber vornehmlich auf die Dimension 3 erweitern. Da es dann drei verschiedene Eigenwerte geben kann, ergibt sich auch eine größere Vielfalt an möglichen Quadriken. Die einfachste Erweiterung der zweidimensionalen Situation entsteht, wenn alle Eigenwerte $\neq 0$ sind. Statt (8.6) erhalten wir dann

$$\tilde{c} = \frac{(\tilde{x}_1 - \tilde{m}_1)^2}{1/\lambda_1} + \frac{(\tilde{x}_2 - \tilde{m}_2)^2}{1/\lambda_2} + \frac{(\tilde{x}_3 - \tilde{m}_3)^2}{1/\lambda_3}. \quad (8.7)$$

Ist $\tilde{c} > 0$ und sind alle Eigenwerte positiv, dann handelt es sich um ein sog. *Ellipsoid*, d. h. um eine geschlossene Fläche um den Ursprung im \mathbb{R}^3 , deren Schnitt mit jeder Ursprungsebene eine Ellipse bildet. Sind dagegen ein oder zwei Eigenwerte negativ, so handelt es sich um ein sog. *Hyperboloid*, das im ersten Fall einschalig (wie etwa die Kühltürme von Kraftwerken) und im zweiten zweischalig ist. Weitere Formen und Sonderfälle entstehen, wenn Eigenwerte 0 sein dürfen, oder wenn $\tilde{c} = 0$ sein sollte. Darunter sind verschiedene Formen von Paraboloiden (einer Fläche, deren Schnitt mit einer oder mehreren Ebenen eine Parabel liefert), Doppelkegeln, elliptische, hyperbolische und parabolische Zylinder (also Zylinder, deren Querschnitte senkrecht zur Zylinderachse Ellipsen, Hyperbeln oder Parabeln darstellen), aber auch ausgeartete Figuren wie parallele Ebenen oder gar nur ein Paar Punkte.

8.3 Beispiele für Quadriken

8.3.1 Eine Ellipse

$$\begin{aligned} Q_1(\mathbf{x}) &:= \frac{1}{36}(8x^2 - 4xy + 5y^2) + \frac{1}{\sqrt{5}}(2x - 6y) + 12 \\ &= \frac{1}{36} \left\langle \mathbf{x} \left| \begin{bmatrix} 8 & -2 \\ -2 & 5 \end{bmatrix} \mathbf{x} \right. \right\rangle + \frac{1}{\sqrt{5}} \left\langle \begin{bmatrix} 2 \\ -6 \end{bmatrix} \left| \mathbf{x} \right. \right\rangle + 12 \end{aligned}$$

Wir bestimmen zunächst die Eigenwerte der Matrix $A := \begin{bmatrix} 8 & -2 \\ -2 & 5 \end{bmatrix}$

$$\det \begin{bmatrix} 8 - \lambda & -2 \\ -2 & 5 - \lambda \end{bmatrix} = (8 - \lambda)(5 - \lambda) - 4 = \lambda^2 - 13\lambda + 36 = (\lambda - 4)(\lambda - 9)$$

zu $\lambda_1 := 4$ und $\lambda_2 := 9$. Der Eigenvektor \mathbf{b}_1 zu λ_1 ist eine Lösung des Gleichungssystems

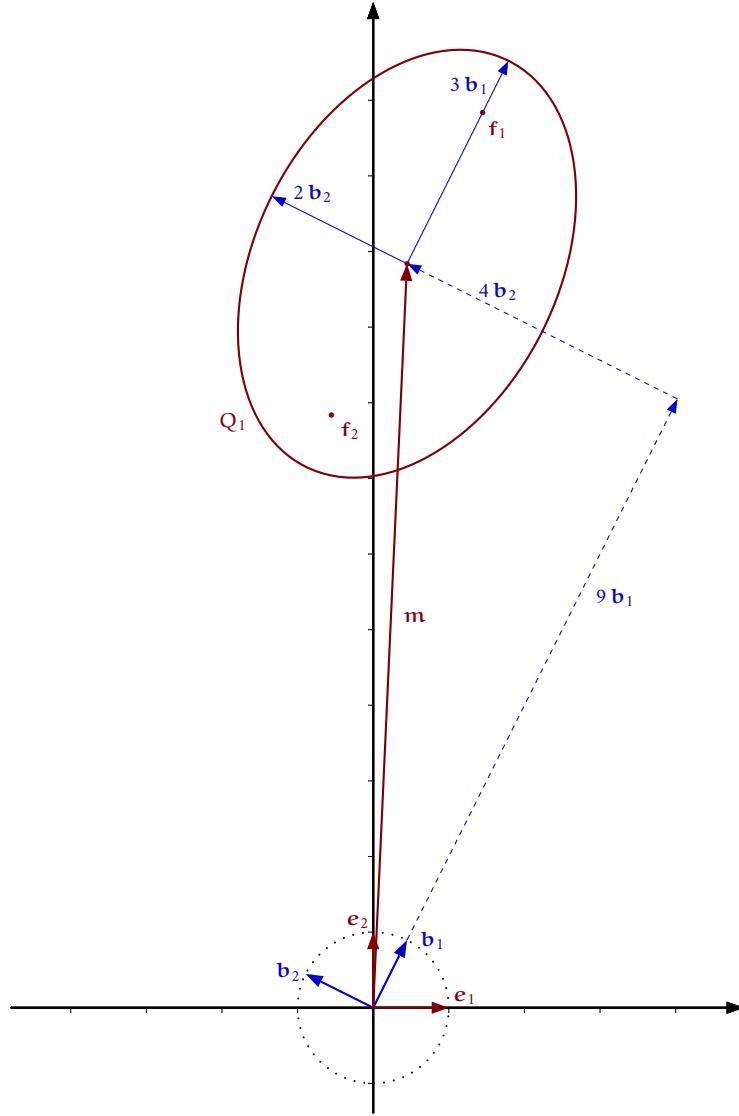
$$\begin{bmatrix} 8 - \lambda_1 & -2 \\ -2 & 5 - \lambda_1 \end{bmatrix} \mathbf{b}_1 = \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix} \mathbf{b}_1 = \mathbf{0},$$

mit der offensichtlichen Lösung $\mathbf{b}_1 \sim \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Den auf 1 normierten Vektor $\frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ wählen wir als Eigenvektor zu λ_1 . Da A selbstadjungiert ist, gibt es eine ONB aus Eigenvektoren. Für den zweiten normierten Eigenvektor \mathbf{b}_2 gibt es daher nur noch die beiden Möglichkeiten $\mathbf{b}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} -2 \\ 1 \end{bmatrix}$ oder $\mathbf{b}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ -1 \end{bmatrix}$. Wir wählen die erste, denn dann wird die ONB

$$\mathcal{B} := \{ \mathbf{b}_1, \mathbf{b}_2 \} = \left\{ \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \frac{1}{\sqrt{5}} \begin{bmatrix} -2 \\ 1 \end{bmatrix} \right\}$$

ein Rechtssystem. Mit der Matrix $B := \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}$ führen wir die Koordinatentransformation $\mathbf{x} = B\mathbf{x}_{\mathcal{B}}$ durch:

$$\begin{aligned} \tilde{Q}_1(\mathbf{x}_{\mathcal{B}}) &= \frac{1}{36} \left\langle B\mathbf{x}_{\mathcal{B}} \left| \begin{bmatrix} 8 & -2 \\ -2 & 5 \end{bmatrix} B\mathbf{x}_{\mathcal{B}} \right. \right\rangle + \frac{1}{\sqrt{5}} \left\langle \begin{bmatrix} 2 \\ -6 \end{bmatrix} \left| B\mathbf{x}_{\mathcal{B}} \right. \right\rangle + 12 \\ &= \frac{1}{36} \left\langle \mathbf{x}_{\mathcal{B}} \left| B^* \begin{bmatrix} 8 & -2 \\ -2 & 5 \end{bmatrix} B\mathbf{x}_{\mathcal{B}} \right. \right\rangle + \frac{1}{5} \left\langle \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -6 \end{bmatrix} \left| \mathbf{x}_{\mathcal{B}} \right. \right\rangle + 12 \\ &= \frac{1}{36} \left\langle \mathbf{x}_{\mathcal{B}} \left| \begin{bmatrix} 4 & 0 \\ 0 & 9 \end{bmatrix} \mathbf{x}_{\mathcal{B}} \right. \right\rangle - \left\langle \begin{bmatrix} 2 \\ 2 \end{bmatrix} \left| \mathbf{x}_{\mathcal{B}} \right. \right\rangle + 12 \\ &= \frac{1}{9} \tilde{x}^2 + \frac{1}{4} \tilde{y}^2 - 2\tilde{x} - 2\tilde{y} + 12 \\ &= \frac{1}{9} (\tilde{x}^2 - 18\tilde{x} + 9^2) + \frac{1}{4} (\tilde{y}^2 - 8\tilde{y} + 4^2) + 12 - 9 - 4 \\ &= \frac{(\tilde{x} - 9)^2}{9} + \frac{(\tilde{y} - 4)^2}{4} - 1. \end{aligned}$$



In der Basis \mathcal{B} handelt es sich bei unserer Quadrik $Q_1 := \{ \mathbf{x} \in \mathbb{R}^2 \mid Q_1(\mathbf{x}) = 0 \}$ also um eine um 9 Einheiten in Richtung \mathbf{b}_1 und um 4 Einheiten in Richtung \mathbf{b}_2 verschobene Ellipse mit der großen Halbachse $a = 3$ und der kleinen $b = 2$, sowie mit $e = \sqrt{3^2 - 2^2} = \sqrt{5}$. Der Mittelpunkt \mathbf{m} liegt bei

$$\mathbf{m} = 9\mathbf{b}_1 + 4\mathbf{b}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 9-8 \\ 18+4 \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ 22 \end{bmatrix}$$

und die beiden Brennpunkte bei $\mathbf{f}_{1/2} = \left[\frac{1}{\sqrt{5}} \pm 1, \frac{22}{\sqrt{5}} \pm 2 \right]^t$ (erhalten aus $\mathbf{m} \pm \sqrt{5}\mathbf{b}_1$).

8.3.2 A Untersuchen Sie die quadratische Form

$$Q_2(\mathbf{x}) := 180x^2 - 84xy + 145y^2 - \sqrt{13}(84x + 334y) + 1261.$$

8.3.3 Eine Hyperbel

$$Q_3(x) := -221x^2 + 174xy + 11y^2 + \sqrt{10} (324x - 28y) - 1640$$

$$= \left\langle x \left| \begin{bmatrix} -221 & 87 \\ 87 & 11 \end{bmatrix} x \right. \right\rangle + \sqrt{10} \left\langle \begin{bmatrix} 324 \\ -28 \end{bmatrix} \middle| x \right\rangle - 1640.$$

Die Eigenwerte:

$$\det \begin{bmatrix} -221 - \lambda & 87 \\ 87 & 11 - \lambda \end{bmatrix} = (11 - \lambda)(-221 - \lambda) - 87^2 = \lambda^2 + 210\lambda - 10000 \\ = (\lambda + 250)(\lambda - 40).$$

Die Eigenwerte sind $\lambda_1 = 40$ und $\lambda_2 = -250$. Der Eigenvektor b_1 zu λ_1 :

$$\begin{bmatrix} -261 & 87 \\ 87 & -29 \end{bmatrix} b_1 = \begin{bmatrix} -3 \cdot 87 & 87 \\ 3 \cdot 29 & -29 \end{bmatrix} b_1 = \mathbf{0}.$$

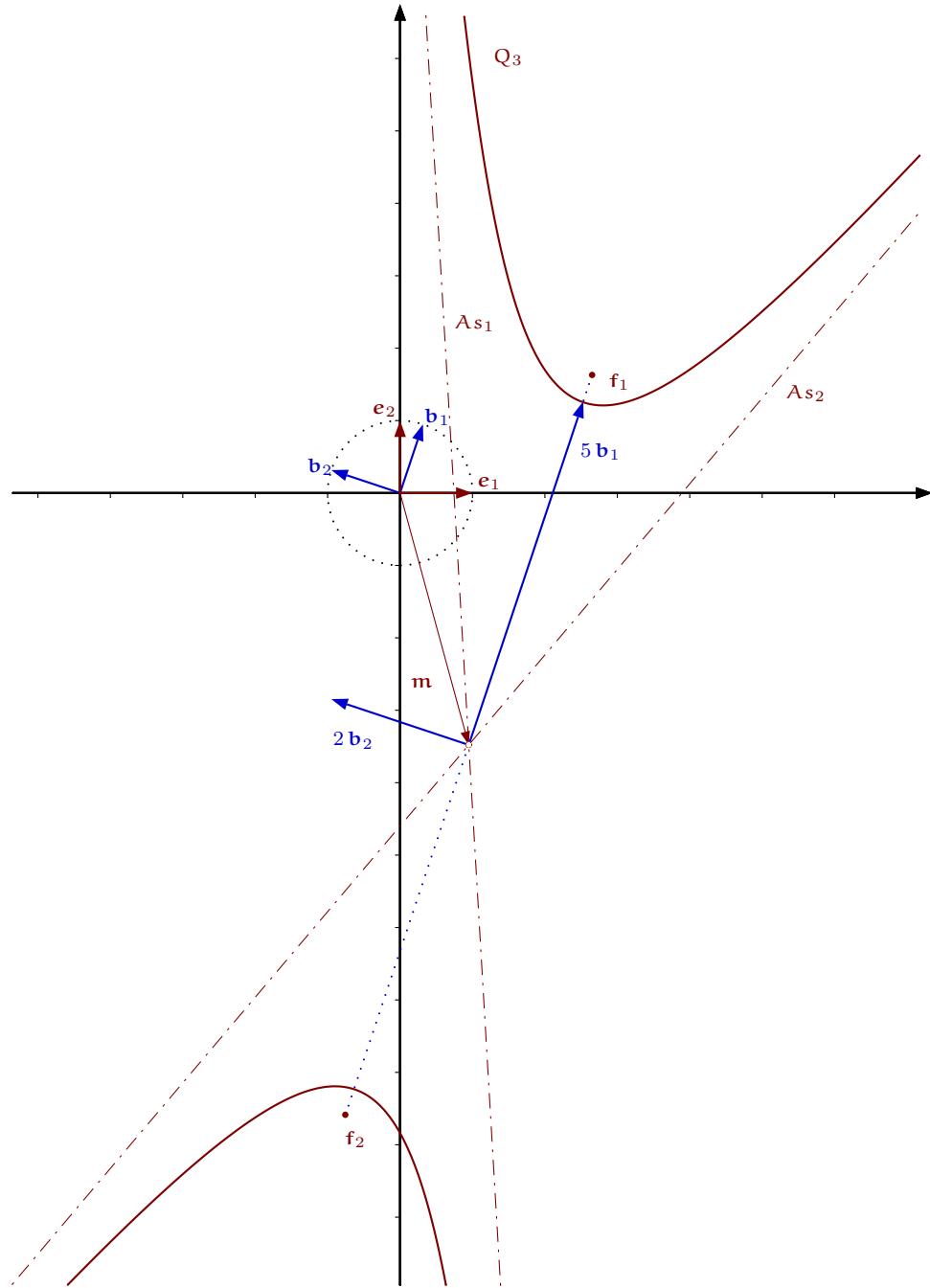
Eine Lösung ist offensichtlich $b_1 = \frac{1}{\sqrt{10}} \begin{bmatrix} 1 \\ 3 \end{bmatrix}$. Damit ist der normierte Eigenvektor zu λ_2 z. B. durch $b_2 = \frac{1}{\sqrt{10}} \begin{bmatrix} -3 \\ 1 \end{bmatrix}$ gegeben. Die Transformationsmatrix B ist daher

$$B := \frac{1}{\sqrt{10}} \begin{bmatrix} 1 & -3 \\ 3 & 1 \end{bmatrix}.$$

Sie vermittelt die Koordinatentransformation $x = Bx_B$:

$$\begin{aligned} \tilde{Q}_3(x_B) &= \left\langle x_B \left| B^* \begin{bmatrix} -221 & 87 \\ 87 & 11 \end{bmatrix} B x_B \right. \right\rangle + \left\langle \begin{bmatrix} 1 & 3 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} 324 \\ -28 \end{bmatrix} \middle| x_B \right\rangle - 1640 \\ &= \left\langle x_B \left| \begin{bmatrix} 40 & 0 \\ 0 & -250 \end{bmatrix} x_B \right. \right\rangle + \left\langle \begin{bmatrix} 240 \\ -1000 \end{bmatrix} \middle| x_B \right\rangle - 1640 \\ &= 40(\tilde{x}^2 + 6\tilde{x} + 3^2) - 250(\tilde{y}^2 + 4\tilde{y} + 2^2) - 360 + 1000 - 1640 \\ &= 40(\tilde{x} + 3)^2 - 250(\tilde{y}^2 + 2)^2 - 1000 \\ &= 1000 \left[\frac{(\tilde{x} + 3)^2}{25} - \frac{(\tilde{y} + 2)^2}{4} - 1 \right]. \end{aligned}$$

In der Basis $\mathcal{B} = \{b_1, b_2\}$ handelt es sich um eine Hyperbel, in b_1 -Richtung um -3 Einheiten und in b_2 -Richtung um -2 Einheiten verschoben. Die Halbachsen sind $a = 5$ und $b = 2$. Damit ist $e = \sqrt{5^2 + 2^2} = \sqrt{29}$. Der Mittelpunkt m liegt bei $m = -3b_1 - 2b_2 = \frac{1}{\sqrt{10}} \begin{bmatrix} 3 \\ -11 \end{bmatrix}$ und die Brennpunkte bei $f_{1/2} = m \pm \sqrt{29} b_1 = \frac{1}{\sqrt{10}} \begin{bmatrix} 3 \pm \sqrt{29} \\ -11 \pm 3\sqrt{29} \end{bmatrix}$.

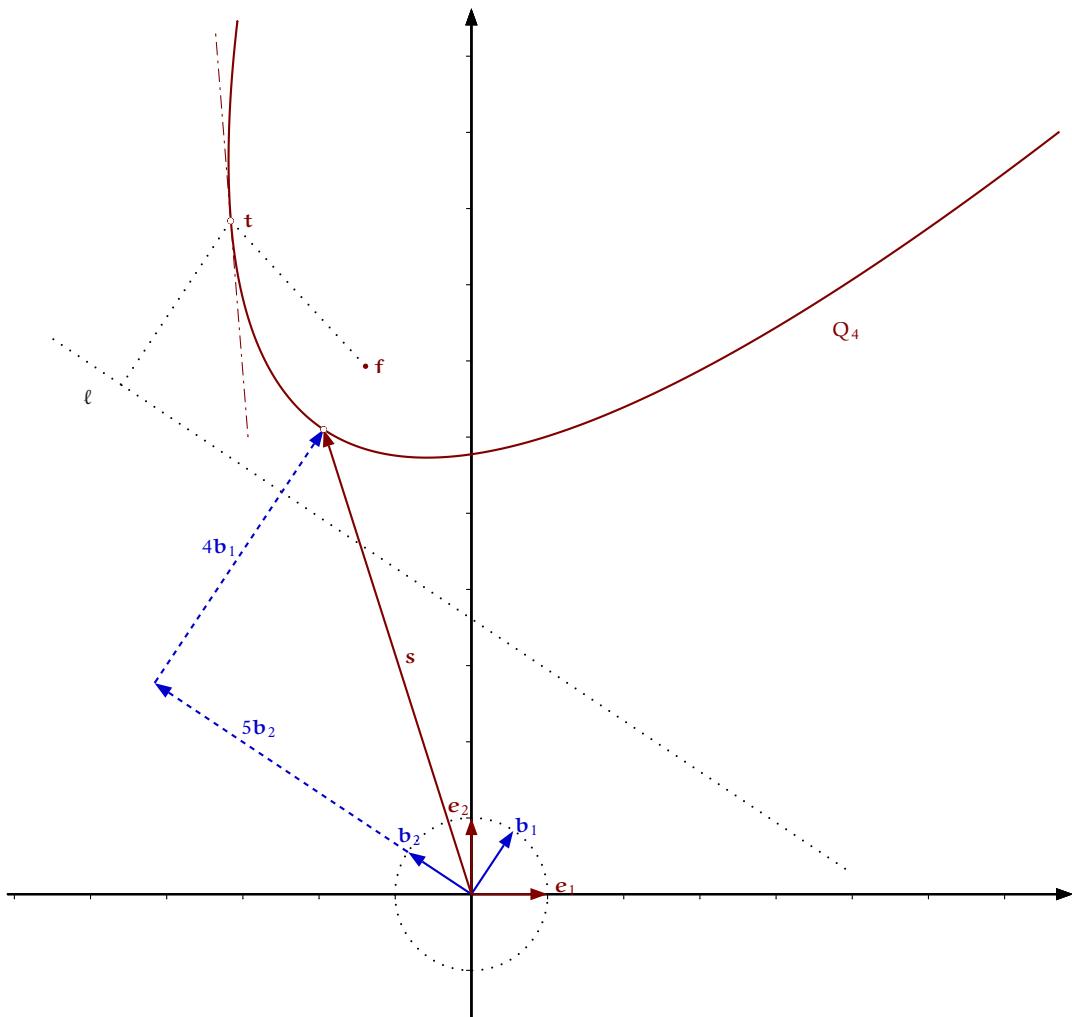


Die Asymptoten haben die Gleichungen $As_1 : 17x+y = 4\sqrt{10}$, $As_2 : 13x-11y = 16\sqrt{10}$. Ein Richtungsvektor z. B. der Geraden As_1 ist durch $5b_1 + 2b_2 \sim \begin{bmatrix} -1 \\ 17 \end{bmatrix}$ gegeben. Daher können wir als Normalenvektor $\begin{bmatrix} 17 \\ 1 \end{bmatrix}$ nehmen. Die Geradengleichung lautet dann $17x + y = c$ mit

einem noch zu bestimmenden c . Der Punkt \mathbf{m} liegt auf beiden Asymptoten. Einsetzen ergibt $c = \frac{1}{\sqrt{10}} (3 \cdot 17 - 11) = \frac{40}{\sqrt{10}} = 4\sqrt{10}$. Dieselbe Rechnung mit dem Richtungsvektor $5\mathbf{b}_1 - 2\mathbf{b}_2$ liefert die Gleichung für A_{S_2} .

8.3.4 Eine Parabel

$$\begin{aligned} Q_4(\mathbf{x}) &:= 9x^2 - 12xy + 4y^2 + \sqrt{13}(22x - 32y) + 533 \\ &= \left\langle \mathbf{x} \left| \begin{bmatrix} 9 & -6 \\ -6 & 4 \end{bmatrix} \mathbf{x} \right. \right\rangle + \sqrt{13} \left\langle \begin{bmatrix} 22 \\ -32 \end{bmatrix} \middle| \mathbf{x} \right\rangle + 533. \end{aligned}$$



Die Eigenwerte der Matrix sind $\lambda_1 = 0$ und $\lambda_2 = 13$, mit den zugehörigen Eigenvektoren $\mathbf{b}_1 = \frac{1}{\sqrt{13}} \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ und $\mathbf{b}_2 = \frac{1}{\sqrt{13}} \begin{bmatrix} -3 \\ 2 \end{bmatrix}$. Die Transformationsmatrix: $B = \frac{1}{\sqrt{13}} \begin{bmatrix} 2 & -3 \\ 3 & 2 \end{bmatrix}$.

$$\tilde{Q}_4(\mathbf{x}_B) = \left\langle \mathbf{x}_B \left| \begin{bmatrix} 0 & 0 \\ 0 & 13 \end{bmatrix}_B \mathbf{x}_B \right. \right\rangle + \left\langle \begin{bmatrix} -52 \\ -130 \end{bmatrix}_B \middle| \mathbf{x}_B \right\rangle + 533$$

$$\begin{aligned}
&= 13\tilde{y}^2 - 52\tilde{x} - 130\tilde{y} + 533 \\
&= 13(\tilde{y}^2 - 10\tilde{y} + 5^2) - 52\tilde{x} + 208 \\
&= 13 \left[(\tilde{y} - 5)^2 - 4(\tilde{x} - 4) \right].
\end{aligned}$$

Da der \tilde{x}^2 -Term fehlt, lässt sich die quadratische Ergänzung nur für \tilde{y} durchführen. Wir erhalten die um 5 Einheiten in \mathbf{b}_2 - und um 4 Einheiten in \mathbf{b}_1 -Richtung verschobene Parabel $\tilde{y}^2 = 4\tilde{x}$. Damit ist $p = 2$. Der Scheitel liegt bei $s = 5\mathbf{b}_2 + 4\mathbf{b}_1 = \frac{1}{\sqrt{13}} \begin{bmatrix} -7 \\ 22 \end{bmatrix}$ und der Brennpunkt bei $s + \frac{p}{2} \mathbf{b}_1 = \frac{1}{\sqrt{13}} \begin{bmatrix} -5 \\ 25 \end{bmatrix}$. Die Leitgerade ℓ geht durch den Punkt $s - \frac{p}{2} \mathbf{b}_1 = \frac{1}{\sqrt{13}} \begin{bmatrix} -9 \\ 19 \end{bmatrix}$ und hat den Normalenvektor \mathbf{b}_1 . Damit ist $\ell = \{ [x, y]^t \mid 2x + 3y = 3\sqrt{13} \}$.

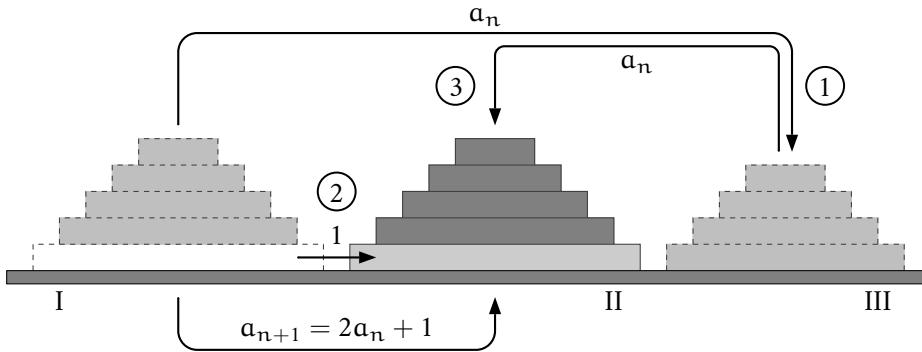
8.3.5 A Untersuchen Sie die quadratischen Formen

$$\begin{aligned}
Q_5(x) &:= 17x^2 - 16xy + 17y^2 - 104x - 4y - 13, \\
Q_6(x) &:= 25x^2 + 120xy + 144y^2 + 286x + 247y + 676.
\end{aligned}$$

8.3.6 A Untersuchen Sie die quadratische Form

$$Q_7(x) := 43x^2 + 36y^2 + 43z^2 + 36xy - 36yz - 22zx - 120x + 288y - 264z + 936.$$

9 Rekurrenzgleichungen



Bei dem Spiel *Turm von Hanoi* geht es darum, einen Turm aus Steinen abnehmender Größe Stein für Stein von Position I in die Position II zu versetzen, ohne einen größeren auf einen kleineren Stein zu legen. Dabei steht der Platz III als Zwischenablage zur Verfügung. a_n sei die Anzahl der Züge, die nötig sind, um einen Turm aus n Steinen zu versetzen. Für einen Turm aus $n+1$ Steinen werden a_{n+1} Züge gebraucht, um die ersten n Steine auf Platz III zu transportieren, einen weiteren, um den letzten Stein nach II zu bringen und anschließend wieder a_n Züge, um die n Steine von Platz III in ihre endgültige Position II zu verschieben. Also ist

$$a_{n+1} = 2a_n + 1. \quad (9.1)$$

Mit der offensichtlichen Anfangsbedingung $a_1 = 1$ haben wir für das Problem eine sog. *Rekurrenzgleichung* (oder *Rekursionsgleichung*) aufgestellt. Das Merkmal einer solchen Gleichung ist, daß die Kenntnis von a_{n+1} auf die Kenntnis von a_n zurückgreift und die wiederum auf die von a_{n-1} usw., bis man bei der Anfangsbedingung a_1 (manchmal auch bei a_0) angekommen ist. Das Ziel ist eine explizite Formel für a_n . Eine mögliche Methode dafür, die bei einfachen Gleichungen erfolgreich sein kann, ist die Rückwärtsentwicklung:

$$\begin{aligned} a_n &= 2a_{n-1} + 1 = 2(2a_{n-2} + 1) + 1 = 4a_{n-2} + 2 + 1 \\ &= 4(2a_{n-3} + 1) + 2 + 1 = 2^3 a_{n-3} + 2^2 + 2^1 + 2^0 \\ &= 2^3(2a_{n-4} + 1) + 2^2 + 2^1 + 2^0 \\ &= 2^4 a_{n-4} + 2^3 + 2^2 + 2^1 + 2^0 \\ &= \dots \\ &= 2^{n-1} a_1 + 2^{n-2} + \dots + 2^2 + 2^1 + 2^0 = \sum_{k=0}^{n-1} 2^k. \end{aligned}$$

In diesem Fall hat sie zwar eine Formel in Form einer Summe geliefert, da die Summenbildung jedoch selbst ein rekursiver Vorgang ist, wird nur dann ein echter Fortschritt erzielt, wenn für

sie eine explizite Formel gefunden werden kann. Das ist meistens nicht der Fall. Glücklicherweise handelt es sich bei diesem Beispiel um die sog. *geometrische Reihe*, für die sehr wohl eine geschlossene Formel bekannt ist. Um sie zu finden setzt man $S_n := \sum_{k=0}^n q^k$ (im Beispiel ist natürlich $q = 2$). Dann ist

$$q \cdot S_n = q + q^2 + q^3 + \cdots + q^n + q^{n+1}.$$

Zieht man davon $S_n = 1 + q + q^2 + \cdots + q^n$ ab, so ergibt sich

$$(q - 1)S_n = q^{n+1} - 1$$

und daraus, für $q \neq 1$,

$$\sum_{k=0}^n q^k = \frac{q^{n+1} - 1}{q - 1}. \quad (9.2)$$

Im vorliegenden Beispiel erhalten wir

$$a_n = \sum_{k=0}^{n-1} 2^k = \frac{2^n - 1}{2 - 1} = 2^n - 1.$$

Demnach sind für einen Turm aus 5 Steinen, bei der vorgestellten Lösungsstrategie, 31 Züge nötig.

9.1 Lineare Rekurrenzgleichungen 1. Ordnung

Wir bringen die Gleichung (9.1) in die systematische Form einer linearen Rekurrenzgleichung 1. Ordnung

$$a_{n+1} - 2a_n = 1. \quad (9.3)$$

Im Hinblick auf unsere Lösungstheorie linearer Gleichungen lösen wir zunächst die homogene lineare Rekurrenzgleichung

$$a_{n+1} - 2a_n = 0 \quad (9.4)$$

durch den Ansatz

$$a_n^{(0)} := \lambda^n, \quad (9.5)$$

mit einem $\lambda \neq 0$, das noch zu bestimmen ist. Durch Einsetzen von (9.5) in (9.4) erhalten wir

$$\lambda^{n+1} - 2\lambda^n = \lambda^n(\lambda - 2) = 0.$$

Das bedeutet $\lambda = 2$, so daß die allgemeine homogene Lösung die Form

$$a_n^{(0)} = t 2^n$$

hat, mit einem noch beliebigen Faktor t . Dieser wird es uns erlauben unseren Ansatz für die allgemeine Lösung an die Anfangsbedingung $a_1 = 1$ anzupassen. Doch zunächst brauchen wir eine spezielle inhomogene Lösung $a_n^{(1)}$ von (9.3). Da die rechte Seite dieser Gleichung einfach

eine Konstante ist, versuchen wir den Ansatz einer konstanten Lösung $a_n^{(1)} = d = \text{konst}$. Wenn sich d bestimmen lässt, ist dieser Ansatz gerechtfertigt. Einsetzen in (9.3) ergibt

$$d - 2d = 1,$$

oder $d = -1$. Damit lautet die allgemeine Lösung von (9.3):

$$a_n = t 2^n - 1.$$

Die freie Konstante t bestimmen wir aus der Anfangsbedingung $a_1 = 1$,

$$1 = 2t - 1,$$

zu $t = 1$. Daher ist die Lösung von (9.3), wie wir bereits gesehen haben, durch $a_n = 2^n - 1$ gegeben.

9.1.1 Konstante Inhomogenität

$$a_n - ca_{n-1} = d, \quad d : \text{konstant}. \quad (9.6)$$

Der übliche Ansatz $a_n^{(0)} = \lambda^n$ für die homogene Lösung führt auf $\lambda = c$ (deshalb wurde das Minuszeichen vor dem c in die Definition von Rekurrenzgleichungen 1. Ordnung mit aufgenommen, da die homogene Lösung sonst $(-c)^n$ wäre).

Für die inhomogene Lösung versuchen wir den Ansatz $a_n^{(1)} = k$, mit einer noch zu bestimmenden Konstanten k . Einsetzen in (9.6) ergibt die Gleichung $k - ck = d$, die nur für $c \neq 1$ nach k auflösbar ist. Wir untersuchen daher vorerst den Fall $c \neq 1$. Dann ist $k = \frac{d}{1-c}$. Damit lautet die allgemeine Lösung von (9.6) bisher

$$a_n = t c^n + \frac{d}{1-c}.$$

t bestimmen wir aus der Anfangsbedingung a_0 :

$$a_0 = t + \frac{d}{1-c}, \quad \text{also} \quad t = a_0 - \frac{d}{1-c}.$$

Das ergibt für den Fall $c \neq 1$ die endgültige Lösung:

$$a_n = a_0 c^n + \frac{d}{1-c} (1 - c^n). \quad (9.7)$$

Für $c = 1$: Der Versuch, die inhomogene Lösung durch eine Konstante zu bestimmen scheitert, wie wir oben gesehen haben. Wir versuchen $a_n^{(1)} = k \cdot n$. In (9.6) eingesetzt:

$$kn - k(n-1) = d,$$

woraus $k = d$ folgt. Damit haben wir durch $a_n^{(1)} = dn$ eine spezielle inhomogene Lösung gefunden. Zusammen mit der homogenen Lösung $a_n^{(0)} = 1$ (s.o.) lautet die allgemeine Lösung $a_n = t + dn$. Die Anfangsbedingung a_0 bestimmt t zu a_0 , so daß die Lösung für $c = 1$ einfach durch

$$a_n = a_0 + dn \quad (9.8)$$

gegeben ist.

9.1.2 Variable Inhomogenität

Wir untersuchen nur Gleichungen der Form

$$a_n - c a_{n-1} = d \cdot n. \quad (9.9)$$

Sei zunächst $c \neq 1$. Die homogene Lösung ist wieder $a_n^{(0)} = c^n$. Für die inhomogene machen wir den Ansatz $a_n^{(1)} = x \cdot n + y$ und versuchen die Unbekannten x und y zu bestimmen. Einsetzen in (9.9) ergibt

$$\begin{aligned} xn + y - cx(n-1) - cy &= dn, \\ x(1-c)n + cx + y(1-c) &= dn. \end{aligned}$$

Da das für alle $n \geq 0$ gelten muß, ergibt sich für $n = 0$

$$y = -\frac{cx}{1-c}.$$

Vergleich der Koeffizienten von n zeigt

$$x = \frac{d}{1-c} \quad \text{und damit} \quad y = -\frac{cd}{(1-c)^2}.$$

Daher ist

$$a_n = tc^n + \frac{d}{1-c} n - \frac{cd}{(1-c)^2}.$$

Berücksichtigen wir noch die Anfangsbedingung a_0 , so erhalten wir endgültig für $c \neq 1$:

$$a_n = a_0 c^n + \frac{cd}{(1-c)^2} (c^n - 1) + \frac{d}{1-c} n. \quad (9.10)$$

Für den Fall $c = 1$ führt der Ansatz $a_n^{(1)} = x \cdot n + y$ auf die unmögliche Bestimmungsgleichung $x = d \cdot n$. Wir versuchen $a_n^{(1)} = x \cdot n^2 + y \cdot n + z$. Einsetzen in (9.9) ergibt

$$2xn + y - x = dn,$$

d. h. $x = \frac{d}{2}$ und $y = \frac{d}{2}$. Da z aus der Bestimmungsgleichung herausfällt, können wir es ohne Bedenken weglassen. Genau wie oben vorgeführt erhält man jetzt unschwer

$$a_n = a_0 + \frac{d}{2} n(n+1). \quad (9.11)$$

9.2 Lineare Rekurrenzgleichungen 2. Ordnung

Wir beginnen mit einem Beispiel, der *FIBONACCI-Folge*. Dabei handelt es sich um ein stark vereinfachtes Populationsmodell einer Kaninchenzucht (etwa aus dem Jahre 1200): Ausgehend von einem jungen Kaninchenpaar bringt jedes erwachsene Paar jährlich ein weiteres zur Welt. Erwachsen ist ein Kaninchenpaar bereits im zweiten Lebensjahr. Außerdem wird davon abgesehen, daß Paare auch sterben könnten. a_n soll die Anzahl der Paare nach n Jahren wiedergeben. Dann ist $a_0 = 1$ und $a_1 = 1$. Im zweiten Jahr hat das Kaninchenpaar Nachkommen, so daß

$a_2 = 2$ ist. Im nächsten Jahr hat wieder nur das erste Paar Nachkommen, denn der Nachwuchs ist noch nicht erwachsen: $a_3 = 3$. Im vierten Jahr sind die Nachkommen des zweiten Jahrs erwachsen und haben zusammen mit dem Elternpaar jeweils ein Kaninchenpaar d. h. a_2 Paare als Nachkommen. Zusammen mit den a_3 Paaren des dritten Jahrs sind inzwischen $a_4 = a_3 + a_2$ Paare vorhanden. Im $n + 1$ -ten Jahr sind die a_n Paare des vorausgehenden Jahrs und die a_{n-1} Nachkommen der in diesem Jahr erwachsen gewordenen Paare vorhanden. Damit haben wir die Rekursionsgleichung

$$a_{n+1} = a_n + a_{n-1}, \quad a_0 = 1, a_1 = 1, \quad (9.12)$$

der FIBONACCI-Folge gefunden. Wenn man alles auf eine Seite bringt und die Indizierung um eins verschiebt, erkennt man, daß es sich um eine homogene Rekurrenzgleichung handelt:

$$a_n - a_{n-1} - a_{n-2} = 0, \quad a_0 = 1, a_1 = 1. \quad (9.13)$$

Sie ist von zweiter Ordnung, da sie auf die zwei direkten Vorläufer a_{n-1} und a_{n-2} der Folge zurückgreift, um das aktuelle Folgenglied a_n zu berechnen. Die ersten 11 Glieder der Folge sind schnell bestimmt:

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89 \dots$$

Der systematische Lösungsansatz dieser homogenen Rekurrenzgleichung 2. Ordnung ist $a_n = \lambda^n$:

$$\lambda^n - \lambda^{n-1} - \lambda^{n-2} = \lambda^{n-2}(\lambda^2 - \lambda - 1) = 0. \quad (9.14)$$

$\lambda^2 - \lambda - 1 = 0$ hat die beiden Lösungen $\lambda_1 = \Phi$ und $\lambda_2 = -\frac{1}{\Phi}$, mit

$$\Phi := \frac{1}{2}(\sqrt{5} + 1) \approx 1.61803398875. \quad (9.15)$$

Es ist eine gute Übung nachzurechnen, daß die zweite Lösung $\lambda_2 = \frac{1}{2}(1 - \sqrt{5})$ tatsächlich mit $-\frac{1}{\Phi} = 1 - \Phi$ übereinstimmt. Die Zahl $\varphi := \frac{1}{\Phi} \approx 0.61803398875$ wird als *goldener Schnitt* bezeichnet. Die allgemeine Lösung der homogenen Gleichung (9.13) ist wegen deren Linearität durch eine Linearkombination der Lösungen Φ^n und $(-\frac{1}{\Phi})^n = (1 - \Phi)^n$ gegeben:

$$a_n = t_1 \Phi^n + t_2 (1 - \Phi)^n. \quad (9.16)$$

t_1 und t_2 werden aus den Anfangsbedingungen $a_0 = 1$ und $a_1 = 1$ bestimmt:

$$\begin{aligned} t_1 + t_2 &= 1, \\ t_1 \Phi + t_2 (1 - \Phi) &= 1. \end{aligned}$$

Dieses lineare Gleichungssystem führt auf $t_2 = 1 - t_1$ und das auf $1 = t_1 \Phi + (1 - t_1)(1 - \Phi) = t_1(2\Phi - 1) + 1 - \Phi = t_1\sqrt{5} + 1 - \Phi$, was leicht nach t_1 aufgelöst werden kann:

$$t_1 = \frac{\Phi}{\sqrt{5}}, \quad t_2 = \frac{\sqrt{5}}{\sqrt{5}} - \frac{\Phi}{\sqrt{5}} = \frac{\frac{1}{2}(\sqrt{5} - 1)}{\sqrt{5}} = -\frac{1 - \Phi}{\sqrt{5}}.$$

In Gleichung (9.16) eingesetzt ergibt das die endgültige Lösung von (9.13),

$$a_n = \frac{1}{\sqrt{5}} (\Phi^{n+1} - (1-\Phi)^{n+1}). \quad (9.17)$$

Da der zweite Summand $(1-\Phi)^{n+1} = (-1)^{n+1}\Phi^{-n-1}$ rasch klein wird, kann (9.17) zur schnellen Berechnung großer FIBONACCI-Zahlen verwendet werden. Dazu bestimmt man nämlich nur

$$a_n \approx \frac{1}{\sqrt{5}} \Phi^{n+1}$$

und runden zur nächsten ganzen Zahl auf bzw. ab. Schon für $n = 10$ und $n = 11$ ergibt sich

$$a_{10} \approx \frac{1}{\sqrt{5}} \Phi^{11} \approx 88.997, \quad a_{11} \approx 144.001,$$

d. h. $a_{10} = 89$ und $a_{11} = 144$. Für $n = 26$ kann man den Unterschied von $a_{26} \approx \frac{1}{\sqrt{5}} \Phi^{27}$ zu $a_{26} = 196418$ kaum noch darstellen. Für $a_{38} \approx \frac{1}{\sqrt{5}} \Phi^{39}$ zeigen die meisten Rechner nur das exakte Ergebnis 63245986 an.

Wir untersuchen nun die allgemeine Rekurrenzgleichung 2. Ordnung mit konstanter Inhomogenität:

$$a a_n + b a_{n-1} + c a_{n-2} = d, \quad a \neq 0. \quad (9.18)$$

Der Ansatz $a_n^{(0)} = \lambda^n$ für die homogene Lösung führt jetzt auf

$$a\lambda^n + b\lambda^{n-1} + c\lambda^{n-2} = \lambda^{n-2}(a\lambda^2 + b\lambda + c) = 0,$$

d. h. auf die gewöhnliche quadratische Gleichung

$$a\lambda^2 + b\lambda + c = 0, \quad (9.19)$$

die bekanntlich die Lösungen

$$\lambda_{1/2} = \frac{1}{2a} \left(-b \pm \sqrt{b^2 - 4ac} \right) \quad (9.20)$$

hat. Wir kümmern uns nicht darum, ob der Ausdruck unter der Wurzel positiv oder negativ ist (im letzteren Fall merken wir uns vor, die Lösung komplex aufzuschreiben). Allerdings müssen wir uns sehr wohl um den Sonderfall $b^2 = 4ac$ Gedanken machen, in dem (9.20) nur die eine Lösung $\lambda = -\frac{b}{2a}$ hat.

$\lambda_1 \neq \lambda_2$ (d. h. $b^2 \neq 4ac$): Die allgemeine homogene Lösung hat die Form

$$a_n^{(0)} = t_1 \lambda_1^n + t_2 \lambda_2^n. \quad (9.21)$$

Für die inhomogene Lösung machen wir, wie bisher immer, zunächst den Ansatz $a_n^{(1)} = x$, mit einer zu bestimmenden Konstanten x . Einsetzen in (9.18) ergibt $x = \frac{1}{a+b+c}$, falls $a+b+c \neq 0$ ist. Für $a+b+c = 0$ führt der Ansatz $a_n^{(1)} = x \cdot n$ zu $x = -\frac{d}{b+2c}$ (b kann wegen $b^2 \neq 4ac$ nicht mit $-2c$ übereinstimmen, so daß hier keine weitere Fallunterscheidung zu treffen

ist). Wie im Beispiel der FIBONACCI-Folge vorgeführt, läßt sich nun die allgemeine Lösung aus der homogenen und der inhomogenen zusammensetzen. Die Anfangsbedingungen a_0 und a_1 führen auf ein lineares Gleichungssystem für die beiden Unbekannten t_1 und t_2 .

$\lambda_1 = \lambda_2$ (d. h. $b^2 = 4ac$ und $\lambda = \lambda_1 = -\frac{b}{2a}$): Jetzt haben wir das Problem, daß uns ein Ansatz der Form (9.21) für die homogene Lösung zunächst nicht möglich ist, da uns nur eine Lösung λ^n der homogenen Rekurrenzgleichung zur Verfügung steht. Um aber einen Lösungsansatz a_n an die beiden Anfangsbedingungen a_0 und a_1 anzupassen, benötigen wir zwei Lösungen der homogenen Rekurrenzgleichung, damit wir sie, wie in (9.21), mit zwei Unbekannten t_1 und t_2 kombinieren können. Erst dann erhalten wir ein Gleichungssystem aus zwei linearen Gleichungen für t_1 und t_2 .

Wir zeigen nun, daß durch $n \cdot \lambda^n$ eine weitere Lösung der homogenen Rekurrenzgleichung gegeben ist. Einsetzen in $aa_n + ba_{n-1} + ca_{n-2}$ müßte Null ergeben. Dabei können wir natürlich $a\lambda^n + b\lambda^{n-1} + c\lambda^{n-2} = 0$ verwenden, denn λ wurde ja gerade so bestimmt, daß diese Gleichung erfüllt ist.

$$\begin{aligned} a n \lambda^n + b(n-1)\lambda^{n-1} + c(n-2)\lambda^{n-2} &= n(a\lambda^n + b\lambda^{n-1} + c\lambda^{n-2}) - b\lambda^{n-1} - 2c\lambda^{n-2} \\ &= -\lambda^{n-2}(b\lambda + c) = -\lambda^{n-2}\left(-\frac{b^2}{2a} + 2c\right) \\ &= -\lambda^{n-2}\left(-\frac{4ac}{2a} + 2c\right) = 0. \end{aligned}$$

Der Ansatz für die homogene Lösung lautet jetzt

$$a_n^{(0)} = (t_1 + t_2 n) \lambda^n, \quad \lambda = -\frac{b}{2a}. \quad (9.22)$$

Nun kann man mit der Bestimmung einer inhomogenen Lösung fortfahren, so wie es oben vorgestellt wurde.

9.2.1 A i) Zeigen Sie (mit Induktion), daß die maximale Anzahl a_n der Teile, in die man eine Ebene durch n Geraden zerlegen kann, der Rekursion $a_{n+1} = a_n + n + 1$, oder $a_n - a_{n-1} = n$, mit $a_0 = 1$ genügt. Geben Sie eine Lösung dieser Rekursionsgleichung an.

ii) Zeigen Sie die sogenannte *Mitternachtsformel*

$$x_{1/2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \quad (9.23)$$

für die Lösungen x_1 und x_2 der quadratischen Gleichung $ax^2 + bx + c = 0$, falls die sogenannte *Diskriminante* $b^2 - 4ac$ nicht negativ ist. Leiten Sie daraus die *pq-Formel*

$$x_{1/2} = -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q} \quad (9.24)$$

für die Lösungen der quadratischen Gleichung in Normalform $x^2 + px + q = 0$ ab.

10 Folgen und Reihen

Eine *Folge* $x = (x_n)_{n \in \mathbb{N}}$ ist eine geordnete Aufzählung von Objekten x_n einer Menge X . Jeder natürlichen Zahl n wird dabei genau ein Objekt $x_n \in X$ zugeordnet. Wir können daher eine Folge auch als eine Funktion x von \mathbb{N} nach X auffassen:

$$x: \mathbb{N} \rightarrow X; n \mapsto x(n) = x_n.$$

Meistens werden wir uns aber eine Folge durch ihre Werte repräsentiert vorstellen

$$x = (x_1, x_2, x_3, \dots, x_n, x_{n+1}, \dots).$$

Die Werte x_n können dabei durch Formeln explizit gegeben sein, wie in den folgenden Beispielen

$$\begin{array}{llll} x_n := \frac{1}{n} & x_n := \left(1 + \frac{1}{n}\right)^n & x_n := n \cdot \sin\left(\frac{1}{n}\right) & x_n := \frac{1}{n} \sin^n \\ x_n := \frac{n^{10}}{2^n} & x_n := \sqrt[n]{n} & x_n := \begin{bmatrix} n^2 e^{-n} \\ \sin(n) \\ n(e^{\frac{1}{n}} - 1) \end{bmatrix} & \dots \end{array}$$

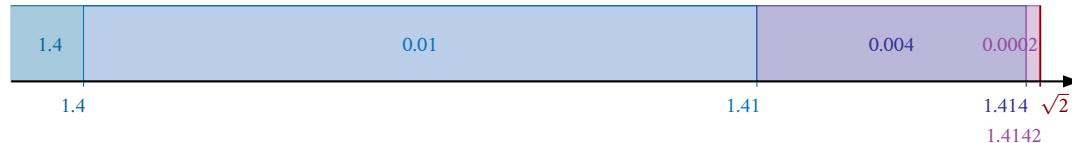
oder rekursiv durch eine Gleichung mit einer oder mehreren Anfangsbedingungen, wie etwa

$$x_{n+1} = x_n + x_{n-1}, \quad x_1 = 1, \quad x_2 = 1, \quad x_{n+1} = \frac{1}{2} \left(x_n + \frac{2}{x_n} \right), \quad x_1 = 1.5, \quad \dots$$

Diese Beispiele zeigen es schon, daß die Menge X normalerweise eine reiche mathematische Struktur aufweist, die es erlaubt, eine Folge genauer zu untersuchen. Wir werden uns hauptsächlich mit $X = \mathbb{R}$, \mathbb{C} , \mathbb{R}^k , oder $X = \mathbb{C}^k$ befassen. Im Rahmen der Analysis sind wir an der Frage interessiert, ob es für eine gegebene Folge $(x_n)_{n \in \mathbb{N}}$ ein Element von X gibt, das durch sie beliebig genau approximiert wird. Wenn es ein solches Element geben sollte, nennen wir es *Grenzwert*, *Grenzelement*, oder *Limes* der Folge und die Folge selbst *konvergent*. Andernfalls bezeichnen wir sie als *divergent*. Um der Frage nach der Existenz eines Grenzwertes nachzugehen zu können, sollte der Raum X wenigstens normiert sein, damit wir die Möglichkeit haben, dem Begriff Approximation durch Abstandsmessungen mit Hilfe der Norm einen Sinn zu verleihen. In den meisten Fällen wird sich diese Norm auf den gewöhnlichen Betrag in \mathbb{R} oder \mathbb{C} reduzieren lassen.

Die Elemente S_n einer *Reihe* $(S_n)_{n \in \mathbb{N}}$ entstehen aus einer Folge $(x_k)_{k \in \mathbb{N}}$ durch Summieren der ersten n Glieder: $S_n = \sum_{k=1}^n x_k$. Eine Reihe ist also eine speziell konstruierte Folge. Hier besteht die Aufgabe normalerweise darin, die Summation bis ins Unendliche fortzusetzen. Das ist eine Situation, der man fast täglich begegnet, ohne sich dessen immer

recht bewußt zu sein. So hat die Zahl $\sqrt{2} = 1.4142135623730951454746218587388\dots$ unendlich viele Nachkommastellen, was hier nichts anderes bedeutet, als daß wir uns die Addition $1 + 0.4 + 0.01 + 0.004 + 0.0002 + 0.00001 + 0.000003 + \dots$ unendlich weit fortgesetzt zu denken haben. Dabei stellen wir uns normalerweise nicht die Frage, ob das überhaupt geht, ob nicht durch unbegrenztes Hinzufügen von Zahlen, die zwar schnell klein werden, die Summe vielleicht doch über alle Grenzen wächst.



Der Grund ist, daß wir eine klare Vorstellung davon haben, daß die Zahl $\sqrt{2}$ zwischen 1.4 und 1.42 liegt. Wir wissen also, die Summation ist möglich, wir wissen aber noch nicht warum (siehe 10.2.5).

Die sogenannte *harmonische Reihe* $\sum_{k=1}^n \frac{1}{k}$ ist ein Beispiel dafür, daß die Frage nach der unbegrenzten Summierbarkeit durchaus berechtigt ist.

$$\begin{aligned} \sum_{k=1}^{2^n} \frac{1}{k} &= 1 + \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}}_{\geq \frac{2}{4} = \frac{1}{2}} + \underbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}_{\geq \frac{4}{8} = \frac{1}{2}} + \underbrace{\frac{1}{9} + \frac{1}{10} + \frac{1}{11} + \frac{1}{12} + \frac{1}{13} + \frac{1}{14} + \frac{1}{15} + \frac{1}{16}}_{\geq \frac{8}{16} = \frac{1}{2}} \\ &\quad + \underbrace{\frac{1}{17} + \dots + \frac{1}{32}}_{\geq \frac{16}{32} = \frac{1}{2}} + \underbrace{\frac{1}{33} + \dots + \frac{1}{64}}_{\geq \frac{32}{64} = \frac{1}{2}} + \dots + \underbrace{\frac{1}{2^{n-1}+1} + \dots + \frac{1}{2^n}}_{\geq \frac{2^{n-1}}{2^n} = \frac{1}{2}} \geq 1 + \frac{n}{2}. \end{aligned}$$

Diese Reihe wächst nämlich, wenn auch sehr langsam (wieso?), über alle Grenzen. Ganz offensichtlich ist es notwendig, daß die Summanden gegen Null streben, damit eine Reihe summierbar ist, aber wie das Beispiel zeigt, ist diese Eigenschaft nicht hinreichend. Wir benötigen verlässlichere Kriterien.

Ein typisches Problem bei der Untersuchung von Reihen besteht darin, daß wir meistens keine Idee dafür vorweisen können, gegen welche Zahl sie konvergiert, falls sie konvergiert. Es wäre demnach wünschenswert, ein Konvergenzkriterium zur Verfügung zu haben, das ohne eine Vermutung über den Grenzwert auskommt. In den Räumen, in denen wir Konvergenzuntersuchungen durchführen, wird uns mit dem *CAUCHY-Kriterium* eines zur Verfügung stehen. Ein prominentes Beispiel für eine Reihe dieser Art ist die *Exponentialreihe*

$$e = \sum_{k=0}^{\infty} \frac{1}{k!}. \quad (10.1)$$

10.0.1 A Die Funktion $f(x) := \frac{1}{x}$ schließt mit dem Teil $[1, \infty)$ der x-Achse eine unendliche Fläche ein. Hat diese einen endlichen, oder einen unendlichen Flächeninhalt?

10.1 Der Grenzwertbegriff

Gegeben sei eine Folge $(x_n)_{n \in \mathbb{N}}$ mit Elementen x_n aus einem normierten Raum X . Gibt es ein $x \in X$, das von der Folge beliebig genau approximiert wird, so nennen wir sie konvergent und bezeichnen x als ihren Grenzwert oder Limes. Aber was soll *beliebig genau approximieren* heißen? Dazu gehen wir erst einmal von einer festen Genauigkeit der Approximation aus, sagen wir von $\varepsilon > 0$. Dann gestehen wir einen Fehler $\|x - x_n\| < \varepsilon$ in der Approximation von x durch die x_n zu. Es ist natürlich nicht ausreichend, wenn diese Genauigkeit für nur einige wenige, oder sogar für unendlich viele Folgenglieder erreicht wird, wenn es immer wieder Elemente x_n geben sollte, die die geforderte Genauigkeit nicht erreichen, d. h. für die $\|x - x_n\| \geq \varepsilon$ gilt. Wir werden also sagen, daß die Folge $(x_n)_{n \in \mathbb{N}}$ das Element x mit der Genauigkeit $\varepsilon > 0$ approximiert, wenn irgendwann einmal alle Folgenglieder sich um weniger als ε von x unterscheiden, d. h., wenn es einen Index $n_\varepsilon \in \mathbb{N}$ gibt, ab dem die geforderte Genauigkeit $\|x - x_n\| < \varepsilon$ erreicht ist. Für alle $n \geq n_\varepsilon$ gilt demnach die Ungleichung $\|x - x_n\| < \varepsilon$. Natürlich wollen wir jede Genauigkeit $\varepsilon > 0$ erreichen und müssen daher für jedes $\varepsilon > 0$ die Existenz eines $n_\varepsilon \in \mathbb{N}$ fordern, ab dem die Abweichungen $\|x - x_n\|$ der Folgenglieder x_n vom Grenzwert x unterhalb der geforderten Genauigkeit ε bleiben.

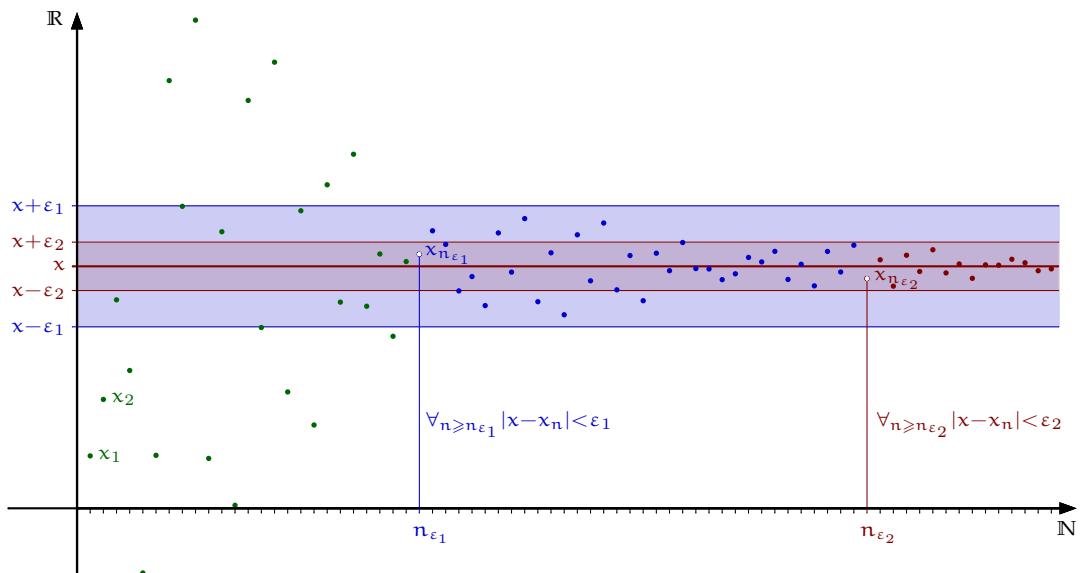


Abb. 10.1 Konvergenz in \mathbb{R}

10.1.1 Definition Eine Folge $(x_n)_{n \in \mathbb{N}}$ aus dem normierten Raum X hat einen Grenzwert $x \in X$, wenn es für jedes $\varepsilon > 0$ einen Index $n_\varepsilon \in \mathbb{N}$ gibt, so daß

$$\|x - x_n\| < \varepsilon \quad \text{für alle } n \geq n_\varepsilon \quad (10.2)$$

gilt. Mit Hilfe der Quantoren \forall und \exists lässt sich das knapper formulieren:

$$\forall \varepsilon > 0 \exists n_\varepsilon \in \mathbb{N} \forall n \geq n_\varepsilon \|x - x_n\| < \varepsilon. \quad (10.3)$$

Dafür schreiben wir

$$x = \lim_{n \rightarrow \infty} x_n, \quad \text{oder} \quad x_n \xrightarrow{n \rightarrow \infty} x. \quad (10.4)$$

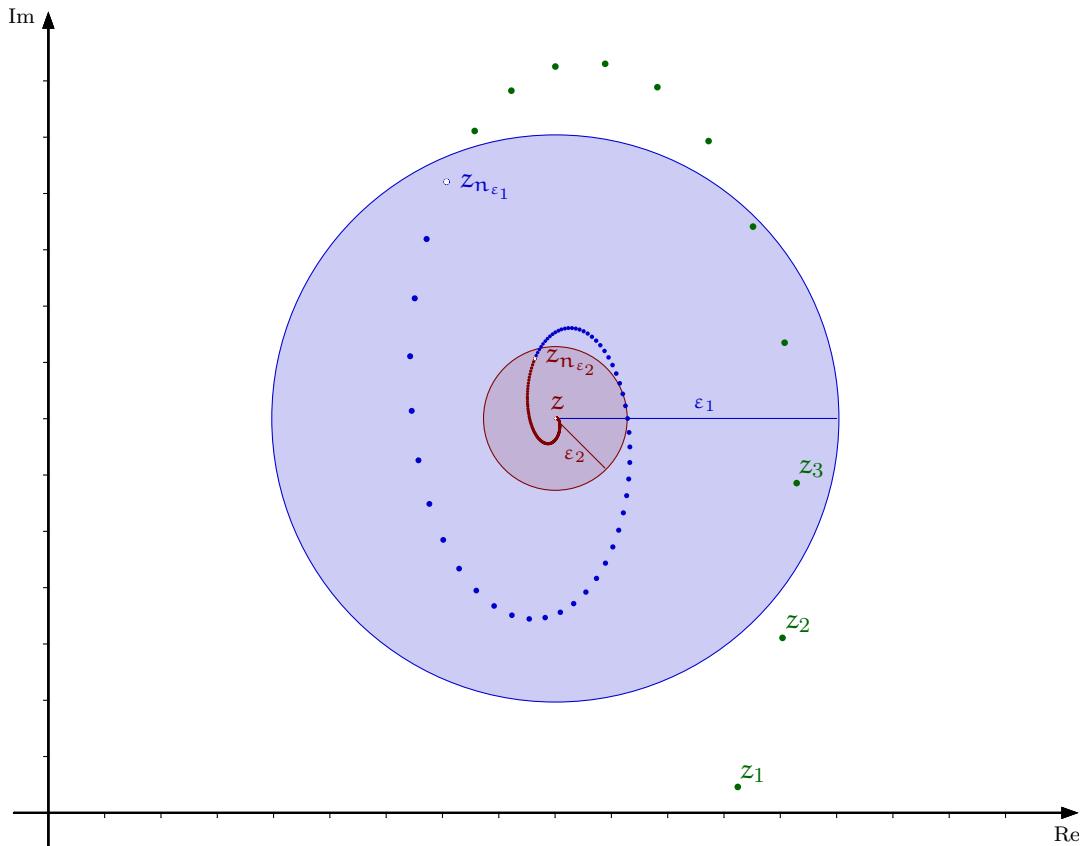


Abb. 10.2 Konvergenz in \mathbb{C}

Für $X = \mathbb{R}$ ist die Norm der gewöhnliche Betrag $|x|$, für $X = \mathbb{C}$ der Betrag $|z_1 + iz_2| = \sqrt{z_1^2 + z_2^2}$ in \mathbb{C} und für $X = \mathbb{R}^3$ ist sie über den Satz des PYTHAGORAS durch die euklidische Länge $\|x\| = \|(x_1, x_2, x_3)\| = \sqrt{x_1^2 + x_2^2 + x_3^2}$ gegeben. Natürlich ist unsere Definition mit diesen Beispielen nicht ausgeschöpft, aber sie stecken das Gebiet ab, in dem wir uns vorwiegend bewegen werden.

Ziehen wir Folgerungen aus unserer Definition. Das erste, was wir von einen sinnvollen Grenzwertbegriff verlangen, ist sicherlich, daß es für eine konvergente Folge maximal *einen* Grenzwert geben darf.

10.1.2 Lemma Der Grenzwert einer konvergenten Folge ist eindeutig.

Beweis. Sei $x = \lim_{n \rightarrow \infty} x_n$ und $x' = \lim_{n \rightarrow \infty} x'_n$. Dann gibt es zu jedem $\varepsilon > 0$ eine $n_\varepsilon \in \mathbb{N}$ und ein

$n'_\varepsilon \in \mathbb{N}$, so daß für alle $n \geq n_\varepsilon$ bzw. $n \geq n'_\varepsilon$

$$\|x - x_n\| < \varepsilon \quad \text{und} \quad \|x' - x_n\| < \varepsilon$$

gilt. Wenn wir den größeren der beiden Indizes n_ε und n'_ε wählen, können wir o. B. d. A. davon ausgehen, daß obige Abschätzungen ab einem gemeinsamen n_ε gelten. Mit Hilfe der Dreiecksungleichung und *Einschieben der Null* in der Form $-x_n + x_n$ können wir dann den Abstand $\|x - x'\|$ von x und x' folgendermaßen abschätzen:

$$\|x - x'\| = \|x - x_n + x_n - x'\| \leq \|x - x_n\| + \|x_n - x'\| < 2\varepsilon$$

für alle $\varepsilon > 0$. Betrachten wir nur den Anfang und das Ende dieser Ungleichung, so haben wir $\|x - x'\| < 2\varepsilon$ für alle $\varepsilon > 0$ gezeigt (die Folgenglieder x_n kommen gar nicht mehr vor). Die einzige nicht negative Zahl, die unter jede noch so kleine Zahl 2ε paßt, ist die Zahl 0. Es muß daher $\|x - x'\| = 0$, wegen der Definitheit der Norm also $x = x'$ gelten. \square

Von einer konvergenten Folge ist zu erwarten, daß sie nicht zu beliebig großen Werten fähig ist, denn ab einem n_ε müssen sich alle Folgenglieder in der Nähe des Grenzwertes x aufhalten. Von den $n_\varepsilon - 1$ Elementen $x_1, \dots, x_{n_\varepsilon-1}$, die einen Abstand $\geq \varepsilon$ von x haben können, hat wenigstens eines die größte Norm.

10.1.3 Lemma *Jede konvergente Folge $(x_n)_{n \in \mathbb{N}}$ ist beschränkt, d.h. es gibt ein $M > 0$, so daß $\|x_n\| \leq M$ für alle $n \in \mathbb{N}$ gilt.*

Beweis. Für jedes $\varepsilon > 0$ gibt es ein $n_\varepsilon \in \mathbb{N}$, so daß für alle $n \geq n_\varepsilon$ die Abschätzung $\|x - x_n\| < \varepsilon$ erfüllt ist, wenn $x \in X$ der Grenzwert der Folge ist. Also haben wir

$$\|x_n\| = \|x_n - x + x\| \leq \|x_n - x\| + \|x\| < \varepsilon + \|x\|$$

für alle $n \geq n_\varepsilon$. Mit $M := \max \{\|x_1\|, \dots, \|x_{n_\varepsilon-1}\|, \varepsilon + \|x\|\}$ ist eine obere Schranke aller Folgenglieder gefunden. \square

Normalerweise wird bei Konvergenzuntersuchungen der Nachweis für die Existenz von n_ε indirekt geführt. Nur selten wird dieser Index explizit angegeben. Die Folge $(\frac{1}{n})_{n \in \mathbb{N}}$ ist eine der wenigen, bei der n_ε direkt berechnet werden kann.

10.1.4 Lemma *Die reelle Folge $(\frac{1}{n})_{n \in \mathbb{N}}$ hat den Grenzwert 0.*

Beweis. Nach dem Archimedischen Axiom (siehe 10.1.23), das wir für den Moment als evident ansehen, gibt es zu jedem $\varepsilon \in \mathbb{R}^+$ eine kleinste natürliche Zahl n_ε mit der Eigenschaft $n_\varepsilon > \frac{1}{\varepsilon}$. Dann gilt für alle $n \geq n_\varepsilon$:

$$n > \frac{1}{\varepsilon} \quad \Rightarrow \quad \frac{1}{n} = \left| \frac{1}{n} - 0 \right| < \varepsilon .$$

\square

Folgen, deren Grenzwert die Zahl Null ist, spielen eine besondere Rolle, denn durch Übergang von einer konvergenten Folge $(x_n)_{n \in \mathbb{N}}$ zur Folge $(x_n - x)_{n \in \mathbb{N}}$ der Differenzen zum Grenzwert x , kann jede Konvergenzuntersuchung auf die einer Folge mit Grenzwert Null zurückgeführt werden. Wir nennen solche Folgen *Nullfolgen*.

Offensichtlich sind mit $(\frac{1}{n})_{n \in \mathbb{N}}$ auch $(\frac{1}{n^2})_{n \in \mathbb{N}}, (\frac{1}{n^3})_{n \in \mathbb{N}}, \dots (\frac{1}{n^k})_{n \in \mathbb{N}}$ Nullfolgen, denn sie streben ja noch viel schneller als $\frac{1}{n}$ gegen Null. Diese Vermutung bestätigt sich als Spezialfall des folgenden Satzes.

10.1.5 Satz (Sandwich-Prinzip) *Für zwei reelle Folgen $(x_n)_{n \in \mathbb{N}}$ und $(y_n)_{n \in \mathbb{N}}$ mit demselben Grenzwert x und einer Folge $(z_n)_{n \in \mathbb{N}}$ mit der Eigenschaft $x_n \leq z_n \leq y_n$ gilt*

$$x = \lim_{n \rightarrow \infty} z_n .$$

Bemerkung: Wenn wir $x = \lim_{n \rightarrow \infty} z_n$ schreiben, meinen wir immer *zwei* Dinge, nämlich erstens, daß der Grenzwert von $(z_n)_{n \in \mathbb{N}}$ existiert und daß er zweitens mit x übereinstimmt.

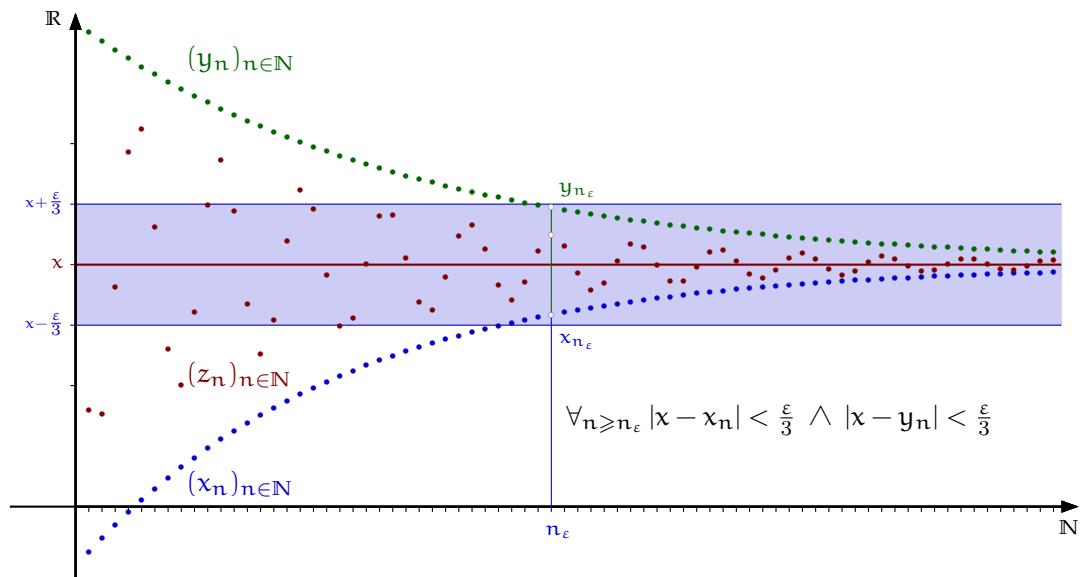


Abb. 10.3 Zum Sandwich-Prinzip

Beweis. Es gibt zu jedem $\varepsilon > 0$ ein $\tilde{n}_\varepsilon \in \mathbb{N}$ und ein $\hat{n}_\varepsilon \in \mathbb{N}$, so daß für alle $n \geq \tilde{n}_\varepsilon$ bzw. für alle $n \geq \hat{n}_\varepsilon$

$$|x - x_n| < \frac{\varepsilon}{3} \quad \text{bzw.} \quad |x - y_n| < \frac{\varepsilon}{3}$$

gilt. Wir wählen für n_ε das Maximum der beiden Werte \tilde{n}_ε und \hat{n}_ε . Dadurch haben wir die beiden Abschätzungen ab diesem gemeinsamen n_ε zu unserer Verfügung. Wir erhalten

$$\begin{aligned} |x - z_n| &= |x - x_n + x_n - z_n| \leq |x - x_n| + |x_n - z_n| = |x - x_n| + z_n - x_n \\ &\leq |x - x_n| + y_n - x_n = |x - x_n| + y_n - x + x - x_n \\ &\leq 2|x - x_n| + |y_n - x| < 3 \cdot \frac{\varepsilon}{3} = \varepsilon \end{aligned}$$

für alle $n \geq n_\varepsilon$. □

Bemerkung: In diesem Beweis haben wir etwas vorgeführt, das man *Beweiskosmetik* nennen könnte. Damit ist folgendes gemeint. In der Definition 10.1.1 des Grenzwertes x einer Folge $(z_n)_{n \in \mathbb{N}}$ verlangen wir, daß ab einem geeigneten $n_\varepsilon \in \mathbb{N}$ die Abschätzung $|x - z_n| < \varepsilon$ erfüllt sein muß. Das ist der Grund, warum wir im eben vorgeführten Beweis zunächst etwas unmotiviert die Forderungen $|x - x_n| < \frac{\varepsilon}{3}$ und $|x - y_n| < \frac{\varepsilon}{3}$ aufgestellt haben (da die Konvergenz von $(x_n)_{n \in \mathbb{N}}$ gegen x den Abstand $|x - x_n|$ schließlich unter *jedes* $\varepsilon > 0$ zwingt, wird sie das sicher auch für $\frac{\varepsilon}{3}$ erreichen). Damit haben wir beweckt, daß am Ende unserer Abschätzungen $|x - z_n|$ nicht kleiner als 3ε , sondern kleiner als ε wird. Bei einem so einfachen Beweis ist das noch leicht zu erreichen. Aber natürlich hätten wir die Behauptung auch gezeigt, wenn unsere Abschätzung nur $|x - z_n| < 3\varepsilon$ ergeben hätte. Denn da sie ja für *alle* $\varepsilon > 0$ erreicht wurde, können wir uns ε auch durch $\frac{\varepsilon}{3}$ ersetzt denken, um damit den formalen Kriterien von Definition 10.1.1 Genüge zu tun. Eine solche Beweiskosmetik wird die Beweisführung meist nur unübersichtlicher machen. Deshalb werden wir sie im Folgenden meist unterlassen. Wenn eine Abschätzung $|x - z_n| < M \cdot \varepsilon$, oder $|x - z_n| < M \cdot \varepsilon^2$ etc. ergeben hat, sehen wir künftig den Konvergenzbeweis als erbracht an und ersparen es uns, den Beweis noch einmal aufzurollen und Änderungen an den Annahmen vorzunehmen, nur um zum Schluß $|x - z_n| < \varepsilon$ schreiben zu dürfen.

10.1.6 A Verneinen Sie die Aussage: $(x_n)_{n \in \mathbb{N}}$ konvergiert gegen x .

10.1.7 A Zeigen Sie: $\lim_{n \rightarrow \infty} \frac{1}{\sqrt[k]{n}} = 0$.

10.1.8 Rechenregeln für Grenzwerte

10.1.9 Satz Es sei $x = \lim_{n \rightarrow \infty} x_n$ und $y = \lim_{n \rightarrow \infty} y_n$. Dann gilt

$$\|x\| = \lim_{n \rightarrow \infty} \|x_n\|, \quad (10.5)$$

$$x \pm y = \lim_{n \rightarrow \infty} x_n \pm y_n, \quad (10.6)$$

$$x \cdot y = \lim_{n \rightarrow \infty} x_n \cdot y_n, \quad (10.7)$$

$$\frac{x}{y} = \lim_{n \rightarrow \infty} \frac{x_n}{y_n} \quad \text{falls } y \neq 0. \quad (10.8)$$

Beweis. Für die kommenden Überlegungen können wir folgendes voraussetzen: Für alle $\varepsilon > 0$ gibt es ein *gemeinsames* $n_\varepsilon \in \mathbb{N}$, so daß $|x - x_n| < \varepsilon$ und $|y - y_n| < \varepsilon$ für alle $n \geq n_\varepsilon$ gilt. Denn andernfalls wählen wir von den beiden Grenzindizes \tilde{n}_ε und \hat{n}_ε für $(x_n)_{n \in \mathbb{N}}$ bzw. $(y_n)_{n \in \mathbb{N}}$ das Maximum.

Der Beweis von (10.6) ist eine Übungsaufgabe.

(10.5) ist eine Folge der umgekehrten Dreiecksungleichung (vgl. Übung 5.5.9)

$$|\|x\| - \|x_n\|| \leq |x - x_n| < \varepsilon$$

für alle $n \geq n_\varepsilon$. Das zeigt die Konvergenz von $(\|x_n\|)_{n \in \mathbb{N}}$ gegen $\|x\|$.

Zu (10.7): Als konvergente Folge ist $(y_n)_{n \in \mathbb{N}}$ beschränkt (Lemma 10.1.3), sagen wir durch eine Zahl $M > 0$. Damit können wir abschätzen:

$$\begin{aligned} |xy - x_n y_n| &= |xy - xy_n + xy_n - x_n y_n| \leq |xy - xy_n| + |xy_n - x_n y_n| \\ &= |x||y - y_n| + |x - x_n||y_n| \leq |x||y - y_n| + |x - x_n| \cdot M \\ &< (|x| + M) \cdot \varepsilon \quad \text{für alle } n \geq n_\varepsilon. \end{aligned}$$

Zu (10.8): Um die Konvergenz der Folge $(\frac{1}{y_n})_{n \in \mathbb{N}}$ zu zeigen, müssen wir $(|y_n|)_{n \in \mathbb{N}}$ nach unten abschätzen, damit die Brüche $\frac{1}{|y_n|}$ beschränkt bleiben. Da $|y| > 0$ ist, gilt

$$|y_n| = |y| - (|y| - |y_n|) \geq |y| - |y - y_n| > |y| - \varepsilon$$

für alle $n \geq n_\varepsilon$. Wählen wir $\varepsilon \leq \frac{|y|}{2}$, dann haben wir $|y_n| > \frac{|y|}{2}$. Jetzt können wir abschätzen:

$$\left| \frac{1}{y} - \frac{1}{y_n} \right| = \left| \frac{y_n - y}{yy_n} \right| = \frac{|y - y_n|}{|y||y_n|} < 2 \frac{|y - y_n|}{|y|^2} < \frac{2}{|y|^2} \cdot \varepsilon$$

für alle $n \geq n_\varepsilon$. Das zeigt $\frac{1}{y} = \lim_{n \rightarrow \infty} \frac{1}{y_n}$. Die eigentliche Behauptung (10.8) folgt nun aus (10.7), da $(\frac{x_n}{y_n})_{n \in \mathbb{N}}$ das Produkt der beiden konvergenten Folgen $(x_n)_{n \in \mathbb{N}}$ und $(\frac{1}{y_n})_{n \in \mathbb{N}}$ ist. \square

10.1.10 A Zeigen Sie (10.6).

10.1.11 A Bestimmen Sie unter der Voraussetzung $b_k \neq 0$

$$\lim_{n \rightarrow \infty} \frac{a_k n^k + a_{k-1} n^{k-1} + \dots + a_1 n + a_0}{b_k n^k + b_{k-1} n^{k-1} + \dots + b_1 n + b_0}.$$

10.1.12 A Zeigen Sie: Für eine konvergente Folge $(x_n)_{n \in \mathbb{N}}$ nicht negativer Zahlen ist der Grenzwert ebenfalls nicht negativ. Ist der Grenzwert auch positiv, wenn alle Folgenglieder positiv sind? Folgern Sie: Aus $a = \lim_{n \rightarrow \infty} a_n$, $b = \lim_{n \rightarrow \infty} b_n$ und $a_n \leq b_n$ für alle $n \in \mathbb{N}$ ergibt sich $a \leq b$.

10.1.13 A Zeigen Sie: Für eine konvergente Folge $(x_n)_{n \in \mathbb{N}}$ nicht negativer Zahlen mit Grenzwert x konvergiert $(\sqrt{x_n})_{n \in \mathbb{N}}$ gegen \sqrt{x} . Betrachten Sie dabei zunächst den Fall $x = 0$ gesondert. Überlegen Sie sich, ob Sie für den Fall $x > 0$ die dritte binomische Formel sinnvoll einsetzen können.

Von den Ergebnissen dieser beiden Übungen werden wir ab jetzt immer wieder Gebrauch machen, ohne sie jedes mal zu erwähnen. Sie nehmen ab jetzt sozusagen den Status erweiterter Grundrechenarten ein.

10.1.14 A Zeigen Sie mit Hilfe vollständiger Induktion die *BERNOULLI-Ungleichung*

$$(1+x)^n \geq 1 + nx \quad \text{für alle } x > -1. \tag{10.9}$$

Zeigen Sie darüber hinaus, daß in dieser Ungleichung für $n \geq 2$ das Gleichheitszeichen genau dann gilt, wenn $x = 0$ ist.

Die Rechenregeln für Grenzwerte sind ein wertvolles Hilfsmittel, um in vielen Fällen bequem den Grenzwert einer Folge zu bestimmen. Wir werden aber immer wieder Folgen begegnen, die einer eingehenderen Untersuchung bedürfen. Immer dann müssen wir uns auf die grundlegende Definition 10.1.1 zurückziehen. Als Beispiel dafür mag der folgende Satz herhalten, für den wir einen kleinen Vorgriff auf die Eigenschaften der n -ten Wurzel $\sqrt[n]{a}$ machen müssen, mit denen wir uns in Satz 10.1.24 ausführlich beschäftigen werden.

10.1.15 Satz

$$\lim_{n \rightarrow \infty} \frac{n^k}{a^n} = 0 \quad \text{für } a \in \mathbb{C}, |a| > 1, k \in \mathbb{N}_0, \quad (10.10)$$

$$\lim_{n \rightarrow \infty} \sqrt[n]{a} = 1 \quad \text{für } a > 0, \quad (10.11)$$

$$\lim_{n \rightarrow \infty} \sqrt[n]{n^k} = 1 \quad \text{für } k \in \mathbb{N}, \quad (10.12)$$

$$\lim_{n \rightarrow \infty} \frac{z^n}{n!} = 0 \quad \text{für } z \in \mathbb{C}. \quad (10.13)$$

Beweis. Wir zeigen (10.10) für $k = 1$. Der allgemeine Fall ist eine Übungsaufgabe. Laut Voraussetzung gilt $|a| = 1 + \delta$ mit einem geeigneten $\delta > 0$. Das führt uns auf die Idee für $|a|^n$ die BERNOULLI-Ungleichung zu versuchen: $|a|^n = (1 + \delta)^n \geq 1 + n\delta$. Damit hätten wir

$$\left| \frac{n}{a^n} \right| = \frac{n}{|a|^n} \leq \frac{n}{1 + n\delta}$$

erreicht – was zu schwach ist, denn $\frac{n}{1 + n\delta}$ konvergiert nicht gegen Null. Wir brauchen im Nenner ein stärkeres Wachstum als im Zähler. Versuchen wir es mit $\sqrt{|a|} > 1$, also mit $\sqrt{|a|} = 1 + \gamma$ und einem geeigneten $\gamma > 0$:

$$\frac{n}{|a|^n} = \frac{n}{(\sqrt{|a|})^n} \leq \frac{n}{(1 + n\gamma)^2}.$$

Nach dem Sandwich-Prinzip 10.1.5 konvergiert die linke Seite gegen Null.

Für (10.11) versuchen wir, zunächst für $a > 1$, über $\sqrt[n]{a} =: 1 + \delta_n$ mit Hilfe der BERNOULLI-Ungleichung etwas über $\delta_n > 0$ zu erfahren:

$$a = (1 + \delta_n)^n \geq 1 + n\delta_n \Rightarrow \frac{a - 1}{n} \geq \delta_n > 0.$$

Damit konvergiert δ_n gegen Null, also $\sqrt[n]{a} = 1 + \delta_n$ gegen 1. Der Fall $0 < a < 1$ ist jetzt eine Anwendung von (10.8).

Für (10.12) und $k = 1$ untersuchen wir $\sqrt[n]{\sqrt{n}} =: 1 + \lambda_n$ und erhalten $\sqrt{n} \geq 1 + n\lambda_n$, also $\lambda_n \leq \frac{\sqrt{n} - 1}{n} \leq \frac{1}{\sqrt{n}}$. Das zeigt, daß $\sqrt[n]{\sqrt{n}}$ gegen 1 konvergiert. Daher konvergiert $\sqrt[n]{n} = \sqrt[n]{\sqrt{n}^2}$ ebenfalls gegen 1. Der Fall $k > 1$ ist jetzt eine Übungsaufgabe.

Zu (10.13): Sei $|z| =: r$ und n_0 die größte ganze Zahl, die kleiner oder gleich r ist. Dann haben wir mit $q := \frac{r}{n_0+1} < 1$:

$$\left| \frac{z^n}{n!} \right| = \frac{r^n}{n!} = \frac{r}{n} \frac{r}{n-1} \cdots \frac{r}{n_0+1} \cdot \frac{r}{n_0} \frac{r}{n_0-1} \cdots \frac{r}{2} r$$

$$= \frac{r}{n} \frac{r}{n-1} \cdots \frac{r}{n_0+1} \cdot \frac{r^{n_0}}{n_0!} \leq q^{n-n_0} \frac{r^{n_0}}{n_0!} = q^n \frac{r^{n_0}}{q^{n_0} n_0!}.$$

Da $(q^n)_{n \in \mathbb{N}}$ für $0 < q < 1$ eine Nullfolge ist (das ergibt sich aus (10.10) für $a = \frac{1}{q}$ und $k = 0$), erhalten wir die Behauptung aus dem Sandwich-Prinzip 10.1.5. \square

10.1.16 A Zeigen Sie die Fälle $k \neq 1$ in Satz 10.1.15.

10.1.17 A Zeigen Sie: $\lim_{n \rightarrow \infty} \sqrt{n^2 + 2} - n = 0$. Was erhalten Sie, wenn Sie 2 durch $a > 0$ ersetzen?

Die Grenzwerte (10.11) und (10.12) werden bei der Untersuchung von Potenzreihen wichtig werden (vgl. S. 252). Allerdings braucht man (10.11) meist in einer etwas flexibleren Form, die wir im folgenden Korollar schon einmal bereitstellen.

10.1.18 Korollar Ist $(a_n)_{n \in \mathbb{N}}$ eine Folge positiver Zahlen, so daß für geeignete Zahlen $M, N > 0$ und für fast alle $n \in \mathbb{N}$ die Abschätzung $M \leq a_n \leq N$ gilt, so folgt

$$\lim_{n \rightarrow \infty} \sqrt[n]{a_n} = 1. \quad (10.14)$$

Das ist insbesondere dann der Fall, wenn $(a_n)_{n \in \mathbb{N}}$ einen Grenzwert $a > 0$ hat.

Beweis. Der Beweis folgt aus der Monotonie der n -ten Wurzel: $\sqrt[n]{M} \leq \sqrt[n]{a_n} \leq \sqrt[n]{N}$. Nach (10.11) konvergieren die Folgen $(\sqrt[n]{M})_{n \in \mathbb{N}}$ und $(\sqrt[n]{N})_{n \in \mathbb{N}}$ beide gegen 1, so daß nach dem Sandwich-Prinzip die Behauptung folgt. Ist $a > 0$ ein Grenzwert von $(a_n)_{n \in \mathbb{N}}$, so ist die Folge beschränkt, d. h. es gibt eine Zahl $N > 0$, wie oben gefordert. Ab einem n_ϵ gilt $a_n \geq a - \epsilon$. Wenn wir $\epsilon := \frac{a}{2}$ wählen, ist $M := \frac{a}{2}$ eine mögliche positive untere Schranke. \square

10.1.19 Beispiel Ein Beispiel soll die Anwendung verdeutlichen. Wir untersuchen die Folge

$$\left(\sqrt[n]{\frac{2n^3 - n + 1}{n + \frac{5}{n}}} \right)_{n \in \mathbb{N}}$$

Ein Blick genügt, um zu sehen, daß n^3 der dominierende Term unter der Wurzel ist. Der Bruch wächst wie n^2 . Wenn wir also n^2 ausklammern, verbleibt eine konvergente Folge unter der Wurzel.

$$\sqrt[n]{\frac{2n^3 - n + 1}{n + \frac{5}{n}}} = \sqrt[n]{n^2 \frac{2n - \frac{1}{n} + \frac{1}{n^2}}{n + \frac{5}{n}}} = \sqrt[n]{n^2} \cdot \sqrt[n]{\frac{2 - \frac{1}{n^2} + \frac{1}{n^3}}{1 + \frac{5}{n^2}}} \xrightarrow{n \rightarrow \infty} 1,$$

denn jetzt haben wir ein Produkt aus zwei Folgen, die jeweils gegen 1 konvergieren, die erste nach (10.12) und die zweite als Ergebnis des Korollars 10.1.18.

10.1.20 Definition Eine Teilmenge A des normierten Raumes X heißt beschränkt, falls es eine Zahl $M > 0$ gibt, so daß $\|x\| \leq M$ für alle $x \in A$ gilt. Eine Teilmenge A von \mathbb{R} heißt nach oben (unten) beschränkt, falls es eine Zahl $s \in \mathbb{R}$ gibt, mit der Eigenschaft $x \leq s$ ($x \geq s$) für alle $x \in A$. s heißt dann obere (untere) Schranke von A . Eine kleinste obere (größte untere) Schranke von A heißt Supremum (Infimum) von A . Sie wird durch $\sup(A)$ bzw. $\inf(A)$ bezeichnet. Ein Supremum (Infimum), das zur Menge A gehört, heißt Maximum (Minimum) von A : $\max(A)$ bzw. $\min(A)$. Wir sagen, eine Folge $(x_n)_{n \in \mathbb{N}}$ sei beschränkt bzw. nach oben (unten) beschränkt, wenn die Menge ihrer Folgenglieder $\{x_n \mid n \in \mathbb{N}\}$ diese Eigenschaft hat. Mit $\sup_{n \in \mathbb{N}} x_n$, manchmal auch nur $\sup_n x_n$, oder $\sup x_n$, mit $\max x_n, \dots$ ($\inf x_n, \min x_n, \dots$) bezeichnen wir die entsprechenden Größen für die Menge der Folgenglieder von $(x_n)_{n \in \mathbb{N}}$.

Eine Folge $(x_n)_{n \in \mathbb{N}}$ heißt monoton wachsend (fallend), falls $x_n \leq x_{n+1}$ ($x_n \geq x_{n+1}$), und sie heißt streng monoton wachsend (fallend), falls $x_n < x_{n+1}$ ($x_n > x_{n+1}$) für alle $n \in \mathbb{N}$ gilt. Wir sagen, sie sei fast überall (f.ü.) monoton wachsend, fallend ..., wenn sie die entsprechende Eigenschaft erst ab einem Index $m \in \mathbb{N}$ hat.

10.1.21 Lemma Falls eine Menge ein Supremum hat, dann ist es eindeutig.

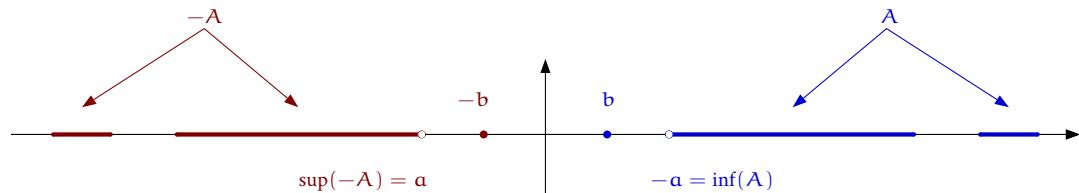
Beweis. a sei das Supremum einer Menge $A \subset \mathbb{R}$, d. h. die kleinste obere Schranke von A . Ein weiteres Element b mit dieser Eigenschaft wäre insbesondere eine obere Schranke, so daß $a \leq b$ gelten müßte. Genauso können wir aber auch $b \leq a$ folgern. Damit bleibt nur $a = b$. \square

Die Menge $(0, 1)$ ist beschränkt, etwa mit der Schranke $M = 1.5$. Natürlich ist das nicht die beste Schranke von A , aber zum Nachweis der Beschränktheit einer Menge verlangt die Definition auch gar nicht nach einer besonders guten Schranke. Für Mengen, die nicht so einfach konstruiert sind wie $(0, 1)$, kann es ziemlich schwer sein, eine Schranke nachzuweisen, so daß man schon zufrieden ist, wenn man irgendeine findet. Ein Beispiel dafür ist $\{(1 + \frac{1}{n})^n \mid n \in \mathbb{N}\}$ mit der Schranke 4 (siehe Übung 10.1.31). Zurück zu $(0, 1)$. Es gilt $\inf(0, 1) = 0$ und $\sup(0, 1) = 1$. Ein Maximum oder Minimum hat diese Menge nicht. Für $[0, 1]$ dagegen ist $\min[0, 1] = 0$ und $\max[0, 1] = 1$. Diese Beispiele sind sehr einfach. Deshalb ist zunächst nicht auszuschließen, daß es kompliziertere Mengen in \mathbb{R} geben könnte, die zwar beschränkt sind, aber kein Supremum oder Infimum haben. Da wir uns \mathbb{R} über geometrische Eigenschaften der Linie als gegeben denken, können wir das nicht entscheiden. Die Menge \mathbb{Q} der rationalen Zahlen können wir uns mit demselben Recht über geometrische Eigenschaften der Linie eingeführt vorstellen. In \mathbb{Q} gibt es aber Mengen, die kein Supremum oder kein Infimum haben. Ein Beispiel ist etwa $W := \{x \in \mathbb{Q} \mid x^2 \leq 2\}$. Eine obere Schranke ist sicher die Zahl 2, denn für jedes $x \geq 2$ gilt $x^2 \geq 4$, also $x \notin W$. Die kleinste obere Schranke a müßte aber die Gleichung $a^2 = 2$ erfüllen (Übung 10.1.32), für die in \mathbb{Q} keine Lösung zu finden ist.

10.1.1 Axiom (Vollständigkeit von \mathbb{R}) Jede nach oben beschränkte Teilmenge von \mathbb{R} hat ein Supremum in \mathbb{R} .

Ist eine Menge $A \subset \mathbb{R}$ nach unten beschränkt, mit einer unteren Schranke b , dann ist $-A := \{x \in \mathbb{R} \mid -x \in A\}$ durch $-b$ nach oben beschränkt und hat daher ein Supremum

$a := \sup(-A)$. $-a$ ist dann eine untere Schranke von A – sie ist sogar die größte untere Schranke von A . Für jede andere untere Schranke b von A ist nämlich $-b$ eine obere Schranke von $-A$. Daher gilt $a \leq -b$ und damit $-a \geq b$. Das bedeutet $-a = \inf(A)$. Aus dem Vollständigkeitsaxiom folgt demnach auch die Existenz des Infimums für Mengen, die nach unten beschränkt sind.



Mit dem folgenden Lemma verschaffen wir uns ein Werkzeug, das bei künftigen Untersuchungen von Suprema oder Infima immer wieder von Nutzen sein wird.

10.1.22 Lemma *Ist a das Supremum einer Menge $A \subset \mathbb{R}$, so gibt es für alle $\varepsilon > 0$ ein Element x von A mit der Eigenschaft $a - \varepsilon \leq x \leq a$. Entsprechend lässt sich für eine Menge $B \subset \mathbb{R}$ mit Infimum b immer ein Element $y \in B$ mit $b \leq y \leq b + \varepsilon$ finden.*

Beweis. Gäbe es ein $\varepsilon > 0$, so daß in dem Intervall $[a - \varepsilon, a]$ kein Element aus A anzutreffen ist, dann müßten alle Elemente von A unterhalb von $a - \varepsilon$ liegen. Daher wäre $a - \varepsilon$ eine kleinere obere Schranke als a , im Widerspruch dazu, daß a die kleinste obere Schranke ist. Also läßt sich immer ein $x \in A$ mit der geforderten Eigenschaft finden. Die Behauptung für das Infimum wird genauso bewiesen. \square

Das Lemma wenden wir gleich auf das sogenannte Archimedische Axiom an und zeigen, daß es bei unserem Zugang zu den reellen Zahlen gar kein Axiom ist, sondern eine Folgerung aus dessen Vollständigkeitsaxiom.

10.1.23 Satz (Archimedisches Axiom) *Für je zwei Zahlen $0 < x < y$ gibt es immer eine natürliche Zahl n mit der Eigenschaft $nx > y$. Insbesondere lässt sich jede positive reelle Zahl durch eine natürliche übertreffen.*

Beweis. Wir nehmen an, es gäbe zwei Zahlen x, y der geforderten Art, so daß $nx \leq y$ für alle $n \in \mathbb{N}$ gilt. Dann wäre $\frac{y}{x}$ eine obere Schranke von \mathbb{N} . Nach dem Vollständigkeitsaxiom von \mathbb{R} hat \mathbb{N} ein Supremum a und nach Lemma 10.1.22 ist für jedes $1 > \varepsilon > 0$ ein Element $n \in \mathbb{N}$ mit $a - \varepsilon \leq n \leq a$ zu finden. Da a das Supremum von \mathbb{N} ist, muß auch $a - \varepsilon \leq n < n + 1 \leq a$ gelten. Das würde bedeuten, daß n und $n + 1$ einen Abstand haben, der kleiner als 1 ist. Dieser offensichtliche Widerspruch zeigt, daß unser Annahme falsch sein muß.

Sei nun $y > 0$. Ist $y \leq 1$, so wird diese Zahl durch $2 \in \mathbb{N}$ übertroffen. Ist $y > 1$, so folgt aus der ersten Behauptung für $x = 1$ die Existenz einer natürlichen Zahl n mit der Eigenschaft $n = n \cdot 1 > y$. \square

Als eine weitere Anwendung zeigen wir die Existenz der n -ten Wurzel $\sqrt[n]{a}$ einer Zahl a , also die Existenz einer Zahl x mit der Eigenschaft $x^n = a$. Genauer:

10.1.24 Satz Für jede Zahl $a \geq 0$ und jedes $n \in \mathbb{N}$ gibt es jeweils genau eine Zahl $x \geq 0$ mit der Eigenschaft $x^n = a$. Wir nennen x die n -te Wurzel von a und bezeichnen sie mit dem Symbol $\sqrt[n]{a}$. Für ungerades n gibt es eine n -te Wurzel auch für $a < 0$, nämlich $\sqrt[n]{a} := -\sqrt[n]{|a|}$.
 $(a_k)_{k \in \mathbb{N}}$ sei eine konvergente Folge mit Grenzwert a , die bei geradem n aus nicht negativen Zahlen besteht. Dann konvergiert die Folge $(\sqrt[n]{a_k})_{k \in \mathbb{N}}$ gegen $\sqrt[n]{a}$.

Beweis. Natürlich ist die Aussage des Satzes nur für $n \geq 2$ interessant. Wir beginnen mit dem Fall $a = 0$ und der Lösung $x = 0$ für die Gleichung $x^n = 0$. Eine weitere kann es nicht geben, denn da \mathbb{R} nullteilerfrei ist, lässt sich kein $x \neq 0$ mit der Eigenschaft $x^n = 0$ finden.

Ab jetzt ist $a > 0$. Wir untersuchen die Menge $W := \{y \in \mathbb{R} \mid y \geq 0 \wedge y^n \leq a\}$. Sie ist nach oben beschränkt, etwa durch $1 + a$. Für $y \geq 1 + a$ gilt nach der BERNOULLI-Ungleichung 10.9 $y^n \geq (1 + a)^n \geq 1 + na > a$, d. h. $y \notin W$. Daher müssen alle $y \in W$ unterhalb von $a + 1$ liegen. Nach dem Vollständigkeitsaxiom von \mathbb{R} existiert $x := \sup W$. Wenn wir $x^n \leq a$ zeigen können, ist $x \in W$ und somit das Maximum von W . Laut Lemma 10.1.22 gibt es für jedes $k \in \mathbb{N}$ ein $y_k \in W$ mit der Eigenschaft $x - \frac{1}{k} \leq y_k \leq x$. Das bedeutet $(x - \frac{1}{k})^n \leq y_k^n \leq a$. Nach (10.7) und Übung 10.1.12 folgt $x^n = \lim_{k \rightarrow \infty} (x - \frac{1}{k})^n \leq a$, mithin $x \in W$, oder $x = \max W$.

Wäre $x^n < a$, dann könnten wir ein $k \in \mathbb{N}$ mit der Eigenschaft $x^n < (x + \frac{1}{k})^n < a$ finden (denn $(x + \frac{1}{k})^n \xrightarrow{k \rightarrow \infty} x^n$). Das hieße aber, daß die Zahl $x + \frac{1}{k} > x$ noch zu W gehört, im Widerspruch zu $x = \max W$. Das zeigt $x^n = a$, also die Existenz der n -ten Wurzel. Für eine weitere n -te Wurzel y von a hätten wir $0 = x^n - y^n = (x - y)(x^{n-1} + x^{n-2}y + x^{n-3}y^2 + \dots + x^{n-k}y^{k-1} + \dots + xy^{n-2} + y^{n-1})$. Die zweite Klammer enthält nur positive Einträge, so daß die erste verschwinden muß. Damit haben wir $x = y$, also die Eindeutigkeit der n -ten Wurzel gezeigt.

Ist n ungerade und $a < 0$, dann löst $x = -\sqrt[n]{|a|}$ die Gleichung $x^n = a$, denn $x^n = (-1)^n(\sqrt[n]{|a|})^n = -|a| = a$. Gäbe es eine weitere Lösung y , so müßte diese ebenfalls negativ sein. Dann wären aber $|x|$ und $|y|$ zwei verschiedene n -te Wurzeln von $|a|$, im Widerspruch zur eben bewiesenen Eindeutigkeit.

Zur Monotonie der Wurzel: Würde für Zahlen $0 \leq x < y$ die Ungleichung $\sqrt[n]{x} \geq \sqrt[n]{y}$ gelten, so hätten wir mit

$$0 < y - x = \sqrt[n]{y}^n - \sqrt[n]{x}^n = (\sqrt[n]{y} - \sqrt[n]{x}) \sum_{\ell=1}^n \sqrt[n]{y}^{n-\ell} \sqrt[n]{x}^{\ell-1} \leq 0$$

sofort einen Widerspruch. Für ungerade n sind noch die Fälle $x < 0 < y$ und $x < y < 0$ zu berücksichtigen. Der erste ist klar, da $\sqrt[n]{x} < 0$ und $\sqrt[n]{y} > 0$ gilt. Den zweiten führen wir auf positive Zahlen zurück: $0 < |y| < |x|$. Dafür haben wir die Monotonie bereits: $\sqrt[n]{|y|} < \sqrt[n]{|x|}$. Das ergibt $\sqrt[n]{x} = -\sqrt[n]{|x|} < -\sqrt[n]{|y|} = \sqrt[n]{y}$.

$(a_k)_{k \in \mathbb{N}}$ sei zunächst eine Nullfolge. Würde $\sqrt[n]{a_k}$ nicht gegen Null konvergieren, dann gäbe es ein $\varepsilon > 0$, so daß für unendlich viele $k \in \mathbb{N}$ die Ungleichung $|\sqrt[n]{a_k}| = \sqrt[n]{|a_k|} \geq \varepsilon$ und damit $|a_k| \geq \varepsilon^n$ gelten müßte, im Widerspruch dazu, daß $(a_k)_{k \in \mathbb{N}}$ eine Nullfolge ist.

Den Fall, daß $(a_k)_{k \in \mathbb{N}}$ gegen eine Zahl $a \neq 0$ konvergiert, behandeln wir für die spezielle Situation, in der n ungerade und $a < 0$ ist. $a > 0$ bzw. gerades n lassen sich leicht an diese Situation anpassen und sind einfacher zu behandeln. Da $a < 0$ ist, müssen fast alle a_k ebenfalls negativ sein. Dann gilt $\sqrt[n]{a}^{n-\ell} \sqrt[n]{a_k}^{\ell-1} = (-1)^{n-\ell} (-1)^{\ell-1} \sqrt[n]{|a|}^{n-\ell} \sqrt[n]{|a_k|}^{\ell-1} = (-1)^{n-1} \sqrt[n]{|a|}^{n-\ell} \sqrt[n]{|a_k|}^{\ell-1} > 0$ für fast alle $k \in \mathbb{N}$, denn $n-1$ ist gerade und daher $(-1)^{n-1} = 1$. Für fast alle k muß auch $|a_k| \geq \frac{1}{2}|a|$ gelten (warum?), so daß wir für diese k folgendermaßen abschätzen können:

$$\sum_{\ell=1}^n \sqrt[n]{a}^{n-\ell} \sqrt[n]{a_k}^{\ell-1} \geq \sum_{\ell=1}^n \sqrt[n]{\frac{1}{2}|a|}^{n-\ell} \sqrt[n]{\frac{1}{2}|a|}^{\ell-1} = n \cdot \sqrt[n]{\frac{1}{2}|a|}^{n-1} =: M > 0.$$

Oberhalb eines geeigneten k_0 haben wir daher

$$|a - a_k| = \left| \sqrt[n]{a} - \sqrt[n]{a_k} \right| \left| \sum_{\ell=1}^n \sqrt[n]{a}^{n-\ell} \sqrt[n]{a_k}^{\ell-1} \right| \geq M \cdot \left| \sqrt[n]{a} - \sqrt[n]{a_k} \right|,$$

oder $\left| \sqrt[n]{a} - \sqrt[n]{a_k} \right| \leq \frac{1}{M} \cdot |a - a_k|$. Daraus ergibt sich $\sqrt[n]{a_k} \xrightarrow{k \rightarrow \infty} \sqrt[n]{a}$ leicht mit Hilfe des Sandwich-Prinzips. \square

10.1.25 A Zeigen Sie die Verallgemeinerung $(x-y) \sum_{k=1}^n x^{n-k} y^{k-1} = x^n - y^n$ der dritten binomischen Formel $(x-y)(x+y) = x^2 - y^2$. Verwenden Sie das Ergebnis, um die Wurzel im Nenner von $\frac{1}{2+3\sqrt{5}}$ zu beseitigen.

10.1.26 A Zeigen Sie die umgekehrte BERNOULLI-Ungleichung

$$\sqrt[n]{1+x} \leq 1 + \frac{x}{n}, \quad x \geq -1. \quad (10.15)$$

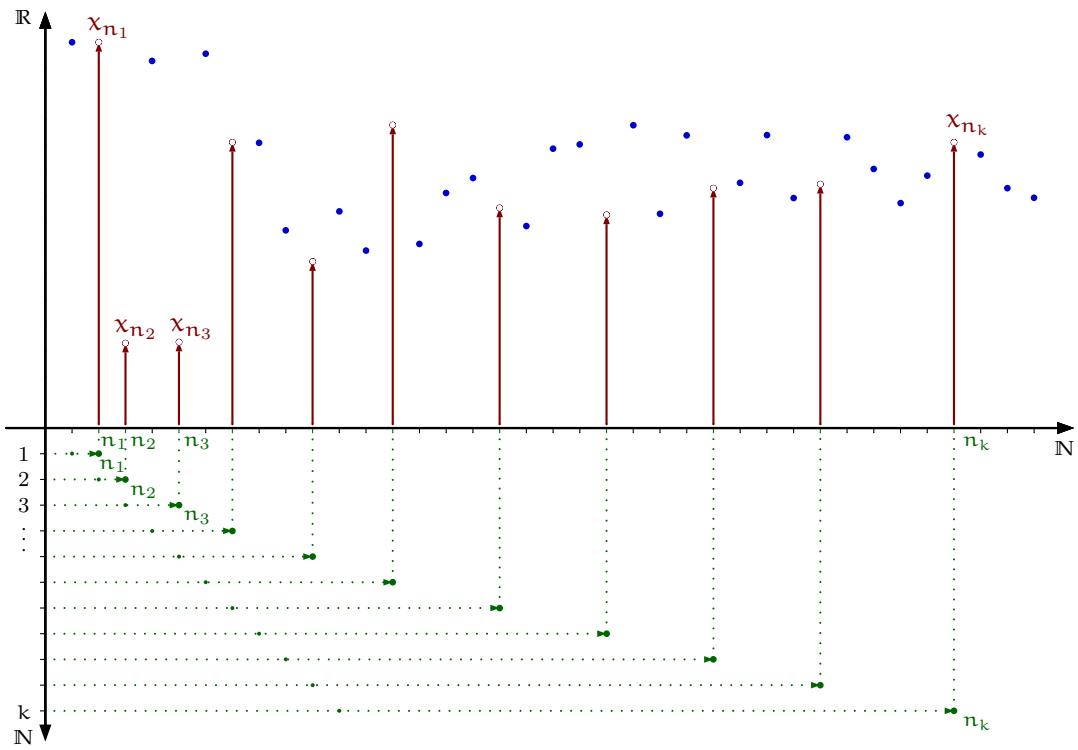


Abb. 10.4 Eine Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$ von $(x_n)_{n \in \mathbb{N}}$ und ihre Indexfolge $(n_k)_{k \in \mathbb{N}}$

10.1.27 Definition Sei $(x_n)_{n \in \mathbb{N}}$ eine Folge in X und $(n_k)_{k \in \mathbb{N}}$ eine streng monoton wachsende Folge natürlicher Zahlen. Dann heißt die Folge $(x_{n_k})_{k \in \mathbb{N}}$ Teilfolge von $(x_n)_{n \in \mathbb{N}}$. $(n_k)_{k \in \mathbb{N}}$ bezeichnen wir als ihre Indexfolge.

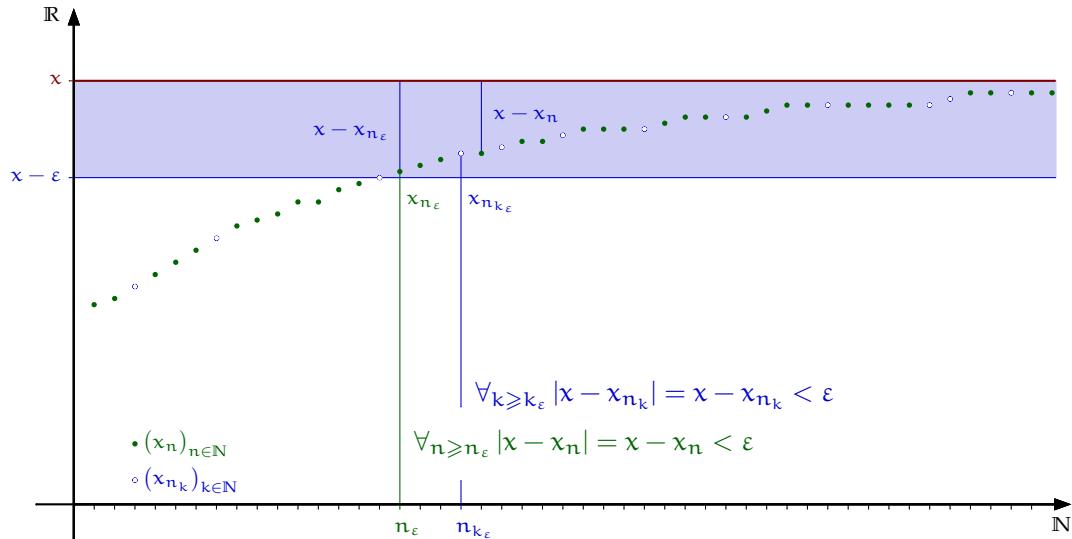
Eine Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$ entsteht aus einer Folge $(x_n)_{n \in \mathbb{N}}$ durch Auswahl unendlich vieler Folgenglieder x_{n_k} . Es ist die Aufgabe der Indexfolge, die Positionen n_k der ausgewählten Elemente zu markieren. In Abbildung 10.4 ist $(n_k)_{k \in \mathbb{N}} = (2, 3, 5, 7, 10, 13, 17, 21, 25, 29, 34, \dots)$ und die zugehörige Teilfolge $(x_{n_k})_{k \in \mathbb{N}} = (x_2, x_3, x_5, x_7, x_{10}, x_{13}, x_{17}, x_{21}, x_{25}, x_{29}, x_{34}, \dots)$. Die strenge Monotonie der Indexfolge gewährleistet, daß kein Folgenglied mehrfach ausgewählt wird. Wenn wir $(x_n)_{n \in \mathbb{N}}$ als eine Funktion $x: \mathbb{N} \rightarrow \mathbb{R}$, $x(n) := x_n$ ansehen, dann ist die Teilfolge eine Verkettung der Funktion $n: \mathbb{N} \rightarrow \mathbb{N}$, $n(k) := n_k$, die die Indexfolge beschreibt, mit der Funktion $x: (x \circ n)(k) = x(n(k)) = x(n_k) = x_{n_k}$.

Eine mitunter nützliche Beobachtung ist $n_k \geq k$, die sich aus der strengen Monotonie der Indexfolge ergibt (vergl. Abb. 10.4).

10.1.28 Satz (Monotone Konvergenz) Eine nach oben (unten) beschränkte monoton wachsende (fallende) Folge $(x_n)_{n \in \mathbb{N}}$ hat einen Grenzwert. Dieser ist durch $\sup x_n$ ($\inf x_n$) gegeben. Hat eine monoton wachsende (fallende) Folge eine konvergente Teilfolge, so ist die Ausgangsfolge ebenfalls konvergent, mit demselben Grenzwert.

Beweis. Sei $(x_n)_{n \in \mathbb{N}}$ eine noch oben beschränkte monoton wachsende Folge. Das Vollständigkeitsaxiom für \mathbb{R} sichert die Existenz von $x := \sup_n x_n$. Dann gibt es zu jedem $\varepsilon > 0$ wenigstens ein Folgenglied x_{n_ε} , so daß $x - \varepsilon < x_{n_\varepsilon} \leq x$, denn andernfalls wäre $x - \varepsilon$ eine kleinere obere Schranke von $(x_n)_{n \in \mathbb{N}}$ als x (vergl. Lemma 10.1.22). Da $(x_n)_{n \in \mathbb{N}}$ monoton wächst, gilt auch $x - \varepsilon < x_{n_\varepsilon} \leq x_n \leq x$, d. h. $|x - x_n| = x - x_n < \varepsilon$ für alle $n \geq n_\varepsilon$. Also ist $x = \lim_{n \rightarrow \infty} x_n$.

Der Beweis für monoton fallende Folgen verläuft analog.



Nun sei x der Grenzwert einer konvergenten Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$ der monoton wachsenden Folge $(x_n)_{n \in \mathbb{N}}$. Da die Teilfolge selbst monoton wächst, ist $x = \sup_k x_{n_k}$. Wir zeigen, daß auch $x = \sup_n x_n$ gilt.

Gäbe es ein Folgenglied $x_n > x$, so würden für alle $k \in \mathbb{N}$ mit $n_k \geq n$ die Abschätzung $x_{n_k} \geq x_n > x$ gelten, im Widerspruch dazu, daß x die kleinste obere Schranke der x_{n_k} ist. Also darf kein x_n oberhalb von x liegen, so daß x auch eine obere Schranke für $(x_n)_{n \in \mathbb{N}}$ ist. Das bedeutet $\sup_n x_n \leq x$. Andererseits gilt $x = \sup_k x_{n_k} \leq \sup_n x_n$, denn das Supremum $\sup_n x_n$ erstreckt sich über mehr Elemente als $\sup_k x_{n_k}$. Daher folgt $x = \sup_n x_n$. Der erste Teil des Satzes zeigt nun die Behauptung $x = \lim_{n \rightarrow \infty} x_n$. \square

Als eine Anwendung stellen wir das **HERON-Verfahren** zur approximativen Berechnung von Quadratwurzeln \sqrt{a} vor. Dabei handelt es sich um die folgende nichtlineare Rekursion:

$$x_{n+1} = \frac{1}{2} \left(x_n + \frac{a}{x_n} \right), \quad x_1 > 0, \quad a > 0. \quad (10.16)$$

Wenn wir für den Moment die Konvergenz der Folge $(x_n)_{n \in \mathbb{N}}$ als gegeben annehmen, können wir uns mit Hilfe der Rechenregeln für Grenzwerte 10.1.9 schnell klar machen, daß als Grenzwert x nur \sqrt{a} in Frage kommt. Die Gleichung (10.16) geht beim Grenzübergang $n \rightarrow \infty$ nämlich in $x = \frac{1}{2} (x + \frac{a}{x})$ über, vorausgesetzt x ist nicht Null (dazu gleich), deren einzige positive Lösung $x = \sqrt{a}$ ist.

Die Rekursionsgleichung produziert bei der positiven Anfangsbedingung $x_1 > 0$ nur positive Folgenglieder x_n . Außerdem

$$x_{n+1} - \sqrt{a} = \frac{1}{2} \left(x_n - 2\sqrt{a} + \frac{a}{x_n} \right) = \frac{1}{2} \left(\sqrt{x_n} - \frac{\sqrt{a}}{\sqrt{x_n}} \right)^2 \geq 0,$$

woraus $x_{n+1} \geq \sqrt{a}$ für alle $n \in \mathbb{N}$, also $x_n \geq \sqrt{a}$ für alle $n \geq 2$ folgt und insbesondere, daß ein möglicher Grenzwert von Null verschieden sein muß. Damit gilt

$$x_{n+1} - x_n = \frac{1}{2} \left(\frac{a}{x_n} - x_n \right) \leq \frac{1}{2} \left(\frac{a}{\sqrt{a}} - x_n \right) = \frac{1}{2} \left(\sqrt{a} - x_n \right) \leq 0.$$

Das bedeutet $x_{n+1} \leq x_n$ für alle $n \geq 2$, d. h. die Folge $(x_n)_{n \in \mathbb{N}}$ ist ab $n = 2$ monoton fallend und nach unten durch \sqrt{a} beschränkt. Nach dem Satz von der monotonen Konvergenz 10.1.28 existiert der Grenzwert des HERON-Verfahrens. Wir wenden es auf $a := 2$ und $x_1 := 1$ an und erhalten

$$x_2 = \frac{3}{2} = 1.5, \quad x_3 = \frac{17}{12} \approx 1.41421568627451, \quad x_4 = \frac{577}{408} \approx 1.41421356237469, \quad \dots$$

Verglichen mit $\sqrt{2} = 1.414213562373095\dots$ ist das schon nach drei Schritten eine gute Näherung (vergl. 11.7.1).

10.1.29 Lemma Eine Teilemenge I von \mathbb{R} ist genau dann ein Intervall, wenn folgendes gilt:
 $\forall a, b \in I, a < b \quad [a, b] \subseteq I$

Beweis. Es ist klar, daß ein Intervall diese Eigenschaft hat: Für je zwei Elemente $a < b$ aus I ist das ganze Intervall $[a, b]$ in I enthalten. Diese Eigenschaft charakterisiert Intervalle. Der Beweis ist eine Anwendung des Vollständigkeitsaxioms und etwas mühsam, weil für jedes der möglichen Intervalle $[a, b]$, (a, b) , $[a, b)$, ..., (a, ∞) , $[a, \infty)$ und $(-\infty, \infty) = \mathbb{R}$ eine eigene Argumentation notwendig ist. Da die Überlegungen aber einem einfachen Schema folgen, führen wir nur zwei typische Fälle vor, stellvertretend für alle anderen. Wir nehmen an, I sei nach unten beschränkt und nach oben unbeschränkt. Daher gibt es $a := \inf I$. Falls a zu I gehört, handelt es sich um den kleinsten Wert von I . Für jedes $n \in \mathbb{N}$ gibt es ein $x \in I$, das größer als $a + n$ ist, denn I ist nach oben unbeschränkt. Daher haben wir $[a, x] \subseteq I$, also insbesondere $[a, a + n] \subseteq I$ für alle $n \in \mathbb{N}$. Das bedeutet $\bigcup_{n \in \mathbb{N}} [a, a + n] = [a, \infty) \subseteq I$. Andererseits kann kein Element von I unterhalb von a liegen. Das bedeutet $[a, \infty)^c \subseteq I^c$, also $I \subseteq [a, \infty)$. Das zeigt $I = [a, \infty)$.

Falls a nicht zu I gehört, gibt es zu jedem $\varepsilon > 0$ ein $x_\varepsilon \in I$, mit $a < x_\varepsilon \leq a + \varepsilon$. Wie oben folgt $[x_\varepsilon, \infty) \subseteq I$, also $\bigcup_{\varepsilon > 0} [x_\varepsilon, \infty) = (a, \infty) \subseteq I$ und $I = (a, \infty)$.

Alle weiteren Fälle ergeben sich durch Fallunterscheidungen über die Beschränktheit nach oben / unten und die Existenz des Maximums / Minimums von I . \square

10.1.30 A Untersuchen Sie die Folgen, die durch die unten aufgeführten Folgenglieder definiert werden, auf Konvergenz und geben Sie gegebenenfalls ihre Grenzwerte an.

$$a_n := \frac{1}{\sqrt[n]{n}}, \quad b_n := (-1)^n, \quad c_n := e^{in\varphi}, \quad \varphi \in (0, 2\pi),$$

$$\begin{aligned} d_n &:= \sqrt[n]{n^3 + 2n + 1}, & e_n &:= \frac{n^4 + 3n^2 - 2^n}{2^n + 3n^4 + 4}, & f_n &:= \frac{3\sqrt[4]{2n^3} - \sqrt[3]{3n^2} + 2\sqrt{10n}}{4\sqrt[4]{8n^3} + 3\sqrt[3]{4n} + 2^{3/2}}, \\ g_n &:= \sqrt[n]{\frac{n^3 + 2n + 1}{n^2 + n + 1}}, & h_n &:= \sqrt{n^2 + n} - n, & i_n &:= \sqrt{n+1} - \sqrt{n-1}. \end{aligned}$$

Hinweise: Für $(a_n)_{n \in \mathbb{N}}$ können Sie zu jedem $\varepsilon > 0$ ein $n_\varepsilon \in \mathbb{N}$ direkt angeben. Gehen Sie dabei genauso vor, wie bei der Folge $(\frac{1}{n})_{n \in \mathbb{N}}$. Verwenden Sie die Monotonie der k-ten Wurzel.

10.1.31 A (Kürzester Beweis zur Existenz von e)

Zeigen Sie, daß die Folge $\left(\left(1 + \frac{1}{n}\right)^{n+1} \right)_{n \in \mathbb{N}}$ streng monoton fallend ist. Rechnen Sie dafür die Beziehung

$$\frac{\left(1 + \frac{1}{n-1}\right)^n}{\left(1 + \frac{1}{n}\right)^{n+1}} = \frac{n-1}{n} \left(1 + \frac{1}{n^2-1}\right)^{n+1}$$

nach und setzen Sie dann die BERNOULLI-Ungleichung ein. Folgern Sie unter Verwendung von $\left(1 + \frac{1}{n}\right)^n = \frac{n}{n+1} \left(1 + \frac{1}{n}\right)^{n+1}$ mit Hilfe der Rechenregeln konvergenter Folgen 10.1.9 die Existenz von $\lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$.

10.1.32 A (*) Zeigen Sie, daß die Menge $W := \{x \in \mathbb{Q} \mid x^2 \leq 2\}$ in \mathbb{Q} kein Supremum hat. Zeigen Sie dafür, daß ein Supremum a die Gleichung $a^2 = 2$ erfüllen müßte.

10.1.33 Cauchy-Folgen

10.1.34 Definition Die Teilmenge $U_\varepsilon(x) := \{ y \in X \mid \|x - y\| < \varepsilon \}$ des normierten Raumes X heißt ε -Umgebung von x .

Für $X = \mathbb{R}$ sind die ε -Umgebungen eines Punktes x die Intervalle $(x - \varepsilon, x + \varepsilon)$, für $X = \mathbb{R}^2$, oder $X = \mathbb{C}$ handelt es sich um Kreisscheiben mit Radius ε und x als Mittelpunkt und für $X = \mathbb{R}^3$ ergeben sich Kugeln mit Radius ε und x als Zentrum.

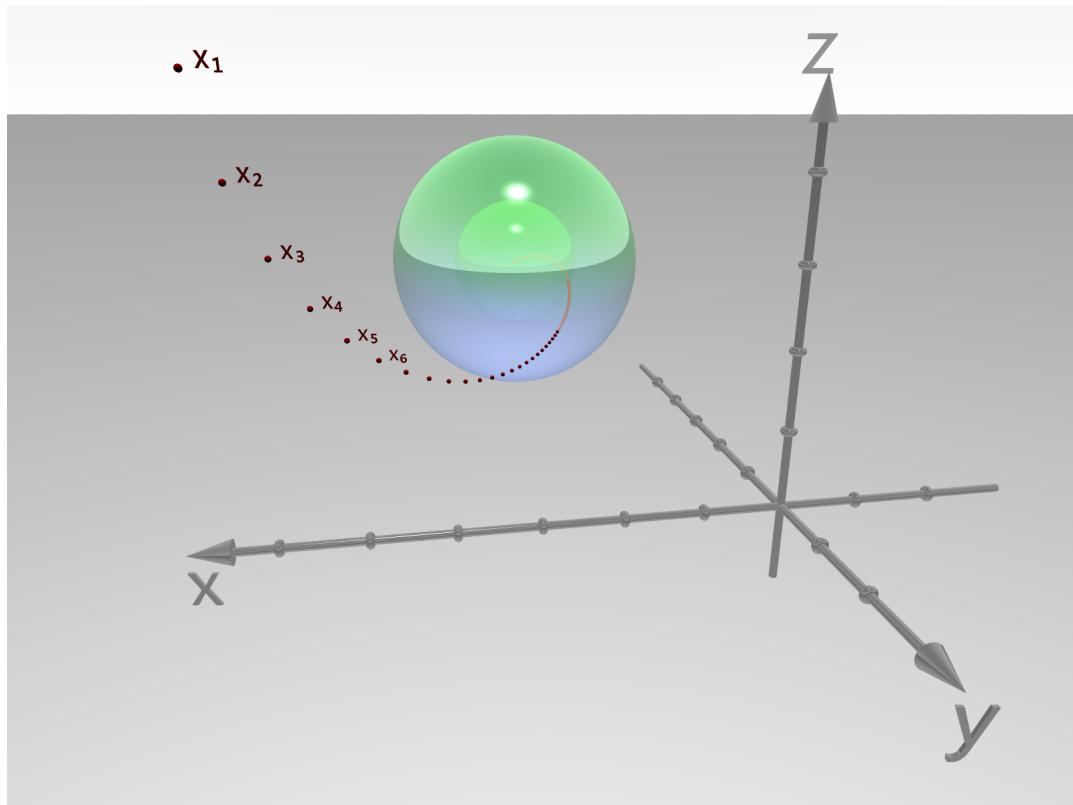


Abbildung 10.5 Zwei ε -Umgebungen im \mathbb{R}^3

Mit dieser Definition lässt sich der Konvergenzbegriff suggestiv fassen:

$$x = \lim_{n \rightarrow \infty} x_n \iff \text{Für jedes } \varepsilon > 0 \text{ befinden sich fast alle } x_n \text{ in } U_\varepsilon(x).$$

Fast alle meint dabei, daß alle Folgenglieder, bis auf endlich viele Ausnahmen, in $U_\varepsilon(x)$ zu finden sind. Es muß also ein letztes Element $x_{n_\varepsilon - 1}$ geben, das nicht in $U_\varepsilon(x)$ liegt, d. h. ab n_ε liegen alle in $U_\varepsilon(x)$, erfüllen also $\|x - x_n\| < \varepsilon$. Das ist die Definition 10.1.1 von $x = \lim_{n \rightarrow \infty} x_n$.

Es gibt eine Abschwächung des Grenzwertbegriffs, der für die weiteren Untersuchungen wichtig werden wird:

10.1.35 Definition x heißt Häufungspunkt einer Folge $(x_n)_{n \in \mathbb{N}}$, wenn in jeder ε -Umgebung $U_\varepsilon(x)$ unendlich viele Folgenglieder zu finden sind.

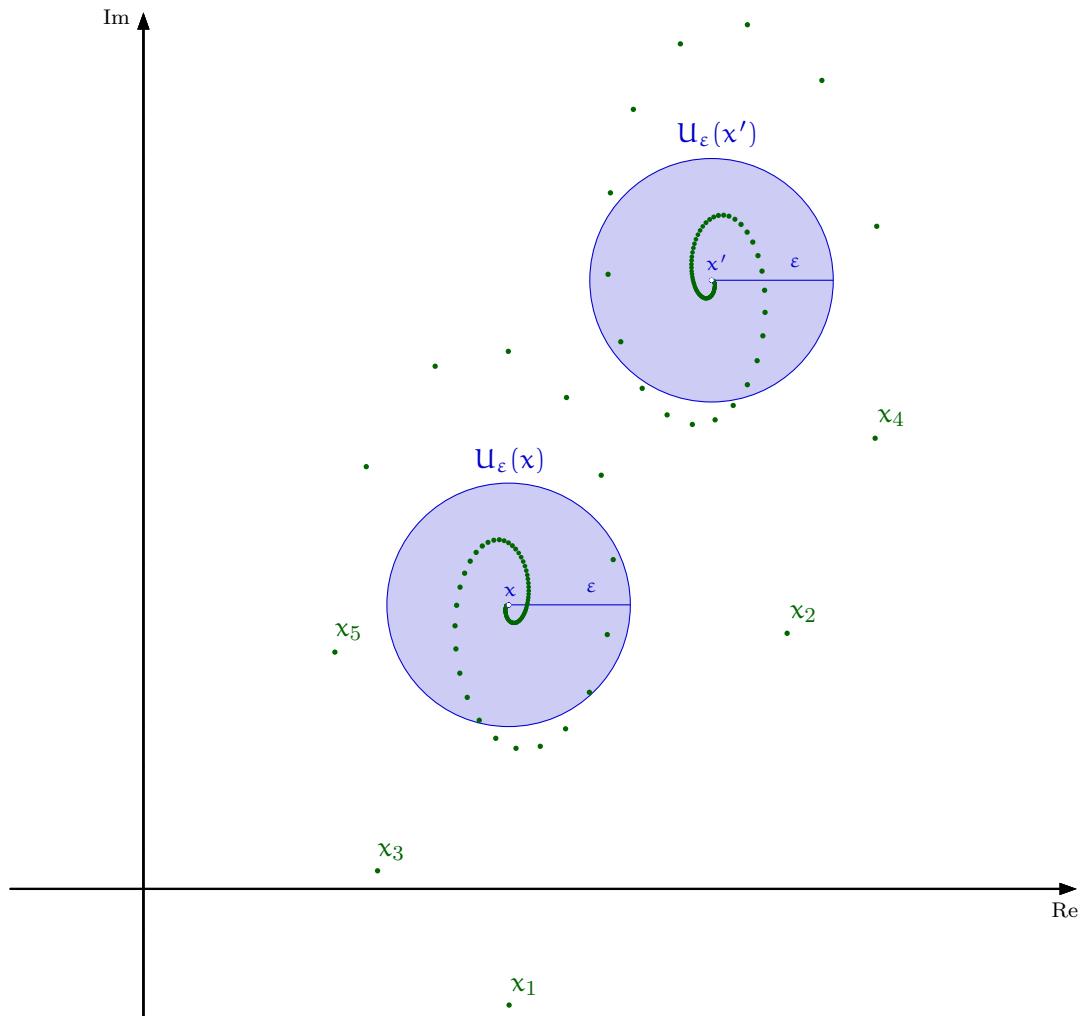


Abbildung 10.6 Eine Folge in \mathbb{C} mit zwei Häufungspunkten

Der Grenzwert x einer Folge $(x_n)_{n \in \mathbb{N}}$ ist insbesondere auch ein Häufungspunkt. Aber nicht jeder Häufungspunkt ist auch ein Grenzwert. Für einen Häufungspunkt wird nämlich nur verlangt, daß unendlich viele Folgenglieder in $U_\varepsilon(x)$ liegen, aber nicht unbedingt fast alle. Es ist durchaus möglich, daß auch außerhalb von $U_\varepsilon(x)$ noch unendlich viele Folgenglieder zu finden sind – im Gegensatz zur Limes-Definition, wo sich außerhalb jeder ε -Umgebung nur *endlich* viele Folgenglieder aufhalten dürfen. Der sog. Vorzeichenflip $(-1)^n$ definiert eine Folge $(x_n)_{n \in \mathbb{N}} = (-1, 1, -1, 1, -1, \dots)$ mit zwei Häufungspunkten. Für $\varepsilon < 1$ liegen in jeder ε -Umgebung $U_\varepsilon(1) = (1 - \varepsilon, 1 + \varepsilon)$ von 1 unendlich viele Folgenglieder (nämlich alle mit geradem Index und alle mit demselben Wert 1), aber auch unendlich viele außerhalb (alle mit ungeradem Index). 1 ist also ein Häufungspunkt der Folge, aber kein Grenzwert.

Die Skizze 10.6 entspricht in etwa dem Bild, das man sich von einer Folge macht, die zwei oder mehrere Häufungspunkte hat. Aber eine Folge kann durchaus unendlich viele Häufungspunkte haben, ja sie kann überhaupt nur aus Häufungspunkten bestehen, wie z. B. die Folge, die aus $\mathbb{Q} \cap [0, 1]$ durch das CANTORSche Abzählverfahren entsteht. Da in jeder ε -Umgebung einer rationalen Zahl unendlich viele andere rationale Zahlen liegen, ist jede rationale Zahl ein Häufungspunkt.

10.1.36 Definition Eine Folge $(I_n)_{n \in \mathbb{N}}$ von Intervallen $I_n = [a_n, b_n]$ heißt Intervallschachtelung, falls die unteren Grenzen a_n eine monoton wachsende und die oberen Grenzen b_n eine monoton fallende Folge mit der Eigenschaft $\lim_{n \rightarrow \infty} b_n - a_n = 0$ bilden.

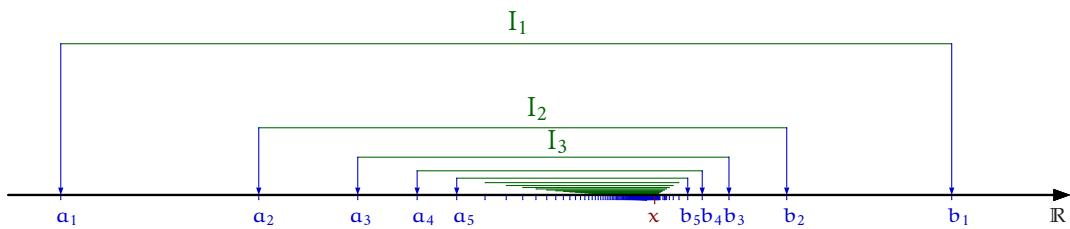


Abbildung 10.7 Eine Intervallschachtelung mit Grenzwert x

10.1.37 Satz Für jede Intervallschachtelung $(I_n)_{n \in \mathbb{N}}$ in \mathbb{R} gibt es genau ein Element $x \in \mathbb{R}$, das in allen Intervallen I_n liegt. Wir nennen x den Grenzwert von $(I_n)_{n \in \mathbb{N}}$.

Beweis. Die oberen Grenzen b_n bilden eine monoton fallende Folge, die nach unten durch a_1 beschränkt ist. Sie hat also nach dem Satz von der monotonen Konvergenz einen Grenzwert x . Analog folgt die Existenz eines Grenzwertes x' der Folge der unteren Grenzen a_n . Wegen

$$x = \lim_{n \rightarrow \infty} b_n + \lim_{n \rightarrow \infty} a_n - b_n = \lim_{n \rightarrow \infty} b_n + a_n - b_n = \lim_{n \rightarrow \infty} a_n = x',$$

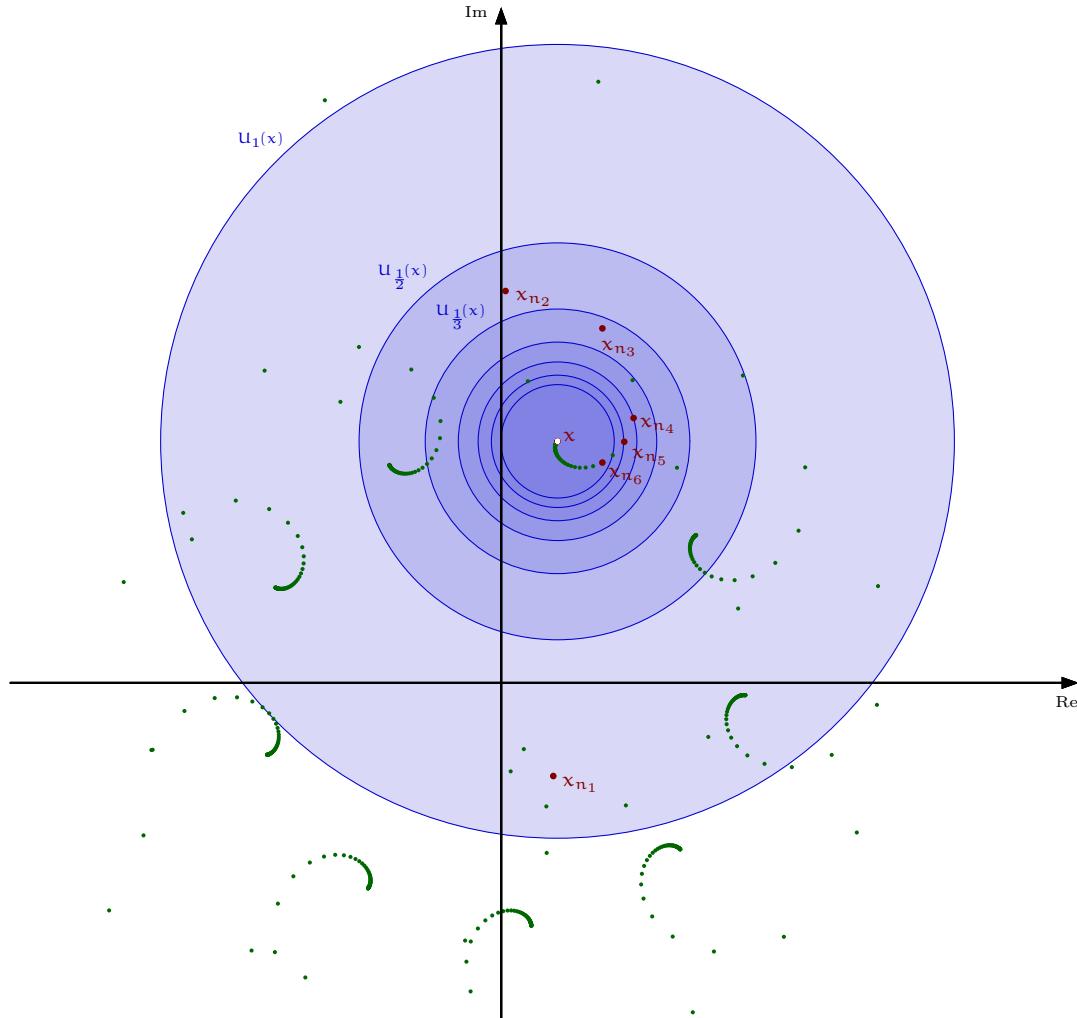
handelt es sich in beiden Fällen um denselben Grenzwert. Aus $a_n \leq x \leq b_n$ folgt $x \in I_n$ für alle $n \in \mathbb{N}$. Gäbe es ein weiteres Element x'' , das in allen Intervallen I_n zu finden ist, so müßte $a_n \leq x'' \leq b_n$ gelten, woraus sich nach dem Sandwich-Prinzip 10.1.5 $x = x''$ ergibt. \square

10.1.38 A Untersuchen Sie die Folge $\left(\left(1 + \frac{(-1)^n}{n} \right)^n \right)_{n \in \mathbb{N}}$. Verwenden Sie dabei die Beziehung $\left(1 - \frac{1}{n} \right)^n = \frac{1 - \frac{1}{n}}{\left(1 + \frac{1}{n-1} \right)^{n-1}}$.

10.1.39 A In der Menge \mathbb{Q} gilt der Satz von der Intervallschachtelung nicht. Können Sie ein Beispiel dafür finden?

10.1.40 Satz Zu jedem Häufungspunkt x einer Folge $(x_n)_{n \in \mathbb{N}}$ gibt es eine Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$, die x als Grenzwert hat. Umgekehrt bestimmt jede konvergente Teilfolge über ihren Grenzwert einen Häufungspunkt von $(x_n)_{n \in \mathbb{N}}$.

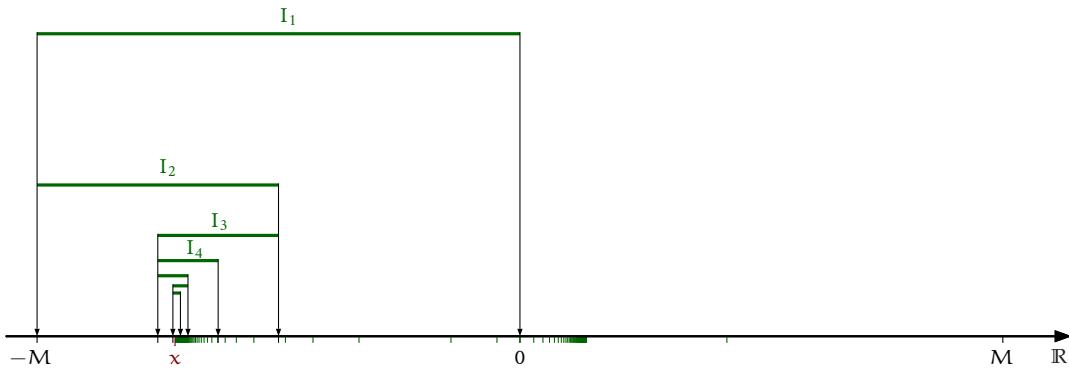
Beweis. In jeder ε -Umgebung $U_\varepsilon(x)$ gibt es unendlich viele Folgenglieder der Folge $(x_n)_{n \in \mathbb{N}}$. Wir wählen für jedes $k \in \mathbb{N}$ in den Umgebungen $U_{\frac{1}{k}}(x)$ ein Folgenglied aus, und zwar für $k = 1$ ein beliebiges Element x_{n_1} aus $U_1(x)$, für $k = 2$ ein beliebiges Element x_{n_2} aus $U_{\frac{1}{2}}(x)$ mit $n_2 > n_1$, dann für $k = 3$ ein $x_{n_3} \in U_{\frac{1}{3}}(x)$ mit $n_3 > n_2$ usw. Das ist jeweils möglich, da in jeder der Umgebungen $U_{\frac{1}{k}}(x)$ unendlich viele Folgenglieder zur Verfügung stehen. Auf diese Weise erhalten wir eine Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$ von $(x_n)_{n \in \mathbb{N}}$, mit der Eigenschaft $\|x - x_{n_\ell}\| < \frac{1}{k}$ für alle $\ell \geq k$. Da wir für jedes $\varepsilon > 0$ ein $k \in \mathbb{N}$ mit $\frac{1}{k} \leq \varepsilon$ finden können, liegen fast alle Folgenglieder dieser Teilfolge in $U_\varepsilon(x)$. Das zeigt $x = \lim_{k \rightarrow \infty} x_{n_k}$.



Nun sei x der Grenzwert einer Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$ von $(x_n)_{n \in \mathbb{N}}$. Dann liegen also in jeder ε -Umgebung von x fast alle Folgenglieder dieser Teilfolge, also insbesondere unendlich viele Folgenglieder der Ausgangsfolge $(x_n)_{n \in \mathbb{N}}$. Damit ist x ein Häufungspunkt von $(x_n)_{n \in \mathbb{N}}$. \square

10.1.41 Satz (BOLZANO-WEIERSTRASS) *Jede beschränkte Folge in \mathbb{R}^n oder \mathbb{C}^n hat wenigstens eine konvergente Teilfolge bzw. einen Häufungspunkt.*

Beweis. Wir führen den Beweis zunächst für \mathbb{R} . Wir gehen von einer beschränkten Folge $(x_n)_{n \in \mathbb{N}}$ aus, deren Folgenglieder also alle in einem geeigneten Intervall $I := [-M, M]$ liegen. Nun teilen wir I in zwei Hälften $[-M, 0], [0, M]$ und wählen diejenige als Intervall I_1 , die unendlich viele Folgenglieder enthält. Sollten beide unendlich viele Folgenglieder enthalten, dann wählen wir das linke. Dasselbe Verfahren wenden wir anschließend auf I_1 an und erhalten ein Intervall I_2 der Länge $\frac{M}{2}$. Auf diese Weise fahren wir fort und erhalten Intervalle I_3, I_4, \dots, I_n der Länge $\frac{M}{4}, \frac{M}{8}, \dots, \frac{M}{2^{n-1}}$. Dieses Intervallteilungsverfahren stellt sicher, daß die Intervalle ineinander liegen: $I_{n+1} \subseteq I_n$. Also bilden die unteren Grenzen dieser Intervalle eine monoton wachsende und die oberen eine monoton fallende beschränkte Folge. Da die Intervalllängen $\frac{M}{2^{n-1}}$ eine Nullfolge bilden, ist durch $(I_n)_{n \in \mathbb{N}}$ eine Intervallschachtelung mit einem Grenzwert x entstanden (Satz 10.1.37). x ist ein Häufungspunkt der Folge $(x_n)_{n \in \mathbb{N}}$, denn für jede ε -Umgebung $U_\varepsilon(x)$ findet man ein Intervall I_n , das ganz in $U_\varepsilon(x)$ enthalten ist (man muß ja nur für $\frac{M}{2^{n-1}} < \varepsilon$ sorgen). Nach Konstruktion der Intervalle I_n enthält die Umgebung dann unendlich viele Folgenglieder. Da es nach Satz 10.1.40 zu dem Häufungspunkt x eine Teilfolge von $(x_n)_{n \in \mathbb{N}}$ gibt, die gegen x konvergiert, ist die Behauptung des Satzes im Falle \mathbb{R} gezeigt.



Untersuchen wir jetzt den Fall einer beschränkten Folge $([x_n, y_n])_{n \in \mathbb{N}}$ im \mathbb{R}^2 . Es gibt also eine Zahl $M > 0$ mit der Eigenschaft $\|[x_n, y_n]\| \leq M$ für alle $n \in \mathbb{N}$. Wegen der Monotonie der Wurzel folgt aus

$$|x_n| = \sqrt{|x_n|^2} \leq \sqrt{|x_n|^2 + |y_n|^2} = \|[x_n, y_n]\| \leq M,$$

dass auch die reelle Folge $(x_n)_{n \in \mathbb{N}}$ beschränkt ist. Daher hat sie eine konvergente Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$, mit einem Grenzwert x . Wir bilden die Teilfolge $([x_{n_k}, y_{n_k}])_{k \in \mathbb{N}}$ der Ausgangsfolge und wenden die eben vorgestellte Überlegung auf die beschränkte Folge $(y_{n_k})_{k \in \mathbb{N}}$ an. Wir erhalten eine konvergente Teilfolge $(y_{n_{k_\ell}})_{\ell \in \mathbb{N}}$ dieser Teilfolge, mit einem Grenzwert y . Diese Teilfolge übertragen wir wieder auf die Ausgangsfolge und erhalten $([x_{n_{k_\ell}}, y_{n_{k_\ell}}])_{\ell \in \mathbb{N}}$. Das nachfolgende Lemma zeigt, dass diese Folge gegen $[x, y]$ konvergiert. Es ist jetzt klar, wie dieses Verfahren auf \mathbb{R}^n , auf \mathbb{C} und dann auch auf \mathbb{C}^n ausgedehnt werden kann. Satz 10.1.40 zeigt die Existenz des Häufungspunktes. \square

10.1.42 Lemma Eine Folge $(x_n)_{n \in \mathbb{N}} = ([x_{1,n}, x_{2,n}, \dots, x_{k,n}]^t)_{n \in \mathbb{N}}$ in \mathbb{R}^k (\mathbb{C}^k) konvergiert genau dann gegen $x = [x_1, x_2, \dots, x_k]^t$, wenn die Koordinatenfolgen $(x_{j,n})_{n \in \mathbb{N}}$ gegen x_j konvergieren ($j = 1, 2, \dots, k$).

Beweis. Es gilt

$$|x_j - x_{j,n}| = \sqrt{|x_j - x_{j,n}|^2} \leq \sqrt{\sum_{i=1}^k |x_i - x_{i,n}|^2} = \|x - x_n\|.$$

Wenn es also für alle $\varepsilon > 0$ ein $n_\varepsilon \in \mathbb{N}$ gibt, das für alle $n \geq n_\varepsilon$ die Abschätzung $\|x - x_n\| < \varepsilon$ ermöglicht, dann gilt das erst recht für $|x_j - x_{j,n}|$. Das zeigt die Konvergenz der Koordinaten: $\lim_{n \rightarrow \infty} x_{j,n} = x_j, j = 1, 2, \dots, k$. Umgekehrt folgt aus dieser Eigenschaft nach den Rechenregeln für konvergente Folgen $\lim_{n \rightarrow \infty} \sum_{j=1}^k |x_j - x_{j,n}|^2 = 0$. Nach Übung 10.1.13 können wir damit auf $\lim_{n \rightarrow \infty} \sqrt{\sum_{j=1}^k |x_j - x_{j,n}|^2} = \lim_{n \rightarrow \infty} \|x - x_n\| = 0$, also auf $x = \lim_{n \rightarrow \infty} x_n$ schließen. \square

10.1.43 Definition Eine Folge $(x_n)_{n \in \mathbb{N}}$ in einem normierten Raum X heißt CAUCHY-Folge, falls für alle $\varepsilon > 0$ ein $n_\varepsilon \in \mathbb{N}$ existiert, so daß für alle $n, m \geq n_\varepsilon$ die Abschätzung $\|x_n - x_m\| < \varepsilon$ gilt. Diese Bedingung wird auch CAUCHY-Bedingung oder CAUCHY-Kriterium genannt.

In der kurzen Notation mit Quantoren:

$$\forall \varepsilon > 0 \exists n_\varepsilon \in \mathbb{N} \forall n, m \geq n_\varepsilon \|x_n - x_m\| < \varepsilon. \quad (10.17)$$

Die CAUCHY-Bedingung beschreibt die Eigenschaft einer konvergenten Folge, daß die Folgenglieder, um dem Grenzwert beliebig nahe zu kommen, sich auch untereinander beliebig annähern müssen. Deshalb ist das folgende Lemma keine Überraschung.

10.1.44 Lemma Jede konvergente Folge ist eine CAUCHY-Folge.

Beweis. $(x_n)_{n \in \mathbb{N}}$ sei eine konvergente Folge mit Grenzwert x . Dann gibt es zu jedem $\varepsilon > 0$ ein $n_\varepsilon \in \mathbb{N}$, so daß für alle $n \geq n_\varepsilon$ die Ungleichung $\|x - x_n\| < \frac{\varepsilon}{2}$ erfüllt ist. Also gilt für alle $n, m \geq n_\varepsilon$:

$$\|x_n - x_m\| \leq \|x_n - x\| + \|x - x_m\| < \varepsilon. \quad \square$$

Das Interessante an der CAUCHY-Bedingung ist, daß sie ohne Vermutung über den möglichen Grenzwert einer Folge auskommt. Bei schwierig zu untersuchenden Folgen ist man meist schon zufrieden, wenn man überhaupt erst einmal weiß, daß es einen Grenzwert gibt, auf dessen genauen Wert man anschließend weitere Anstrengungen verwenden kann. Das HERON-Verfahren (10.16) zur Bestimmung der Quadratwurzel ist ein solches Beispiel. Erst nachdem die Existenz des Grenzwertes gesichert war, konnte man diesen dann über die Rekursionsgleichung (10.16) leicht bestimmen. Hier kam mit dem Satz von der monotonen Konvergenz schon ein Konvergenzkriterium zum Einsatz, das auch ohne genaue Kenntnis des möglichen Grenzwertes auskommt, allerdings die Monotonie der Folge voraussetzt. Wenn das CAUCHY-Kriterium für die Konvergenz einer Folge nicht nur notwendig, sondern auch hinreichend ist, dann haben wir ein Kriterium zur Verfügung, das an die Folge keine weiteren Anforderungen stellt. Sicher ist das nicht in allen Räumen so, denn in \mathbb{Q} wird jede Folge, die $\sqrt{2}$ approximiert, eine CAUCHY-Folge sein, aber keinen Grenzwert haben. Tatsächlich sind in allen Räumen, in denen der Satz von BOLZANO-WEIERSTRASS gilt, alle CAUCHY-Folgen konvergent. Da der Satz von BOLZANO-WEIERSTRASS eine direkte Folge des Vollständigkeitsaxioms von \mathbb{R} ist, kann man sich überlegen, daß die Konvergenz von CAUCHY-Folgen äquivalent zur Vollständigkeit von \mathbb{R} ist. Das ist

der Grund, warum mitunter auch die Konvergenz von CAUCHY-Folgen als Kriterium für die Vollständigkeit eines Raumes verwendet wird.

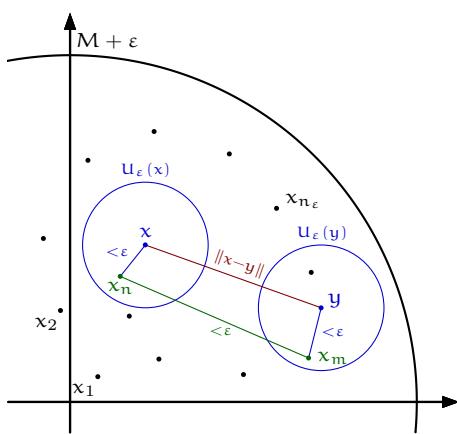
10.1.45 Satz In \mathbb{R}^n und \mathbb{C}^n haben alle CAUCHY-Folgen einen Grenzwert.

Beweis. Der Beweis verläuft in zwei Schritten. Im ersten Schritt weisen wir die Beschränktheit einer CAUCHY-Folge $(x_n)_{n \in \mathbb{N}}$ nach, die für die Konvergenz natürlich unabdingbar ist (Lemma 10.1.3). Damit wird im zweiten Schritt durch den Satz von BOLZANO-WEIERSTRASS 10.1.41 die Existenz eines Häufungspunktes x bzw. die Konvergenz einer Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$ gegen einen Grenzwert x sichergestellt. Die Arbeit besteht dann nur noch darin, zu zeigen, daß x auch der Grenzwert der Ausgangsfolge $(x_n)_{n \in \mathbb{N}}$ ist. Dafür geben wir zwei Beweise an, einmal unter Verwendung des Häufungspunktes und einmal mit Hilfe der Teilfolge, die der Satz von BOLZANO-WEIERSTRASS jeweils bereitstellt.

1) $(x_n)_{n \in \mathbb{N}}$ sei eine CAUCHY-Folge. Für alle $\varepsilon > 0$ gibt es also ein $n_\varepsilon \in \mathbb{N}$, so daß für alle $m, n \geq n_\varepsilon$ die Abschätzung $\|x_m - x_n\| < \varepsilon$ gilt. Die endlich vielen Folgenglieder $x_1, \dots, x_{n_\varepsilon}$ sind in ihrer Norm natürlich beschränkt, sagen wir durch $M > 0$. Die unendlich vielen, die den Rest der Folge ausmachen, können höchstens noch einen Beitrag ε zu M hinzufügen: Für alle $n \geq n_\varepsilon$ gilt nämlich

$$\|x_n\| \leq \|x_n - x_{n_\varepsilon}\| + \|x_{n_\varepsilon}\| < M + \varepsilon.$$

Damit haben wir für alle Folgenglieder $\|x_n\| < \varepsilon + M$.



2) Nach dem Satz von BOLZANO-WEIERSTRASS gibt es einen Häufungspunkt x für die Folge $(x_n)_{n \in \mathbb{N}}$. Wir nehmen an, es gäbe einen weiteren $y \neq x$. Dann liegen in den beiden ε -Umgebungen $U_\varepsilon(x)$ bzw. $U_\varepsilon(y)$ jeweils unendlich viele Folgenglieder, o. B. d. A. also ein $x_n \in U_\varepsilon(x)$ und ein $x_m \in U_\varepsilon(y)$, mit $n, m \geq n_\varepsilon$. Wir wählen $\varepsilon < \frac{1}{3} \|x - y\|$. Daraus würde

$$\begin{aligned} \|x - y\| &\leq \|x - x_n\| + \|x_n - x_m\| + \|x_m - y\| \\ &< 3 \cdot \frac{1}{3} \|x - y\| = \|x - y\| \end{aligned}$$

folgen, ein offensichtlicher Widerspruch. Damit gibt es nur den einen Häufungspunkt x für $(x_n)_{n \in \mathbb{N}}$. Der

ist sogar ein Grenzwert, denn in jeder ε -Umgebung von x müssen fast alle Folgenglieder liegen. Gäbe es außerhalb dieser Umgebung auch noch unendlich viele, dann müßte es einen weiteren Häufungspunkt y außerhalb von $U_\varepsilon(x)$ geben, was, wie wir gerade gezeigt haben, nicht möglich ist.

2a) Nach dem Satz von BOLZANO-WEIERSTRASS gibt es eine konvergente Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$ von $(x_n)_{n \in \mathbb{N}}$, mit einem Grenzwert x . Wir haben damit einerseits die Konvergenzbedingung für die Teilfolge zur Verfügung, also

für alle $\varepsilon > 0$ gibt es ein $k_\varepsilon \in \mathbb{N}$, so daß für alle $k \geq k_\varepsilon$ die Ungleichung
 $\|x - x_{n_k}\| < \frac{\varepsilon}{2}$ gilt.

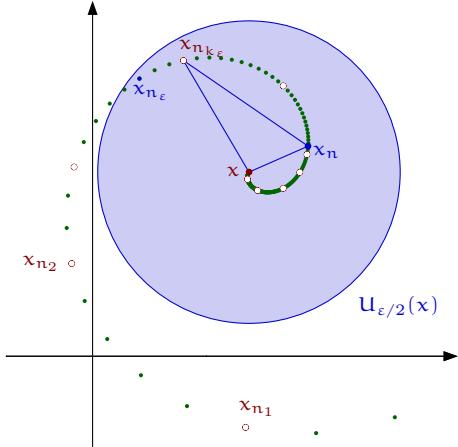
Andererseits gilt die CAUCHY-Bedingung für die gesamte Folge:

Für alle $\varepsilon > 0$ gibt es ein $n_\varepsilon \in \mathbb{N}$, so daß für alle $m, n \geq n_\varepsilon$ die Ungleichung $\|x_m - x_n\| < \frac{\varepsilon}{2}$ gilt.

Wir dürfen o. B. d. A. annehmen, daß $n_{k_\varepsilon} \geq n_\varepsilon$ gilt, denn andernfalls müssen wir nur für $k_\varepsilon \geq n_\varepsilon$ sorgen. Die Konvergenzbedingung für die Teilfolge bleibt dabei weiterhin gültig. Für alle $n \geq n_\varepsilon$ können wir dann folgendermaßen abschätzen:

$$\begin{aligned}\|x - x_n\| &\leq \|x - x_{n_{k_\varepsilon}}\| + \|x_{n_{k_\varepsilon}} - x_n\| \\ &< 2 \cdot \frac{\varepsilon}{2} = \varepsilon.\end{aligned}$$

Das bedeutet $x = \lim_{n \rightarrow \infty} x_n$. □



10.1.46 Vertauschung von Grenzwerten

$$\begin{array}{ccccccccc}
 & & & & & & & & \xrightarrow{n \rightarrow \infty} \\
 a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} & \dots \rightarrow & \alpha_1 = \lim_{n \rightarrow \infty} a_{1n} \\
 a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} & \dots \rightarrow & \alpha_2 = \lim_{n \rightarrow \infty} a_{2n} \\
 a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} & \dots \rightarrow & \alpha_3 = \lim_{n \rightarrow \infty} a_{3n} \\
 a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} & \dots \rightarrow & \alpha_4 = \lim_{n \rightarrow \infty} a_{4n} \\
 a_{51} & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} & \dots \rightarrow & \alpha_5 = \lim_{n \rightarrow \infty} a_{5n} \\
 a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & a_{66} & \dots \rightarrow & \alpha_6 = \lim_{n \rightarrow \infty} a_{6n} \\
 \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & a_{mn} & \vdots \\
 \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \nearrow & \downarrow \\
 \beta_1 = \lim_{m \rightarrow \infty} a_{m1} & \beta_2 = \lim_{m \rightarrow \infty} a_{m2} & \beta_3 = \lim_{m \rightarrow \infty} a_{m3} & \beta_4 = \lim_{m \rightarrow \infty} a_{m4} & \beta_5 = \lim_{m \rightarrow \infty} a_{m5} & \beta_6 = \lim_{m \rightarrow \infty} a_{m6} & \dots \rightarrow ? &
 \end{array}$$

Eine unendliche Matrix $(a_{mn})_{m \in \mathbb{N}, n \in \mathbb{N}}$ von Elementen a_{mn} eines normierten Raumes bezeichnen wir als *Doppelfolge*. Es ist nicht offensichtlich, was unter der Konvergenz einer solchen Folge zu verstehen ist. Zwei naheliegende Möglichkeiten sind, jeweils zwei Grenzwerte zu betrachten, nämlich zunächst den Grenzwert $n \rightarrow \infty$ und dann den Grenzwert $m \rightarrow \infty$, bzw. zuerst $m \rightarrow \infty$ und dann $n \rightarrow \infty$:

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} a_{mn}, \quad \lim_{n \rightarrow \infty} \lim_{m \rightarrow \infty} a_{mn}.$$

Allerdings ist nicht klar, ob diese sogenannten *iterierten Grenzwerte* überhaupt existieren, oder gar übereinstimmen. Wenn sie übereinstimmen, dann heißt das, wir können die Reihenfolge der Grenzwerte vertauschen. Zwei einfache Beispiele zeigen, daß das normalerweise nicht zu erwarten ist. Für $a_{mn} := \frac{n-m}{n+m}$ gilt

$$\begin{aligned}
 \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{n-m}{n+m} &= \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1 - \frac{m}{n}}{1 + \frac{m}{n}} = \lim_{m \rightarrow \infty} 1 = 1, \\
 \lim_{n \rightarrow \infty} \lim_{m \rightarrow \infty} \frac{n-m}{n+m} &= \lim_{n \rightarrow \infty} \lim_{m \rightarrow \infty} \frac{\frac{n}{m} - 1}{\frac{n}{m} + 1} = \lim_{n \rightarrow \infty} -1 = -1,
 \end{aligned}$$

und für $b_{mn} := \frac{n+m}{n}$

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{n+m}{n} = \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} 1 + \frac{m}{n} = \lim_{m \rightarrow \infty} 1 = 1,$$

während der Grenzwert $m \rightarrow \infty$ offensichtlich nicht existiert. Selbst wenn es die iterierten Grenzwerte gibt, ist ohne weitere Voraussetzungen also nicht zu erwarten, daß sie übereinstimmen, und aus der Existenz eines der beiden kann nicht ohne Weiteres auf die Existenz des anderen geschlossen werden.

Wir können auch unsere Grenzwertdefinition 10.1.1 an die Gegebenheiten einer Doppelfolge anpassen. a heißt *Doppelimes* der Doppelfolge $(a_{mn})_{m,n \in \mathbb{N}}$, falls

$$\forall \varepsilon > 0 \exists m_\varepsilon \in \mathbb{N}, n_\varepsilon \in \mathbb{N} \forall m \geq m_\varepsilon, n \geq n_\varepsilon \|a_{mn} - a\| < \varepsilon \quad (10.18)$$

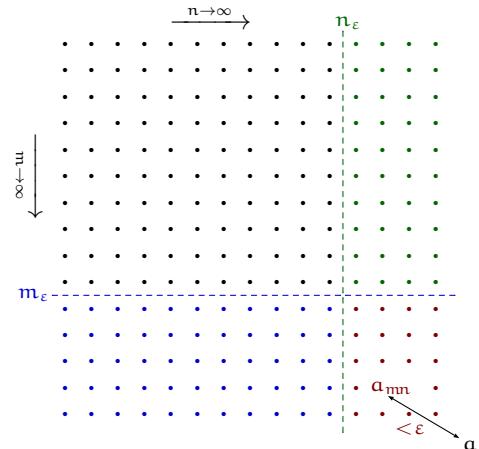
erfüllt ist. Wir schreiben dafür

$$a = \lim_{m,n \rightarrow \infty} a_{mn} \quad (10.19)$$

Allerdings folgt auch aus der Existenz dieses Doppellimes nicht automatisch die der iterierten Grenzwerte. Etwa für die Doppelfolge

$$c_{mn} := \frac{(-1)^m}{n} + \frac{(-1)^n}{m}$$

gibt es weder $\lim_{n \rightarrow \infty} c_{mn}$ noch $\lim_{m \rightarrow \infty} c_{mn}$, aber der Doppellimes ist 0, denn wegen $|c_{mn}| \leq \frac{1}{n} + \frac{1}{m}$ ist $|c_{mn}| < 2\epsilon$, für $m, n \geq m_\epsilon = n_\epsilon = [\frac{1}{\epsilon}] + 1$.



10.1.47 Satz (1. Vertauschbarkeitssatz)

Existieren für eine Doppelfolge $(a_{mn})_{m,n \in \mathbb{N}}$ die Grenzwerte

$$i) \lim_{n \rightarrow \infty} a_{mn} \quad ii) \lim_{m \rightarrow \infty} a_{mn} \quad iii) \lim_{m,n \rightarrow \infty} a_{mn},$$

dann gibt es auch die iterierten Grenzwerte und diese sind gleich:

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} a_{mn} = \lim_{n \rightarrow \infty} \lim_{m \rightarrow \infty} a_{mn} = \lim_{m,n \rightarrow \infty} a_{mn}.$$

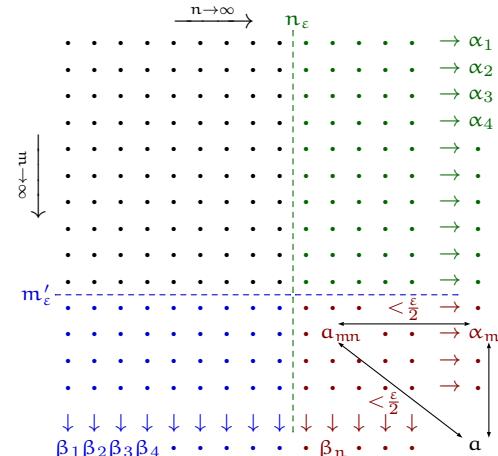
Beweis.

Für $a := \lim_{m,n \rightarrow \infty} a_{mn}$, $\alpha_m := \lim_{n \rightarrow \infty} a_{mn}$, sowie $\beta_n := \lim_{m \rightarrow \infty} a_{mn}$ gilt

$$\|a - \alpha_m\| \leq \|a - a_{mn}\| + \|a_{mn} - \alpha_m\|.$$

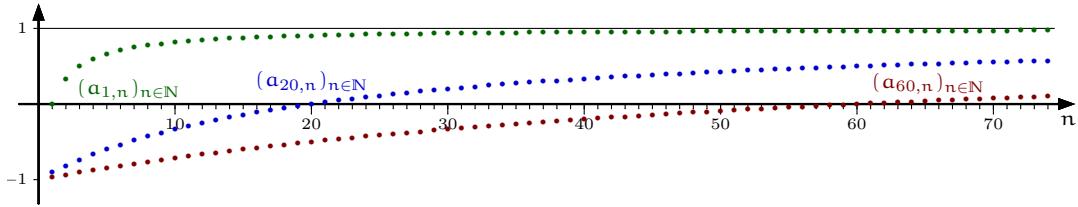
Dann gibt es m'_ϵ und n_ϵ , so daß $\|a - a_{mn}\| < \frac{\epsilon}{2}$ für alle $m \geq m'_\epsilon$ und $n \geq n_\epsilon$ gilt. Ein solches n fixieren wir und wählen anschließend ein m''_ϵ , mit $\|a_{mn} - \alpha_m\| < \frac{\epsilon}{2}$ für alle $m \geq m''_\epsilon$. Dann folgt $\|a - \alpha_m\| < \epsilon$ für alle $m \geq \max\{m'_\epsilon, m''_\epsilon\}$, also $a = \lim_{m \rightarrow \infty} \alpha_m = \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} a_{mn}$.

Genauso zeigen wir auch die Existenz des anderen iterierten Grenzwerts und daß er mit a übereinstimmt: $a = \lim_{n \rightarrow \infty} \beta_n = \lim_{n \rightarrow \infty} \lim_{m \rightarrow \infty} a_{mn}$. □



Wir schauen uns noch einmal die Doppelfolge $(a_{mn})_{m,n \in \mathbb{N}}$ von Seite 235 an und versuchen zu verstehen, woran die Vertauschung der Grenzwerte scheitert. Die Zahlen $a_{mn} = \frac{n-m}{n+m}$ konvergieren für $n \rightarrow \infty$ gegen 1. Aber a_{mn} ist für $n < m$ offensichtlich negativ und daher weit vom Grenzwert 1 entfernt. Das bedeutet, das n_ϵ , ab dem $|1 - a_{mn}| < \epsilon$ ist, muß sicher größer als m sein. Mit wachsendem m wird die Konvergenz für $n \rightarrow \infty$ daher beliebig schlecht, weil wir immer länger warten müssen, bis die gewünschte Genauigkeit ϵ erreicht wird. Das

Verhalten für $n \rightarrow \infty$ hängt also sehr empfindlich von dem zweiten Index m ab. Dasselbe gilt auch für den Grenzwert $a_{mn} \xrightarrow{m \rightarrow \infty} -1$, denn $a_{mn} > 0$, solange m noch kleiner als n ist.



Sehen wir uns die Grenzwertbedingung für die Grenzwerte $a_{mn} \xrightarrow{n \rightarrow \infty} 1$ an:

$$\forall m \in \mathbb{N} \forall \varepsilon > 0 \exists n_{m,\varepsilon} \forall n \geq n_{m,\varepsilon} |a_{mn} - 1| < \varepsilon.$$

Der Grenzindex $n_{m,\varepsilon}$ für die Genauigkeit ε hängt normalerweise von dem zweiten Index m ab. So kann es passieren, daß er mit wachsendem m über alle Grenzen hinaus strebt. Wir können dieses Verhalten unterbinden, wenn wir eine stärkere Konvergenzbedingung fordern, bei der der Index n_ε nicht von m abhängt. Wir sagen, die Folgen $(a_{mn})_{m,n \in \mathbb{N}}$ konvergieren für $n \rightarrow \infty$ gleichmäßig bzgl. m gegen α_m , wenn

$$\forall \varepsilon > 0 \exists n_\varepsilon \forall n \geq n_\varepsilon \forall m \in \mathbb{N} \|a_{mn} - \alpha_m\| < \varepsilon \quad (10.20)$$

erfüllt ist. Entsprechend ist die gleichmäßige Konvergenz $m \rightarrow \infty$ bzgl. $n \in \mathbb{N}$ definiert.

10.1.48 Satz (2. Vertauschbarkeitssatz)

Für eine Doppelfolge $(a_{mn})_{m,n \in \mathbb{N}}$ seien die Grenzwerte

$$i) \lim_{n \rightarrow \infty} a_{mn} \quad \text{und} \quad ii) \lim_{m \rightarrow \infty} a_{mn}$$

vorhanden. Wenn der Grenzwert i) bzgl. m , oder ii) bzgl. n gleichmäßig existiert, dann gibt es die iterierten Grenzwerte, den Doppelimes und alle diese Grenzwerte sind gleich:

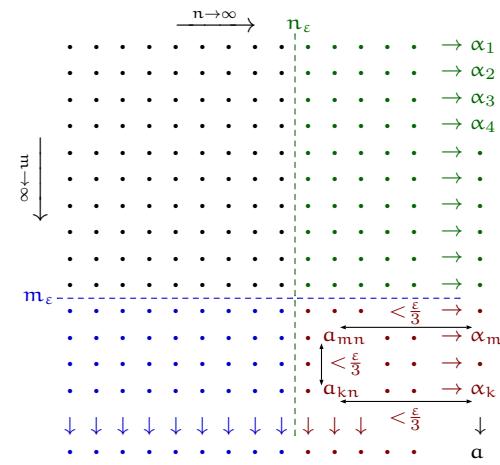
$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} a_{mn} = \lim_{n \rightarrow \infty} \lim_{m \rightarrow \infty} a_{mn} = \lim_{m,n \rightarrow \infty} a_{mn}.$$

Beweis. Wir gehen o. B. d. A. davon aus, daß die Folgen $a_{mn} \xrightarrow{n \rightarrow \infty} \alpha_m$ gleichmäßig bzgl. $m \in \mathbb{N}$ konvergieren. Wir zeigen, daß der Doppelimes $\lim_{m,n \rightarrow \infty} a_{mn}$ existiert. Dadurch haben wir den ersten Vertauschbarkeitssatz 10.1.47 zur Verfügung, aus dem die Behauptung folgt. Dafür weisen wir zuerst die CAUCHY-Bedingung für $(\alpha_m)_{m \in \mathbb{N}}$ nach und haben damit, daß diese Folge konvergiert. Anschließend zeigen wir, daß ihr Grenzwert der Doppelimes von $(a_{mn})_{m,n \in \mathbb{N}}$ ist.

Wir schätzen folgendermaßen ab:

$$\begin{aligned}\|\alpha_m - \alpha_k\| &\leqslant \|\alpha_m - a_{mn}\| \\ &+ \|a_{mn} - a_{kn}\| \\ &+ \|a_{kn} - \alpha_k\|.\end{aligned}$$

Der erste Summand $\|\alpha_m - a_{mn}\|$ und der letzte $\|a_{kn} - \alpha_k\|$ kann jeweils unabhängig von m bzw. k ab einem Index n_ε kleiner als $\frac{\varepsilon}{3}$ gemacht werden. Das folgt aus der gleichmäßigen Konvergenz bzgl. m von $(a_{mn})_{n \in \mathbb{N}}$. Wir halten ein $n \geq n_\varepsilon$ fest und wählen ein m_ε , so daß (für dieses eine n) $\|a_{mn} - a_{kn}\| < \frac{\varepsilon}{3}$ für alle $m, k \geq m_\varepsilon$ gilt. Das geht, denn die Folge $(a_{mn})_{m \in \mathbb{N}}$ ist ja laut Voraussetzung konvergent und erfüllt damit die CAUCHY-Bedingung. Zusammengefaßt haben wir daher $\|\alpha_m - \alpha_k\| < \varepsilon$ für alle $m, k \geq m_\varepsilon$, also die CAUCHY-Bedingung für die Folge $(\alpha_m)_{m \in \mathbb{N}}$. Ihr Grenzwert sei a .



Jetzt können wir erneut abschätzen: $\|a - a_{mn}\| \leq \|\alpha_m - a_m\| + \|\alpha_m - a_{mn}\|$.

Der erste Summand $\|\alpha_m - a_m\|$ wird ab einem geeigneten m'_ε kleiner als $\frac{\varepsilon}{2}$ und der zweite $\|\alpha_m - a_{mn}\|$ für alle $n \geq n_\varepsilon$ unabhängig von m . Insgesamt haben wir $\|a - a_{mn}\| < \varepsilon$ für alle $m \geq m'_\varepsilon$ und $n \geq n_\varepsilon$, also $a = \lim_{m, n \rightarrow \infty} a_{mn}$. Damit ist alles gezeigt. \square

10.1.49 Gleichmäßige Konvergenz Wir werden immer wieder der Situation begegnen, daß eine Folge von einem Parameter abhängt. Eine Doppelfolge $(a_{mn})_{m, n \in \mathbb{N}}$ etwa bildet bezüglich ihrem zweiten Index eine Folge $(a_{mn})_{m \in \mathbb{N}}$ mit dem Parameter $n \in \mathbb{N}$. Oder eine Folge von Funktionen $(f_n)_{n \in \mathbb{N}}$ auf einem gemeinsamen Definitionsbereich D . Die Funktionswerte $(f_n(x))_{n \in \mathbb{N}}$ bilden eine Folge mit dem Parameter $x \in D$. Das Konvergenzverhalten solcher Folgen wird im Allgemeinen von dem Parameter abhängen. Von besonderem Interesse sind aber oft die Situationen, in denen das nicht der Fall ist, wenn die Konvergenz sozusagen unempfindlich ist gegenüber ihrem Parameter. Solche Folgen bezeichnen wir als *gleichmäßig konvergent* (als Verallgemeinerung von 10.20). Eine wichtige Anwendung ist die gleichmäßige Konvergenz von stetigen Funktionen (vergl. 11.1.29 und Satz 11.1.30).

10.1.50 Definition Haben wir für jedes $x \in X$ einer Menge X eine Folge $(a_n(x))_{n \in \mathbb{N}}$ aus einem normierten Raum V gegeben, dann sprechen wir von einer Folge mit Parameter $x \in X$. Mitunter bezeichnen wir die Folge auch einfach durch $(a_n)_{n \in \mathbb{N}}$, wenn die Parametermenge X aus dem Kontext heraus bekannt ist. Sie heißt punktweise konvergent bezüglich x , falls es für jedes $x \in X$ ein Element $a(x) \in V$ mit der Eigenschaft

$$\forall \varepsilon > 0 \exists n_{\varepsilon, x} \in \mathbb{N} \forall n \geq n_{\varepsilon, x} \|a(x) - a_n(x)\| < \varepsilon \quad (10.21)$$

gibt. Sie heißt gleichmäßig konvergent bezüglich x , falls

$$\forall \varepsilon > 0 \exists n_\varepsilon \in \mathbb{N} \forall n \geq n_\varepsilon \forall x \in D \|a(x) - a_n(x)\| < \varepsilon \quad (10.22)$$

erreicht werden kann. Wir bezeichnen die Folge als gleichmäßige CAUCHY-Folge bezüglich $x \in X$, wenn die CAUCHY-Bedingung gleichmäßig bezüglich x erfüllt ist:

$$\forall \varepsilon > 0 \exists n_\varepsilon \in \mathbb{N} \forall m, n \geq n_\varepsilon \forall x \in D \|a_m(x) - a_n(x)\| < \varepsilon. \quad (10.23)$$

10.1.51 Lemma Jede gleichmäßige CAUCHY-Folge auf \mathbb{C}^k ist gleichmäßig konvergent.

Beweis. Eine gleichmäßige CAUCHY-Folge $(a_n(x))_{n \in \mathbb{N}}$ mit Parametermenge X hat für jedes $x \in X$ einen Grenzwert $a(x) \in \mathbb{C}^k$, da in \mathbb{C}^k alle CAUCHY-Folgen konvergent sind. $a(x)$ ist der gleichmäßige Grenzwert der Folge $(a_n(x))_{n \in \mathbb{N}}$. Für alle $\varepsilon > 0$ gibt es nämlich ein n_ε , so daß für alle $m, n \geq n_\varepsilon$ und für alle $x \in X$ die Abschätzung $\|a_n(x) - a_m(x)\| < \frac{\varepsilon}{2}$ erfüllt ist. Das ist die gleichmäßige CAUCHY-Bedingung. $a(x)$ ist aber auch der gewöhnliche Grenzwert der Folge $(a_n(x))_{n \in \mathbb{N}}$. Daher gibt es ein $\tilde{n}_{\varepsilon, x}$, mit $\|a(x) - a_m(x)\| < \frac{\varepsilon}{2}$ für alle $m \geq \tilde{n}_{\varepsilon, x}$. Wir wählen ein solches m , das o. B. d. A. größer als n_ε angenommen werden darf. Dann gilt für alle $n \geq n_\varepsilon$

$$\|a(x) - a_n(x)\| \leq \|a(x) - a_m(x)\| + \|a_m(x) - a_n(x)\| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Das zeigt die gleichmäßige Konvergenz von $(a_n(x))_{n \in \mathbb{N}}$ gegen $a(x)$, denn der Grenzindex n_ε konnte unabhängig von $x \in X$ gewählt werden. \square

10.2 Reihen

10.2.1 Definition Sei $(x_n)_{n \in \mathbb{N}}$ eine Folge in dem normierten Raum X . Dann ist die zugehörige Reihe $(S_n)_{n \in \mathbb{N}}$ die Folge der sogenannten Partialsummen

$$S_n := \sum_{k=1}^n x_k \quad (10.24)$$

die durch Addition der ersten n Folgenglieder x_1, x_2, \dots, x_n entstehen. Meist wird eine Reihe einfach nur durch die Partialsumme in (10.24) angegeben. Falls die Reihe einen Grenzwert hat, wird dieser durch eine unendliche Summe symbolisiert:

$$\sum_{k=1}^{\infty} x_k := \lim_{n \rightarrow \infty} \sum_{k=1}^n x_k \quad (10.25)$$

Den Grenzwert einer Reihe erhalten wir, wenn es gelingt, die unendlich vielen Folgenglieder zu einer endlichen Größe zu addieren. Dafür ist sicher notwendig, daß die Folgenglieder eine Nullfolge bilden. Mit der harmonischen Reihe (Seite 210) haben wir allerdings schon eine Reihe kennengelernt, die zeigt, daß diese Eigenschaft nicht hinreichend ist. Zwar streben die Summanden $\frac{1}{k}$ dieser Reihe gegen Null, aber offensichtlich nicht schnell genug. Mit der sogenannten *geometrischen Reihe* stellen wir ein erstes und im Folgenden wichtiges Beispiel einer konvergenten Reihe vor. Für eine Zahl $q \in \mathbb{C}$ ist sie durch

$$S_n := \sum_{k=1}^n q^k$$

definiert. Sie ist eine der wenigen Reihen, für die es eine geschlossene Formel für die Partialsummen gibt. Für $q \neq 1$ gilt

$$\begin{aligned} (1 - q)S_n &= \sum_{k=1}^n q^k - \sum_{k=1}^n q^{k+1} = \sum_{k=1}^n q^k - \sum_{\ell=2}^{n+1} q^\ell \\ &= q + \sum_{k=2}^n q^k - \sum_{\ell=2}^n q^\ell - q^{n+1} = q - q^{n+1}. \end{aligned}$$

Auflösen nach S_n ergibt

$$\sum_{k=1}^n q^k = \frac{q - q^{n+1}}{1 - q}. \quad (10.26)$$

Ist $|q| < 1$, dann konvergiert dieser Ausdruck nach (10.10). Wir erhalten

$$\sum_{k=1}^{\infty} q^k = \frac{q}{1 - q}. \quad (10.27)$$

Für die meisten Reihen gibt es jedoch keine geschlossene Formeln für die Partialsummen S_n , und meist hat man auch keine konkrete Vorstellung von einem möglichen Grenzwert. Wir werden sehen, daß die Reihe $\sum_{k=1}^n \frac{1}{k^2}$ konvergiert, aber es ist sicher nicht unsere erste Idee, daß der

Grenzwert $\frac{\pi^2}{6}$ sein könnte (was er ist, was aber mit unseren bisherigen Methoden nur schwerlich nachzuweisen wäre). Mit dem CAUCHY-Kriterium 10.1.43 steht uns aber *das* Hilfsmittel zur Untersuchung von Reihen zur Verfügung. Es lautet für eine Reihe $(S_n)_{n \in \mathbb{N}}$ folgendermaßen:

Für alle $\varepsilon > 0$ gibt es ein $n_\varepsilon \in \mathbb{N}$, so daß für alle $n > m \geq n_\varepsilon$

$$\begin{aligned}\|S_n - S_m\| &= \left\| \sum_{k=1}^n x_k - \sum_{k=1}^m x_k \right\| = \left\| \sum_{k=1}^m x_k + \sum_{k=m+1}^n x_k - \sum_{k=1}^m x_k \right\| \\ &= \left\| \sum_{k=m+1}^n x_k \right\| < \varepsilon\end{aligned}$$

gilt. Für den Spezialfall $m = n - 1$ erhalten wir daraus immerhin, daß die Summanden x_k tatsächlich eine Nullfolge bilden müssen: $\|x_n\| < \varepsilon$ für alle $n \geq n_\varepsilon$. Aber im Allgemeinen ist in dieser Form das CAUCHY-Kriterium nicht sehr hilfreich, da die letzte Summe nicht leicht zu kontrollieren ist. Wünschenswert ist es, diese durch $\sum_{k=m+1}^n \|x_k\|$ zu ersetzen, denn das ist eine Summe nicht negativer Zahlen. Reihen, für die die Summe der Normen

$$\sum_{k=1}^n \|x_k\|$$

konvergiert, zeichnen sich durch ein besonders gutes Konvergenzverhalten aus. Solche Reihen nennt man *absolut konvergent*. Zunächst machen wir uns klar, daß die absolute Konvergenz auch die normale Konvergenz der Reihe nach sich zieht. Dafür formulieren wir zunächst die CAUCHY-Bedingung für die Reihe der Normen: Für alle $\varepsilon > 0$ gibt es ein $n_\varepsilon \in \mathbb{N}$, so daß für alle $n > m \geq n_\varepsilon$ gilt:

$$\left| \sum_{k=1}^n \|x_k\| - \sum_{k=1}^m \|x_k\| \right| = \sum_{k=1}^n \|x_k\| - \sum_{k=1}^m \|x_k\| = \sum_{k=m+1}^n \|x_k\| < \varepsilon.$$

Eine Anwendung der Dreiecksungleichung zeigt, daß dann auch die CAUCHY-Bedingung für die ursprüngliche Reihe erfüllt ist:

$$\left\| \sum_{k=1}^n x_k - \sum_{k=1}^m x_k \right\| = \left\| \sum_{k=m+1}^n x_k \right\| \leq \sum_{k=m+1}^n \|x_k\| < \varepsilon$$

für alle $n > m \geq n_\varepsilon$.

10.2.2 Satz Eine absolut konvergente Reihe ist konvergent.

Wir werden am Beispiel der LEIBNIZ-Reihen (10.38) konvergente Reihen kennenlernen, die nicht absolut konvergent sind.

Der offensichtliche Vorteil absoluter Konvergenz besteht darin, daß man es nur mit Reihen nicht negativer Zahlen zu tun hat, die meistens viel einfacher zu untersuchen sind, als die eigentliche Reihe. Hauptsächlich liegt das daran, daß die Reihe der Normen $\sum_{k=1}^n \|x_k\|$ insbesondere eine monoton wachsende Folge darstellt und, nach dem Satz 10.1.28 von der monotonen Konvergenz, nur einer oberen Schranke bedarf, um einen Grenzwert zu haben. Meistens bemüht man dazu das CAUCHY-Kriterium nicht direkt, sondern schätzt die Summanden $\|x_k\|$ gegen die Summanden $b_k \geq 0$ einer bekannten, schon konvergenten Reihe $\sum_{k=1}^n b_k$ ab.

10.2.3 Satz (Majorantenkriterium) Lassen sich fast alle Normen $\|x_k\|$ einer Reihe $\sum_{k=1}^n x_k$ gegen die Summanden b_k einer konvergenten Reihe $\sum_{k=1}^n b_k$ abschätzen, d. h., gilt ab einem geeigneten $k_0 \in \mathbb{N}$

$$\|x_k\| \leq b_k,$$

so ist sie absolut konvergent. $\sum_{k=1}^n b_k$ heißt dann (konvergente) Majorante von $\sum_{k=1}^n x_k$.

Beweis. Sei $M := \sum_{k=k_0}^{\infty} b_k$. Dann folgt $\sum_{k=k_0}^n \|x_k\| \leq \sum_{k=k_0}^n b_k \leq \sum_{k=k_0}^{\infty} b_k = M$. Daher ist die monoton wachsende Folge $(\sum_{k=k_0}^n \|x_k\|)_{n \in \mathbb{N}}$ nach oben durch M beschränkt. Nach dem Satz von der monotonen Konvergenz hat sie einen Grenzwert, oder anders gesagt: Die Reihe $\sum_{k=1}^n x_k$ ist absolut konvergent. \square

Um das Majorantenkriterium anwenden zu können, benötigen wir konvergente Reihen zur Abschätzung. Die geometrische Reihe mit $q \in (0, 1)$ wird dafür oft verwendet. Ein anderes Beispiel ist $\sum_{k=1}^n \frac{1}{k^2}$, die von der Reihe $\sum_{k=2}^n \frac{1}{k(k-1)}$ ab $k = 2$ majorisiert wird. Bei dieser Reihe stellt jede Partialsumme eine sogenannte *Teleskop-Summe* dar,

$$\sum_{k=2}^n \frac{1}{k(k-1)} = \sum_{k=2}^n \frac{1}{k-1} - \frac{1}{k} = 1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \frac{1}{3} - \cdots - \frac{1}{n-1} + \frac{1}{n-1} - \frac{1}{n} = 1 - \frac{1}{n},$$

bei der sich, bis auf den ersten und den letzten, alle Summanden herausheben. Diese Reihe positiver Zahlen hat offensichtlich die obere Schranke 1 und ist daher konvergent (mit Grenzwert 1). Laut Majorantenkriterium ist dann auch die Reihe $\sum_{k=1}^n \frac{1}{k^2}$ konvergent und hat einen Grenzwert ≤ 2 .

Nun, da diese Reihe als konvergent identifiziert ist, können wir sie als Majorante für andere Reihen verwenden, etwa für $\sum_{k=1}^n \frac{1}{k^p}$, $p \geq 2$.

Ein weiteres wichtiges Beispiel, das wir für spätere Verwendung untersuchen wollen, ist die sogenannte *Exponentialreihe* (dieser Name wird sich später rechtfertigen, siehe 10.2.24)

$$\sum_{k=0}^n \frac{z^k}{k!}, \quad z \in \mathbb{C}. \quad (10.28)$$

Da wir nach einer Majorante suchen, können wir die Summanden durch ihre Beträge $\frac{r^k}{k!}$ ersetzen ($r := |z|$). Für ein $k_0 \in \mathbb{N}$ mit der Eigenschaft $r < k_0$, $q := \frac{r}{k_0+1} < 1$ und $k \geq k_0 + 1$ gilt

$$\begin{aligned} \frac{r^k}{k!} &= \frac{r}{k} \cdot \frac{r}{k-1} \cdot \frac{r}{k-2} \cdots \frac{r}{k_0+1} \cdot \frac{r}{k_0} \cdot \frac{r}{k_0-1} \cdots \frac{r}{2} \cdot r \\ &< q \cdot q \cdot q \cdots q \cdot \frac{r^{k_0}}{k_0!} = q^{k-k_0} \cdot \frac{r^{k_0}}{k_0!} = q^k \cdot \frac{r^{k_0}}{q^{k_0} k_0!}. \end{aligned}$$

Die mit dem Faktor $\frac{r^{k_0}}{q^{k_0} k_0!}$ multiplizierte geometrische Reihe $\sum_{k=0}^n q^k$ stellt also eine konvergente Majorante für die Exponentialreihe dar. Damit ist deren absolute Konvergenz für jedes $z \in \mathbb{C}$ nachgewiesen.

10.2.4 Satz Für absolut konvergente Reihen $\sum_{k=1}^n a_k$ und $\sum_{k=1}^n b_k$ gelten die Rechenregeln:

$$i) \sum_{k=1}^{\infty} (ta_k + sb_k) = t \sum_{k=1}^{\infty} a_k + s \sum_{k=1}^{\infty} b_k \text{ ist für alle } t, s \in \mathbb{K} \text{ absolut konvergent.}$$

ii) Jede Aufspaltung der Folge (a_k) der Summanden in zwei Teilfolgen (a_{n_k}) und (a_{m_k}) führt zu einer Aufspaltung der Reihe in zwei absolut konvergente Teilreihen:

$$\sum_{k=1}^{\infty} a_k = \sum_{k=1}^{\infty} a_{n_k} + \sum_{k=1}^{\infty} a_{m_k} \quad (10.29)$$

iii) Jede Umordnung $(a_{\rho(n)})$ der Summanden ergibt eine absolut konvergente Reihe, mit demselben Wert wie die Ausgangsreihe:

$$\sum_{k=1}^{\infty} a_{\rho(k)} = \sum_{k=1}^{\infty} a_k \quad (10.30)$$

Dabei ist eine Umordnung durch eine bijektive Abbildung $\rho : \mathbb{N} \rightarrow \mathbb{N}$ der Indizes gegeben.

Eine Aufspaltung in zwei Teilsummen ist meist von der Art, daß man über alle Summanden mit geraden und mit ungeraden Indizes getrennt summieren will. Durch ii) sind aber auch viel allgemeinere Situationen möglich. Z. B. könnte $n_k := 10^k$ sein, also $(a_{n_k}) = (a_{10}, a_{100}, a_{1000}, \dots)$. Die Indexfolge (m_k) der zweiten Teilfolge (a_{m_k}) , die (a_{n_k}) zur ganzen Folge (a_k) ergänzt, ist jetzt nicht ganz leicht durch Formeln wiederzugeben. Aber es ist klar, was sie zu leisten hat. Sie muß nämlich die Lücken zwischen n_k und n_{k+1} auffüllen. Das heißt $m_k = k$, für $k = 1, \dots, 9$. Dann geht es mit den Indizes 11 bis 99 weiter: $m_{10} = 11, m_{11} = 12, \dots, m_{98} = 99$, also $m_k = k + 1$, für $k = 10, \dots, 10^2 - 2$. Anschließend müssen die Zahlen 101, ..., 999 von m_k produziert werden: $m_{99} = 101, m_{100} = 102, \dots, m_{997} = 999$, also $m_k = k + 2$ für $k = 10^2 - 1, \dots, 10^3 - 3$. Wie kann das fortgesetzt werden? Die nächste Zehnerpotenz sollte uns das Bildungsgesetz verraten: $m_{998} = 1001, m_{999} = 1002, \dots, m_{9996} = 9999$, also $m_k = k + 3$ für $k = 10^3 - 2, \dots, 10^4 - 4$. Das bedeutet: $m_k = 10^r + r$ für $k = 10^r - (r-1), \dots, 10^{r+1} - (r+1)$ und $r \in \mathbb{N}_0$. Obwohl wir bei diesem Beispiel eine Formel für die Teilfolge (a_{m_k}) gefunden haben, ist nicht zu erwarten, daß das immer gelingen kann. Wenn die Indexfolge (n_k) aus einer Aufzählung aller Primzahlen besteht, müßte (m_k) durch eine Formel gegeben sein, die alle Nichtprimzahlen in aufsteigender Reihenfolge wiedergibt. Die wird uns wohl so schnell nicht einfallen. Das ist aber auch gar nicht nötig, da die zweite Teilsumme berechnet werden kann, auch wenn die Teilfolge nur implizit über die Bedingung gegeben ist, den Rest der Ausgangsfolge darzustellen.

Beweis. Zu i): Eine einfache Anwendung der Rechenregeln konvergenter Folgen.

Zu ii): Wegen $\sum_{k=1}^p \|a_{n_k}\| \leq \sum_{k=1}^{\infty} \|a_k\|$ und $\sum_{k=1}^p \|a_{m_k}\| \leq \sum_{k=1}^{\infty} \|a_k\|$ sind beide Teilreihen $\sum_{k=1}^p a_{n_k}$ und $\sum_{k=1}^p a_{m_k}$ absolut konvergent. In der Summe $\sum_{k=1}^p a_{n_k} + \sum_{k=1}^p a_{m_k}$ kommen alle Summanden zwischen 1 und $\min\{n_p, m_p\}$ vor und einige weitere, die vom unterschiedlichen Voranschreiten der Teilfolgen stammen (man mache sich das am oben vorgestellten Beispiel klar). Die Differenz $\sum_{k=1}^{\infty} a_k - \sum_{k=1}^p a_{n_k} - \sum_{k=1}^p a_{m_k}$ enthält also weniger

Summanden als $\sum_{k=\min\{n_p, m_p\}}^{\infty} a_k$. Wegen der absoluten Konvergenz der Ausgangsreihe, gibt es zu jedem $\varepsilon > 0$ ein p_ε , so daß für alle $p \geq p_\varepsilon$

$$\left\| \sum_{k=1}^{\infty} a_k - \sum_{k=1}^p a_{n_k} - \sum_{k=1}^p a_{m_k} \right\| \leq \sum_{k=\min\{n_p, m_p\}}^{\infty} \|a_k\| \leq \sum_{k=p}^{\infty} \|a_k\| < \varepsilon$$

gilt. Dabei ist $n_p \geq p$ und $m_p \geq p$ der Grund für die letzte Ungleichung. Das zeigt (10.29).

Zu iii): Die Reihe $\sum_{k=1}^n a_{\rho(k)}$ ist absolut konvergent, da $\sum_{k=1}^{\infty} \|a_k\|$ eine obere Schranke für $\sum_{k=1}^n \|a_{\rho(k)}\|$ darstellt. Wegen der Bijektivität von ρ gibt es für jedes $p \in \mathbb{N}$ ein $n_p \in \mathbb{N}$, so daß für alle $n \geq n_p$ die Inklusion $\{1, \dots, p\} \subseteq \{\rho(1), \dots, \rho(n)\}$ gilt. Daher enthält $\sum_{k=1}^{\infty} a_k - \sum_{k=1}^n a_{\rho(k)}$ weniger Summanden als $\sum_{k=p}^{\infty} a_k$. Für jedes $\varepsilon > 0$ wählen wir jetzt ein p_ε , so daß für alle $p \geq p_\varepsilon$ die Abschätzung $\sum_{k=p}^{\infty} \|a_k\| < \varepsilon$ erfüllt ist. Dann haben wir für alle $n \geq n_{p_\varepsilon}$

$$\left\| \sum_{k=1}^{\infty} a_k - \sum_{k=1}^n a_{\rho(k)} \right\| \leq \sum_{k=p_\varepsilon}^{\infty} \|a_k\| < \varepsilon.$$

Das zeigt (10.30). □

10.2.5 Satz (Dezimalentwicklung) Für jede positive reelle Zahl x gibt es eine Folge $(x_k)_{k \in \mathbb{N}_0}$ aus \mathbb{N}_0 , so daß $x_k \in \{0, 1, 2, \dots, 9\}$ für alle $k \in \mathbb{N}$ und

$$x = \sum_{k=0}^{\infty} \frac{x_k}{10^k} \tag{10.31}$$

gilt. Diese Reihenentwicklung wird Dezimalentwicklung von x genannt. Statt (10.31) schreibt man üblicherweise $x = x_0 . x_1 x_2 x_3 x_4 x_5 \dots$. Die Entwicklung heißt abbrechend, wenn es ein kleinstes $n \in \mathbb{N}$ gibt, so daß $x_k = 0$ für alle $k > n$ gilt. Dafür schreibt man $x = x_0 . x_1 x_2 \dots x_{n-1} x_n$.

Die Dezimalentwicklung einer Zahl ist genau dann nicht eindeutig, wenn sie abbrechend ist: $x = x_0 . x_1 x_2 \dots x_{n-1} x_n$. In diesem Fall ist auch $x = x_0 . x_1 x_2 \dots x_{n-1} (x_n - 1) 999 \dots$ eine Entwicklung von x .

Eine Zahl x ist genau dann rational, wenn ihre Dezimalentwicklung abbrechend, oder periodisch ist. Letzteres bedeutet, daß es eine Stelle $n \in \mathbb{N}_0$ gibt, ab der die Dezimalentwicklung aus der endlosen Wiederholung einer endlichen Ziffernfolge $x_{n+1} \dots x_{n+p}$ besteht:

$x = x_0 . x_1 \dots x_n x_{n+1} \dots x_{n+p} x_{n+1} \dots x_{n+p} \dots$ Dafür ist die folgende Schreibweise vorgesehen:

$$x_0 . x_1 \dots x_n \overline{x_{n+1} \dots x_{n+p}}. \tag{10.32}$$

Beweis. Zunächst überlegen wir uns, daß jede Dezimalentwicklung konvergiert. Wir können o. B. d. A. $x_0 = 0$ annehmen. Da alle Summanden in $\sum_{k=1}^n \frac{x_k}{10^k}$ positiv sind, brauchen wir nur eine konvergente Majorante. Die ist schnell gefunden: $\sum_{k=1}^n \frac{x_k}{10^k} \leq \sum_{k=1}^n \frac{9}{10^k} \leq 9 \sum_{k=1}^{\infty} \left(\frac{1}{10}\right)^k = 9 \frac{\frac{1}{10}}{1 - \frac{1}{10}} = 1$. Nach kurzem Nachdenken erkennt man, daß dieses Ergebnis zu erwarten war, denn eine Dezimalentwicklung $0.x_1 x_2 x_3 \dots$ sollte natürlich eine Zahl ergeben, die nicht größer als 1 ist.

Nun zur Existenz der Entwicklung: Dafür brauchen wir die sogenannte *GAUSS-Klammer* $\lfloor y \rfloor$, die für eine reelle Zahl y als die größte ganze Zahl definiert ist, die kleiner als y oder gleich y ist: $\lfloor y \rfloor := \max\{n \in \mathbb{Z} \mid n \leq y\}$. Um die Idee hinter der Konstruktion der Dezimalentwicklung zu verstehen, führen wir die ersten Schritte an dem konkreten Zahlenbeispiel $x = \pi$ vor. Hier ist natürlich $x_0 = 3, x_1 = 1, x_2 = 4, x_3 = 1, x_4 = 5$, etc. x_0 erhalten wir einfach durch $\lfloor \pi \rfloor = 3$. Für die erste Nachkommastelle müssen wir x_0 von π abziehen, um die Zahl $\tilde{x}_1 := \pi - x_0 = 0.1415 \dots < 1$ zu erhalten. Jetzt können wir die erste Nachkommastelle $x_1 = 1$ über $\lfloor 10\tilde{x}_1 \rfloor = \lfloor 1.415 \dots \rfloor = 1$ gewinnen. Das wiederholen wir für die zweite Nachkommastelle: $\tilde{x}_2 := \tilde{x}_1 - \frac{x_1}{10} = \pi - x_0 - \frac{x_1}{10} = 0.0415 \dots$ führt auf $x_2 := \lfloor 100\tilde{x}_2 \rfloor = \lfloor 4.15 \dots \rfloor = 4$.

Das sollte reichen, um das allgemeine Verfahren zu formulieren:

0. $x_0 := \lfloor x \rfloor$, dann gilt $0 \leq \tilde{x}_1 := x - x_0 < 1$.
1. $x_1 := \lfloor 10\tilde{x}_1 \rfloor$. Wir erhalten $0 \leq 10\tilde{x}_1 - x_1 < 1$, also $0 \leq \tilde{x}_1 - \frac{x_1}{10} < \frac{1}{10}$, oder ausführlich $0 \leq \tilde{x}_2 := x - x_0 - \frac{x_1}{10} < \frac{1}{10}$.
2. $x_2 := \lfloor 10^2 \tilde{x}_2 \rfloor$, mit $0 \leq 10^2 \tilde{x}_2 - x_2 < 1$, also

$$0 \leq \tilde{x}_3 := \tilde{x}_2 - \frac{x_2}{10^2} = x - x_0 - \frac{x_1}{10} - \frac{x_2}{10^2} = x - \sum_{k=0}^2 \frac{x_k}{10^k} < \frac{1}{10^2}.$$

- n. Wir gehen davon aus, daß \tilde{x}_{n-1} bereits konstruiert und die Abschätzung $\tilde{x}_n := x - \sum_{k=0}^{n-1} \frac{x_k}{10^k} < \frac{1}{10^{n-1}}$ erfüllt ist. Dann definieren wir $x_n := \lfloor 10^n \tilde{x}_n \rfloor$ und erhalten

$$0 \leq 10^n \tilde{x}_n - x_n < 1, \text{ also}$$

$$0 \leq \tilde{x}_{n+1} := \tilde{x}_n - \frac{x_n}{10^n} = x - \sum_{k=0}^{n-1} \frac{x_k}{10^k} - \frac{x_n}{10^n} = x - \sum_{k=0}^n \frac{x_k}{10^k} < \frac{1}{10^n}.$$

Damit ist die Dezimalentwicklung von x induktiv konstruiert worden. Die letzte Abschätzung zeigt, daß die Reihe $\sum_{k=0}^n \frac{x_k}{10^k}$ gegen x konvergiert.

Zur Eindeutigkeit der Entwicklung: Es seien $x = \sum_{k=0}^{\infty} \frac{x_k}{10^k}$ und $y = \sum_{k=0}^{\infty} \frac{y_k}{10^k}$ zwei Dezimalentwicklungen von x und n die Position, an der sie sich das erste mal unterscheiden. Wir können o. B. d. A. $x_n > y_n$ annehmen. Dann gilt folgende Kette von Abschätzungen

$$\frac{1}{10^n} \leq \frac{x_n - y_n}{10^n} = \sum_{k=n+1}^{\infty} \frac{y_k - x_k}{10^k} \leq \sum_{k=n+1}^{\infty} \frac{9}{10^k} = \frac{9}{10^{n+1}} \sum_{\ell=0}^{\infty} \frac{1}{10^\ell} = \frac{9}{10^{n+1}} \cdot \frac{10}{9} = \frac{1}{10^n}.$$

Würde für irgendein $k \geq n + 1$ die Ungleichung $y_k - x_k < 9$ gelten, so müßte in obiger Abschätzung das zweite \leq durch $<$ ersetzt werden, was sofort einen Widerspruch ergäbe. Also gilt $y_k - x_k = 9$, d. h. $y_k = 9$ und $x_k = 0$ für alle $k \geq n + 1$. Genauso folgt auch $y_n = x_n - 1$. Damit ist die eine Dezimalentwicklung von x abbrechend und die andere periodisch mit einer um 1 kleineren Ziffer an der Position n : $x = x_0 . x_1 \dots x_{n-1} x_n = x_0 . x_1 \dots x_{n-1} (x_n - 1)\bar{9}$. Man kann mit Hilfe der geometrischen Reihe auch leicht die Umkehrung zeigen, daß also $x_0 . x_1 \dots x_{n-1} (x_n - 1)\bar{9}$ gleich $x_0 . x_1 \dots x_{n-1} x_n$ ist.

Als nächstes beweisen wir, daß jede periodische Dezimalentwicklung eine rationale Zahl darstellt:

$$x_0 . x_1 x_2 \dots x_n \overline{x_{n+1} \dots x_{n+p}} = \sum_{k=0}^n \frac{x_k}{10^k} + 0.00 \dots 0 \overline{x_{n+1} \dots x_{n+p}}$$

$$\begin{aligned}
&= \frac{1}{10^n} \sum_{k=0}^n 10^{n-k} x_k + \frac{1}{10^n} \sum_{k=1}^p 10^{n+p-k} x_{n+k} \cdot 0 \cdot \underbrace{00\dots01}_{p \text{ Stellen}} \\
&= \frac{1}{10^n} \sum_{k=0}^n 10^{n-k} x_k + \frac{1}{10^n} \sum_{k=1}^p 10^{n+p-k} x_{n+k} \cdot \sum_{\ell=1}^{\infty} \left(\frac{1}{10^p} \right)^\ell \\
&= \frac{1}{10^n} \sum_{k=0}^n 10^{n-k} x_k + \frac{1}{10^n} \sum_{k=1}^p 10^{n+p-k} x_{n+k} \cdot \frac{1}{10^p - 1} \in \mathbb{Q}.
\end{aligned}$$

Für die Umkehrung müssen wir klären, was es bedeutet, die Dezimalentwicklung eines Bruchs $\frac{p}{q}$ zu berechnen. Das Verfahren beruht auf dem *Teilen mit Rest*, das wir in 3.1.2 kennengelernt haben. Es gilt $p = x_0 q + r_0$, mit $0 \leq x_0 \leq p$ und $0 \leq r_0 < q$. x_0 bestimmt die Stellen vor dem Dezimalpunkt. Falls $r_0 = 0$ gelten sollte, sind wir fertig. Andernfalls teilen wir $10r_0$ durch q : $10r_0 = x_1 q + r_1$. Wegen $x_1 q + r_1 = 10r_0 < 10q$ gilt $x_1 q < 10q - r_1 \leq 10q$, d.h. $x_1 \leq 9$. Auf diese Weise fahren wir fort. Für die Zahl $\frac{p}{q}$ bedeutet das

$$\begin{aligned}
\frac{p}{q} &= \frac{x_0 q + r_0}{q} = x_0 + \frac{1}{10} \frac{10r_0}{q} = x_0 + \frac{1}{10} \frac{x_1 q + r_1}{q} = x_0 + \frac{x_1}{10} + \frac{1}{100} \frac{10r_1}{q} \\
&= x_0 + \frac{x_1}{10} + \frac{1}{100} \frac{x_2 q + r_2}{q} = x_0 + \frac{x_1}{10} + \frac{x_2}{100} + \frac{1}{10^3} \frac{10r_2}{q} = \dots
\end{aligned}$$

Wenn das Verfahren abbricht, weil ein Rest r_n verschwindet, dann ist die Dezimalentwicklung $x_0 \cdot x_1 x_2 \dots x_n$. Andernfalls kann man es unbegrenzt fortsetzen. Allerdings fängt es an sich zu wiederholen, sobald einer der Reste das erste Mal erneut auftritt. Da diese bei nicht abbrechender Entwicklung zwischen 1 und $q - 1$ liegen, muß spätestens nach $q - 1$ Schritten einer ein weiteres Mal vorkommen. Damit ist die Dezimalentwicklung von $\frac{p}{q}$ in diesem Fall periodisch. \square

Es ist ganz instruktiv, die Dezimalentwicklung eines Bruches an einem konkreten Zahlenbeispiel durchzuführen – und wenn wir schon mal dabei sind, stellen wir dieser Rechnung das übliche Rechenverfahren mit Papier und Bleistift gegenüber, das man in der Grundschule lernt, um es in der Schule mit Hilfe des Taschenrechners wieder zu vergessen.

$$\begin{aligned}
\frac{64}{27} &= \frac{2 \cdot 27 + 10}{27} = 2 + \frac{1}{10} \frac{100}{27} & 64 & : 27 = 2.\overline{370} \\
&= 2 + \frac{1}{10} \frac{3 \cdot 27 + 19}{27} = 2 + \frac{3}{10} + \frac{1}{100} \frac{190}{27} & 54 \\
&= 2.3 + \frac{1}{100} \frac{7 \cdot 27 + 1}{27} = 2.3 + \frac{7}{100} + \frac{1}{1000} \frac{10}{27} & 81 \\
&= 2.37 + \frac{1}{10^3} \frac{0 \cdot 27 + 10}{27} = 2.37 + \frac{0}{10^3} + \frac{1}{10000} \frac{100}{27} & 190 \\
&= 2.370 + \frac{1}{10^4} \frac{3 \cdot 27 + 19}{27} = \dots & 189 \\
&&& \vdots
\end{aligned}$$

Man kann das noch einmal an der Zahl $\frac{1}{81}$ üben. Auch die Umkehrung sollte man an einem Beispiel gesehen haben. Natürlich ist der Bruch zu, sagen wir 1.234, schnell gefunden: $1.234 =$

$\frac{1234}{1000} = \frac{617}{500}$. Interessanter ist es, eine periodische Entwicklung in den zugehörigen Bruch zu verwandeln:

$$\begin{aligned} 0.\overline{31707} &= 31707 \cdot 0.\overline{00001} = 31707 \cdot 0.00001\ 00001\ 00001\ 00001\dots \\ &= 31707 \cdot (0.00001 + 0.00000\ 00001 + 0.00000\ 00000\ 00001 + \dots) \\ &= 31707 \cdot (10^{-5} + 10^{-10} + 10^{-15} + \dots) = 31707 \cdot \sum_{k=1}^{\infty} (10^{-5})^k \\ &= 31707 \cdot \frac{10^{-5}}{1 - 10^{-5}} = \frac{31707}{10^5 - 1} = \frac{31707}{99999} = \frac{13}{41}. \end{aligned}$$

Hat man das soweit verstanden, dann wird man für $0.\overline{216}$ diese Rechnung nicht noch einmal durchführen, sondern gleich $0.\overline{216} = \frac{216}{999} = \frac{8}{37}$ finden. Ein letztes Beispiel:

$$0.56\overline{1956} = \frac{56}{100} + \frac{1}{100} 0.\overline{1956} = \frac{56}{100} + \frac{1}{100} \frac{1956}{9999} = \frac{561900}{100 \cdot 9999} = \frac{1873}{3333} = 0.\overline{5619}.$$

10.2.6 Definition Eine Reihe der Form

$$\sum_{k=0}^n (-1)^k a_k \quad (10.33)$$

heißt alternierend, wenn die Folge $(a_k)_{k \in \mathbb{N}}$ aus positiven Zahlen besteht. Ist diese Folge eine monoton Nullfolge, so wird (10.33) als LEIBNIZ-Reihe bezeichnet.

10.2.7 Satz Jede LEIBNIZ-Reihe $S_n = \sum_{k=0}^n (-1)^k a_k$ ist konvergent. Die Folge der Partialsummen (S_{2n}) bzw. (S_{2n+1}) ist monoton fallend bzw. monoton steigend. Es gilt für alle $n \geq 0$

$$a_{2n} \leq S_{2n}, \quad (10.34)$$

$$0 \leq S_{2n+1} \leq a_0, \quad (10.35)$$

$$S_{2n+1} \leq S \leq S_{2n} \quad (10.36)$$

$$|S - S_n| \leq a_n \quad (10.37)$$

Dabei ist $S := \sum_{k=0}^{\infty} (-1)^k a_k$ der Grenzwert der Reihe.

Beweis. Der Beweis besteht im Wesentlichen daraus, die Summanden in S_{2n} bzw. S_{2n+1} auf verschiedene Weise zusammenzufassen.

$$S_{2n} = (a_0 - a_1) + (a_2 - a_3) + \dots + (a_{2n-2} - a_{2n-1}) + a_{2n} \geq a_{2n} \quad (i)$$

$$\begin{aligned} S_{2n} &= a_0 - a_1 + a_2 - a_3 + \dots + a_{2n-2} - (a_{2n-1} - a_{2n}) \\ &= S_{2n-2} - (a_{2n-1} - a_{2n}) \leq S_{2(n-1)}. \end{aligned} \quad (ii)$$

(i) und (ii) ergeben (10.34). Nach dem Satz von der monotonen Konvergenz ist die Folge (S_{2n}) konvergent, mit einem Grenzwert \bar{S} .

Dieselben Überlegungen für S_{2n+1} :

$$S_{2n+1} = a_0 - (a_1 - a_2) - (a_3 - a_4) - \dots - (a_{2n-1} - a_{2n}) - a_{2n+1} \leq a_0, \quad (iii)$$

$$S_{2n+1} = (a_0 - a_1) + (a_2 - a_3) + \cdots + (a_{2n-2} - a_{2n-1}) + (a_{2n} - a_{2n+1}) \geq 0, \quad (\text{iv})$$

$$S_{2n+1} = S_{2n-1} + (a_{2n} - a_{2n+1}) \geq S_{2n-1}, \quad (\text{v})$$

$$S_{2n+1} = S_{2n} - a_{2n+1} \leq S_{2n}. \quad (\text{vi})$$

(iii), (vi) und (v) ergeben (10.35). Nach dem Satz von der monotonen Konvergenz gibt es auch einen Grenzwert \underline{S} , gegen den (S_{2n+1}) von unten konvergiert. Aus (vi) erhalten wir in der Grenze $n \rightarrow \infty$ die Abschätzung $\underline{S} \leq \bar{S}$. Aber $0 \leq S_{2n} - S_{2n+1} = a_{2n+1}$ zeigt, daß beide Folgen gegen denselben Grenzwert konvergieren und zwar S_{2n} von oben und S_{2n+1} von unten: $S := \underline{S} = \bar{S}$. Damit ist (10.36) bewiesen. Schließlich (10.37):

$$|S_{2n} - S| = S_{2n} - S \leq S_{2n} - S_{2n+1} = a_{2n+1} \leq a_{2n},$$

$$|S_{2n+1} - S| = S - S_{2n+1} \leq S_{2n} - S_{2n+1} = a_{2n+1}.$$

In jedem Fall erhalten wir $|S_n - S| \leq a_n$. □

Ein wichtiges Beispiel ist die (eigentliche) LEIBNIZ-Reihe

$$\sum_{k=0}^n (-1)^k \frac{1}{k+1} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} - \cdots + (-1)^n \frac{1}{n+1}. \quad (10.38)$$

Sie erfüllt ganz augenscheinlich die Bedingungen obigen Satzes. Da die Summe der Beträge dieser Reihe aber die harmonische Reihe ergibt (Seite 210), kann sie nicht absolut konvergent sein. An diesem einfachen Beispiel kann man den Mechanismus studieren, der eine solche Reihe zum Konvergieren bringt.

10.2.8 A Untersuchen Sie die Teilfolge $\left(\sum_{k=0}^{2n-1} (-1)^k \frac{1}{k+1} \right)_{n \in \mathbb{N}}$ der LEIBNIZ-Reihe, indem Sie die Summanden paarweise auswerten. Welche Reihe erhalten Sie und warum kann man deren Konvergenz erkennen?

Aber natürlich gibt es auch LEIBNIZ-Reihen, die absolut konvergent sind, so wie etwa $\sum_{k=0}^n (-1)^k \frac{1}{(k+1)^2}$.

10.2.9 Definition Es sei $(x_n)_{n \in \mathbb{N}}$ eine Folge in \mathbb{R} . Ist sie nach oben beschränkt, dann wird, falls sie Häufungspunkte hat, der größte als Limes Superior $\limsup_{n \rightarrow \infty} x_n$ bzw., falls sie nach unten beschränkt ist, der kleinste als Limes Inferior $\liminf_{n \rightarrow \infty} x_n$ bezeichnet. Die Schreibweisen $\overline{\lim}_{n \rightarrow \infty} x_n$ und $\underline{\lim}_{n \rightarrow \infty} x_n$ sind ebenfalls gebräuchlich. Sollte die Folge nach oben, bzw. unten unbeschränkt sein, so gilt der Limes Superior bzw. Inferior als nicht existent.

Zunächst einmal muß eine nach oben beschränkte Folge nicht unbedingt einen Häufungspunkt haben. Sie könnte nach unten unbeschränkt und monoton fallend sein, wie etwa die (uninteressante) Folge $(-n)_{n \in \mathbb{N}}$. Hat sie allerdings Häufungspunkte, so müssen diese ebenfalls nach oben beschränkt sein und daher ein Supremum s haben (die Menge der Häufungspunkte könnten dabei nach unten durchaus unbeschränkt sein, wie das Beispiel der rationalen Zahlen ≤ 1 zeigt).

Wir wollen uns überlegen, daß s selbst wieder ein Häufungspunkt der Folge ist und damit das Maximum aller Häufungspunkte darstellt: In jeder ε -Umgebung $(s - \varepsilon, s + \varepsilon)$ von s gibt es wenigstens einen Häufungspunkt x von $(x_n)_{n \in \mathbb{N}}$. Falls $x = s$ gelten sollte, sind wir fertig, andernfalls gibt es in jeder noch so kleinen δ -Umgebung $(x - \delta, x + \delta)$ von x unendlich viele Folgenglieder. Wir wählen δ so klein, daß $(x - \delta, x + \delta)$ in $(s - \varepsilon, s + \varepsilon)$ enthalten ist. Damit enthält letztere unendlich viele Folgenglieder, und da ε beliebig war, enthält jede ε -Umgebung von s unendlich viele Folgenglieder. s ist also ebenfalls ein Häufungspunkt von $(x_n)_{n \in \mathbb{N}}$. Dieselben Überlegungen kann man für den kleinsten Häufungspunkt anstellen. Damit ist die Definition 10.2.9 nicht leer.

Bemerkung: Man kann den Limes Superior und den Limes Inferior durch Formeln wiedergeben: $\overline{\lim}_{n \rightarrow \infty} x_n = \inf \{ \sup \{ x_k \mid k \geq n \} \mid n \geq 0 \}$ und entsprechend für den Limes Inferior $\underline{\lim}_{n \rightarrow \infty} x_n = \sup \{ \inf \{ x_k \mid k \geq n \} \mid n \geq 0 \}$. Da wir dafür keine direkte Anwendung haben, wollen wir sie nicht beweisen.

10.2.10 Beispiel Die Folge $(\sqrt[n]{(-4)^n + 0.5((-1)^n + 1)6^n})$ hat den Limes Superior 6 und den Limes Inferior -4: Für ungerades n sind die Folgenglieder einfach durch $\sqrt[n]{-4^n} = -4$ gegeben und für gerades durch $\sqrt[n]{4^n + 6^n} = 6\sqrt[n]{(\frac{2}{3})^n + 1}$. Nach Korollar 10.1.18 konvergiert das gegen 6. Somit gibt es genau die zwei Häufungspunkte -4 und 6.

Damit haben wir die Vorarbeiten abgeschlossen, um das folgende hinreichende Kriterium für die absolute Konvergenz einer Reihe zu formulieren.

10.2.11 Satz (Quotienten-/Wurzelkriterium) *Die Reihe $\sum_{k=0}^n a_k$ ist absolut konvergent, falls eine der folgenden Bedingungen erfüllt ist:*

$$\overline{\lim}_{n \rightarrow \infty} \frac{\|a_{n+1}\|}{\|a_n\|} = q < 1, \quad (\text{Quotientenkriterium}) \quad (10.39)$$

$$\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{\|a_n\|} = q < 1. \quad (\text{Wurzelkriterium}) \quad (10.40)$$

Die Reihe ist sicher nicht konvergent, falls eine der folgenden Ungleichungen gelten sollte:

$$\underline{\lim}_{n \rightarrow \infty} \frac{\|a_{n+1}\|}{\|a_n\|} = q > 1, \quad (10.41)$$

$$\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{\|a_n\|} = q > 1. \quad (10.42)$$

Beweis. Zu (10.40): Wir wählen ein $\varepsilon > 0$, so daß $\tilde{q} := q + \varepsilon < 1$ gilt. Dann liegen ab einem geeigneten n_ε alle Folgenglieder $\sqrt[n]{\|a_n\|}$ unterhalb \tilde{q} , denn gäbe es oberhalb von \tilde{q} unendlich viele, so müßte dort auch noch ein Häufungspunkt zu finden sein, der dann größer als der größte Häufungspunkt q wäre. Wir haben also $\sqrt[n]{\|a_n\|} \leq \tilde{q}$, oder $\|a_n\| \leq \tilde{q}^n$ für alle $n \geq n_\varepsilon$. Das bedeutet, daß die geometrische Reihe zu \tilde{q} eine Majorante für $\sum_{k=1}^n a_k$ darstellt. Diese Reihe ist somit absolut konvergent. Für $q > 1$ wählen wir ein $\varepsilon > 0$, so daß $\tilde{q} := q - \varepsilon > 1$ gilt. Da q ein Häufungspunkt ist, müssen unendlich viele Folgenglieder in der Umgebung $(q - \varepsilon, q + \varepsilon)$ von q liegen. Unendlich oft gilt demnach $\sqrt[n]{\|a_n\|} \geq \tilde{q}$, oder $\|a_n\| \geq \tilde{q}^n$. Daher ist $(a_n)_{n \in \mathbb{N}}$ keine Nullfolge und die Reihe nicht konvergent.

Zu (10.39): Wir wählen ein $\varepsilon > 0$ in der Weise, daß $\tilde{q} := q + \varepsilon < 1$ gilt. Dann liegen ab einem geeigneten n_ε alle Folgenglieder $\frac{\|a_{n+1}\|}{\|a_n\|}$ unterhalb \tilde{q} (mit derselben Begründung, wie oben). Für alle $n \geq n_\varepsilon$ haben wir demnach

$$\frac{\|a_{n+1}\|}{\|a_n\|} \leq \tilde{q}, \text{ also } \|a_{n+1}\| \leq \tilde{q} \|a_n\|.$$

Diese Ungleichung lässt sich rückwärts entwickeln:

$$\|a_{n+1}\| \leq \tilde{q} \|a_n\| \leq \tilde{q}^2 \|a_{n-1}\| \leq \dots \leq \tilde{q}^{n+1-n_\varepsilon} \|a_{n_\varepsilon}\| = \tilde{q}^{n+1} \tilde{q}^{-n_\varepsilon} \|a_{n_\varepsilon}\|.$$

Also werden die Summanden die Reihe $\sum_{k=1}^n \|a_k\|$ ab $k = n_\varepsilon + 1$ durch ein Vielfaches der konvergenten geometrischen Reihe $\sum_{k=1}^n \tilde{q}^k$ majorisiert. Nach dem Majorantenkriterium 10.2.3 ist die Reihe absolut konvergent.

Sollte (10.41) gelten, dann lässt sich ein $\varepsilon > 0$ mit $\tilde{q} := q - \varepsilon > 1$ wählen. Jetzt müssen fast alle Quotienten oberhalb von \tilde{q} liegen, denn andernfalls gäbe es noch einen Häufungspunkt $\leq \tilde{q} < q$, im Widerspruch dazu, daß q der kleinste Häufungspunkt aller Quotienten $\frac{\|a_{n+1}\|}{\|a_n\|}$ ist. Stellen wir nun dieselben Überlegungen wie oben unter diesem neuen Gesichtspunkt an, dann erhalten wir $\|a_{n+1}\| \geq \tilde{q}^{n+1} \tilde{q}^{-n_\varepsilon} \|a_{n_\varepsilon}\|$. Da $\tilde{q} > 1$ ist, konvergiert dieser Ausdruck nicht gegen Null, so daß die Minimalvoraussetzung für die Konvergenz einer Reihe, nämlich daß ihre Summanden wenigstens eine Nullfolge bilden, nicht erfüllt ist. Die Reihe ist demnach nicht konvergent. Man mache sich klar, daß dieser Schluß mit dem größten Häufungspunkt q_1 nicht funktioniert, da oberhalb von $q_1 - \varepsilon$ zwar unendlich viele Quotienten liegen müssen, aber im Allgemeinen nicht fast alle, was für den Beweis zwingend gebraucht wurde (vergl. auch Beispiel 10.2.12, vii). \square

Das Quotienten- und das Wurzelkriterium sind hinreichend für absolute Konvergenz. Wie der Beweis zeigt, bedeuten sie ein Konvergenzverhalten, das mindestens so gut wie das einer geometrischen Reihe ist. Das ist natürlich nicht bei allen konvergenten Reihen der Fall. Trotzdem stellt die Anwendung dieser Kriterien oft den ersten Versuch dar, eine gegebene Reihe auf Konvergenz hin zu überprüfen. Verläuft der nicht erfolgreich (z. B., wenn man einen Grenzwert $q = 1$ erhält), dann muß man sich eben nach einer schwächeren Majorante umsehen, oder die Reihe nach allen Regeln der Kunst untersuchen. Man kann jedenfalls aus dem Versagen der beiden Kriterien nicht auf die Divergenz der Reihe schließen.

Auffallend ist die Abweichung des Quotientenkriteriums gegenüber dem Wurzelkriterium, was die Divergenz der Reihe angeht. Man darf dabei nicht aus den Augen verlieren, daß es sich auch hier um ein hinreichendes Kriterium handelt. Das heißt, die Folge kann auch divergieren, falls diese Bedingung nicht erfüllt ist. Wir untersuchen einige Beispiele, um auszuloten, was alles möglich ist.

10.2.12 Beispiel

i) $\sum_{k=0}^n \frac{z^k}{k!}$: Das Quotienten-Kriterium ergibt $q = 0$, denn

$$\frac{\frac{|z|^{k+1}}{(k+1)!}}{\frac{|z|^k}{k!}} = |z| \frac{k!}{(k+1)!} = \frac{|z|}{k+1} \xrightarrow{k \rightarrow \infty} 0.$$

Damit ist die Exponentialreihe für jedes $z \in \mathbb{C}$ absolut konvergent.

ii) $\sum_{k=1}^n \frac{1}{k^2}$: Das Quotienten- und das Wurzelkriterium versagen, obwohl die Reihe konvergiert:

$$\frac{k^2}{(k+1)^2} = \left(\frac{k}{k+1}\right)^2 \xrightarrow{k \rightarrow \infty} 1, \quad \sqrt[k]{\frac{1}{k^2}} = \frac{1}{\sqrt[k]{k^2}} \xrightarrow{k \rightarrow \infty} 1.$$

iii) $\sum_{k=0}^n a_k$, mit $a_k = \frac{1}{3^k}$ für gerade und $a_k = \frac{1}{2^k}$ für ungerade k . Das Quotientenkriterium versagt, da die Folge $\frac{a_{k+1}}{a_k}$ unbeschränkt ist. Für gerades k ist dieser Bruch nämlich $\frac{1}{2} \left(\frac{3}{2}\right)^k$ und für ungerades $\frac{1}{3} \left(\frac{2}{3}\right)^k$. Das Wurzelkriterium:

$$\sqrt[k]{a_k} = \begin{cases} \frac{1}{3}, & k \text{ gerade} \\ \frac{1}{2}, & k \text{ ungerade} \end{cases}$$

Der größte Häufungspunkt dieser Folge ist offensichtlich $\frac{1}{2}$. Das Wurzelkriterium ist erfüllt und die Reihe konvergiert.

iv) $\sum_{k=1}^n \frac{k^k}{k!}$ ist divergent, denn das Quotientenkriterium ergibt $q = e > 1$:

$$\frac{(k+1)^{k+1} k!}{(k+1)! k^k} = \frac{(k+1)^{k+1}}{(k+1) k^k} = \left(\frac{k+1}{k}\right)^k = \left(1 + \frac{1}{k}\right)^k \xrightarrow{k \rightarrow \infty} e.$$

v) $\sum_{k=1}^n \frac{k!}{k^k}$ ist konvergent, denn das Quotientenkriterium ist mit $q = e^{-1}$ erfüllt:

$$\frac{(k+1)! k^k}{(k+1)^{k+1} k!} = \frac{(k+1) k^k}{(k+1)^{k+1}} = \left(\frac{k}{k+1}\right)^k = \frac{1}{\left(1 + \frac{1}{k}\right)^k} \xrightarrow{k \rightarrow \infty} \frac{1}{e} < 1.$$

vi) $\sum_{k=1}^n \frac{z^k}{k}$, $z \in \mathbb{C}$. Wir untersuchen, für welche z die Reihe absolut konvergiert:

$$\sqrt[k]{\frac{1}{k} |z|^k} = \frac{|z|}{\sqrt[k]{k}} \xrightarrow{k \rightarrow \infty} |z|.$$

Das Wurzelkriterium ist also für $|z| < 1$ erfüllt, die Reihe damit absolut konvergent. Für $|z| > 1$ erhalten wir aber ebenfalls eine Antwort, nämlich, daß die Reihe sicher nicht konvergiert. Für $|z| = 1$ erfahren wir mit dem Wurzelkriterium nichts. Wir wissen aber, daß die Reihe noch für $z = -1$ konvergiert (allerdings nicht absolut) und für $z = 1$ divergiert. Eine erschöpfende Behandlung findet man im Beispiel 11.5.7.

vii) Die triviale Reihe $\sum_{k=1}^n a_k$, mit $a_k = 1$ für gerades k und $a_k = 2$ für ungerades k , ist natürlich divergent. $\frac{a_{k+1}}{a_k} = 2$ für gerades k und $\frac{a_{k+1}}{a_k} = \frac{1}{2}$ für ungerades, zeigt $\lim_{k \rightarrow \infty} \frac{a_{k+1}}{a_k} = \frac{1}{2}$ und $\overline{\lim}_{k \rightarrow \infty} \frac{a_{k+1}}{a_k} = 2$. D. h., weder (10.39) noch (10.41) läßt sich anwenden, so daß das Quotientenkriterium die Reihe nicht als divergent identifizieren kann. Wegen $\overline{\lim}_{k \rightarrow \infty} \sqrt[k]{a_k} = 1$ versagt allerdings auch das Wurzelkriterium. Natürlich benötigt diese Reihe keine Untersuchung. Sie soll auch nur demonstrieren, daß dieses Kriterium, wenn es um die Divergenz der Reihe geht, oft von begrenztem Nutzen ist. Meistens muß die Folge (a_k) bereits gut sichtbar divergieren, damit (10.41) oder (10.42) erfüllt ist.

10.2.13 Potenzreihen

10.2.14 Definition Sei (a_k) eine reelle oder komplexe Folge. Für ein festes $z_0 \in \mathbb{R} (\mathbb{C})$ und jedes $z \in \mathbb{R} (\mathbb{C})$ wird durch $\sum_{k=0}^n a_k(z - z_0)^k$ eine Reihe definiert. Der Konvergenzbereich $U(z_0)$ dieser Reihe ist die Menge aller z , für die sie konvergiert. Die Funktion f auf $U(z_0)$, die durch

$$f(z) := \sum_{k=0}^{\infty} a_k(z - z_0)^k \quad (10.43)$$

definiert wird, heißt Potenzreihe. z_0 wird als Entwicklungspunkt von f bezeichnet. Wir sagen, f ist um z_0 herum in eine Potenzreihe entwickelt.

10.2.15 Satz Für jede Potenzreihe f gibt es entweder eine Zahl $R \geq 0$, den sogenannten Konvergenzradius, so daß für den Konvergenzbereich $\{z \mid |z - z_0| < R\} \subseteq U(z_0) \subseteq \{z \mid |z - z_0| \leq R\}$ gilt, oder $U(z_0) = \mathbb{R} (\mathbb{C})$. R läßt sich durch

$$R = \frac{1}{\limsup_{k \rightarrow \infty} \sqrt[k]{|a_k|}}. \quad (10.44)$$

bestimmen, falls der \limsup existiert und einen Wert aus $(0, \infty)$ liefert. Sollte er die Zahl 0 ergeben, so ist $U(z_0) = \mathbb{R} (\mathbb{C})$. Falls der \limsup nicht existiert, weil die Folge $(\sqrt[k]{|a_k|})$ unbeschränkt ist, wird $U(z_0) = \{z_0\}$ einpunkig. Über die Konvergenz auf dem Rand $\{z \mid |z - z_0| = R\}$ von $U(z_0)$ sagt das Wurzelkriterium nichts aus. Auf der offenen Kreisscheibe $\{z \mid |z - z_0| < R\}$ ist die Potenzreihe absolut konvergent und auf $\{z \mid |z - z_0| > R\}$ divergent.

Man kann den Konvergenzradius auch mit Hilfe des Quotientenkriteriums bestimmen, falls es den Grenzwert $\lim_{k \rightarrow \infty} \frac{|a_{k+1}|}{|a_k|}$ gibt:

$$R = \frac{1}{\lim_{k \rightarrow \infty} \frac{|a_{k+1}|}{|a_k|}}. \quad (10.45)$$

Dabei gilt, wie beim Wurzelkriterium, $U(z_0) = \mathbb{R} (\mathbb{C})$, sollte der Grenzwert 0 sein. Existiert nur noch $\overline{\lim}_{k \rightarrow \infty} \frac{|a_{k+1}|}{|a_k|}$ (und damit auch $\underline{\lim}_{k \rightarrow \infty} \frac{|a_{k+1}|}{|a_k|}$), dann wird durch

$$R_1 := \frac{1}{\overline{\lim}_{k \rightarrow \infty} \frac{|a_{k+1}|}{|a_k|}} \quad (10.46)$$

der Radius einer Kreisscheibe bestimmt, in deren Inneren absolute Konvergenz vorliegt und durch

$$R_2 := \frac{1}{\underline{\lim}_{k \rightarrow \infty} \frac{|a_{k+1}|}{|a_k|}} \quad (10.47)$$

der Radius einer weiteren Kreisscheibe, in deren Äußerem die Potenzreihe divergiert. Über die Konvergenz auf dem Kreisring $\{z \mid R_1 \leq |z| \leq R_2\}$ wird dabei keine Aussage gemacht.

Es gilt $R_1 \leq R \leq R_2$.

Es ist bequem, für die Sonderfälle in diesem Satz die Sprachregelung $R = \infty$ zuzulassen, obwohl ∞ natürlich keine Zahl ist. Wir vereinbaren dann $\frac{1}{0} := \infty$ und $\frac{1}{\infty} := 0$. Mit dieser Absprache wird der Konvergenzradius ohne Ausnahmen durch die Formeln (10.45) oder (10.44) wieder-gegeben. Der Satz besagt dann, daß reellwertige Potenzreihen immer auf symmetrischen Intervallen $(z_0 - R, z_0 + R)$ und komplexwertige auf Kreisscheiben $\{ z \in \mathbb{C} \mid |z - z_0| < R \}$ absolut konvergieren. Dabei ist mit $\mathbb{R} = \infty$ auch das unendlich ausgedehnte Intervall $(-\infty, \infty) = \mathbb{R}$ bzw. die unendlich ausgedehnte Kreisscheibe \mathbb{C} möglich.

Der Satz sagt nichts über die Konvergenz der Potenzreihe auf den Randpunkten des Konvergenzbereichs aus.

Beweis. Wir wenden das Wurzelkriterium (10.40) auf die Reihe $\sum_{k=0}^n a_k(z - z_0)^k$ an: Die Ungleichung

$$\overline{\lim}_{k \rightarrow \infty} \sqrt[k]{|a_k||z - z_0|^k} = |z - z_0| \overline{\lim}_{k \rightarrow \infty} \sqrt[k]{|a_k|} < 1.$$

muß erfüllt sein, damit nach dem Wurzelkriterium absolute Konvergenz vorliegt. Zunächst der Sonderfall, daß der Limes Superior (gemäß obiger Sprachregelung) ∞ ergibt. In diesem Fall muß $|z - z_0| = 0$, also $z = z_0$ sein, damit die Reihe noch konvergiert. Von diesem Fall abgesehen, erhalten wir die folgende Bedingung an z

$$|z - z_0| < \frac{1}{\overline{\lim}_{k \rightarrow \infty} \sqrt[k]{|a_k|}},$$

also $|z - z_0| < R$, die absolute Konvergenz garantiert. Das bedeutet, daß alle Elemente der Menge $\{ z \mid |z - z_0| < R \}$ zum Konvergenzbereich $U(z_0)$ gehören. Für ein z mit $|z - z_0| > R$ besagt das Wurzelkriterium, daß die Reihe sicher nicht konvergiert. Eine notwendige Bedingung für absolute Konvergenz ist also, daß z in der Menge $\{ z \mid |z - z_0| \leq R \}$ liegen muß: $U(z_0) \subseteq \{ z \mid |z - z_0| \leq R \}$.

Die Überlegungen für das Quotientenkriterium (10.45) verlaufen völlig analog, sollte der Grenzwert in (10.45) existieren. Gibt es dagegen nur noch den Limes Superior, dann liefert das Quotientenkriterium zwei verschiedene Radien R_1 und R_2 für Konvergenz ($|z - z_0| < R_1$) bzw. Divergenz ($|z - z_0| > R_2$). Die Ungleichung $R_1 \leq R \leq R_2$ ist eine Folgerung aus dem nächsten Lemma. \square

10.2.16 Lemma $(a_n)_{n \in \mathbb{N}}$ sei eine Folge positiver Zahlen.

i) Wenn $\left(\frac{a_{n+1}}{a_n} \right)_{n \in \mathbb{N}}$ beschränkt ist, existieren $\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n}$, sowie $\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{a_n}$, und es gilt

$$\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} \leq \overline{\lim}_{n \rightarrow \infty} \sqrt[n]{a_n} \leq \overline{\lim}_{n \rightarrow \infty} \frac{a_{n+1}}{a_n}. \quad (10.48)$$

Falls $(\sqrt[n]{a_n})_{n \in \mathbb{N}}$ beschränkt ist, gilt zumindest der linke Teil der Ungleichung.

ii) Sollte $(\sqrt[n]{a_n})_{n \in \mathbb{N}}$ unbeschränkt sein, dann gilt das auch für $\left(\frac{a_{n+1}}{a_n} \right)_{n \in \mathbb{N}}$.

Falls $\left(\frac{a_{n+1}}{a_n} \right)_{n \in \mathbb{N}}$ keine Häufungspunkte hat, ist auch $(\sqrt[n]{a_n})_{n \in \mathbb{N}}$ unbeschränkt.

Beweis. i) Wir gehen davon aus, daß die Folge $(\frac{a_{n+1}}{a_n})_{n \in \mathbb{N}}$ beschränkt ist, so daß $\overline{\lim}_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} =: q$ existiert. Natürlich gibt es dann auch $\lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} =: q_1$. Für jedes $\varepsilon > 0$ müssen fast alle Quotienten unterhalb von $q + \varepsilon$ liegen, d. h. es gibt ein $n_\varepsilon \in \mathbb{N}$, so daß $\frac{a_{n+1}}{a_n} < q + \varepsilon$ für alle $n \geq n_\varepsilon$ gilt. Diese Ungleichung läßt sich rückwärts entwickeln (vergl. den Beweis zu Satz 10.2.11): $a_{n+1} < (q + \varepsilon)^{n+1-n_\varepsilon} a_{n_\varepsilon}$ für alle $n \geq n_\varepsilon$, also $a_n < (q + \varepsilon)^n \frac{a_{n_\varepsilon}}{(q + \varepsilon)^{n_\varepsilon}}$ und damit

$$\sqrt[n]{a_n} < (q + \varepsilon) \sqrt[n]{\frac{a_{n_\varepsilon}}{(q + \varepsilon)^{n_\varepsilon}}}$$

für alle $n > n_\varepsilon$. Da der Wurzelausdruck gegen 1 konvergiert, ist die rechte Seite der Ungleichung nach oben beschränkt, so daß $\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{a_n}$ existiert und nicht größer als $q + \varepsilon = (q + \varepsilon) \overline{\lim}_{n \rightarrow \infty} \sqrt[n]{\frac{a_{n_\varepsilon}}{(q + \varepsilon)^{n_\varepsilon}}}$ ist. Da $\varepsilon > 0$ beliebig war, folgt $\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{a_n} \leq q = \overline{\lim}_{n \rightarrow \infty} \frac{a_{n+1}}{a_n}$.

Für den linken Teil der Ungleichung (10.48) gehen wir davon aus, daß $(\sqrt[n]{a_n})_{n \in \mathbb{N}}$ beschränkt ist, daß also $\overline{\lim}_{n \rightarrow \infty} \sqrt[n]{a_n} = q$ existiert. Für die Existenz von $q_1 = \lim_{n \rightarrow \infty} \frac{a_{n+1}}{a_n}$ brauchen wir nur einen Häufungspunkt der Folge $(\frac{a_{n+1}}{a_n})_{n \in \mathbb{N}}$, denn sie ist automatisch nach unten durch 0 beschränkt. Wir nehmen an, sie hätte keinen einzigen Häufungspunkt. Dann können im Intervall $[0, m]$ nur endlich viele Folgenglieder liegen, bei beliebigem $m > 0$. Für alle n oberhalb eines geeigneten $n_m \in \mathbb{N}$ gilt also $\frac{a_{n+1}}{a_n} > m$, oder $a_{n+1} > ma_n > m^2 a_{n-1} > \dots > m^{n+1} \frac{a_{n_m}}{m^{n_m}}$. Das bedeutet $\sqrt[n]{a_n} > m \sqrt[n]{\frac{a_{n_m}}{m^{n_m}}}$ für all $n > n_m$. Da der Wurzelausdruck gegen 1 konvergiert, wird er, eventuell ab einem weiteren $\tilde{n}_m \geq n_m$, größer als $\frac{1}{2}$ werden: $\sqrt[n]{a_n} > \frac{m}{2}$ für alle $n \geq \tilde{n}_m$. Daher wäre die Folge $(\sqrt[n]{a_n})_{n \in \mathbb{N}}$ unbeschränkt, im Widerspruch zu unserer Annahme (damit haben wir übrigens schon den zweiten Teil von ii) bewiesen).

Bleibt der linke Teil der Ungleichung (10.48) nachzuweisen: $q_1 \leq q$. Das geschieht ähnlich, wie für den rechten Teil. Falls $q_1 = 0$ gelten sollte, gibt es nichts zu zeigen. Wir können also von $q_1 > 0$ ausgehen. Unterhalb von $q_1 - \varepsilon > 0$ können dann nur endlich viele $\frac{a_{n+1}}{a_n}$ liegen, so daß $a_{n+1} \geq (q_1 - \varepsilon)a_n$ für fast alle $n \in \mathbb{N}$ gelten muß. Daraus schließen wir, wie oben, auf $\sqrt[n]{a_n} > (q - \varepsilon) \sqrt[n]{\frac{a_{n_\varepsilon}}{(q - \varepsilon)^{n_\varepsilon}}}$, für alle n oberhalb eines geeigneten $n_\varepsilon \in \mathbb{N}$. Das bedeutet $q = \overline{\lim}_{n \rightarrow \infty} \sqrt[n]{a_n} \geq q_1 - \varepsilon$ und, da $\varepsilon > 0$ beliebig war, $q \geq q_1$.

ii) Wir müssen nur noch zeigen, daß aus der Unbeschränktheit von $(\sqrt[n]{a_n})_{n \in \mathbb{N}}$ auch die von $(\frac{a_{n+1}}{a_n})_{n \in \mathbb{N}}$ folgt. Das haben wir aber schon erledigt. Denn wäre $(\frac{a_{n+1}}{a_n})_{n \in \mathbb{N}}$ beschränkt, dann wäre es auch $(\sqrt[n]{a_n})_{n \in \mathbb{N}}$, wie wir oben gezeigt haben. \square

Einige wichtige Potenzreihen lassen sich nicht einfach in der Standardform $\sum_{k=0}^{\infty} a_k(z - z_0)^k$ behandeln. So ist z. B. $\sum_{k=0}^{\infty} \frac{z^{2k}}{k!} = \exp(z^2)$ nur durch die Konstruktion

$$a_k := \begin{cases} \frac{1}{(\frac{k}{2})!} & , k \text{ gerade} \\ 0 & , k \text{ ungerade} \end{cases}$$

in die Standardform zu bringen, für die dann das Wurzel- oder Quotientenkriterium zur Bestimmung von R nur mühsam anzuwenden ist. Abhilfe schafft das folgende Korollar, das die meisten interessanten Fälle abdeckt.

10.2.17 Korollar Für eine Potenzreihe $f(z) = \sum_{k=0}^{\infty} a_k z^{p+k+q}$, $p \in \mathbb{N}$, $q \in \mathbb{Z}$, bestimmt

$$R = \frac{1}{\sqrt[p]{\limsup_{k \rightarrow \infty} \sqrt[k]{|a_k|}}} \quad (10.49)$$

$$R_1 = \frac{1}{\sqrt[p]{\limsup_{k \rightarrow \infty} \frac{|a_{k+1}|}{|a_k|}}} \quad R_2 = \frac{1}{\sqrt[p]{\limsup_{k \rightarrow \infty} \frac{|a_{k+1}|}{|a_k|}}} \quad (10.50)$$

den Konvergenzradius R , also den Bereich $\{z \in \mathbb{C} \mid |z - z_0| < R\}$ absoluter Konvergenz und den Bereich $\{z \in \mathbb{C} \mid |z - z_0| > R\}$, auf dem die Potenzreihe divergiert. Bei Verwendung des Quotientenkriteriums sind diese Bereiche $\{z \in \mathbb{C} \mid |z - z_0| < R_1\}$ bzw. $\{z \in \mathbb{C} \mid |z - z_0| > R_2\}$ (vergl. Satz 10.2.15).

Beweis. Der Beweis verläuft fast völlig analog zum Beweis von 10.2.15. Deshalb führen wir nur die entscheidende Abweichung vor. Das Wurzelkriterium, auf die Reihe $\sum_{k=0}^{\infty} a_k z^{p+k+q}$ angewandt, ergibt

$$\begin{aligned} \overline{\limsup_{k \rightarrow \infty}} \sqrt[k]{|a_k||z - z_0|^{p+k+q}} &= |z - z_0|^p \overline{\limsup_{k \rightarrow \infty}} \sqrt[k]{|a_k|} \sqrt[k]{|z - z_0|^q} \\ &= |z - z_0|^p \overline{\limsup_{k \rightarrow \infty}} \sqrt[k]{|a_k|} < 1, \end{aligned}$$

denn $\overline{\limsup_{k \rightarrow \infty}} \sqrt[k]{|z - z_0|^q} = 1$. Auflösen nach $|z - z_0|$ zeigt (10.49). (10.50) zeigt man genauso. \square

10.2.18 Beispiel Den Konvergenzradius der Reihe $f(z) := \sum_{k=0}^{\infty} \frac{1}{k+1} \binom{2k}{k} z^k$ bestimmen wir mit dem Quotientenkriterium:

$$\frac{\frac{1}{k+2} \binom{2k+2}{k+1}}{\frac{1}{k+1} \binom{2k}{k}} = \frac{(k+1)(2k+2)! k!^2}{(k+2)(k+1)!^2 (2k)!} = \frac{(2k+2)(2k+1)}{(k+1)(k+2)} = \frac{2(2k+1)}{k+2} \xrightarrow{k \rightarrow \infty} 4.$$

Der Konvergenzradius ist demnach $R = \frac{1}{4}$.

Bemerkung: Die Zahlen $\frac{1}{k+1} \binom{2k}{k}$ werden CATALAN-Zahlen genannt. Sie tauchen z. B. bei der diskreten BROWNSchen Bewegung auf.

10.2.19 A Untersuchen Sie die folgenden Reihen auf Konvergenz.

- i) $\sum_{k=0}^{\infty} \frac{3^k + 2^k}{4^k + 7}$
- ii) $\sum_{k=1}^{\infty} \frac{k^2 - 2k + 4}{2k + 3k^2 + k^4}$
- iii) $\sum_{k=0}^{\infty} ((-4)^k + 0.5((-1)^k + 1)6^k)z^k$
- iv) $\sum_{k=0}^{\infty} a_k z^k, \quad a_k := \begin{cases} 2 & , k \text{ ungerade} \\ 1 & , k \text{ gerade} \end{cases}.$

10.2.20 A Sei $(a_n)_{n \in \mathbb{N}}$ die FIBONACCI-Folge (vgl. 9.12).

i) Zeigen Sie, daß die Folge $\left(\frac{a_{n+1}}{a_n}\right)_{n \in \mathbb{N}}$ gegen den goldenen Schnitt $\Phi = \frac{1}{2}(\sqrt{5} + 1)$ konvergiert. Verwenden Sie dafür Darstellung (9.17) von a_n und beweisen Sie damit $\frac{a_n}{\Phi^{n+1}} \xrightarrow{n \rightarrow \infty} \frac{1}{\sqrt{5}}$.

ii) Bestimmen Sie den Konvergenzradius der Reihe $\sum_{k=0}^{\infty} a_k z^k$.

iii) Bestimmen Sie die Funktion $f(z) := \sum_{k=0}^{\infty} a_k z^k$

10.2.21 Satz (Doppelreihensatz) *Für eine Doppelfolge $(a_{k\ell})_{k,\ell \in \mathbb{N}}$ aus \mathbb{C}^n gebe es eine Konstante $K > 0$ mit der Eigenschaft*

$$\sum_{k=1}^m \sum_{\ell=1}^n \|a_{k\ell}\| \leq K$$

für alle $m, n \in \mathbb{N}$. Dann gelten die folgenden äquivalenten Aussagen:

i) Die Doppelreihe $\sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} a_{k\ell}$ ist absolut konvergent.

ii) Die Doppelreihe $\sum_{\ell=1}^{\infty} \sum_{k=1}^{\infty} a_{k\ell}$ ist absolut konvergent.

iii) $\sum_{k,\ell=1}^{\infty} a_{k\ell} := \lim_{m,n \rightarrow \infty} \sum_{k=1}^m \sum_{\ell=1}^n a_{k\ell}$ ist absolut konvergent.

iv) Eine Umordnung $\psi : \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}$ ergibt eine absolut konvergente Reihe $\sum_{k=1}^{\infty} a_{\psi(k)}$.

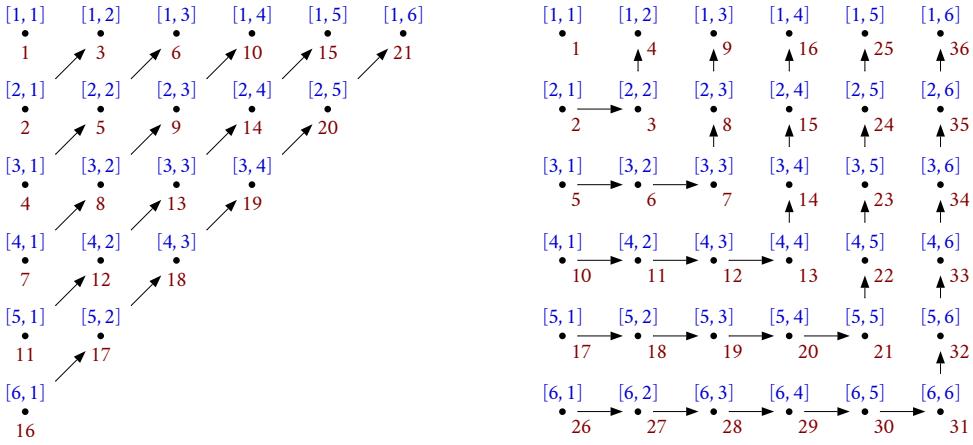
Der Wert all dieser Reihen ist gleich:

$$\sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} a_{k\ell} = \sum_{\ell=1}^{\infty} \sum_{k=1}^{\infty} a_{k\ell} = \sum_{k,\ell=1}^{\infty} a_{k\ell} = \sum_{k=1}^{\infty} a_{\psi(k)}. \quad (10.51)$$

Unter der *absoluten Konvergenz einer Doppelreihe* verstehen wir hier, daß der iterierte Grenzwert $\sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} \|a_{k\ell}\|$ existiert. Unter iii) meint die absolute Konvergenz, daß $\sum_{k,\ell=1}^{\infty} \|a_{k\ell}\|$ existiert (vergl. (10.19)). Anders als bei Einfachreihen, ist nicht sofort ersichtlich, daß daraus die Existenz von $\sum_{k,\ell=1}^{\infty} a_{k\ell}$ folgt.

Die *Umordnung* einer Doppelreihe wird durch eine bijektive Abbildung ψ des eindimensionalen Indexraumes \mathbb{N} auf den zweidimensionalen $\mathbb{N} \times \mathbb{N}$ beschrieben. Anschaulich bedeutet das, einen Weg durch das zweidimensionale Zahlenschema $\mathbb{N} \times \mathbb{N}$ zu legen, der jedes Zahlenpaar $[k, \ell]$ genau einmal trifft. Entlang dieses Weges werden dann die Ausdrücke $a_{k,\ell}$ aufsummiert und ergeben so eine Einfachreihe $\sum_{n=1}^{\infty} a_{\psi(n)}$. Eine häufig verwendete Umordnung ist δ , der Weg entlang der Diagonalen, wie er im linken Teil von Abbildung 10.8 zu sehen ist. Er verläuft entlang aller Paare $[i, j]$, für die $i + j - 1$ die Nummer der Diagonalen ergibt. Diese Umordnung tritt in natürlicher Weise beim Produkt von Potenzreihen auf. Die quadratische Umordnung χ auf der rechten Seite von Abbildung 10.8 werden wir für den Beweis des Doppelreihensatzes verwenden.

Bemerkung: Wenn es die Lesbarkeit erfordert, schreiben wir auch $a_{k,\ell}$ für $a_{k\ell}$.



$$\delta(k) := \left[\frac{1}{2}n(n+1) + 1 - k, k - \frac{1}{2}n(n-1) \right], \quad \chi(k) := \begin{cases} \min\{n^2 + 1 - k, n\}, & k - (n-1)^2 \leq n^2 \\ \min\{k - (n-1)^2, n\}, & n^2 < k \leq n^2 \end{cases}$$

Abb. 10.8 Diagonal-Abzählung δ und quadratische Abzählung χ

Beweis. Aus $\sum_{\ell=1}^n \|a_{k\ell}\| \leq \sum_{k=1}^m \sum_{\ell=1}^n \|a_{k\ell}\| \leq K$ folgt die absolute Konvergenz von $\sum_{\ell=1}^{\infty} a_{k\ell}$ für alle $k \in \mathbb{N}$. Außerdem gilt auch $\sum_{k=1}^m \|\sum_{\ell=1}^{\infty} a_{k\ell}\| \leq \lim_{n \rightarrow \infty} \sum_{k=1}^m \sum_{\ell=1}^n \|a_{k\ell}\| \leq K$. Daher ist $\sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} a_{k\ell}$ absolut konvergent, d.h., der iterierte Grenzwert $\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \sum_{k=1}^m \sum_{\ell=1}^n a_{k\ell}$ existiert.

Zu i) \Rightarrow ii) \Rightarrow iii): Da die Doppelreihe $\sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} a_{k\ell}$ absolut konvergent ist, muß $\sum_{k=1}^m \sum_{\ell=1}^n \|a_{k\ell}\| \leq \sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} \|a_{k\ell}\| =: K$ für alle $n, m \in \mathbb{N}$ gelten. Das bedeutet insbesondere $\sum_{k=1}^m \|a_{k\ell}\| \leq \sum_{k=1}^m \sum_{\ell=1}^n \|a_{k\ell}\| \leq K$ für alle $m, \ell \in \mathbb{N}$, so daß natürlich auch $\sum_{k=1}^{\infty} a_{k\ell}$ absolut konvergiert. Außerdem gilt $\sum_{\ell=1}^n \|\sum_{k=1}^m a_{k\ell}\| \leq \sum_{\ell=1}^n \sum_{k=1}^m \|a_{k\ell}\| = \lim_{m \rightarrow \infty} \sum_{\ell=1}^n \sum_{k=1}^m \|a_{k\ell}\| \leq K$, woraus die absolute Konvergenz der zweiten Doppelsumme $\sum_{\ell=1}^{\infty} \sum_{k=1}^{\infty} a_{k\ell}$ folgt. Das zeigt ii). Damit haben wir bisher die Existenz der beiden iterierten Grenzwerte $a := \lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \sum_{k=1}^m \sum_{\ell=1}^n a_{k\ell} = \sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} a_{k\ell}$ und $b := \lim_{n \rightarrow \infty} \lim_{m \rightarrow \infty} \sum_{k=1}^m \sum_{\ell=1}^n a_{k\ell} = \sum_{\ell=1}^{\infty} \sum_{k=1}^{\infty} a_{k\ell}$ nachgewiesen.

Für alle $\varepsilon > 0$ gibt es ein m_{ε} , so daß für alle $m > m' \geq m_{\varepsilon}$ folgende Abschätzung erfüllt ist:

$$\begin{aligned} \left\| \sum_{k=1}^m \sum_{\ell=1}^n a_{k\ell} - \sum_{k=1}^{m'} \sum_{\ell=1}^n a_{k\ell} \right\| &\leq \sum_{k=m'+1}^m \sum_{\ell=1}^n \|a_{k\ell}\| \leq \sum_{k=m'+1}^m \sum_{\ell=1}^{\infty} \|a_{k\ell}\| \\ &= \sum_{k=1}^m \sum_{\ell=1}^{\infty} \|a_{k\ell}\| - \sum_{k=1}^{m'} \sum_{\ell=1}^{\infty} \|a_{k\ell}\| < \varepsilon. \end{aligned}$$

Das liegt daran, daß $(\sum_{k=1}^m \sum_{\ell=1}^{\infty} \|a_{k\ell}\|)_{m \in \mathbb{N}}$ eine konvergente Folge bildet. Daher ist $(\sum_{k=1}^m \sum_{\ell=1}^n a_{k\ell})_{m \in \mathbb{N}}$ eine bezüglich $n \in \mathbb{N}$ gleichmäßige CAUCHY-Folge, die nach Lemma 10.1.51 gleichmäßig bezüglich $n \in \mathbb{N}$ konvergiert. Damit sind die Voraussetzungen des zweiten

Vertauschbarkeitssatzes 10.1.48 gegeben. Wir haben daher $a = b = \lim_{m,n \rightarrow \infty} \sum_{k=1}^m \sum_{\ell=1}^n a_{k\ell} = \sum_{k,\ell=1}^{\infty} a_{k\ell}$, natürlich als absolut konvergente Reihe, denn unsere bisherigen Überlegungen lassen sich auch auf die Reihen über $\|a_{k\ell}\|$ anwenden. Das ergibt iii).

Zu iii) \Rightarrow iv): Wir zeigen zunächst nur, daß aus iii) die absolute Konvergenz der Reihe $\sum_{k=1}^{\infty} a_{\chi(k)}$ mit der quadratischen Abzählung χ folgt. Laut Voraussetzung gibt es eine Zahl $\alpha \geq 0$, die Grenzwert der Doppelfolge $(\sum_{k=1}^m \sum_{\ell=1}^n \|a_{k\ell}\|)_{m,n \in \mathbb{N}}$ ist. Für alle $\varepsilon > 0$ gibt es demnach $\tilde{m}_\varepsilon, \tilde{n}_\varepsilon \in \mathbb{N}$, so daß für alle $m \geq \tilde{m}_\varepsilon$ und $n \geq \tilde{n}_\varepsilon$ die Abschätzung $0 \leq \alpha - \sum_{k=1}^m \sum_{\ell=1}^n \|a_{k\ell}\| < \varepsilon$ erfüllt ist. Daher ist α eine obere Schranke K. Aus unseren Vorüberlegungen am Anfang des Beweises wissen wir, daß damit i) und alle Folgerungen, die wir gezogen haben, gültig sind. Insbesondere existiert $a = \sum_{k,\ell=1}^{\infty} a_{k\ell}$, und dieser Wert stimmt mit den iterierten Grenzwerten überein. Wegen $\sum_{k=1}^n \|a_{\chi(k)}\| \leq \sum_{k=1}^{n^2} \|a_{\chi(k)}\| = \sum_{k=1}^n \sum_{\ell=1}^n \|a_{k\ell}\| \leq \alpha$, existiert $b := \sum_{k=1}^{\infty} a_{\chi(k)}$. Für alle $\varepsilon > 0$ gibt es demnach m_ε und n_ε , sowie q_ε , so daß für alle $m \geq m_\varepsilon$, $n \geq n_\varepsilon$ und $q \geq q_\varepsilon$ die Abschätzungen $\|a - \sum_{k=1}^m \sum_{\ell=1}^n a_{k\ell}\| < \frac{\varepsilon}{2}$ und $\|b - \sum_{p=1}^q a_{\chi(p)}\| < \frac{\varepsilon}{2}$ erfüllt sind. Wir wählen ein $n \geq \max\{n_\varepsilon, m_\varepsilon, q_\varepsilon\}$ und erhalten damit

$$\|a - b\| \leq \left\| a - \sum_{k=1}^n \sum_{\ell=1}^n a_{k\ell} \right\| + \left\| \sum_{p=1}^{n^2} a_{\chi(p)} - b \right\| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

Da $\varepsilon > 0$ beliebig ist, muß $a = b$ gelten.

Jetzt sei ψ eine beliebige Umordnung der Doppelreihe. Dann ist $\psi = \chi \circ \lambda$, mit der bijektiven Abbildung $\lambda := \chi^{-1} \circ \psi: \mathbb{N} \rightarrow \mathbb{N}$. Also ist $\sum_{k=1}^n a_{\psi(k)} = \sum_{k=1}^n a_{\chi(\lambda(k))}$ eine Umordnung der absolut konvergenten Einfachreihe $\sum_{p=1}^n a_{\chi(p)}$ und daher, nach Satz 10.2.4, selbst absolut konvergent, mit dem gleichen Grenzwert a.

Zu iv) \Rightarrow i): $\sum_{k=1}^{\infty} a_{\psi(k)}$ ist absolut konvergent. Also ist auch $\sum_{p=1}^{\infty} a_{\chi(p)} = \sum_{p=1}^{\infty} a_{\psi(\lambda^{-1}(p))}$ als Umordnung dieser Reihe absolut konvergent. Das bedeutet insbesondere $\sum_{p=1}^m \|a_{\chi(p)}\| \leq K$, für eine geeignete Schranke $K > 0$. Für $m = n^2$ heißt das $\sum_{p=1}^{n^2} \|a_{\chi(p)}\| = \sum_{k=1}^n \sum_{\ell=1}^n \|a_{k\ell}\| \leq K$. Daraus folgt für alle $m, n \in \mathbb{N}$ auch $\sum_{k=1}^m \sum_{\ell=1}^n \|a_{k\ell}\| \leq \sum_{k=1}^p \sum_{\ell=1}^p \|a_{k\ell}\| \leq K$, mit $p := \max\{m, n\}$. Aus dieser Eigenschaft haben wir zu Beginn des Beweises auf die absolute Konvergenz der Doppelreihe $\sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} a_{k\ell}$ schließen können, also auf i). \square

10.2.22 A Zeigen Sie, daß durch

$$\delta: \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}, \quad (10.52)$$

$$k \mapsto \left[\frac{1}{2}n(n+1) + 1 - k, k - \frac{1}{2}n(n-1) \right], \quad \frac{1}{2}n(n-1) < k \leq \frac{1}{2}n(n+1),$$

$$\delta^{-1}: [i, j] \mapsto \frac{1}{2}(i+j-1)(i+j) + 1 - i,$$

$$\chi: \mathbb{N} \rightarrow \mathbb{N} \times \mathbb{N}, \quad (10.53)$$

$$k \mapsto \left[\min\{n^2 + 1 - k, n\}, \min\{k - (n-1)^2, n\} \right], \quad (n-1)^2 < k \leq n^2,$$

$$\chi^{-1}: [i, j] \mapsto \max\{i, j\}^2 - \max\{i, j\} + j - i + 1,$$

jeweils eine bijektive Abbildung definiert wird.

10.2.23 A $\sum_{k=1}^{\infty} a_k$ und $\sum_{\ell=1}^{\infty} b_{\ell}$ seien absolut konvergente Reihen. Zeigen Sie, daß dann die Voraussetzungen des Doppelreihensatzes 10.2.21 für das Produkt $(\sum_{k=1}^{\infty} a_k)(\sum_{\ell=1}^{\infty} b_{\ell}) = \sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} a_k b_{\ell}$ erfüllt sind.

Unter den Voraussetzungen des Doppelreihensatzes gilt

$$\sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} a_{k\ell} = \sum_{n=1}^{\infty} \sum_{k=1}^n a_{n+1-k,k}, \quad \text{bzw.} \quad \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} a_{k\ell} = \sum_{n=0}^{\infty} \sum_{k=0}^n a_{n-k,k}. \quad (10.54)$$

Denn

$$\sum_{k=1}^{\infty} \sum_{\ell=1}^{\infty} a_{k\ell} = \sum_{k=1}^{\infty} a_{\delta(k)} = \sum_{n=1}^{\infty} \sum_{k=\frac{1}{2}n(n-1)+1}^{\frac{1}{2}n(n+1)} a_{\frac{1}{2}n(n+1)+1-k, k-\frac{1}{2}n(n-1)} = \sum_{n=1}^{\infty} \sum_{\ell=1}^n a_{n+1-\ell, \ell}.$$

Dabei haben wir die Indextransformation $\ell = k - \frac{1}{2}n(n-1)$ verwendet. Für die zweite Doppelsumme setzen wir $b_{k,\ell} := a_{k-1,\ell-1}$. Dann ist

$$\begin{aligned} \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} a_{k\ell} &= \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} a_{n-1,m-1} = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} b_{n,m} = \sum_{n=1}^{\infty} \sum_{m=1}^n b_{n+1-m,m} \\ &= \sum_{n=1}^{\infty} \sum_{m=1}^n a_{n-m,m-1} = \sum_{k=0}^{\infty} \sum_{m=1}^{k+1} a_{k-m+1,m-1} = \sum_{k=0}^{\infty} \sum_{\ell=0}^k a_{k-\ell,\ell}. \end{aligned}$$

Die dabei verwendeten Indextransformationen kann sich jeder selbst überlegen.

10.2.24 Die Exponentialfunktion Wir führen die Exponentialfunktion $\mathbb{C} \ni z \mapsto \exp(z)$ über ihre Reihe

$$\exp(z) := \sum_{k=0}^{\infty} \frac{z^k}{k!} \quad (10.55)$$

ein. Von der absoluten Konvergenz dieser Reihe für alle $z \in \mathbb{C}$ haben wir uns bereits überzeugt (Seite 242 und Beispiel 10.2.12 i)). Wir müssen zeigen, daß diese Funktion, wenigstens für $x \in \mathbb{Q}$, tatsächlich die Funktionswerte e^x hat, wenn wir $e := \exp(1)$ setzen. Der Schlüssel dafür ist die sogenannte *Funktionalgleichung der Exponentialfunktion*:

$$\exp(z+w) = \exp(z)\exp(w) \quad (10.56)$$

für alle $z, w \in \mathbb{C}$. Der Beweis dieser Formel ist eine einfache Anwendung des Doppelreihensatzes auf $\exp(z)\exp(w) = \sum_{k=0}^{\infty} \frac{z^k}{k!} \exp(w) = \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} \frac{z^k}{k!} \frac{w^{\ell}}{\ell!}$. Die Voraussetzungen sind natürlich erfüllt (vergl. Übung 10.2.23). Damit können wir (10.56) einfach ausrechnen. Wir verwenden die Diagonalabzählung δ und (10.54):

$$\begin{aligned} \exp(z)\exp(w) &= \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} \frac{z^k}{k!} \frac{w^{\ell}}{\ell!} = \sum_{n=0}^{\infty} \sum_{k=0}^n \frac{z^{n-k}}{(n-k)!} \frac{w^k}{k!} = \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^n \frac{n!}{k!(n-k)!} z^{n-k} w^k \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} \sum_{k=0}^n \binom{n}{k} z^{n-k} w^k = \sum_{n=0}^{\infty} \frac{(z+w)^n}{n!} = \exp(z+w). \end{aligned}$$

10.2.25 Lemma Die reelle Exponentialfunktion $\exp : \mathbb{R} \rightarrow \mathbb{R}^+$ hat keine Nullstellen, ist streng monoton wachsend und damit injektiv. Sie wächst für $x \rightarrow \infty$ gegen Unendlich, und für $x \rightarrow -\infty$ strebt sie gegen Null.

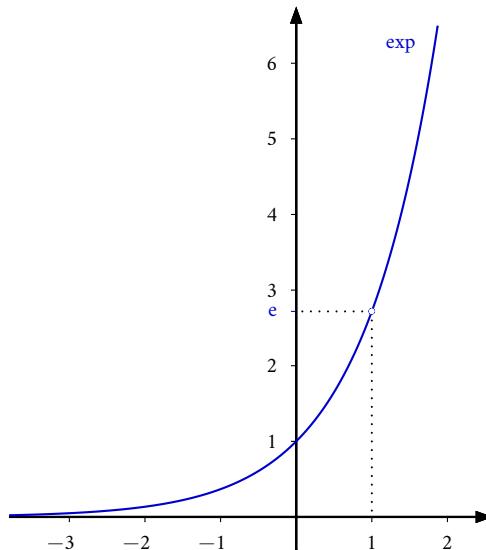
Beweis. Für $h > 0$ ist $\exp(h) = \sum_{k=0}^{\infty} \frac{h^k}{k!} > 1 + h > 1$, insbesondere also positiv. Aus dieser Abschätzung erkennt man auch, daß $\exp(h)$ für $h \rightarrow \infty$ beliebig groß wird. Aus $1 = \exp(0) = \exp(h-h) = \exp(h)\exp(-h)$ folgt $\exp(-h) = \frac{1}{\exp(h)} > 0$ und $\exp(-h) < \frac{1}{1+h} < 1$. Das zeigt $\exp(-h) \rightarrow 0$ für $h \rightarrow \infty$. Damit nimmt die Exponentialfunktion nur positive Werte an. Jetzt folgt die strenge Monotonie aus der Funktionalgleichung: Für $x < y$ gilt $\exp(y) = \exp(x + (y-x)) = \exp(x)\exp(y-x) > \exp(x)$, denn $y-x > 0$ und $\exp(y-x) > 1$. Das zeigt auch die Injektivität von \exp , denn für $x \neq y$, o. B. d. A. $x < y$, folgt $\exp(x) < \exp(y)$, also $\exp(x) \neq \exp(y)$. \square

10.2.26 Korollar Für alle $x \in \mathbb{Q}$ gilt $e^x = \exp(x)$.

Beweis. Zunächst haben wir $e^0 = 1 = \exp(0)$ und für $n \in \mathbb{N}$:

$$\begin{aligned} e^n &= \underbrace{\exp(1) \exp(1) \cdots \exp(1)}_{n \text{ mal}} = \exp(\underbrace{1+1+\cdots+1}_{n \text{ mal}}) = \exp(n), \\ e &= \exp\left(\frac{1}{n} \cdot n\right) = \exp\left(\underbrace{\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}}_{n \text{ mal}}\right) = \exp\left(\frac{1}{n}\right)^n. \end{aligned}$$

Die letzte Gleichung zeigt $\exp\left(\frac{1}{n}\right) = e^{\frac{1}{n}}$. Genauso sieht man $e^{\frac{m}{n}} = \exp\left(\frac{m}{n}\right)$, so daß die Behauptung bereits für alle $0 \leq x \in \mathbb{Q}$ stimmt. Aus $1 = \exp(x-x) = \exp(x)\exp(-x) = e^x \exp(-x)$ erhalten wir schließlich auch $\exp(-x) = e^{-x}$. \square



Für $x \in \mathbb{R} \setminus \mathbb{Q}$ ist der Ausdruck e^x durch keine der elementaren Rechenvorschriften abgedeckt. Wenn man e^x auch für solche x definieren will, muß man es mit einem Grenzwert versuchen. Man approximiert x durch eine Folge $(x_n)_{n \in \mathbb{N}}$ rationaler Zahlen und definiert dann e^x durch den Grenzwert der Folge (e^{x_n}) . Allerdings muß man dafür diesen Grenzwert erst einmal nachweisen. Wir machen uns die Sache hier etwas leichter, indem wir e^x für irrationale x einfach durch $\exp(x)$ definieren, denn die Exponentialfunktion haben wir ja sogar für alle komplexen Zahlen zur Verfügung. Wir werden sehen, daß beide Zugänge äquivalent sind (11.1.32).

10.2.27 Lemma e ist irrational.

Beweis. Wir gehen vom Gegenteil aus: $e = \frac{p}{q} \in \mathbb{Q}$, mit teilerfremden natürlichen Zahlen p und q . Dann ergibt eine Multiplikation der Gleichung $e = \frac{p}{q} = \sum_{k=0}^{\infty} \frac{1}{k!} = \sum_{k=0}^q \frac{1}{k!} + \sum_{k=q+1}^{\infty} \frac{1}{k!}$ mit $q!$:

$$p(q-1)! - \sum_{k=0}^q \frac{q!}{k!} = \sum_{k=q+1}^{\infty} \frac{q!}{k!} = \frac{1}{q+1} + \sum_{k=q+2}^{\infty} \frac{1}{k} \cdot \frac{1}{k-1} \cdots \frac{1}{q+1} > 0.$$

Die linke Seite der Gleichung bestimmt eine ganze Zahl > 0 , also ein Element aus \mathbb{N} . Die Summe der rechten Seite wird sicher größer, wenn wir nur die ersten beiden Faktoren der Summanden behalten

$$\begin{aligned} p(q-1)! - \sum_{k=0}^q \frac{q!}{k!} &< \frac{1}{q+1} + \sum_{k=q+2}^{\infty} \frac{1}{k} \cdot \frac{1}{k-1} = \frac{1}{q+1} + \lim_{n \rightarrow \infty} \sum_{k=q+2}^n \frac{1}{k-1} - \frac{1}{k} \\ &= \frac{1}{q+1} + \lim_{n \rightarrow \infty} \frac{1}{q+1} - \frac{1}{n} = \frac{2}{q+1} \leq 1, \end{aligned}$$

da $q \geq 1$. Dabei haben wir verwendet, daß es sich bei $\sum_{k=q+2}^n \frac{1}{k-1} - \frac{1}{k} = \frac{1}{q+1} - \frac{1}{n}$ um eine Teleskop-Summe handelt, bei der sich die Summanden fast vollständig gegenseitig aufheben. Damit wäre die linke Seite eine natürliche Zahl < 1 , was nicht möglich ist. Also ist unsere Annahme falsch und daher $e \notin \mathbb{Q}$. \square

10.2.28 A Die Exponentialfunktion ist sicher eine der wichtigsten Funktionen in der Mathematik. Daher ist es kein Fehler, wenn man noch einen weiteren Zugang zu dieser Funktion kennt. Wir skizzieren in dieser Aufgabe einen Weg zur reellen Exponentialfunktion, der als Hilfsmittel nur die elementaren Rechenregeln konvergenter Folgen und die BERNOULLI-Ungleichung benötigt.

- i) Zeigen Sie: Für alle $x \in \mathbb{R}$ und für $\mathbb{N} \ni n_0 > |x| + 1$ ist $\left(1 + \frac{x}{n}\right)^n$, wenigstens ab n_0 , monoton wachsend. Weisen Sie dafür $\frac{(1 + \frac{x}{n+1})^{n+1}}{(1 + \frac{x}{n})^n} \geq 1$ für $n \geq n_0$ nach.
- ii) Folgern Sie daraus, daß $\left(1 - \frac{x}{n}\right)^{-n}$ ab n_0 monoton fallend, positiv und daher konvergent ist. Den Grenzwert bezeichnen Sie mit $\exp(x)$.
- iii) Zeigen Sie $\lim_{n \rightarrow \infty} \left(1 - \frac{x^2}{n^2}\right)^n = 1$ und folgern Sie daraus $\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n = \exp(x)$.
- iv) $(x_n)_{n \in \mathbb{N}}$ sei eine reelle Folge, die gegen x konvergiert. Zeigen Sie, daß ab einem geeigneten $n_1 \in \mathbb{N}$ die folgenden Ungleichungen gelten:

$$1 + \frac{x_n - x}{1 + \frac{x}{n}} \leq \frac{\left(1 + \frac{x_n}{n}\right)^n}{\left(1 + \frac{x}{n}\right)^n} \leq \frac{1}{1 + \frac{x - x_n}{1 + \frac{x_n}{n}}}.$$

- v) Folgern Sie daraus: $\lim_{n \rightarrow \infty} \frac{\left(1 + \frac{x_n}{n}\right)^n}{\left(1 + \frac{x}{n}\right)^n} = 1$ und $\lim_{n \rightarrow \infty} \left(1 + \frac{x_n}{n}\right)^n = \exp(x)$.

- vi) Zeigen Sie: $\exp(x+y) = \exp(x)\exp(y)$.

11 Funktionen

11.1 Stetige Funktionen

11.1.1 Definition X und Y seien normierte Räume. Eine Funktion $f : D_f \subseteq X \rightarrow Y$ mit dem Definitionsbereich D_f heißt an einer Stelle $x_0 \in D_f$ stetig, falls für jede Folge $(x_n)_{n \in \mathbb{N}}$ in D_f , die den Grenzwert x_0 hat, die Bildfolge $(f(x_n))_{n \in \mathbb{N}}$ in Y gegen den Funktionswert $f(x_0)$ konvergiert. Diese Eigenschaft fassen wir durch die Schreibweise

$$\lim_{x \rightarrow x_0} f(x) = f(x_0) \quad (11.1)$$

zusammen. Folgen $(x_n)_{n \in \mathbb{N}}$ mit der Eigenschaft $x_n \in D_f$ für alle $n \in \mathbb{N}$ nennen wir zulässig. Eine Funktion, die in jedem Punkt einer Menge $A \subseteq D_f$ stetig ist, heißt stetig auf A .

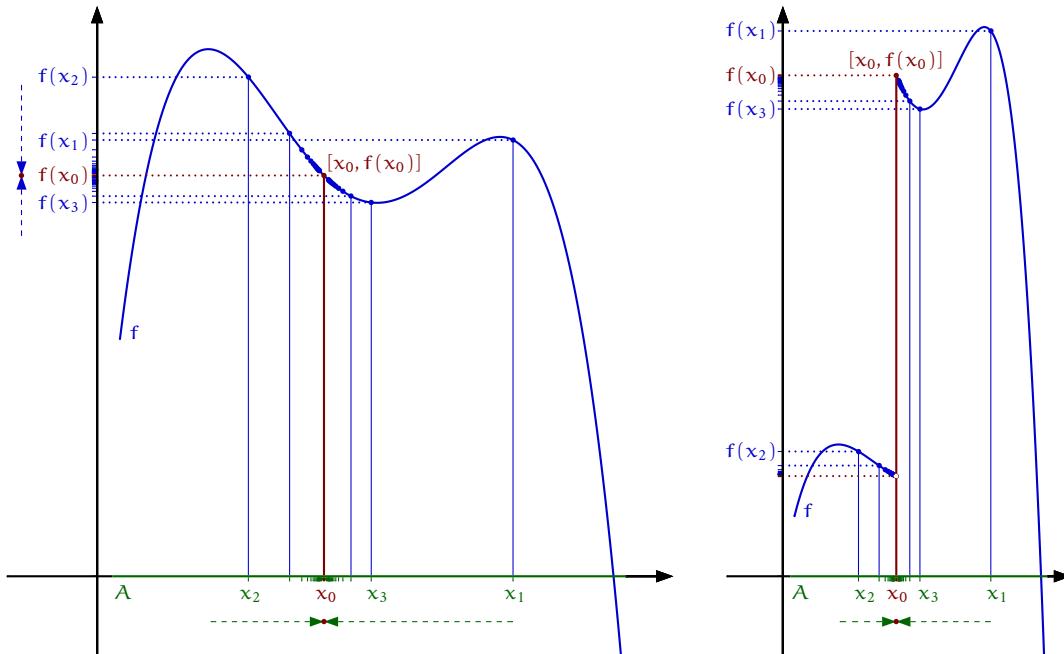


Abb. 11.1 Eine stetige Funktion

Eine unstetige Funktion

Für reellwertige Funktionen ist in der linken Skizze der Stetigkeitsbegriff veranschaulicht. Die rechte Skizze zeigt eine an der Stelle x_0 mit einer Sprungstelle versehene Funktion. Nähert sich eine zulässige Folge $(x_n)_{n \in \mathbb{N}}$, wie in der Skizze, der Zahl x_0 von beiden Seiten, so ist die Bildfolge $(f(x_n))$ nicht konvergent. Nähert sie sich von links, so ist die Bildfolge zwar konvergent,

aber sie konvergiert nicht gegen den Funktionswert $f(x_0)$. Diese Funktion ist in x_0 also sicher unstetig. Wir werden noch andere Arten von Unstetigkeitsstellen kennenlernen. In diesem Zusammenhang werden wir auch die Notation $\lim_{x \rightarrow x_0} f(x) = a$ verwenden, die eine sinngemäße Erweiterung der Definition (11.1) darstellt:

11.1.2 Definition Für eine Funktion $f : D \rightarrow Y$ mit Werten in einem normierten Raum Y und Definitionsbereich $D_f \subseteq \mathbb{R}$ sind die Grenzwerte

$$\lim_{x \rightarrow x_0} f(x) = a \quad (11.2)$$

$$\lim_{x \rightarrow x_0^-} f(x) = a \quad (11.3)$$

$$\lim_{x \rightarrow x_0^+} f(x) = a \quad (11.4)$$

folgendermaßen definiert: Für alle zulässigen Folgen $(x_n)_{n \in \mathbb{N}}$, die gegen x_0 konvergieren, gilt $\lim_{n \rightarrow \infty} f(x_n) = a$. Für den linksseitigen und den rechtsseitigen Grenzwert (11.3) bzw. (11.4) wird von den zulässigen Folgen noch $x_n \leq x_0$ bzw. $x_n \geq x_0$ verlangt.

Ist $x_0 \in D_f$ und $a = f(x_0)$, dann bedeutet (11.2) die Stetigkeit von f in x_0 .

Gilt $\lim_{x \rightarrow x_0^-} f(x) = f(x_0)$ oder $\lim_{x \rightarrow x_0^+} f(x) = f(x_0)$, so heißt f an der Stelle x_0 linksseitig bzw. rechtsseitig stetig.

Man beachte, daß x_0 jetzt nicht unbedingt zum Definitionsbereich von f gehören muß. Ein Beispiel ist $\lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 1$, vergl. (11.21). Für die einseitigen Grenzwerte haben die zulässigen Folgen $(x_n)_{n \in \mathbb{N}}$ dann $x_n < x_0$ bzw. $x_n > x_0$ zu erfüllen. Für $D_f = (x_0, \infty)$ etwa kann es natürlich keinen linksseitigen Grenzwert geben, wohl aber einen rechtsseitigen. Für $D_f = [x_0, \infty)$ ist der linksseitige Grenzwert trivial, da die einzige zulässige Folge $(x_n)_{n \in \mathbb{N}}$, die gegen x_0 konvergiert, konstant sein muß. Die rechtsseitige Stetigkeit von f an der Stelle x_0 ist in diesem Fall gleichbedeutend mit der Stetigkeit in x_0 .

11.1.3 Bemerkung Man findet auch Definitionen für die Stetigkeit an der Stelle x_0 , die von unserer in einem Aspekt abweicht: Von den zulässigen Folgen $(x_n)_{n \in \mathbb{N}}$ wird verlangt, daß $x_n \neq x_0$ für alle $n \in \mathbb{N}$ zu gelten hat. Das bedeutet dann, daß für Stetigkeitspunkt x_0 nur Häufungspunkte von D_f in Frage kommen. Das sind Punkte, die aus D_f heraus beliebig genau durch Folgen approximiert werden können. Für isolierte Stellen, wie etwa $x_0 = 2$ in $D_f = [0, 1] \cup \{2\}$, stellt sich die Frage nach der Stetigkeit dann gar nicht. In unserer Definition sind Funktionen an isolierten Punkten x_0 automatisch stetig, da die einzigen zulässigen Folgen, die gegen x_0 konvergieren können, konstant sein müssen. Abgesehen von solchen wenig wichtigen Stellen sind die beiden resultierenden Stetigkeitsbegriffe äquivalent. Unsere Definition erfordert nur etwas weniger Fallunterscheidungen.

Wir sind dem Stetigkeitsbegriff schon im Zusammenhang mit den Rechenregeln konvergenter Folgen begegnet, ohne daß wir das an dieser Stelle schon thematisieren konnten. Wir haben z. B.

$\lim_{n \rightarrow \infty} \sqrt{x_n} = \sqrt{x}$ für konvergente positive Folgen $(x_n)_{n \in \mathbb{N}}$ mit Grenzwert x (vergl. Übung 10.1.13). Im Lichte von Definition 11.1.1 heißt das einfach, daß die Wurzel stetig ist. So gesehen stellt jede stetige Funktion eine Rechenregel für konvergente Folgen dar. Man kann das sehr einprägsam durch

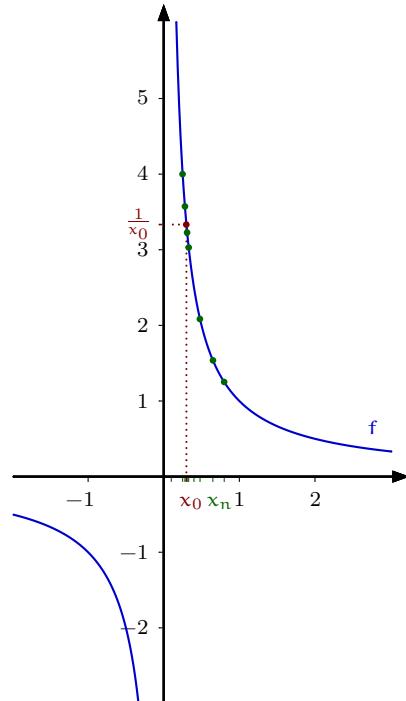
$$\lim_{n \rightarrow \infty} f(x_n) = f\left(\lim_{n \rightarrow \infty} x_n\right) \quad (11.5)$$

wiedergeben – natürlich nur für zulässige Folgen $(x_n)_{n \in \mathbb{N}}$ mit $\lim_{n \rightarrow \infty} x_n \in D_f$.

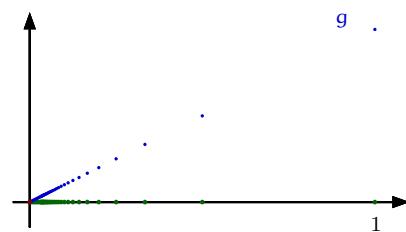
11.1.4 Beispiele stetiger Funktionen Abbildung 11.1

veranschaulicht, was man sich normalerweise unter stetigen Funktionen vorstellt. Das Bild trifft den Sachverhalt auch ziemlich gut, wenn D_f ein abgeschlossenes Intervall ist. Allerdings ist der glatte Kurvenverlauf nicht unbedingt ein typisches Merkmal stetiger Funktionen. Abbildung 11.8 zeigt eine stetige Funktion, die sozusagen nur aus Knicken besteht: Egal wie stark man einen Kurventeil vergrößert, man wird immer eine gezackte Linie finden (auch wenn die Skizze das nicht wirklich wiedergeben kann). Hier zeigt sich bereits, daß es mehr stetige Funktionen gibt, als nach landläufiger Meinung zu vermuten wäre, die von Vorstellungen stammen, wie etwa, daß eine stetige Funktion *ohne abzusetzen zu zeichnen* sein sollte. Sehen wir uns weitere Beispiele an.

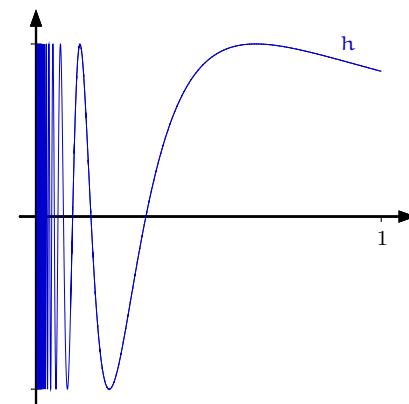
i) $f(x) := \frac{1}{x}$, mit $D_f := \mathbb{R} \setminus \{0\}$ ist eine stetige Funktion auf D_f . Sie ist aber nicht ohne abzusetzen zu zeichnen, denn erstens hat niemand einen ausreichend langen Arm und zweitens ist da noch die Definitionslücke bei 0. Gemäß unserer Definition ist sie trotzdem stetig, denn Stetigkeit ist zunächst eine lokale Eigenschaft, die sich in kleinen Umgebungen der einzelnen Stetigkeitspunkte entscheidet. Dabei ist die Stelle $x_0 = 0$ völlig unproblematisch. Da sie nicht zum Definitionsbereich gehört, stellt sich die Frage nach der Stetigkeit von f an dieser Stelle überhaupt nicht. Um jede andere Stelle $x_0 \neq 0$ finden wir immer ausreichend Platz, um sie durch zulässige konvergente Folgen $(x_n)_{n \in \mathbb{N}}$ zu approximieren. Die Stetigkeit ergibt sich hier einfach aus den Rechenregeln konvergenter Folgen: $\lim_{n \rightarrow \infty} f(x_n) = \lim_{n \rightarrow \infty} \frac{1}{x_n} = \frac{1}{x_0} = f(x_0)$.



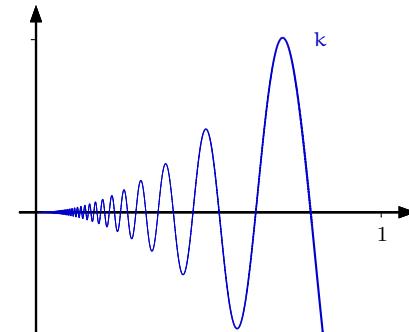
ii) Sei $D_g := \{\frac{1}{n} \mid n \in \mathbb{N}\} \cup \{0\}$ und $g(x) := \frac{x}{2}$. Dann ist g stetig. Die Stetigkeit ist nur für $x = 0$ zu überprüfen. Eine zulässige Nullfolge ist von der Art $(\frac{1}{m_k})_{k \in \mathbb{N}}$, mit einer Folge $(m_k)_{k \in \mathbb{N}}$ natürlicher Zahlen, die bestimmt gegen ∞ divergiert. D. h., für jedes $n \in \mathbb{N}$ gibt es ein $k_n \in \mathbb{N}$, so daß $m_k \geq n$ für alle $k \geq k_n$ gilt. Daher haben wir $g(\frac{1}{m_k}) = \frac{1}{2m_k} \leq \frac{1}{2n}$ für alle $k \geq k_n$ und alle $n \in \mathbb{N}$. Das zeigt $g(\frac{1}{m_k}) \xrightarrow{k \rightarrow \infty} 0 = g(0)$.



iii) Die Funktion $h(x) := 0.5 \sin(\frac{1}{x})$ mit dem Definitionsbereich $D_h := (0, 1]$ ist stetig, denn h ist die Verkettung der beiden stetigen Funktionen \sin und $x \mapsto \frac{1}{x}$ (vergl. Beispiel 11.1.12 iv) und Abschnitt 11.1.14). Für $x \rightarrow 0$ durchläuft $\frac{1}{x}$ die Werte aus $[1, \infty)$. Die unendlich vielen Oszillationen der Sinusfunktion auf diesem Intervall werden von h in das Intervall $(0, 1]$ gepresst. Trotzdem ist die Funktion stetig, denn in einer ausreichend kleinen Umgebung eines jeden $x_0 \in (0, 1]$ gibt es nur endlich viele Oszillationen. Natürlich kann es den einseitigen Grenzwert $\lim_{x \rightarrow 0^+} h(x)$ nicht geben. So erzeugt etwa die zulässige Folge $x_n := \frac{2}{(2n+1)\pi}$ eine Bildfolge $h(x_n) = 0.5 \sin((2n+1)\frac{\pi}{2}) = \frac{(-1)^n}{2}$, die offensichtlich nicht konvergiert. Demnach gibt es keine Möglichkeit, die Funktion stetig auf den Randpunkt 0 fortzusetzen.



iv) Die Funktion $k(x) := x^2 \sin(\frac{10}{x})$ mit dem Definitionsbereich $D_k := (0, 1]$ ist stetig, denn auch k ist die Verkettung von stetigen Funktionen. Die unendlich vielen Oszillationen auf $(0, 1]$ sind weiterhin vorhanden. Im Gegensatz zu h werden sie aber für $x \rightarrow 0$ mit der Amplitudenfunktion $x \mapsto x^2$ auf 0 herunter gedämpft. Darauf kann man diese Funktion stetig auf den Randpunkt 0 durch $k(0) := 0$ fortsetzen: Für eine zulässige Nullfolge $(x_n)_{n \in \mathbb{N}}$ gilt nämlich $|k(x_n)| = x_n^2 |\sin(\frac{10}{x_n})| \leq x_n^2 \xrightarrow{n \rightarrow \infty} 0$.



Um für eine Funktion f , die auf $D_f \subseteq \mathbb{R}$ definiert ist, aber durchaus Werte in einem normierten Raum annehmen darf (üblicherweise \mathbb{R}, \mathbb{C} , oder \mathbb{C}^n), die Existenz eines Grenzwertes $\lim_{x \rightarrow x_0} f(x)$ nachzuweisen, ist es mitunter hilfreich, wenn man sich dabei auf die Existenz und die Gleichheit der beiden *einseitigen Grenzwerte* $\lim_{x \rightarrow x_0^-} f(x)$ und $\lim_{x \rightarrow x_0^+} f(x)$ beschränken kann.

11.1.5 Satz Eine Funktion $f : D \rightarrow Y$ mit Werten in einem normierten Raum Y und einem Definitionsbereich $D_f \subseteq \mathbb{R}$ ist genau dann in $x_0 \in D_f$ stetig, wenn sie dort rechtsseitig und linksseitig stetig ist:

$$\lim_{x \rightarrow x_0^-} f(x) = \lim_{x \rightarrow x_0^+} f(x) = f(x_0) \quad (11.6)$$

Beweis. Ist f in x_0 stetig, dann gilt natürlich (11.6), da $\lim_{x \rightarrow x_0} f(x) = f(x_0)$ laut Definition ja $\lim_{n \rightarrow \infty} f(x_n) = f(x_0)$ für alle zulässigen und gegen x_0 konvergenten Folgen $(x_n)_{n \in \mathbb{N}}$ bedeutet. Das umfaßt natürlich auch die Folgen, die sich nur von einer Seite dem Wert x_0 nähern.

Für die Umkehrung gehen wir davon aus, daß es sich bei x_0 um einen Häufungspunkt von D_f handelt, der von beiden Seiten durch zulässige Folgen approximierbar ist, denn andernfalls ist die Aussage des Satzes trivial. Wir nehmen an, daß f eine Funktion ist, für die (11.6) gilt, die aber in x_0 nicht stetig ist. Dann muß es eine zulässige und gegen x_0 konvergente Folge $(x_n)_{n \in \mathbb{N}}$ geben, für die $(f(x_n))$ nicht gegen $f(x_0)$ konvergiert. Es gibt also ein $\varepsilon > 0$, so daß $\|f(x_n) - f(x_0)\| \geq \varepsilon$ für unendlich viele $n \in \mathbb{N}$ gilt. Das heißt, wir können eine Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$ von $(x_n)_{n \in \mathbb{N}}$ bilden (die natürlich ebenfalls x_0 zum Grenzwert hat), für die $(f(x_{n_k}))_{k \in \mathbb{N}}$ nicht gegen $f(x_0)$ konvergiert. Sollte $x_{n_k} \leq x_0$, oder $x_{n_k} \geq x_0$ für fast alle $k \in \mathbb{N}$ gelten, so haben wir einen Widerspruch zu (11.6). Also müssen auf beiden Seiten von x_0 unendlich viele x_{n_k} anzutreffen sein. Wir erhalten sofort den selben Widerspruch, wenn wir die Teilfolge in zwei weitere Teilstufen aufteilen, eine mit Folgengliedern $x_{n_k} \leq x_0$ und eine mit $x_{m_k} > x_0$. \square

Der Satz bleibt auch in der Version $\lim_{x \rightarrow x_0} f(x) = a \Leftrightarrow \lim_{x \rightarrow x_0^-} f(x) = a = \lim_{x \rightarrow x_0^+} f(x)$ richtig. In diesem Fall muß x_0 nicht zu D_f gehören, aber es muß von beiden Seiten beliebig genau durch zulässige Folgen approximiert werden können. Eine solche Situation tritt typischerweise dann auf, wenn f links und rechts von x_0 definiert ist und stetig auf x_0 fortgesetzt werden soll. Die Funktion $f(x) := \frac{\sin(x)}{x}$ für $x \neq x_0 := 0$ ist ein Beispiel dafür. Sie läßt sich durch den Wert $f(0) := 1$ stetig auf 0 fortsetzen (vergl. (11.21)).

11.1.6 Satz Für eine Funktion $f : D \rightarrow Y$, mit Werten in einem normierten Raum Y und einem Definitionsbereich $D_f \subseteq \mathbb{R}$ ist die Existenz von $\lim_{x \rightarrow x_0} f(x) = a$ äquivalent zu

$$\lim_{x \rightarrow x_0^-} f(x) = \lim_{x \rightarrow x_0^+} f(x) = a. \quad (11.7)$$

Beweis. Der Beweis verläuft wie der für Satz 11.1.5. Man hat nur $f(x_0)$ durch a zu ersetzen. \square

11.1.7 Das ε - δ -Kriterium Wir haben die Definition der Stetigkeit auf dem Folgenbegriff aufgebaut, weil wir uns mit diesem bereits weitgehend vertraut gemacht haben, und weil die Definition dadurch sehr anschaulich wird: *Die Folge der Funktionswerte einer konvergenten Folge konvergiert gegen den Funktionswert ihres Grenzwerts.* Mitunter ist es aber vorteilhaft, wenn man sich nur mit den Umgebungen möglicher Stetigkeitspunkte auseinandersetzen muß. Das wird durch die folgende äquivalente Version des Stetigkeitsbegriffs ermöglicht.

11.1.8 Satz (ε - δ -Kriterium) Unter den Voraussetzungen von Definition 11.1.1 ist eine Funktion f an der Stelle $x_0 \in D_f$ genau dann stetig, wenn es zu jedem $\varepsilon > 0$ ein $\delta > 0$ gibt, so daß für alle $x \in D_f$ mit $\|x - x_0\| < \delta$ die Abschätzung $\|f(x) - f(x_0)\| < \varepsilon$ gilt.

Mit Quantoren läßt sich das wie folgt formulieren:

$$\forall \varepsilon > 0 \exists \delta > 0 \forall_{x \in D_f} \|x - x_0\| < \delta \quad \|f(x) - f(x_0)\| < \varepsilon \quad (11.8)$$

Anschaulich heißt das:

Zu jeder noch so kleinen ε -Umgebung $U_\varepsilon(f(x_0)) = \{y \in Y \mid \|f(x_0) - y\| < \varepsilon\}$ des Bildpunktes $f(x_0)$ läßt sich jeweils eine passende δ -Umgebung $U_\delta(x_0) = \{x \in D_f \mid \|x - x_0\| < \delta\}$ des Urbildpunktes x_0 finden, die durch f ganz in $U_\varepsilon(f(x_0))$ abgebildet wird.

Beweis. Wir führen den Beweis in folgender Form

$$\lim_{x \rightarrow x_0} f(x) = a \Leftrightarrow \quad (11.9)$$

$$\forall \varepsilon > 0 \exists \delta > 0 \forall_{x \in D_f} \quad \|f(x) - a\| < \varepsilon,$$

die für $a = f(x_0)$ die Behauptung ergibt.

Zunächst die Richtung \Leftarrow :

Dafür sei $(x_n)_{n \in \mathbb{N}}$ eine zulässige Folge, die gegen x_0 konvergiert. Zu $\varepsilon > 0$ sei $\delta > 0$

nach dem ε - δ -Kriterium gegeben. Dann finden wir ein $n_\delta \in \mathbb{N}$, so daß für alle $n \geq n_\delta$ die Abschätzung $\|x_0 - x_n\| < \delta$ erfüllt ist. Das bedeutet $\|f(x_n) - a\| < \varepsilon$ für alle $n \geq n_\delta$. Damit haben wir bereits $\lim_{n \rightarrow \infty} f(x_n) = a$ gezeigt. Da die gegen x_0 konvergente Folge $(x_n)_{n \in \mathbb{N}}$ beliebig war, haben wir auch $\lim_{x \rightarrow x_0} f(x) = a$ bewiesen.

Jetzt die andere Richtung: Wir gehen davon aus, daß für jede Folge $(x_n)_{n \in \mathbb{N}}$ aus D_f , die gegen x_0 konvergiert, der Grenzwert $\lim_{n \rightarrow \infty} f(x_n)$ existiert und mit a übereinstimmt. Wir nehmen an, die rechte Seite sei nicht erfüllt, d. h., wir nehmen an, es gäbe ein $\varepsilon > 0$, so daß für alle $\delta > 0$ jeweils ein $x \in D_f$ mit der Eigenschaft $\|x - x_0\| < \delta$, aber $\|f(x) - a\| \geq \varepsilon$ zu finden ist.

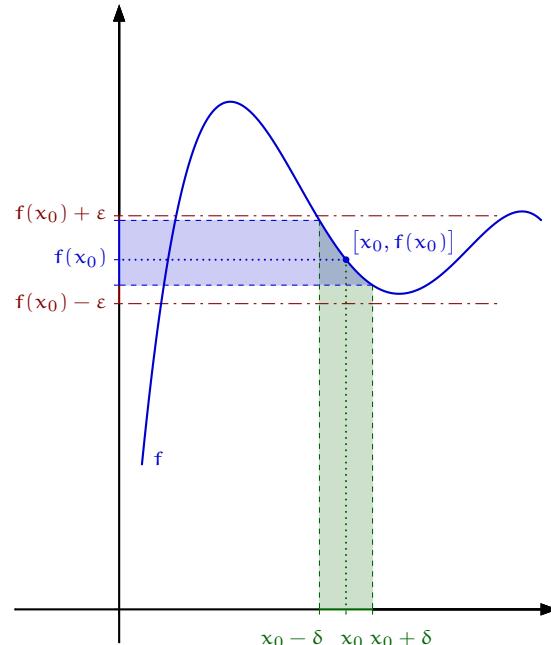
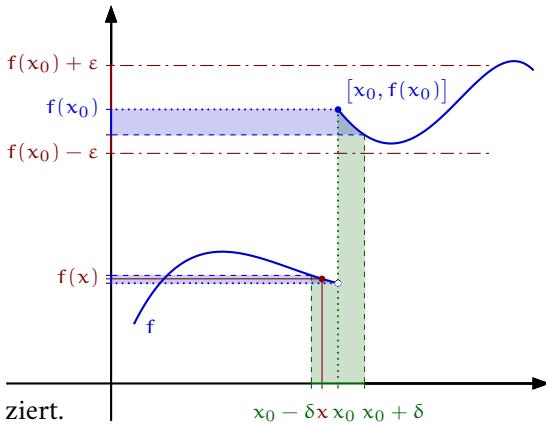


Abb. 11.2 Eine stetige Funktion



Wir wählen eine Folge immer kleiner werdender $\delta_n > 0$, etwa $\delta_n := \frac{1}{n}$. Dann gibt es $x_n \in D_f$ mit $\|x_n - x_0\| < \frac{1}{n}$ und $\|f(x_n) - a\| \geq \varepsilon$. Nach dem Sandwich-Prinzip 10.1.5 konvergiert $(x_n)_{n \in \mathbb{N}}$ gegen x_0 . Laut Voraussetzung müßte $\lim_{n \rightarrow \infty} f(x_n) = a$ gelten. Das steht im Widerspruch dazu, daß $\|f(x_n) - a\|$ keine Nullfolge ist. Damit haben wir nachgewiesen, daß die linke Seite unserer Behauptung 11.9 auch die rechte impliziert. \square

Abb. 11.3 Eine im Punkt x_0 unstetige Funktion

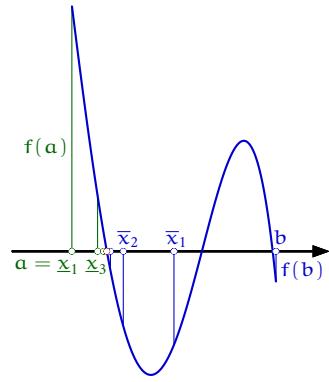
11.1.9 Satz (Satz vom Maximum) Jede auf einem abgeschlossenen, beschränkten Intervall $[a, b]$ stetige reellwertige Funktion hat dort ein Maximum und ein Minimum.

Beweis. Wir führen den Beweis durch Widerspruch. Wir nehmen zunächst an, daß es eine stetige, o. B. d. A. nach oben unbeschränkte Funktion f auf dem Intervall $[a, b]$ gibt. Dann finden wir für jedes $n \in \mathbb{N}$ ein $x_n \in [a, b]$ mit $f(x_n) \geq n$. Die so gebildete Folge $(x_n)_{n \in \mathbb{N}}$ von Urbildern aus $[a, b]$ ist natürlich beschränkt und hat daher nach dem Satz von BOLZANO-WEIERSTRAS eine konvergente Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$ mit Grenzwert $x \in [a, b]$. Da f stetig ist, ist auch die Bildfolge $(f(x_{n_k}))_{k \in \mathbb{N}}$ konvergent, mit Grenzwert $f(x)$, und daher insbesondere beschränkt. Das steht im Widerspruch zu der Eigenschaft, daß laut Konstruktion $f(x_{n_k}) \geq n_k \geq k$ für alle $k \in \mathbb{N}$ gelten muß. f ist also nach oben beschränkt und nach unten ebenfalls, da wir unsere Überlegungen ja nur auf $-f$ anwenden müssen. Das bedeutet, die Wertemenge $W := \{f(x) \mid x \in [a, b]\}$ hat nach dem Vollständigkeitsaxiom ein Supremum w . Wir zeigen, daß w sogar ein Maximum von f ist. Dazu wählen wir in jedem Intervall $[w - \frac{1}{n}, w]$ einen Funktionswert $f(\tilde{x}_n)$ aus. Laut Konstruktion konvergiert die Folge $(f(\tilde{x}_n))_{n \in \mathbb{N}}$ gegen w . Sie erzeugt eine Folge $(\tilde{x}_n)_{n \in \mathbb{N}}$ in $[a, b]$, für die es eine konvergente Teilfolge $(\tilde{x}_{n_k})_{k \in \mathbb{N}}$ mit Grenzwert $\tilde{x} \in [a, b]$ gibt. Die Stetigkeit von f bedeutet $f(\tilde{x}) = \lim_{k \rightarrow \infty} f(\tilde{x}_{n_k}) = w$, d. h. w ist selbst ein Funktionswert und ist daher nicht nur das Supremum, sondern das Maximum aller Funktionswerte. Ein Minimum ergibt sich, wenn wir unsere Überlegungen mit dem Infimum von W wiederholen. \square

Bemerkung: Bei genauem Lesen des Beweises stellt man fest, daß wir eigentlich nirgends von der Eigenschaft Gebrauch gemacht haben, daß f auf einem Intervall definiert ist. Tatsächlich mußte nur der Satz von BOLZANO-WEIERSTRAS zur Verfügung stehen. Deshalb läßt sich die Aussage des Satzes auch für Funktionen $f: \mathbb{R}^n \rightarrow \mathbb{R}$ beweisen, sofern sie auf einer abgeschlossenen Menge definiert sind, auf der der Satz von BOLZANO-WEIERSTRAS gilt.

11.1.10 Satz (Zwischenwertsatz) Eine reelle stetige Funktion auf dem abgeschlossenen Intervall $[a, b]$ nimmt alle Werte zwischen $f(a)$ und $f(b)$ als Funktionswerte an.

Beweis: Wir zeigen zunächst eine scheinbar speziellere Situation, nämlich, daß zwischen a und b eine Nullstelle von f zu finden sein muß, wenn o. B. d. A. $f(a) > 0$ und $f(b) < 0$ ist. Dazu halbieren wir $[a, b]$. Ist der Funktionswert am Mittelpunkt des Intervalls Null, dann sind wir fertig. Andernfalls wählen wir dasjenige Teilintervall $I_1 := [\underline{x}_1, \bar{x}_1]$, dessen Randpunkte Funktionswerte mit verschiedenen Vorzeichen aufweist. Dabei ist $f(\underline{x}_1) > 0$ und $f(\bar{x}_1) < 0$. Mit dem so erhaltenen Intervall verfahren wir auf gleiche Weise und erhalten entweder eine Nullstelle von f , oder ein weiteres Intervall $I_2 := [\underline{x}_2, \bar{x}_2]$, das wiederum demselben Verfahren unterzogen wird usw. Dieser Vorgang ergibt entweder direkt eine Nullstelle und bricht dann ab, oder erzeugt eine konvergente Intervallteilung $(I_n)_{n \in \mathbb{N}}$ mit einem Grenzwert $x_* \in [a, b]$. Letzteres bedeutet für die Funktionswerte an den Randpunkten \underline{x}_n und \bar{x}_n von I_n dann $f(\underline{x}_n) > 0$ und $f(\bar{x}_n) < 0$. Die Stetigkeit von f und $x_* = \lim_{n \rightarrow \infty} \underline{x}_n = \lim_{n \rightarrow \infty} \bar{x}_n$ bedeutet $f(x_*) = \lim_{n \rightarrow \infty} f(\underline{x}_n) \geq 0$ und $f(x_*) = \lim_{n \rightarrow \infty} f(\bar{x}_n) \leq 0$, also $f(x_*) = 0$. x_* ist die gesuchte Nullstelle. Nun zur allgemeinen Situation. Falls $f(a) = f(b)$ gelten sollte, ist nichts zu zeigen. Wir dürfen also $f(a) > f(b)$ annehmen (andernfalls ersetzen wir f durch $-f$) und wählen ein $y \in (f(b), f(a))$. Die Funktion $g(x) := f(x) - y$ hat die Eigenschaft $g(a) = f(a) - y > 0$ und $g(b) = f(b) - y < 0$. Nach unseren bisherigen Überlegungen gibt es ein $x_* \in [a, b]$, mit $g(x_*) = 0$, also $f(x_*) = y$. \square

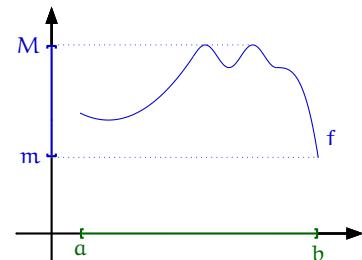


Der Zwischenwertsatz ist die mathematische Fassung der gängigen Vorstellung, daß man eine stetige Funktion auf einem abgeschlossenen Intervall $[a, b]$ ohne abzusetzen zeichnen können sollte.

11.1.11 Korollar Sei f eine stetige Funktion auf einem Intervall I . Dann ist das Bild $f(I)$ ebenfalls ein Intervall. Ist I abgeschlossen und beschränkt, so gilt das auch für das Bild, d. h. $f(I) = [m, M]$, mit geeigneten $m \leq M$.

Beweis. Wir können davon ausgehen, daß $f(I)$ nicht einpunktig ist. Für zwei Punkte $y_1 < y_2$ aus $f(I)$ ist jedes $y \in (y_1, y_2)$ nach dem Zwischenwertsatz ein Funktionswert und gehört daher zu $f(I)$. Also gilt $[y_1, y_2] \subseteq f(I)$. Nach Lemma 10.1.29 ist $f(I)$ ein Intervall.

Auf einem abgeschlossenen und beschränkten Intervall $I := [a, b]$ hat f , nach dem Satz vom Maximum 11.1.9, ein Minimum $m \in f(I)$ und ein Maximum $M \in f(I)$. Daher gilt $[m, M] \subseteq f(I)$. Kein Wert außerhalb von $[m, M]$ kann zu $f(I)$ gehören, denn das würde der Minimalität von m oder der Maximalität von M widersprechen. Folglich ist $f(I) = [m, M]$. \square



11.1.12 Beispiel

- i) Die Potenzfunktionen $x \mapsto x^k$ sind für alle $k \in \mathbb{Z}$ stetig, denn aus $x_n \xrightarrow{n \rightarrow \infty} x$ folgt nach den Rechenregeln 10.1.9 für konvergente Folgen $x_n^k \xrightarrow{n \rightarrow \infty} x^k$ (natürlich nur für $x \neq 0$, falls $k < 0$ sein sollte).

ii) Alle Polynome $p(x) = a_0 + a_1x + \dots + a_kx^k$ sind stetig, was mit i) und 10.1.9 leicht zu sehen ist.

iii) Die Wurzelfunktion $x \mapsto \sqrt[k]{x}$ ist stetig. Sie ist die Umkehrfunktion von $x \mapsto x^k$ mit dem Definitionsbereich \mathbb{R}_0^+ für gerade und \mathbb{R} für ungerade k . Da diese Funktion stetig ist, folgt die Behauptung aus Satz 11.1.24.

iv) Sind f und g auf einem gemeinsamen Definitionsbereich stetig, dann gilt das auch für $f \pm g$, $f \cdot g$, $\frac{f}{g}$ (an jeder Stelle x mit $g(x) \neq 0$). Lassen sich f und g verketten, dann ist auch $f \circ g$ stetig. Bis auf die letzte Aussage wird das wieder von Satz 10.1.9 abgedeckt. Die Stetigkeit der Verkettung: Aus $x \xrightarrow{n \rightarrow \infty} x$ folgt wegen der Stetigkeit von g : $g(x_n) \xrightarrow{n \rightarrow \infty} g(x)$. Die Stetigkeit von f ergibt dann $f \circ g(x_n) = f(g(x_n)) \xrightarrow{n \rightarrow \infty} f(g(x)) = f \circ g(x)$, d.h. die Stetigkeit von $f \circ g$. Das erweitert die als stetig erkannten Funktionen beträchtlich. Dazu gehören jetzt auch $x \mapsto \frac{x^2-1}{x^2+1}$, $x \mapsto \sqrt{x^2-1}$, $x \mapsto \frac{1}{\sqrt{x^2+1}}$, etc.

11.1.13 Lemma Eine Funktion f auf $[a, b]$ sei an der Stelle $x_0 \in [a, b]$ stetig und es gelte $f(x_0) \neq 0$. Dann gibt es eine Umgebung $U_\varepsilon(x_0) = (x_0 - \varepsilon, x_0 + \varepsilon)$ von x_0 , so daß $f(x) \neq 0$ für alle $x \in U_\varepsilon(x_0) \cap [a, b]$ gilt.

Beweis. Gäbe es kein $\varepsilon > 0$ mit der genannten Eigenschaft, so ließe sich für jedes $n \in \mathbb{N}$ eine Nullstelle $x_n \in (x_0 - \frac{1}{n}, x_0 + \frac{1}{n}) \cap [a, b]$ von f finden. Die Folge $(x_n)_{n \in \mathbb{N}}$ konvergiert offensichtlich gegen x_0 , so daß die Stetigkeit von f auf den Widerspruch $f(x_0) = \lim_{n \rightarrow \infty} f(x_n) = 0$ führt. \square

Für diesen Beweis benötigen wir die Eigenschaft, daß f eine reellwertige Funktion ist, eigentlich gar nicht. Für eine stetige Funktion $f: X \rightarrow Y$ zwischen den normierten Räumen X und Y ist $f(x) \neq 0$ für alle $x \in D_f \cap U_\varepsilon(x_0)$, falls $f(x_0) \neq 0$ gilt. Dabei ist die ε -Umgebung $U_\varepsilon(x_0)$ jetzt durch $U_\varepsilon := \{x \in X \mid \|x - x_0\| < \varepsilon\}$ definiert.

11.1.14 Die trigonometrischen Funktionen Sinus, Kosinus und Tangens

In einem rechtwinkligen Dreieck ist nach dem Strahlensatz das Verhältnis zweier Seiten unabhängig vom Maßstab, in dem es gezeichnet ist. Daher sind die vier Verhältnisse $\frac{b}{c}$, $\frac{a}{c}$, $\frac{b}{a}$ und $\frac{a}{b}$ nur noch eine Funktion des Winkels $x \in (0, \frac{\pi}{2})$:

$$\sin(x) := \frac{b}{c}, \quad (11.10)$$

$$\cos(x) := \frac{a}{c}, \quad (11.11)$$

$$\tan(x) := \frac{b}{a} = \frac{b \cdot \frac{1}{c}}{a \cdot \frac{1}{c}} = \frac{\sin(x)}{\cos(x)}, \quad (11.12)$$

$$\cot(x) := \frac{a}{b} = \frac{1}{\tan(x)}. \quad (11.13)$$

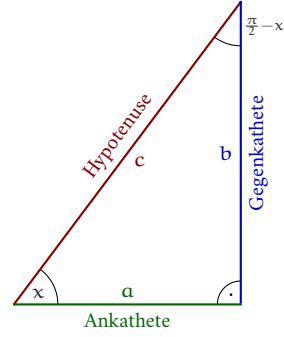


Abb. 11.4 Sinus, Kosinus und Tangens

Nach dem Satz von PYTHAGORAS ist $b^2 + a^2 = c^2$, also auch $\frac{b^2}{c^2} + \frac{a^2}{c^2} = \frac{c^2}{c^2} = 1$. Diese einfache Beobachtung ergibt den zentralen Zusammenhang zwischen sin und cos

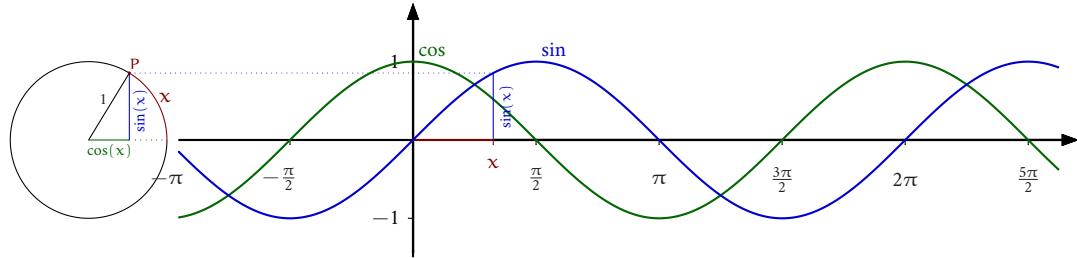
$$\sin^2(x) + \cos^2(x) = 1. \quad (11.14)$$

Da für den Gegenwinkel $\frac{\pi}{2} - x$ von x die Rollen von Gegen- und Ankathete vertauscht sind, gilt

$$\sin(x) = \cos\left(\frac{\pi}{2} - x\right), \quad (11.15)$$

$$\tan(x) = \cot\left(\frac{\pi}{2} - x\right). \quad (11.16)$$

Man kann sich $\sin(x)$ und $\cos(x)$ als die Länge der Gegen- bzw. Ankathete eines Dreiecks veranschaulichen, wenn man von der freien Wahl des Maßstabs Gebrauch macht und die Länge der Hypotenuse 1 wählt. Dann ist der Winkel x die Länge eines Bogens auf einem Kreis mit Radius 1, dessen Endpunkt P die Ecke eines rechtwinkligen Dreiecks markiert. Die Koordinaten von P lauten dann $[\cos(x), \sin(x)]$.



Diese Beobachtung kann man nutzen, um die Definition von $\sin(x)$ und $\cos(x)$ über den Bereich $(0, \frac{\pi}{2})$ hinaus fortzusetzen, indem man vereinbart, unter $\sin(x)$ die y-Koordinate und unter $\cos(x)$ die x-Koordinate eines Punktes P zu verstehen, der die Bogenlänge x vom Schnittpunkt des Einheitskreises mit der positiven x-Achse zurückgelegt hat. Dabei wird x positiv gerechnet, wenn die Bewegung im mathematisch positiven Sinne (also entgegen dem Uhrzeigersinn), und negativ wenn sie im mathematisch negativen Sinne erfolgt. Auf diese Weise werden sin und cos auf ganz \mathbb{R} definierte, 2π -periodische Funktionen: $\sin(x + 2\pi) = \sin(x)$, $\cos(x + 2\pi) = \cos(x)$ für alle $x \in \mathbb{R}$.

11.1.15 Die Additionssätze der trigonometrischen Funktionen

Aus Abb. 11.5 lässt sich $g = \sin(x)\sin(y)$, $f = \cos(x)\sin(y)$, $\sin(x+y) = \sin(x)\cos(y)+f$ und $\cos(x+y) = \cos(x)\cos(y)-g$ ablesen. Eingesetzt ergeben sich die Additionssätze für Sinus und Kosinus:

$$\sin(x+y) = \sin(x)\cos(y) + \cos(x)\sin(y), \quad (11.17)$$

$$\cos(x+y) = \cos(x)\cos(y) - \sin(x)\sin(y). \quad (11.18)$$

Daraus folgen die Additionssätze für Tangens und Kotangens:

$$\tan(x+y) = \frac{\tan(x) + \tan(y)}{1 - \tan(x)\tan(y)}, \quad (11.19)$$

$$\cot(x+y) = \frac{\cot(x)\cot(y) - 1}{\cot(x) + \cot(y)}, \quad (11.20)$$

denn

$$\tan(x+y) = \frac{\sin(x+y)}{\cos(x+y)} = \frac{\sin(x)\cos(y) + \cos(x)\sin(y)}{\cos(x)\cos(y) - \sin(x)\sin(y)} \cdot \frac{\frac{1}{\cos(x)\cos(y)}}{\frac{1}{\cos(x)\cos(y)}}$$

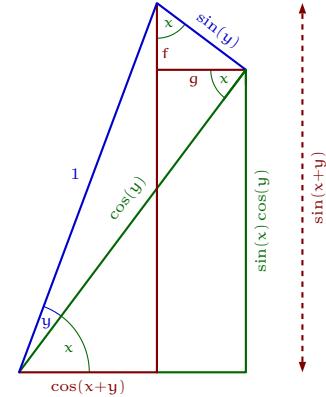


Abb. 11.5 Zum Additionssatz

$$= \frac{\frac{\sin(x)\cos(y)}{\cos(x)\cos(y)} + \frac{\cos(x)\sin(y)}{\cos(x)\cos(y)}}{\frac{\cos(x)\cos(y)}{\cos(x)\cos(y)} - \frac{\sin(x)\sin(y)}{\cos(x)\cos(y)}} = \frac{\tan(x) + \tan(y)}{1 - \tan(x)\tan(y)}.$$

Eine ähnliche Rechnung ergibt 11.20.

11.1.16 Die Stetigkeit der trigonometrischen Funktionen

Abb. 11.6 entnimmt man

$$0 \leq |\sin(x)| \leq d \leq |x|, \\ 0 \leq 1 - \cos(x) \leq |x|.$$

Das zeigt die Stetigkeit von \sin und \cos an der Stelle $x = 0$:

$$\lim_{x \rightarrow 0} \sin(x) = 0 = \sin(0), \quad \lim_{x \rightarrow 0} \cos(x) = 1 = \cos(0).$$

Mit Hilfe von (11.17) und (11.18) folgt daraus bereits die Stetigkeit von \sin und \cos an jeder Stelle x :

$$\begin{aligned} \sin(x+h) - \sin(x) &= \sin(x)\cos(h) + \cos(x)\sin(h) - \sin(x) \\ &= \sin(x)(\cos(h) - 1) + \cos(x)\sin(h) \xrightarrow{h \rightarrow 0} 0, \\ \cos(x+h) - \cos(x) &= \cos(x)\cos(h) - \sin(x)\sin(h) - \cos(x) \\ &= \cos(x)(\cos(h) - 1) + \sin(x)\sin(h) \xrightarrow{h \rightarrow 0} 0. \end{aligned}$$

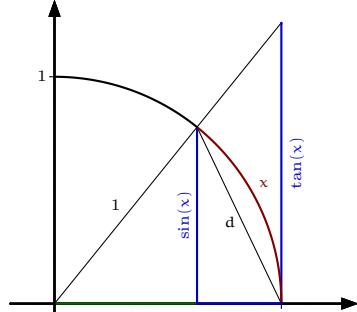


Abb. 11.6 Zur Stetigkeit von \sin und \cos

Aus den Rechenregeln für stetige Funktionen (Beispiel 11.1.12 iv)) erhalten wir auch die Stetigkeit von $x \mapsto \tan(x) = \frac{\sin(x)}{\cos(x)}$ und $x \mapsto \cot(x) = \frac{1}{\tan(x)}$.

Aus Abbildung 11.6 lässt sich eine weitere wichtige Abschätzung gewinnen. Dazu vergleichen wir die Flächeninhalte $\frac{1}{2}\sin(x)\cos(x)$ bzw. $\frac{1}{2}\tan(x)$ der beiden rechtwinkligen Dreiecke mit den Katheten $\cos(x)$ und $\sin(x)$, bzw. 1 und $\tan(x)$ mit dem Flächeninhalt $\frac{x}{2}$ des Kreissektors zum Winkel $x > 0$:

$$\frac{1}{2}\sin(x)\cos(x) \leq \frac{x}{2} \leq \frac{1}{2}\tan(x) = \frac{1}{2}\frac{\sin(x)}{\cos(x)}.$$

Nach Division durch $\frac{1}{2}\sin(x) > 0$ erhalten wir, zunächst nur für $x > 0$, da es sich bei $x \mapsto \frac{x}{\sin(x)}$ aber um eine gerade Funktion handelt, auch für $x < 0$:

$$\cos(x) \leq \frac{x}{\sin(x)} \leq \frac{1}{\cos(x)}.$$

Die Stetigkeit von \cos und das Sandwich-Prinzip 10.1.5 zeigen $\lim_{x \rightarrow 0} \frac{x}{\sin(x)} = 1$, (vergl. Übung 11.9.3). Daher gilt

$$\lim_{x \rightarrow 0} \frac{\sin(x)}{x} = 1 \tag{11.21}$$

$$\lim_{x \rightarrow 0} \frac{1 - \cos(x)}{x} = 0 \tag{11.22}$$

Dabei ist (11.22) die Konsequenz folgender Überlegung:

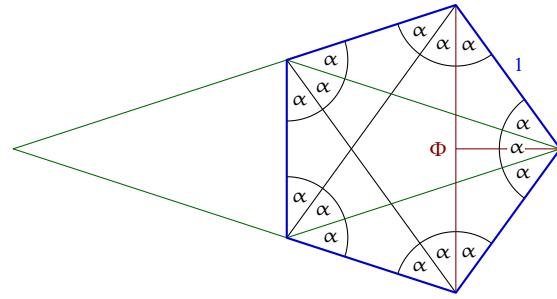
$$\frac{1 - \cos(x)}{x} = \frac{(1 - \cos(x))(1 + \cos(x))}{x(1 + \cos(x))} = \frac{1 - \cos^2(x)}{x(1 + \cos(x))} = \sin(x) \frac{\sin(x)}{x} \frac{1}{1 + \cos(x)} \xrightarrow{x \rightarrow 0^+} 0 \cdot 1 \cdot \frac{1}{2}.$$

11.1.17 A Bestimmen Sie mit dem Intervallteilungsverfahren eine Nullstelle im Intervall $[1.2, 2.25]$ für die Funktion

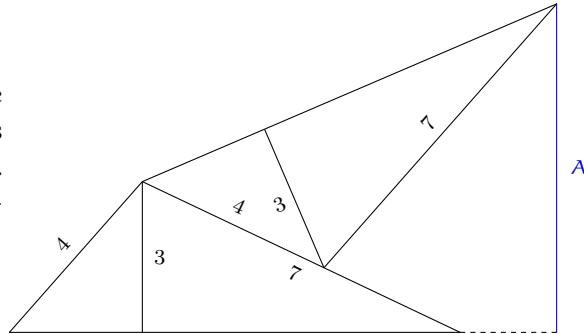
$$f(x) := x^7 - 5.3x^6 + 10.3x^5 - 11.066x^4 + 10.22x^3 - 6.2896x^2 + 0.92x - 0.5236.$$

Fertigen Sie eine Skizze an.

11.1.18 A Machen Sie sich anhand der Skizze klar, daß es sich bei Φ um den goldenen Schnitt $\frac{1}{2}(\sqrt{5} + 1)$ handelt. Berechnen Sie anschließend $\cos\left(\frac{\pi}{5}\right)$, $\sin\left(\frac{\pi}{5}\right)$, $\cos\left(\frac{2\pi}{5}\right)$, $\sin\left(\frac{2\pi}{5}\right)$, $\cos\left(\frac{\pi}{10}\right)$ und $\sin\left(\frac{\pi}{10}\right)$.



11.1.19 A Bestimmen Sie die Länge der Strecke A und zeigen Sie, daß es sich dabei um eine rationale Zahl handelt. Verwenden Sie möglicherweise den sin-Additionssatz (11.17).



11.1.20 Umkehrfunktionen

11.1.21 Definition Eine Funktion $f : X \supseteq D_f \rightarrow Y$ heißt injektiv, falls für $x_1, x_2 \in D_f$ aus $x_1 \neq x_2$ auch $f(x_1) \neq f(x_2)$ folgt.

11.1.22 Lemma Eine injektive Funktion $f : X \supseteq D_f \rightarrow Y$ ist auf ihrer Wertemenge $\{f(x) \mid x \in D_f\}$ umkehrbar, d.h. es gibt eine Funktion $g : Y \supseteq \text{im } f \rightarrow X$ mit der Eigenschaft $f(g(y)) = y$ für alle $y \in \text{im } f$ und $g(f(x)) = x$ für alle $x \in D_f$.

Beweis. Für jedes Element $y \in \text{im } f$ gibt es genau ein $x \in D_f$ mit der Eigenschaft $y = f(x)$. Damit definiert die Vorschrift g , die jedem $y \in \text{im } f$ dieses eindeutig bestimmte $x \in D_f$ zuordnet, eine Funktion von $\text{im } f$ nach X . $g(y)$ ist also der Wert aus D_f , der durch f auf $y \in \text{im } f$ abgebildet wird: $f(g(y)) = y$. Laut Definition von g ist $g(f(x)) = x$ für alle $x \in D_f$. \square

Die Funktion g aus Lemma 11.1.22 heißt *Umkehrfunktion* von f und wird mit f^{-1} bezeichnet. Dabei ist der Exponent -1 als Inverse der Verkettung zweier Funktionen zu verstehen und nicht als Inverse der Multiplikation:

$$f^{-1} \circ f = \text{id}_{|D_f}, \quad (11.23)$$

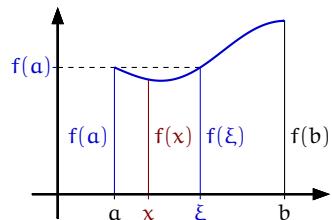
$$f^{-1}(f(x)) = x, \quad (11.24)$$

$\text{id}_{|D_f}$ bzw. $\text{id}_{|\text{im } f}$ sind die Einschränkungen der *identischen Funktion* $x \mapsto x$ auf D_f bzw. $y \mapsto y$ auf $\text{im } f$. So ist die Umkehrfunktion f^{-1} von $f : x \mapsto x^3$ durch die Wurzelfunktion $f^{-1} : y \mapsto \sqrt[3]{y}$ gegeben und nicht etwa durch $y \mapsto \frac{1}{y^3}$.

Bestimmt wird die Umkehrfunktion f^{-1} von f , indem die Zuordnung $x \mapsto f(x)$ umgekehrt wird, also durch Auflösen der Gleichung $y = f(x)$ nach x : $x = f^{-1}(y)$.

11.1.23 Lemma Eine stetige reellwertige Funktion auf einem Intervall ist genau dann injektiv, wenn sie entweder streng monoton wachsend, oder streng monoton fallend ist. Die Umkehrfunktion ist dann ebenfalls entweder streng monoton wachsend, oder streng monoton fallend.

Beweis: Wir können davon ausgehen, daß das Intervall nicht einpunktig ist, denn sonst ist die Aussage des Lemmas trivial. f sei zunächst auf dem Intervall $[a, b]$ stetig und injektiv ($a < b$). Insbesondere ist $f(a) \neq f(b)$. Wir dürfen o. B. d. A. $f(a) < f(b)$ annehmen (sonst ersetzen wir f durch $-f$). Daraus ergibt sich bereits, daß f auf $[a, b]$ streng monoton wächst. Zunächst folgt $f(a) < f(x) < f(b)$ für jedes $x \in (a, b)$. Gäbe es nämlich ein $x \in (a, b)$ mit $f(a) > f(x)$ oder $f(b) < f(x)$, so wäre im ersten Fall $f(a)$ ein Zwischenwert von f auf dem Teilintervall $[x, b]$ und im zweiten $f(b)$ einer für f auf $[a, x]$. In beiden Fällen gäbe es nach dem Zwischenwertsatz ein $\xi \in (x, b)$ bzw. ein $\xi \in [a, x]$, mit $f(a) = f(\xi)$, bzw. $f(b) = f(\xi)$. Das würde jeweils der Injektivität von f widersprechen. Wenden wir dieses Ergebnis auf das Teilintervall $[a, x]$ an, auf das die Voraussetzungen ja ebenfalls zutreffen, so erhalten wir $f(a) < f(y) < f(x)$ für alle $y \in (a, x)$. Also folgt aus $y < x$ die



Ungleichung $f(y) < f(x)$. Damit ist f auf $[a, b]$ streng monoton wachsend.

Ist das Intervall offen, oder halboffen, möglicherweise unbeschränkt, so müssen wir unsere Überlegungen ergänzen. Betrachten wir etwa ein Intervall $[a, d)$. Auf einem Teilintervall $[a, b] \subset [a, d)$ greifen unsere bisherigen Überlegungen. Wir können annehmen, daß f dort streng monoton wächst. Dann muß f auf jedem weiteren Teilintervall $[a, c]$, das $[a, b]$ umfaßt, ebenfalls streng monoton wachsen. Dafür muß nur $f(a) < f(c)$ gelten, wie wir oben gesehen haben. Wäre $f(a) > f(c)$ ($f(a) = f(c)$ ist wegen der Injektivität ausgeschlossen), so könnten wir wie oben darauf schließen, daß f auf $[a, c]$ streng monoton fällt, im Widerspruch dazu, daß f auf dem Teilintervall $[a, b]$ monoton wächst.

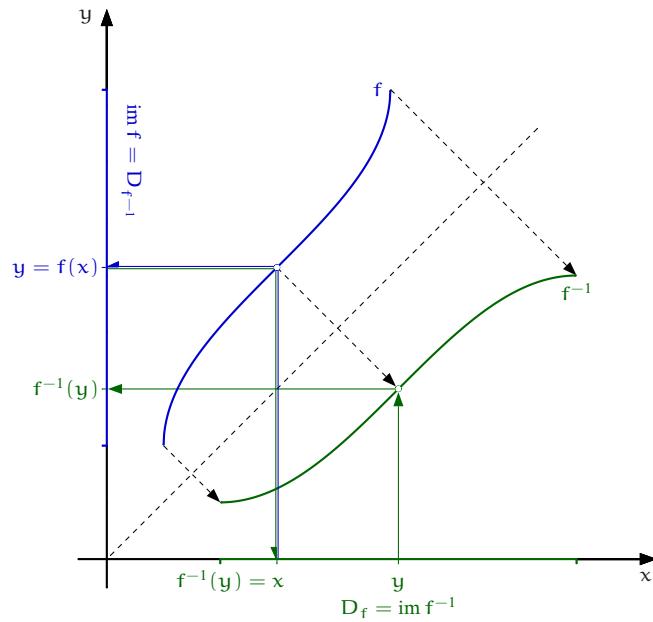
Zwei beliebige Werte $x < y$ aus $[a, d)$ liegen immer in einem geeigneten Intervall $[a, c]$, das $[a, b]$ umfaßt. Daher gilt $f(x) < f(y)$. Also ist f auf dem gesamten Intervall streng monoton wachsend. Für alle anderen Intervalle lassen sich unsere Gedankengänge jetzt leicht anpassen. Daß aus der strengen Monotonie die Injektivität folgt ist eine einfache Übung.

Für zwei Funktionswerte $y_1 = f(x_1) < y_2 = f(x_2)$ muß $x_1 < x_2$ gelten, denn aus $x_1 > x_2$ ($x_1 = x_2$ kommt natürlich nicht in Frage) würde sich der Widerspruch $f(x_1) > f(x_2)$ ergeben, da f ja streng monoton wächst. Also ist $f^{-1}(y_1) = x_1 < x_2 = f^{-1}(y_2)$. Das zeigt, daß auch f^{-1} streng monoton wächst. \square

11.1.24 Satz Ist die reellwertige Funktion f auf einem Intervall injektiv und stetig, so ist ihre Wertmenge ein Intervall, auf dem die Umkehrfunktion ebenfalls stetig ist.

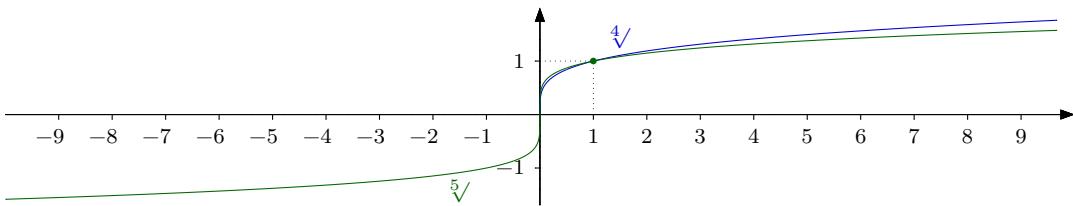
Beweis. Wir gehen natürlich davon aus, daß das Intervall nicht einpunktig ist. Nach Lemma 11.1.23 dürfen wir annehmen, daß f streng monoton wächst. Nach Korollar 11.1.11 ist im f ein Intervall. Wir betrachten eine konvergente Folge $(y_n)_{n \in \mathbb{N}}$ aus im f mit einem Grenzwert $y \in$ im f und wollen $x_n := f^{-1}(y_n) \xrightarrow{n \rightarrow \infty} f^{-1}(y) =: x$ zeigen. Als konvergente Folge ist $(y_n)_{n \in \mathbb{N}}$ beschränkt und liegt in einem geeigneten Intervall $[f(a), f(b)] \subseteq$ im f (man mache sich das klar). Da f^{-1} ebenfalls streng monoton wächst, liegen alle Urbilder x_n in $[a, b]$. Wir nehmen an, daß die Folge $(x_n)_{n \in \mathbb{N}}$ nicht gegen x konvergiert. Da sie beschränkt ist, läßt sich eine konvergente Teilfolge $(x_{n_k})_{k \in \mathbb{N}}$ gewinnen, die einen Grenzwert $\tilde{x} \neq x$ in $[a, b]$ hat. $(x_n)_{n \in \mathbb{N}}$ muß nämlich wenigstens einen Häufungspunkt \tilde{x} haben, der von x verschieden ist, wenn $(x_n)_{n \in \mathbb{N}}$ nicht gegen x konvergieren soll. Die Teilfolge $(y_{n_k})_{k \in \mathbb{N}}$ konvergiert ebenfalls gegen y . Da f stetig ist, erhalten wir einen Widerspruch zur Injektivität von f : $f(\tilde{x}) = \lim_{k \rightarrow \infty} f(x_{n_k}) = \lim_{k \rightarrow \infty} y_{n_k} = y = f(x)$. Also muß für jede der gegen y konvergenten Folgen $(y_n)_{n \in \mathbb{N}}$ der Grenzwert von $(f^{-1}(y_n))_{n \in \mathbb{N}}$ existieren und mit $f^{-1}(y)$ übereinstimmen. Damit ist f^{-1} stetig. \square

Wenn wir den Graphen von f^{-1} in der Weise zeichnen, wie f^{-1} wirkt, d.h., indem wir jedem Punkt der y -Achse, der einen Funktionswert darstellt, waagrecht daneben seinen x -Wert zuordnen, so erhalten wir den Graphen von f . Eine Funktion auf diese Weise zu zeichnen ist natürlich etwas ungewohnt. Um sie, wie jede andere Funktion auch, von der x -Achse aus zu zeichnen, haben wir die Rollen von y - und x -Achse zu vertauschen. Wir können uns das folgendermaßen vorstellen: Wir denken uns das Koordinatensystem und f in zwei identischen Kopien auf zwei durchsichtige Folien gezeichnet, die deckungsgleich übereinander liegen. Wir nehmen nun die obere Folie, blättern sie um und bringen anschließend die Koordinatenachsen wieder zur Deckung (auch in ihrem Durchlaufsinn von links nach rechts bzw. von unten nach oben), wobei nun aber die x -Achse auf der y -Achse und die y -Achse auf der x -Achse des unteren Bildes liegt – das ist eine Drehung um den Ursprung mit dem Winkel $\frac{\pi}{2}$ im Uhrzeigersinn. Als Ergebnis erhalten wir den Graphen von f^{-1} durch Spiegelung des Graphen von f an der ersten Winkelhalbierenden.



11.1.25 Korollar Die Funktionen $\text{id}^n : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$ für gerades $n \in \mathbb{N}$ und $\text{id}^n : \mathbb{R} \rightarrow \mathbb{R}$ für ungerades $n \in \mathbb{N}$, sind stetig, streng monoton wachsend und haben die stetigen, streng monoton wachsenden Umkehrfunktionen $\sqrt[n]{\cdot} : \mathbb{R}_0^+ \rightarrow \mathbb{R}_0^+$, $x \mapsto \sqrt[n]{x}$, bzw. $\sqrt[n]{\cdot} : \mathbb{R} \rightarrow \mathbb{R}$, $x \mapsto \sqrt[n]{x}$.

Beweis. Die Injektivität der Funktionen id^n auf ihren jeweiligen Definitionsbereichen $\mathbb{R}_0^+ = [0, \infty)$ bzw. $\mathbb{R} = (-\infty, \infty)$ haben wir uns schon im Beispiel 2.5.9 überlegt (siehe aber auch Beispiel 11.4.3). Ihre Wertebereiche sind die Intervalle \mathbb{R}_0^+ bzw. \mathbb{R} . Nach Lemma 11.1.23 und Satz 11.1.24 sind die Umkehrfunktionen $\sqrt[n]{\cdot}$ auf diesen Intervallen streng monoton wachsend und stetig. \square

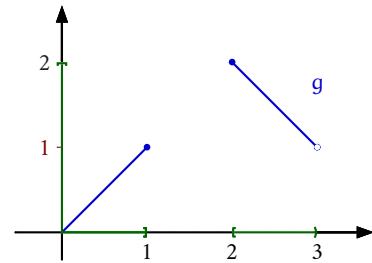
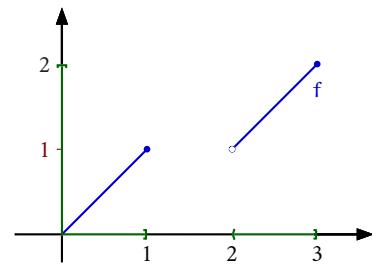


Bemerkung: Wir haben den Begriff Umkehrfunktion schon in Kapitel 2 im Abschnitt 2.5.5 unter dem Gesichtspunkt funktionaler Relationen behandelt.

11.1.26 Beispiel Die Funktion $f(x) := \begin{cases} x & , 0 \leq x \leq 1 \\ x - 1 & , 2 < x \leq 3 \end{cases}$ ist stetig und streng monoton steigend, also stetig und injektiv. Das Bild ist das Intervall $[0, 2]$.

Bis auf die Forderung, daß der Definitionsbereich von f ein Intervall ist, sind die Voraussetzungen von Satz 11.1.24 erfüllt. Trotzdem hat die Umkehrfunktion $f^{-1}(y) = \begin{cases} y & , 0 \leq y \leq 1 \\ y + 1 & , 1 < y \leq 2 \end{cases}$ an der Position $y = 1$ eine Sprungstelle und ist daher nicht mehr stetig. Auf die Forderung des Satzes, daß f auf einem Intervall gegeben sein muß, kann daher im Allgemeinen nicht verzichtet werden.

Die Funktion $g(x) := \begin{cases} x & , 0 \leq x \leq 1 \\ 4 - x & , 2 \leq x < 3 \end{cases}$ ist stetig und injektiv, allerdings nicht streng monoton. Bis auf das Intervall als Definitionsbereich, sind die Bedingungen von Lemma 11.1.23 erfüllt, ohne daß dessen Behauptung gilt. Auch hier ist auf die Intervalleigenschaft des Definitionsbereichs im Allgemeinen nicht zu verzichten.



11.1.27 Funktionenfolgen

11.1.28 Definition Eine Folge $(f_n)_{n \in \mathbb{N}}$ von Funktionen mit einem gemeinsamen Definitionsbereich D konvergiert gegen die Funktion f , falls für alle $x \in D$

$$f(x) = \lim_{n \rightarrow \infty} f_n(x) \quad (11.25)$$

gilt. f heißt (punktweise) Grenzfunktion der Folge $(f_n)_{n \in \mathbb{N}}$.

Die minimale Forderung, die wir an die Konvergenz einer Folge $(f_n)_{n \in \mathbb{N}}$ stellen, ist die Konvergenz aller Funktionswerte $(f_n(x))_{n \in \mathbb{N}}$ – eben die punktweise Konvergenz von $(f_n)_{n \in \mathbb{N}}$. Das nebenstehende Beispiel zeigt, daß dabei im Allgemeinen die Stetigkeit der Funktionen f_n nicht an die Grenzfunktion weitergegeben wird. Wir wählen $D := [0, 1]$ und $f_n(x) := x^n$. Diese Funktionswerte konvergieren für $x \in [0, 1)$ gegen Null und für $x = 1$ gegen 1. Die Grenzfunktion

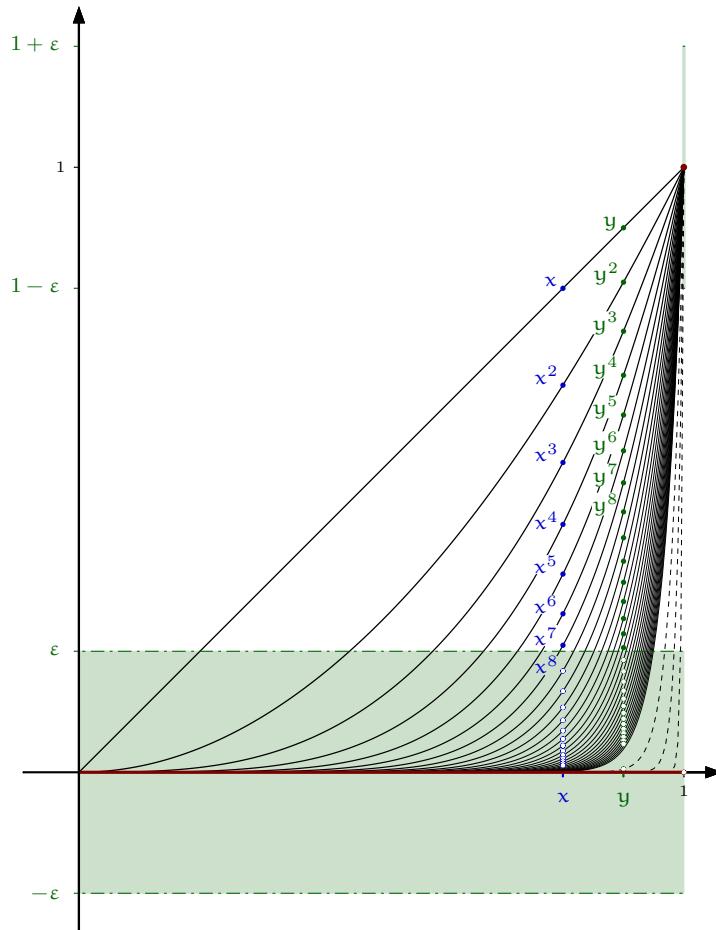
$$f(x) = \begin{cases} 0, & 0 \leq x < 1 \\ 1, & x = 1 \end{cases}$$

ist offensichtlich unstetig. Der Grund für dieses Verhalten ist, daß $f_n(x)$ gegen $f(x)$ um so langsamer konvergiert, je näher sich $x \in [0, 1)$ bei 1 befindet. Das können wir präziser fassen, wenn wir die Konvergenzbedingung (11.25) ausführlich ausschreiben:

Für alle $x \in D$ und jedes $\varepsilon > 0$ gibt es ein $n_{\varepsilon, x} \in \mathbb{N}$, so daß für alle $n \geq n_{\varepsilon, x}$ die Abschätzung $\|f(x) - f_n(x)\| < \varepsilon$ gilt.

Mit Hilfe der Quantoren läßt sich das kürzer formulieren:

$$\forall x \in D \forall \varepsilon > 0 \exists n_{\varepsilon, x} \in \mathbb{N} \forall n \geq n_{\varepsilon, x} \|f(x) - f_n(x)\| < \varepsilon. \quad (11.26)$$



Man sieht, daß der Grenzindex, ab dem die Genauigkeit ε unterschritten wird, im Allgemeinen nicht nur von ε , sondern auch noch von der Stelle x abhängen kann, an der die Funktionenfolge konvergieren soll. In unserem Beispiel muß $n_{\varepsilon,x}$ immer größer gewählt werden, je mehr sich x der Zahl 1 nähert, ohne daß es ein größtes n_ε gibt, das für alle x gleichermaßen gelten würde (sobald wir die Umkehrfunktion \ln der Exponentialfunktion zur Verfügung haben (Seite 306), zeigt eine einfache Rechnung $n_{\varepsilon,x} > \frac{\ln(\varepsilon)}{\ln(x)}$ für $\varepsilon, x \in (0, 1)$).

Wenn wir die Unabhängigkeit des Grenzindexes $n_{\varepsilon,x}$ von x fordern, erhalten wir (als Spezialfall von 10.1.50) eine schärfere Konvergenzbedingung für Funktionenfolgen:

11.1.29 Definition (Gleichmäßige Konvergenz) Eine Folge $(f_n)_{n \in \mathbb{N}}$ von Funktionen konvergiert auf dem gemeinsamen Definitionsbereich D gleichmäßig gegen f , falls

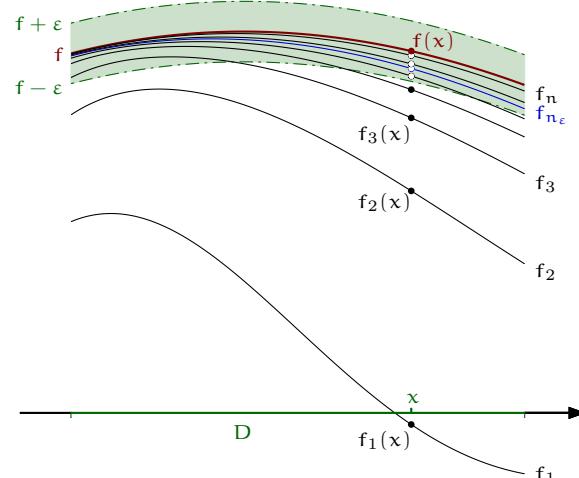
$$\forall \varepsilon > 0 \exists n_\varepsilon \in \mathbb{N} \forall x \in D \forall n \geq n_\varepsilon \|f(x) - f_n(x)\| < \varepsilon \quad (11.27)$$

gilt. Wir sagen auch, daß $f_n(x)$ gleichmäßig bzgl. $x \in D$ gegen $f(x)$ konvergiert. Eine Folge $(f_n)_{n \in \mathbb{N}}$ heißt gleichmäßige CAUCHY-Folge, falls die CAUCHY-Bedingung gleichmäßig bzgl. $x \in D$ erfüllt ist:

$$\forall \varepsilon > 0 \exists n_\varepsilon \in \mathbb{N} \forall x \in D \forall m, n \geq n_\varepsilon \|f_m(x) - f_n(x)\| < \varepsilon \quad (11.28)$$

Zu beachten ist die Position des Allquantors $\forall x \in D$. In $\exists n_\varepsilon \in \mathbb{N} \forall x \in D$ wird behauptet, daß es ein $n_\varepsilon \in \mathbb{N}$ gibt, das die Abschätzung für alle $x \in D$ ermöglicht.

Den Bereich aller Punkte $[x, y]$, die von den Punkten $[x, f(x)]$ des Graphen von f einen Abstand kleiner als ε haben, also $\{[x, y] \mid \|f(x) - y\| < \varepsilon\}$, hat die Form eines Schlauchs, dessen Mitte durch den Graphen von f vorgegeben wird und der sich in vertikaler Richtung nach oben und unten jeweils ε Einheiten weit ausdehnt. Dieser Schlauch wird ε -Schlauch um f genannt. Gleichmäßige Konvergenz bedeutet, daß sich fast alle Funktionen in diesem Schlauch befinden. Ein Grund für diese strengere Konvergenzbedingung ist, daß sie die Stetigkeit der approximierenden Folge auf die Grenzfunktion überträgt. Das ist der Inhalt des nächsten Satzes.

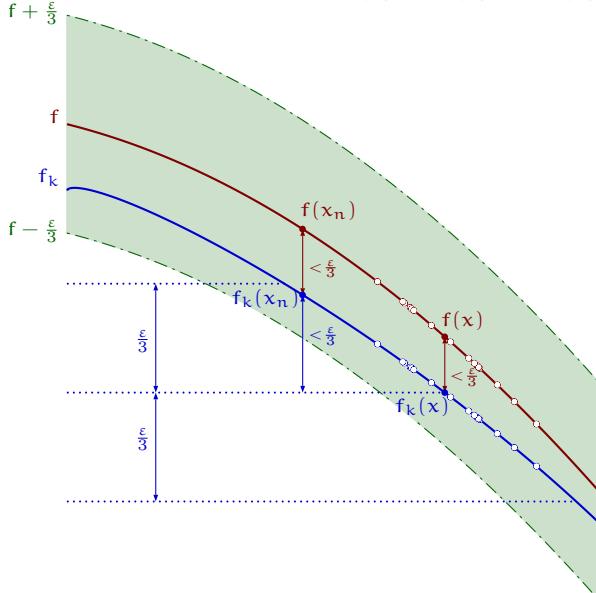


11.1.30 Satz $(f_k)_{k \in \mathbb{N}}$ sei eine Folge stetiger Funktionen, die auf einer Menge $D \subseteq \mathbb{R}$ oder $D \subseteq \mathbb{C}$ gleichmäßig gegen eine Funktion f konvergiert. Dann ist f auf D stetig.

Beweis. Zu zeigen ist, daß für jede Folge $(x_n)_{n \in \mathbb{N}}$ in D , die gegen ein Element $x \in D$ konvergiert, auch die Bildfolge $(f(x_n))_{n \in \mathbb{N}}$ gegen $f(x)$ konvergiert. Für jede der approximierenden

Funktionen f_k ist das wahr, denn die sind laut Voraussetzung stetig. Die Idee besteht nun darin, die Stetigkeit von f an einer der Funktionen f_k zu testen, die sich von f ausreichend wenig unterscheidet. Zunächst spielen wir die Differenz $|f(x) - f(x_n)|$ auf die Differenz $|f_k(x) - f_k(x_n)|$ zurück:

$$\begin{aligned}|f(x) - f(x_n)| &= |f(x) - f_k(x) + f_k(x) - f_k(x_n) + f_k(x_n) - f(x_n)| \\ &\leq |f(x) - f_k(x)| + |f_k(x) - f_k(x_n)| + |f_k(x_n) - f(x_n)|\end{aligned}$$



Dann wählen wir ein festes $k \in \mathbb{N}$, so daß $|f(x) - f_k(x)| < \frac{\varepsilon}{3}$ und $|f_k(x_n) - f(x_n)| < \frac{\varepsilon}{3}$ für alle $n \in \mathbb{N}$ gilt. Die erste Ungleichung ist natürlich leicht zu erfüllen, denn $f_k(x)$ konvergiert ja gegen $f(x)$. Die zweite Ungleichung gilt, weil die Funktionen f_k gleichmäßig gegen f konvergieren. Bei dem nun festgehaltenen k kann ein $n_\varepsilon \in \mathbb{N}$ gefunden werden, so daß $|f_k(x) - f_k(x_n)| < \frac{\varepsilon}{3}$ für alle $n \geq n_\varepsilon$ gilt. Das ist die Stetigkeit von f_k . Zusammengenommen haben wir damit

$$|f(x) - f(x_n)| < 3 \cdot \frac{\varepsilon}{3} = \varepsilon$$

für alle $n \geq n_\varepsilon$ erreicht. Das zeigt die Stetigkeit von f . \square

11.1.31 Korollar Eine Potenzreihe $f(z) = \sum_{k=0}^{\infty} a_k(z - z_0)^k$ mit einem Konvergenzradius $R > 0$ ist auf der offenen Kreisscheibe $\{z \in \mathbb{C} \mid |z - z_0| < R\}$ um z_0 mit Radius R stetig und auf jeder abgeschlossenen Kreisscheibe $U_r := \{z \in \mathbb{C} \mid |z - z_0| \leq r\}$ mit einem kleineren Radius $r < R$ gleichmäßig konvergent.

Beweis. Wir zeigen, daß die Folge (f_n) , die durch $f_n(z) := \sum_{k=0}^n a_k(z - z_0)^k$ definiert wird, für jedes $r \in (0, R)$ auf der Menge U_r gleichmäßig gegen f konvergiert:

$$\begin{aligned}\left|f(z) - \sum_{k=0}^n a_k(z - z_0)^k\right| &= \left|\sum_{k=n+1}^{\infty} a_k(z - z_0)^k\right| \leq \sum_{k=n+1}^{\infty} |a_k| |z - z_0|^k \leq \sum_{k=n+1}^{\infty} |a_k| r^k \\ &= \sum_{k=0}^{\infty} |a_k| r^k - \sum_{k=0}^n |a_k| r^k < \varepsilon\end{aligned}$$

gilt ab einem geeigneten $n_\varepsilon \in \mathbb{N}$, denn die Reihe $\sum_{k=0}^n |a_k| r^k$ ist laut Voraussetzung konvergent. Die Abschätzung weist daher die Existenz eines von $z \in U_r$ unabhängigen Grenzindexen n_ε nach, ab dem die Genauigkeit ε der Approximation gewährleistet ist. Das zeigt die gleichmäßige Konvergenz auf U_r . Nach Satz 11.1.30 ist f als gleichmäßiger Grenzwert der stetigen Funktionen f_n auf U_r stetig. Da $r < R$ beliebig war, ist f in jedem Punkt aus $\{z \in \mathbb{C} \mid |z - z_0| < R\}$ stetig. Dafür muß $r < R$ nur als $r = \frac{1}{2}(|z - z_0| + R)$ gewählt werden. Dann liegt z in U_r und damit in einer Menge, auf der f stetig ist. \square

11.1.32 A Die Exponentialfunktion $\exp(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!}$ ist auf ganz \mathbb{C} stetig, da die Exponentialreihe nach Beispiel 10.2.12 i) einen unendlichen Konvergenzradius hat. Für jede rationale Approximation $(x_n)_{n \in \mathbb{N}}$ einer irrationalen Zahl x konvergiert die Folge (e^{x_n}) also gegen $\exp(x)$. Daher sollte sich die Exponentialfunktion auch direkt, ohne Verwendung der Exponentialreihe, durch $e^x := \lim_{n \rightarrow \infty} e^{x_n}$ definieren lassen. Das ist auch tatsächlich möglich, nur erfordert es einige Überlegungen. Bei diesem Zugang ist die Basis e vor keiner anderen ausgezeichnet, so daß wir gleich eine beliebige Basis $a > 0$ verwenden können. Es genügt aber von $a > 1$ auszugehen, denn für $a < 1$ kann über den Kehrwert auf die Ergebnisse von $a > 1$ zurückgegriffen werden. Wir definieren also $a^x := \lim_{n \rightarrow \infty} a^{x_n}$.

- i) Zeigen Sie dafür zunächst, daß $\mathbb{Q} \ni x \mapsto a^x$ streng monoton wachsend ist.
- ii) Folgern Sie, daß $(a^{x_n})_{n \in \mathbb{N}}$ beschränkt ist.
- iii) Folgern Sie aus $a^{x_n} - a^{x_m} = a^{x_n}(1 - a^{x_m-x_n})$, daß nur der Klammerausdruck kontrolliert werden muß. Verwenden Sie die CAUCHY-Bedingung für $(x_n)_{n \in \mathbb{N}}$ in der Form $-\frac{1}{k} \leq x_m - x_n \leq \frac{1}{k}$ für alle n, m oberhalb eines geeigneten n_k . Folgern Sie daraus $\frac{1}{\sqrt[k]{a}} \leq a^{x_m-x_n} \leq \sqrt[k]{a}$. Mit Hilfe von (10.11) können Sie die CAUCHY-Bedingung für $(a^{x_n})_{n \in \mathbb{N}}$ jetzt nachweisen.
- iv) Zeigen Sie, daß $\exp_a(x) := \lim_{n \rightarrow \infty} a^{x_n}$ wohldefiniert ist, d. h., für alle rationale Folgen $(x_n)_{n \in \mathbb{N}}$ und $(z_n)_{n \in \mathbb{N}}$, die gegen x konvergieren, gilt $\lim_{n \rightarrow \infty} a^{x_n} = \lim_{n \rightarrow \infty} a^{z_n}$.
- v) Folgern Sie $\exp_a(x) = a^x$ für $x \in \mathbb{Q}$.
- vi) Jetzt definieren wir $a^x := \exp_a(x)$ auch für $x \in \mathbb{R}$. Zeigen Sie, daß weiterhin die Potenzrechengesetze gelten.
- vii) Zeigen Sie, daß $\mathbb{R} \ni x \mapsto a^x$ stetig ist.
- viii) Zeigen Sie, als kleiner Vorgriff auf die ln-Funktion (siehe Seite 306), daß der hier vorgestellte Zugang zur allgemeinen Potenzfunktion mit der in diesem Buch bevorzugten Version äquivalent ist:

$$a^x = e^{x \ln(a)}, \quad x \in \mathbb{R}.$$

11.1.33 A Wenn Sie schon mal bei dem Thema Potenzrechengesetze sind, überlegen Sie sich doch, wie die gewöhnlichen, mit rationalen Exponenten, eigentlich zustande kommen.

11.2 Differentialrechnung

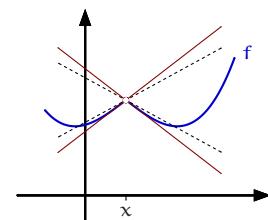
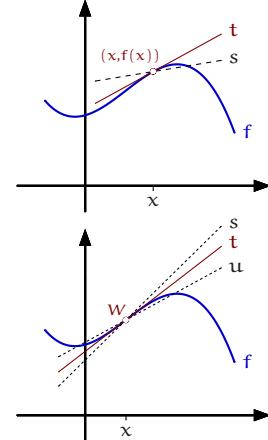
Der Begriff *Steigung* ist für Geraden $f(x) = mx + b$ eine anschauliche Größe. Sie beschreibt den linearen Zuwachs $\Delta f(x) = f(x+1) - f(x) =: m$ von $f(x)$ je Einheitsschritt $\Delta x = 1$ in x -Richtung. Auch wenn das bei Geraden meist nicht im Vordergrund steht, ist die eigentliche Natur der Steigung ein *Verhältnis*, nämlich

$$m = \frac{\Delta f(x)}{\Delta x} = \frac{f(x+h) - f(x)}{h},$$

also der Funktionswertzuwachs $\Delta f(x) := f(x+h) - f(x)$ pro Argumentzuwachs $\Delta x := h$. Dieses Konzept, das sich bei Geraden bewährt hat, möchte man auch für (weitgehend) beliebige Funktionen zur Verfügung haben. Natürlich muß man sich erst einmal klar werden, was Steigung in diesem allgemeineren Kontext bedeuten soll. Offensichtlich wird die Steigung für eine Funktion f von der betrachteten Stelle x abhängen. Die einfachste Idee besteht dann darin, den neuen Begriff der Steigung auf den schon bekannten von Geraden zurückzuführen. Wir vereinbaren, daß eine Funktion f an der Stelle x genau so steil sein soll, wie ihre Tangente an dieser Stelle (genauer: an der Stelle $[x, f(x)]$). Dieses Konzept können wir dann einlösen, wenn es uns gelingt, genau zu fassen, was wir unter Tangenten einer Funktion verstehen wollen.

Eine gängige Vorstellung besteht darin, von einer Tangente zu verlangen, daß sie die Kurve im Kurvenpunkt $[x, f(x)]$ nur berührt, aber nicht schneidet. Diese Vorstellung wird noch dadurch gestützt, daß sie bei vielen Funktionen für die meisten Kurvenpunkte auch tatsächlich zutrifft. Aber wie steht es mit einem Wendepunkt W einer Kurve, also einem Punkt, in dem sie z. B. von einer Rechtskurve in eine Linkskurve übergeht? Jede Gerade durch W muß die Kurve in diesem Punkt schneiden. Trotzdem wird vermutlich jeder, der unter den Geraden s, t und u zu wählen hat, zu dem Schluß gelangen, daß als Tangente nur t in Frage kommt. Das liegt daran, daß t die Kurve in einer kleinen Umgebung von W besser annähert, als die anderen Geraden. Diese Eigenschaft steht also im Vordergrund und wird deshalb als Ausgangspunkt der Tangentendefinition gewählt. Eine Tangente in einem Punkt der Kurve nähert sie in einer Umgebung dieses Punktes besser als alle anderen Geraden an. Damit meinen wir, daß der Unterschied der Funktionswerte von Kurve und Tangente in einer beliebig kleinen Umgebung rechts und links des Kurvenpunktes kleiner ist, als der Unterschied zwischen Kurve und jeder anderen Geraden durch diesen Punkt. Damit kann es in einem Kurvenpunkt nur eine Tangente geben – vorausgesetzt, es gibt überhaupt eine.

Man mache sich klar, daß es z. B. für einen Punkt, in dem die Kurve einen Knick hat, keine Gerade geben kann, die unsere Forderungen an eine Tangente erfüllt. Die strenge Bedingung, daß eine Tangente die Kurve bestmöglich approximieren soll, führt also dazu, daß es Kurvenpunkte geben kann, in denen keine Tangente möglich ist. Tatsächlich ist das aber kein Fehler unserer Tangentendefinition. Wir dürfen nämlich nicht aus den Augen verlieren, daß unser primäres Ziel nicht die

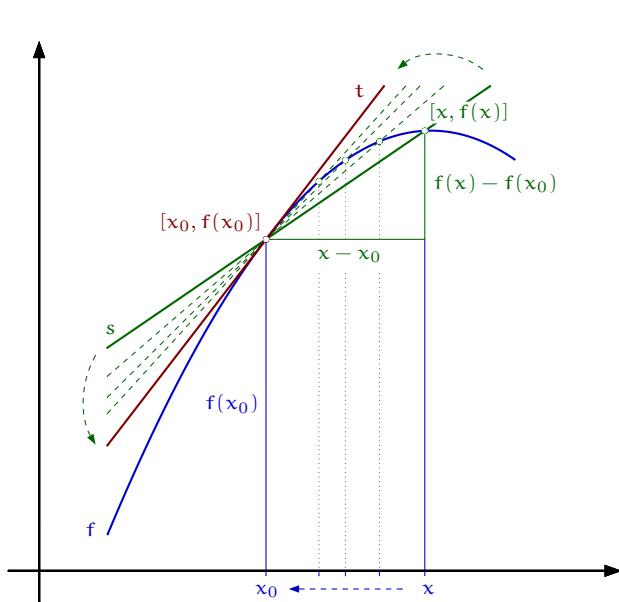


Tangente, sondern ihre Steigung ist. Und wenn wir von *der* Steigung einer Funktion an einer Stelle x reden wollen, dann darf es auch nur eine geben.

11.2.1 Definition Eine reelle Funktion auf einem Intervall I heißt an einer Stelle $x_0 \in I$ differenzierbar, falls der Grenzwert

$$f'(x_0) := \lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} \quad (11.29)$$

existiert. Die Zahl $f'(x_0)$ wird als Ableitung von f an der Stelle x_0 bezeichnet. Ist f an jeder Stelle in I differenzierbar, dann heißt f auf I differenzierbar und f' Ableitung von f .



Gemäß unserer Vereinbarung 11.1.1 ist $\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$ eine Abkürzung dafür, daß für jede zulässige und gegen x_0 konvergente Folge $(x_n)_{n \in \mathbb{N}}$ in I , der Bruch $\frac{f(x_n) - f(x_0)}{x_n - x_0}$ gegen die Zahl $f'(x_0)$ konvergiert. Die Zulässigkeit der Folge meint hier, daß sich die Folgenglieder x_n im Definitionsbereich der Funktion $x \mapsto \frac{f(x) - f(x_0)}{x - x_0}$ befinden und daher insbesondere von x_0 verschieden sind. Manchmal ist es bequem, diesen Grenzwert in der Form

$$f'(x_0) = \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

zu untersuchen. Setzen wir $x = x_0 + h$, so ergibt sich die ursprüngliche Definition.

Mit den Symbolen $\Delta f(x_0) := f(x) - f(x_0)$ und $\Delta x_0 := x - x_0$ für die Differenzen der Funktionswerte bzw. der Argumente läßt sich der Bruch in (11.29) als $\frac{\Delta f(x_0)}{\Delta x_0}$ schreiben. Daher röhren die Namen *Differenzenquotient* für ihn, *Differentialquotient* für seinen Grenzwert $f'(x_0)$ und die Schreibweise

$$\frac{d}{dx} f(x_0) := f'(x_0). \quad (11.30)$$

Dabei ist der Bruch $\frac{d}{dx}$ symbolisch zu verstehen, als Erinnerung an die Entstehungsweise von $f'(x_0)$. Den Symbolen d und dx kommen dabei keine eigenständigen Bedeutungen zu, denn dx wäre sonst als $\lim_{\Delta x \rightarrow 0} \Delta x$ zu verstehen, was natürlich einfach 0 ergibt. Normalerweise wird die Stelle x_0 , an der die Ableitung zu bilden ist, nicht besonders hervorgehoben. Daher werden wir auch die Schreibweise $\frac{df(x)}{dx}$ für $f'(x)$ verwenden. Das ist bequem, denn so kann man von der Ableitung einer Funktion sprechen, die durch eine Formel gegeben ist, ohne ihr erst einen Namen geben zu müssen. Etwa durch

$$\frac{dx^2}{dx} = 2x$$

wird mit der linken Seite die Aufgabe gestellt, die Ableitung der Funktion $x \mapsto x^2$ anzugeben und mit der rechten Seite die Lösung $x \mapsto 2x$ mitgeteilt.

Der Skizze entnimmt man, daß durch den Differenzenquotienten die Steigung einer Sekante s durch $[x_0, f(x_0)]$ und $[x, f(x)]$ gegeben ist. Der Grenzwert (11.29) präzisiert dann die Vorstellung, die Steigung $f'(x_0)$ einer Tangente t an der Stelle $[x_0, f(x_0)]$ an f durch Sekantensteigungen zu approximieren. Die *Tangentengleichung* ist bei Kenntnis von $f'(x_0)$ leicht anzugeben:

$$t(x) = f'(x_0)(x - x_0) + f(x_0). \quad (11.31)$$

Denn das ist offensichtlich die Gleichung einer Geraden mit der Steigung $f'(x_0)$, also der Tangentensteigung in $[x_0, f(x_0)]$, die an der Stelle $x = x_0$ den Wert $f(x_0)$ annimmt und somit durch den Kurvenpunkt $[x_0, f(x_0)]$ geht. Es muß sich also um die Tangente selbst handeln.

Wir müssen uns noch davon überzeugen, daß (11.31) in einer Umgebung von x_0 tatsächlich die beste Approximation von f durch eine Gerade ist.

11.2.2 Lemma *f sei an der Stelle x_0 differenzierbar und $s(x) := m(x - x_0) + f(x_0)$ eine Gerade durch $[x_0, f(x_0)]$ mit einer Steigung $m \neq f'(x_0)$. Dann gibt es eine Umgebung $(x_0 - \delta, x_0 + \delta)$ von x_0 , in der $|f(x) - s(x)| > |f(x) - t(x)|$ gilt.*

Beweis. Die Differenzierbarkeit in x_0 bedeutet, daß es ein $\delta > 0$ mit folgender Eigenschaft gibt (vergl. (11.9)):

$$\left| \frac{f(x) - f(x_0)}{x - x_0} - f'(x_0) \right| < \varepsilon$$

für alle x mit $|x - x_0| < \delta$. Wir wählen $\varepsilon := \frac{1}{3} |f'(x_0) - m|$. Damit erhalten wir

$$|f(x) - t(x)| = \left| \frac{f(x) - f(x_0)}{x - x_0} - f'(x_0) \right| |x - x_0| < \frac{1}{3} |f'(x_0) - m| |x - x_0|$$

und

$$\begin{aligned} |f(x) - s(x)| &= \left| \left(\frac{f(x) - f(x_0)}{x - x_0} - f'(x_0) \right) (x - x_0) + (f'(x_0) - m)(x - x_0) \right| \\ &\geq |f'(x_0) - m| |x - x_0| - \left| \frac{f(x) - f(x_0)}{x - x_0} - f'(x_0) \right| |x - x_0| \\ &> |f'(x_0) - m| |x - x_0| - \frac{1}{3} |f'(x_0) - m| |x - x_0| \\ &= \frac{2}{3} |f'(x_0) - m| |x - x_0| > |f(x) - t(x)|. \end{aligned} \quad \square$$

Der folgende Satz führt die Definition 11.2.1 in eine äquivalente Formulierung ohne Bruch über. Sie wird sich in folgenden Beweisen als bequem erweisen und gestattet es, den Begriff der Ableitung später auch auf Funktionen $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ auszudehnen.

11.2.3 Satz *Eine Funktion f auf einem Intervall I ist an einer Stelle $x_0 \in I$ genau dann differenzierbar, wenn es eine in x_0 stetige Funktion Θ mit der Eigenschaft*

$$f(x) - f(x_0) = \Theta(x)(x - x_0) \quad (11.32)$$

gibt. Ist eine dieser äquivalenten Bedingungen erfüllt, so ist $\Theta(x_0) = f'(x_0)$.

Beweis. f sei an der Stelle x_0 differenzierbar. Dann definieren wir Θ durch

$$\Theta(x) := \begin{cases} \frac{f(x) - f(x_0)}{x - x_0} & \text{für } x \neq x_0 \\ f'(x_0) & \text{für } x = x_0. \end{cases}$$

Für jede Folge $(x_n)_{n \in \mathbb{N}}$ in I , die gegen x_0 konvergiert, erhalten wir damit

$$|\Theta(x_n) - \Theta(x_0)| = \begin{cases} \left| \frac{f(x_n) - f(x_0)}{x_n - x_0} - f'(x_0) \right| & \text{für } x_n \neq x_0 \\ 0 & \text{für } x_n = x_0 \end{cases} \xrightarrow{n \rightarrow \infty} 0.$$

d. h. Θ ist an der Stelle x_0 stetig.

Gehen wir nun von der Existenz der Funktion Θ aus, so gilt für alle $x \neq x_0$

$$\frac{f(x) - f(x_0)}{x - x_0} = \Theta(x).$$

Die Stetigkeit von Θ in x_0 bedeutet dann

$$\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = \Theta(x_0).$$

Das zeigt, daß f an der Stelle x_0 differenzierbar ist und daß $f'(x_0) = \Theta(x_0)$ gilt. \square

11.2.4 Satz Ist eine Funktion an einer Stelle differenzierbar, so ist sie dort auch stetig.

Beweis. f sei an der Stelle x_0 differenzierbar. Nach Satz 11.2.3 gibt es eine in x_0 stetige Funktion Θ mit der Eigenschaft $f(x) - f(x_0) = \Theta(x)(x - x_0)$. Als Produkt zweier in x_0 stetiger Funktionen ist $x \mapsto f(x) - f(x_0)$ und daher auch f an der Stelle x_0 stetig. \square

Die Umkehrung dieses Satzes gilt nicht. Die Funktion $f(x) := |x|$ zum Beispiel ist auf ganz \mathbb{R} stetig, aber an der Stelle 0 nicht differenzierbar. Für die Nullfolge $(x_n)_{n \in \mathbb{N}} := ((-1)^n \frac{1}{n})_{n \in \mathbb{N}}$ ist die Folge der Differenzenquotienten

$$\frac{|x_n|}{x_n} = \begin{cases} 1 & \text{für gerades } n \\ -1 & \text{für ungerades } n \end{cases}$$

nämlich nicht konvergent.

11.2.5 Satz Die beiden Funktionen f und g seien auf einem gemeinsamen Definitionsbereich D gegeben und an der Stelle $x \in D$ differenzierbar. Dann ist auch die Linearkombination $tf + sg$, $t, s \in \mathbb{R}$, das Produkt fg und der Quotient $\frac{f}{g}$ an dieser Stelle differenzierbar (letzteres, falls $g(x) \neq 0$). Dabei gelten die folgenden Ableitungsregeln.

$$(tf + sg)'(x) = t f'(x) + s g'(x), \quad (\text{Summenregel}) \quad (11.33)$$

$$(fg)'(x) = f'(x)g(x) + g'(x)f(x), \quad (\text{Produktregel}) \quad (11.34)$$

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)}. \quad (\text{Quotientenregel}) \quad (11.35)$$

Lassen sich f und g verketten und ist g an der Stelle x , f an der Stelle $g(x)$ differenzierbar, so ist auch $f \circ g$ an der Stelle x differenzierbar, mit der Ableitung

$$\frac{d}{dx} f(g(x)) = f'(g(x))g'(x). \quad (\text{Kettenregel}) \quad (11.36)$$

Beweis. (11.33) ergibt sich einfach aus den Rechenregeln für konvergente Folgen, angewandt auf die Differenzenquotienten für $tf + sg$.

Zu (11.34):

$$\lim_{y \rightarrow x} \frac{f(y)g(y) - f(x)g(x)}{y - x} = \lim_{y \rightarrow x} \frac{f(y) - f(x)}{y - x} g(y) + f(x) \lim_{y \rightarrow x} \frac{g(y) - g(x)}{y - x} = f'(x)g(x) + f(x)g'(x),$$

denn g ist nach Satz 11.2.4 an der Stelle x stetig, so daß der erste Summand gegen $f'(x)g(x)$ konvergiert. Der zweite Summand verlangt nur die Differenzierbarkeit von g in x .

Zu (11.35): g ist stetig und an der Stelle x nicht Null. Nach Lemma 11.1.13 gibt es eine ganze Umgebung von x , auf der $g(y) \neq 0$ erfüllt ist. Auf dieser Umgebung gilt

$$\frac{\left(\frac{1}{g}\right)(y) - \left(\frac{1}{g}\right)(x)}{y - x} = \frac{\frac{1}{g(y)} - \frac{1}{g(x)}}{y - x} = -\frac{1}{g(y)g(x)} \frac{g(y) - g(x)}{y - x} \xrightarrow{y \rightarrow x} -\frac{g'(x)}{g^2(x)}.$$

Das Ergebnis folgt nun aus der Produktregel (11.34):

$$\left(\frac{f}{g}\right)'(x) = \left(f \cdot \frac{1}{g}\right)'(x) = f'(x) \cdot \frac{g(x)}{g^2(x)} - f(x) \frac{g'(x)}{g^2(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{g^2(x)}.$$

Den Beweis der Kettenregel (11.36) führen wir in der linearisierten Form 11.2.3 der Ableitung. Nach Voraussetzung gibt es eine an der Stelle $g(x)$ stetige Funktion Θ mit der Eigenschaft $f(z) - f(g(x)) = \Theta(z)(z - g(x))$ und $\Theta(g(x)) = f'(g(x))$. Außerdem gibt es eine in x stetige Funktion Γ mit der Eigenschaft $g(y) - g(x) = \Gamma(y)(y - x)$ und $\Gamma(x) = g'(x)$. Daraus erhalten wir

$$f(g(y)) - f(g(x)) = \Theta(g(y))(g(y) - g(x)) = \Theta(g(y))\Gamma(y)(y - x).$$

Die Funktion $y \mapsto \tilde{\Theta}(y) := \Theta(g(y))\Gamma(y)$ ist als Verkettung und Produkt stetiger Funktionen wieder stetig an der Stelle x . Nach Satz 11.2.3 ist $f \circ g$ dort differenzierbar und hat die Ableitung $(f \circ g)'(x) = \tilde{\Theta}(x) = \Theta(g(x))\Gamma(x) = f'(g(x))g'(x)$. \square

Die Produkt- und die Quotientenregel merkt man sich in der Form

$$(uv)' = u'v + uv', \quad \left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2}. \quad (11.37)$$

In der Kettenregel wird $f'(g(x))$ als *äußere Ableitung* und $g'(x)$ als *innere Ableitung* bezeichnet, so daß die Kettenregel durch *äußere Ableitung mal innere Ableitung* im Gedächtnis bleiben sollte.

11.2.6 Satz Es gilt

$$\frac{dx^n}{dx} = nx^{n-1}, \quad n \in \mathbb{Q}, \quad (11.38)$$

$$\sin'(x) = \cos(x), \quad \cos'(x) = -\sin(x), \quad (11.39)$$

$$\frac{de^x}{dx} = e^x. \quad (11.40)$$

Beweis. (11.38) zeigen wir zunächst für $n \in \mathbb{N}$ und $n \geq 2$ (für $n = 0$ und $n = 1$ ist die Behauptung elementar):

$$\begin{aligned} \frac{(x+h)^n - x^n}{h} &= \frac{1}{h} \left(\sum_{k=0}^n \binom{n}{k} x^{n-k} h^k - x^n \right) = \frac{1}{h} \left(nx^{n-1} h + \sum_{k=2}^n \binom{n}{k} x^{n-k} h^k \right) \\ &= nx^{n-1} + \sum_{k=2}^n \binom{n}{k} x^{n-k} h^{k-1} \xrightarrow{h \rightarrow 0} nx^{n-1}, \end{aligned}$$

denn in der letzten Summe enthält jeder Summand mindestens einmal den Faktor h . Sie verschwindet daher für $h \rightarrow 0$. Für Potenzen x^{-n} mit negativem Exponenten setzen wir die Quotientenregel ein:

$$\frac{d}{dx} x^{-n} = \frac{d}{dx} \frac{1}{x^n} = -\frac{nx^{n-1}}{x^{2n}} = -nx^{-n-1}.$$

Damit ist (11.38) für alle $n \in \mathbb{Z}$ bestätigt. Für Exponenten der Form $\frac{1}{n}$ verwenden wir die Beziehung

$$\frac{a^n - b^n}{a - b} = a^{n-1} + a^{n-2}b + a^{n-3}b^2 + \dots + ab^{n-2} + b^{n-1} = \sum_{k=0}^{n-1} a^{n-1-k} b^k,$$

die man leicht durch Ausmultiplizieren oder direkt durch Polynomdivision bestätigt. Wir setzen $a := y^{\frac{1}{n}}$ und $b := x^{\frac{1}{n}}$ und bilden den Kehrwert obiger Gleichung:

$$\frac{y^{\frac{1}{n}} - x^{\frac{1}{n}}}{y - x} = \frac{1}{\sum_{k=0}^{n-1} (y^{\frac{1}{n}})^{n-1-k} (x^{\frac{1}{n}})^k} \xrightarrow{y \rightarrow x} \frac{1}{\sum_{k=0}^{n-1} (x^{\frac{1}{n}})^{n-1-k} (x^{\frac{1}{n}})^k} = \frac{1}{n} x^{\frac{1}{n}-1}.$$

Dieser Grenzwert existiert, da die Bildung der n -ten Wurzel und des Kehrwertes stetige Abbildungen sind (vergl. Satz 10.1.24). Für $x^{-\frac{1}{n}}$ gewinnt man die Behauptung wieder mit Hilfe der Quotientenregel. Für die allgemeine Form $\frac{m}{n}$ des Exponenten verwenden wir die Kettenregel:

$$\frac{d}{dx} x^{\frac{m}{n}} = \frac{d}{dx} (x^{\frac{1}{n}})^m = m(x^{\frac{1}{n}})^{m-1} \cdot \frac{1}{n} x^{\frac{1}{n}-1} = \frac{m}{n} x^{\frac{m-1}{n} + \frac{1}{n}-1} = \frac{m}{n} x^{\frac{m}{n}-1}.$$

Die Ableitungen (11.39) von sin und cos beruhen auf den Additionssätzen (5.50) bzw. (5.49) der trigonometrischen Funktionen, sowie den Grenzwerten $\lim_{h \rightarrow 0} \frac{\sin(h)}{h} = 1$ und $\lim_{h \rightarrow 0} \frac{1-\cos(h)}{h} = 0$ (vgl. (11.21) bzw. (11.22)).

$$\begin{aligned} \frac{\sin(x+h) - \sin(x)}{h} &= \frac{\sin(x)\cos(h) + \sin(h)\cos(x) - \sin(x)}{h} \\ &= \sin(x) \frac{\cos(h)-1}{h} + \frac{\sin(h)}{h} \cos(x) \xrightarrow{h \rightarrow 0} \cos(x), \end{aligned}$$

$$\begin{aligned}\frac{\cos(x+h) - \cos(x)}{h} &= \frac{\cos(x)\cos(h) - \sin(x)\sin(h) - \cos(x)}{h} \\ &= \cos(x) \frac{\cos(h) - 1}{h} - \frac{\sin(h)}{h} \sin(x) \xrightarrow{h \rightarrow 0} -\sin(x).\end{aligned}$$

Wir erhalten $\sin' = \cos$ und $\cos' = -\sin$.

Für die Ableitung (11.40) der Exponentialfunktion verwenden wir ihre Potenzreihen Darstellung (10.55).

$$\frac{e^{x+h} - e^x}{h} = \frac{e^x e^h - e^x}{h} = e^x \frac{e^h - 1}{h} = e^x \sum_{k=1}^{\infty} \frac{h^k}{k!} = e^x \sum_{k=1}^{\infty} \frac{h^{k-1}}{k!} = e^x \left(1 + \sum_{k=2}^{\infty} \frac{h^{k-1}}{k!}\right) \xrightarrow{h \rightarrow 0} e^x,$$

denn die letzte Summe beschreibt eine Potenzreihe, mit dem Funktionswert 0 an der Stelle $h = 0$. Da Potenzreihen stetig sind (Korollar 11.1.31), folgt die behauptete Konvergenz in der letzten Zeile. \square

11.2.7 A Zeigen Sie $\tan'(x) = \frac{1}{\cos^2(x)} = \tan^2(x) + 1$.

11.2.8 A Leiten Sie die folgenden Funktionen ab.

$$\begin{array}{ll} f(x) = \frac{1}{40}(x^4 - 26x^2 + 48x - 23) & g(x) = \frac{x^3 - 5x^2 - x + 5}{3x^2} \\ h(x) = \frac{3x^3}{3x^2 - 4} & k(x) = x^2 e^x \\ l(x) = e^{-3x+2} & m(x) = e^{-2x^2} \\ n(x) = x e^{-x} & p(x) = \sin(e^{2x^2}) \\ q(x) = \frac{3(x-1)^2}{x^3 - 2} & \cosh(x) := \frac{1}{2}(e^x + e^{-x}) \\ \sinh(x) := \frac{1}{2}(e^x - e^{-x}) & \tanh(x) := \frac{e^x - e^{-x}}{e^x + e^{-x}} \\ \cot(x) = \frac{\cos(x)}{\sin(x)} & r(x) = \frac{(x^2 + 1)e^x}{x - 1} \\ s(x) = (x^3 - 8x^2)^8 & t(x) = \frac{1}{(2x^3 + x)^3} \\ u(x) = \sin(2x) \cos(x) & v(x) = e^{-3x} \cos(x) \\ w(x) = \frac{e^x \cos(x)}{e^x + \sin(x)} & z(x) = \frac{e^{-3x}}{\cos(2x)} \end{array}$$

11.2.9 Lemma Sei f auf dem Intervall (a, b) differenzierbar. Ist $x_0 \in (a, b)$ eine Stelle, an der f ein lokales Maximum oder ein lokales Minimum annimmt, so gilt dort $f'(x_0) = 0$.

Beweis. O. B. d. A. sei bei $x_0 \in (a, b)$ ein lokales Maximum von f . Es gibt eine in x_0 stetige Funktion Θ mit der Eigenschaft $f(x) - f(x_0) = \Theta(x)(x - x_0)$ und $\Theta(x_0) = f'(x_0)$. Wäre $f'(x_0) \neq 0$, sagen wir $f'(x_0) > 0$, so gäbe es nach Lemma 11.1.13 eine ganze Umgebung $U(x_0) \subseteq (a, b)$ von x_0 , in der $\Theta(x) > 0$ gilt. Das bedeutet

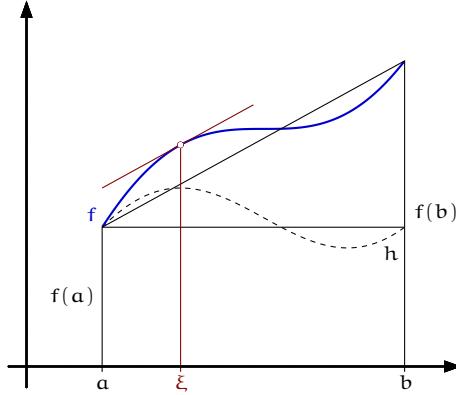
$$f(x) = f(x) - f(x_0) + f(x_0) = \Theta(x)(x - x_0) + f(x_0) > f(x_0)$$

für alle $x > x_0$ aus $U(x_0)$. Wäre $f'(x_0) < 0$, so folgte analog, daß $f(x) > f(x_0)$ für alle $x < x_0$ aus $U(x_0)$ gilt. Beide Fälle widersprechen der Eigenschaft, daß $f(x_0)$ in einer kleinen Umgebung von x_0 der größte Funktionswert von f sein soll. \square

11.2.10 A Warum gilt Lemma 11.2.9 im Allgemeinen nur für Stellen x_0 im Inneren von $[a, b]$?

11.2.11 Satz (Mittelwertsatz) Sei f auf dem Intervall $[a, b]$ stetig und in (a, b) differenzierbar. Dann gibt es eine Stelle $\xi \in (a, b)$ mit der Eigenschaft

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}. \quad (11.41)$$



Beweis: Wir zeigen zunächst den Spezialfall $f(a) = f(b)$. Ist f auf $[a, b]$ konstant, so ist nichts zu zeigen, denn für jedes $\xi \in (a, b)$ gilt $f'(\xi) = 0$. f sei also nicht konstant. Dann nimmt f als stetige Funktion auf $[a, b]$ nach Satz 11.1.9 an einer Stelle $x_1 \in [a, b]$ sein Maximum und an einer Stelle $x_2 \in [a, b]$ sein Minimum an. Wäre sowohl x_1 als auch x_2 auf dem Rand von $[a, b]$, so müßte f konstant sein, da dann $f(x_1) = f(x_2)$ gelten würde. Also muß wenigstens eine der beiden Stellen x_1 oder x_2 in (a, b) liegen. Diese Stelle wählen wir für ξ . Nach Lemma 11.2.9 muß dort $f'(\xi) = 0$ gelten. Für den speziellen Fall $f(a) = f(b)$ ist das (11.41).

Nun sei $f(a) \neq f(b)$. Wir bilden die Hilfsfunktion h durch

$$h(x) := f(x) - \frac{f(b) - f(a)}{b - a}(x - a),$$

für die offensichtlich $h(a) = f(a)$ und $h(b) = f(b)$ gilt. Wir finden demnach eine Stelle $\xi \in (a, b)$ mit $h'(\xi) = 0$, also

$$0 = f'(\xi) - \frac{f(b) - f(a)}{b - a}.$$

Das zeigt (11.41) im allgemeinen Fall. □

Eine geringfügige Anpassung der Beweisidee liefert folgende Verallgemeinerung.

11.2.12 Satz (Verallgemeinerter Mittelwertsatz) f und g seien auf dem Intervall $[a, b]$ stetig und in (a, b) differenzierbar. Außerdem gelte $g'(x) \neq 0$ für alle $x \in (a, b)$. Dann ist $g(a) \neq g(b)$, und es gibt ein $\xi \in (a, b)$ mit der Eigenschaft

$$\frac{f'(\xi)}{g'(\xi)} = \frac{f(b) - f(a)}{g(b) - g(a)}. \quad (11.42)$$

Beweis. Würde $g(a) = g(b)$ gelten, so gäbe es nach dem Mittelwertsatz 11.2.11 eine Stelle $\xi \in (a, b)$ mit $g'(\xi) = 0$, im Widerspruch zu den Voraussetzungen an g . Also läßt sich die rechte Seite in (11.42) bilden. Wir verallgemeinern die Hilfsfunktion h aus dem Beweis zum Mittelwertsatz:

$$h(x) := f(x) - \frac{f(b) - f(a)}{g(b) - g(a)} (g(x) - g(a)).$$

Durch Einsetzen bestätigt man leicht $h(a) = f(a) = h(b)$, so daß es nach dem Mittelwertsatz ein $\xi \in (a, b)$ mit $h'(\xi) = 0$ geben muß. Das bedeutet

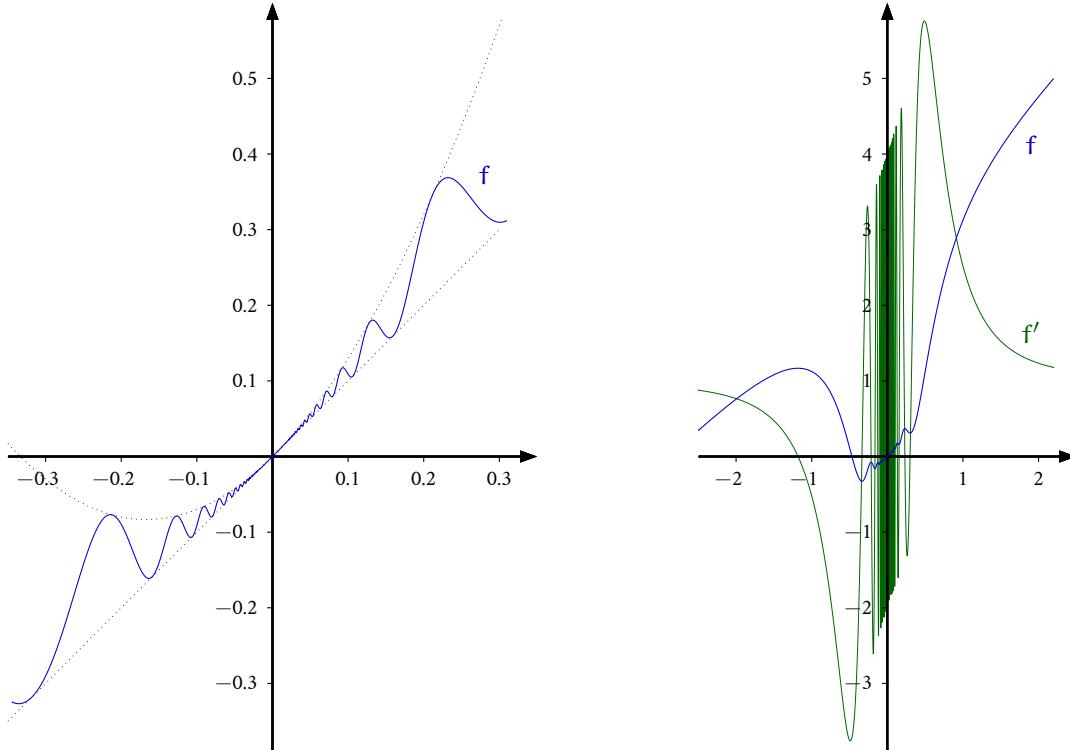
$$0 = f'(\xi) - \frac{f(b) - f(a)}{g(b) - g(a)} g'(\xi),$$

woraus (11.42) unmittelbar folgt. \square

11.2.13 Korollar Ist f in (a, b) differenzierbar und gilt $f'(x) = 0$ für alle $x \in (a, b)$, so muß f auf (a, b) konstant sein.

Beweis. Wir wählen ein beliebiges $\beta \in (a, b)$. Für jedes $x \in (a, \beta)$ gibt es nach dem Mittelwertsatz dann jeweils ein $\xi \in (x, \beta)$, mit $0 = f'(\xi) = \frac{f(\beta) - f(x)}{\beta - x}$. Daraus folgt $f(\beta) = f(x)$ für alle $x \in (a, \beta)$, d. h., f ist auf (a, β) konstant. Da $\beta < b$ beliebig war, ist f auch auf (a, b) konstant. \square

11.2.14 Beispiel Funktionen wie $x \mapsto x^n$, \sin , \cos usw. sind unendlich oft ableitbar. Das bedeutet insbesondere, daß ihre Ableitungen stetig sind. Es ist aber durchaus möglich, daß eine Funktion überall differenzierbar, die Ableitung jedoch nicht überall stetig ist. Funktionen, wie $x \mapsto \frac{1}{x}$, oder $x \mapsto |x|$, die einem dabei zuerst in den Sinn kommen könnten, sind keine Beispiele, denn die erste ist beliebig oft und die zweite ist an der Stelle 0 nicht differenzierbar. Ein Beispiel ist



$$f(x) := \begin{cases} 3x^2 \sin^2\left(\frac{1}{x}\right) + x & , x \neq 0, \\ 0 & , x = 0, \end{cases}$$

$$f'(x) = \begin{cases} 6x \sin^2\left(\frac{1}{x}\right) - 3 \sin\left(\frac{2}{x}\right) + 1 & , x \neq 0, \\ 1 & , x = 0. \end{cases}$$

f ist offensichtlich stetig und für $x \neq 0$ mit Hilfe der Produkt- und Kettenregel ableitbar. Einzig die Stelle $x = 0$ muß direkt mit Hilfe des Differenzenquotienten untersucht werden: $\frac{f(x)}{x} = 3x \sin^2\left(\frac{1}{x}\right) + 1 \xrightarrow{x \rightarrow 0} 1 = f'(0)$. Die Werte $f'(x)$ oszillieren auf dem Intervall $(0, 1)$ bzw. $(-1, 0)$ unendlich oft. Das liegt am Term $-3 \sin\left(\frac{2}{x}\right)$ in der Ableitung, der alle Oszillationen des Sinus oberhalb von $x = 2$ in den Bereich $(0, 1)$ transformiert. f' ist augenscheinlich an der Stelle $x = 0$ nicht stetig.

Dieser Eindruck bestätigt sich mit der Nullfolge $x_n := \frac{4}{(2n+1)\pi}$, denn $f(x_n) = 6x_n \sin^2\left(\frac{1}{x_n}\right) - 3 \sin((2n+1)\frac{\pi}{2}) + 1 = 6x_n \sin^2\left(\frac{1}{x_n}\right) - 3(-1)^n + 1$. Der Ausdruck $6x_n \sin^2\left(\frac{1}{x_n}\right)$ konvergiert gegen Null, aber $-3(-1)^n + 1$ wechselt beständig zwischen -2 und 4 . Die Folge $(f'(x_n))_{n \in \mathbb{N}}$ hat also zwei Häufungspunkte und ist somit nicht konvergent. Das Beispiel zeigt auch, daß aus $f'(x) > 0$ nicht unbedingt darauf geschlossen werden kann, daß f in einer noch so kleinen Umgebung von x streng monoton wächst, wenn man die Stetigkeit von f' an dieser Stelle nicht zur Verfügung hat.

Die Unstetigkeit von f' bei $x = 0$ ist derart, daß der linksseitige und der rechtsseitige Grenzwert nicht existiert. Es handelt sich dabei also nicht um eine Sprungstelle. Wir werden gleich sehen, daß f' auch keine Sprungstelle haben kann.

11.2.15 Satz $f: [a, b] \rightarrow \mathbb{R}$ sei eine stetige Funktion, die auf (a, b) differenzierbar ist. Dann gilt auf jedem abgeschlossenen Intervall $[\alpha, \beta] \subset (a, b)$ der Zwischenwertsatz für f' .

Beweis. Wir definieren die stetigen Funktionen Θ und Λ auf $[\alpha, \beta]$ durch

$$\Theta(x) := \begin{cases} \frac{f(\beta) - f(x)}{\beta - x} & , x \neq \beta, \\ f'(\beta) & , x = \beta, \end{cases} \quad \Lambda(x) := \begin{cases} \frac{f(x) - f(\alpha)}{x - \alpha} & , x \neq \alpha, \\ f'(\alpha) & , x = \alpha. \end{cases}$$

Nach dem Mittelwertsatz gibt es ein $\xi \in (\alpha, \beta)$ mit $\Theta(\alpha) = \Lambda(\beta) = f'(\xi)$. Nach dem Zwischenwertsatz nehmen Θ und Λ auf $[\alpha, \beta]$ alle Werte von $f'(\xi)$ bis $f'(\beta)$ bzw. von $f'(\xi)$ bis $f'(\alpha)$ an. Dabei handelt es sich nach dem Mittelwertsatz um Funktionswerte von f' . Also nimmt f' auf $[\alpha, \beta]$ alle Werte von $f'(\alpha)$ bis $f'(\beta)$ an. \square

11.2.16 Beispiel Bei der Funktion $f: [-3, 3] \rightarrow \mathbb{R}$, $f(x) := x^3$ und $[\alpha, \beta] := [-1, 2]$ kann man den Beweis des Satzes noch einmal nachzuvollziehen.

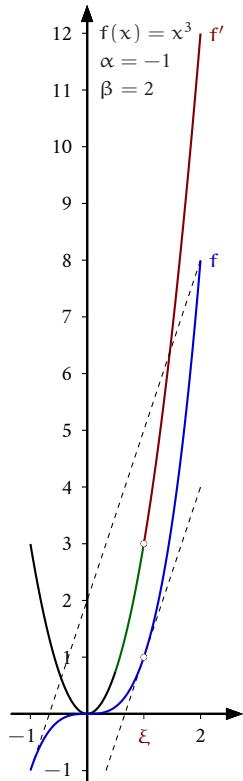
$$\Lambda(x) = \frac{f(x) - f(-1)}{x + 1} = \frac{x^3 + 1}{x + 1} = x^2 - x + 1 = (x - \frac{1}{2})^2 + \frac{3}{4},$$

$$\Theta(x) = \frac{f(2) - f(x)}{2 - x} = \frac{x^3 - 8}{x - 2} = x^2 + 2x + 4 = (x + 1)^2 + 3,$$

$$f'(\xi) = 3\xi^2 = \frac{f(2) - f(-1)}{2 + 1} = 3$$

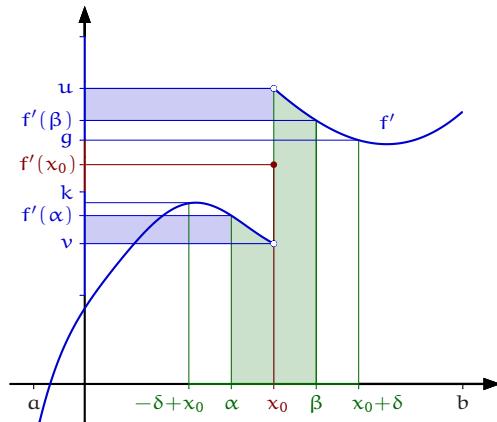
führt auf $\xi = 1$. $\Theta(2) = f'(2) = 12$, $\Lambda(-1) = f'(-1) = 3$. Aus dem Mittelwertsatz folgt $\Lambda(x) = f'(\eta_x)$. Dabei hängt $\eta_x \in (-1, x)$ natürlich von der Stelle x ab. Der Satz macht keine Aussage darüber, ob $x \mapsto \eta_x$ als gewöhnliche Funktion zu bestimmen ist, aber bei diesem einfachen Beispiel ist $\eta_x := \sqrt{\frac{1}{3}(x - \frac{1}{2})^2 + \frac{1}{4}}$ eine Möglichkeit. Der Wertebereich von Λ ist offensichtlich $[\frac{3}{4}, 3]$.

$\Theta(x) = f'(\gamma_x)$ wird durch $\gamma_x := \sqrt{\frac{1}{3}(x + 1)^2 + 1}$ erfüllt. Der Wertebereich von Θ ist das Intervall $[3, 12]$. Damit haben wir nachgerechnet, daß $f'(x)$ alle Werte aus $[\frac{3}{4}, 12]$ annimmt. Das deckt sogar einen größeren Bereich ab, als der von Satz 11.2.15 sicher vorhergesagte $[f'(-1), f'(2)] = [3, 12]$.



11.2.17 Korollar Für eine reelle Funktion $f: [a, b] \rightarrow \mathbb{R}$, die auf (a, b) differenzierbar ist, kann f' in (a, b) keine Sprungstelle haben.

Beweis. Wäre $x_0 \in (a, b)$ eine Sprungstelle von f' , so müßten der linksseitige Grenzwert $v := \lim_{x \rightarrow x_0^-} f'(x)$ von f' und der rechtsseitige Grenzwert $u := \lim_{x \rightarrow x_0^+} f'(x)$ existieren und verschieden sein: $\varepsilon := u - v \neq 0$. Wir können o. B. d. A. von $\varepsilon > 0$ ausgehen. Dann gibt es ein $\delta > 0$, so daß für alle $x \in (x_0 - \delta, x_0)$ die Abschätzung $|f'(x) - v| < \frac{\varepsilon}{3}$, oder besser $-\frac{\varepsilon}{3} < f'(x) - v < \frac{\varepsilon}{3}$ gilt und $-\frac{\varepsilon}{3} < f'(y) - u < \frac{\varepsilon}{3}$ für alle $y \in (x_0, x_0 + \delta)$. Für $x \in (x_0, x_0 + \delta)$ haben wir daher



$$f'(x) = u + (f'(x) - u) > u - \frac{\varepsilon}{3} =: g = \frac{2}{3}u + \frac{1}{3}v.$$

Das gilt insbesondere für $x = \beta := x_0 + \frac{\delta}{2}$. Genauso erhalten wir für $y \in (x_0 - \delta, x_0)$ die Abschätzung $f'(y) < v + \frac{\varepsilon}{3} =: k = \frac{1}{3}u + \frac{2}{3}v = \frac{1}{3}u + \frac{1}{3}v + \frac{1}{3}v < \frac{1}{3}u + \frac{1}{3}u + \frac{1}{3}v = \frac{2}{3}u + \frac{1}{3}v = g < f'(x)$, insbesondere für $y = \alpha := x_0 - \frac{\delta}{2}$. Die Werte aus dem Intervall $[k, g]$, mit möglicherweise einer Ausnahme, nämlich dem Wert $f'(x_0)$, kommen also nicht als Funktionswerte von

f' auf dem Intervall $[\alpha, \beta]$ in Frage. Laut Satz 11.2.15 müßten aber alle Werte aus $[f'(\alpha), f'(\beta)]$ von f' auf $[\alpha, \beta]$ angenommen werden, im Widerspruch zu $[f'(\alpha), f'(\beta)] \supseteq [k, g]$. \square

11.2.18 Beispiel Die Vorstellung, die man von einem lokalen Maximum $H = [x_0, f(x_0)]$ einer stetig differenzierbaren Funktion f hat, entspricht etwa dem linken Teil der Abbildung 11.7. Das heißt, man erwartet, daß die Funktionswerte auf einem ausreichend kleinen Intervall $[x_0 - \varepsilon, x_0]$ links der Stelle x_0 monoton wächst und auf $[x_0, x_0 + \delta]$ rechts von ihr monoton fällt. Das trifft auch normalerweise zu, jedenfalls dann, wenn sich in einer Umgebung des Punktes nur endlich viele weitere Extrema befinden. Der rechte Teil von Abbildung 11.7 zeigt, daß es andernfalls Maxima gibt, so daß f in keinem der Intervalle links oder rechts von x_0 monoton wachsend bzw. fallend ist.

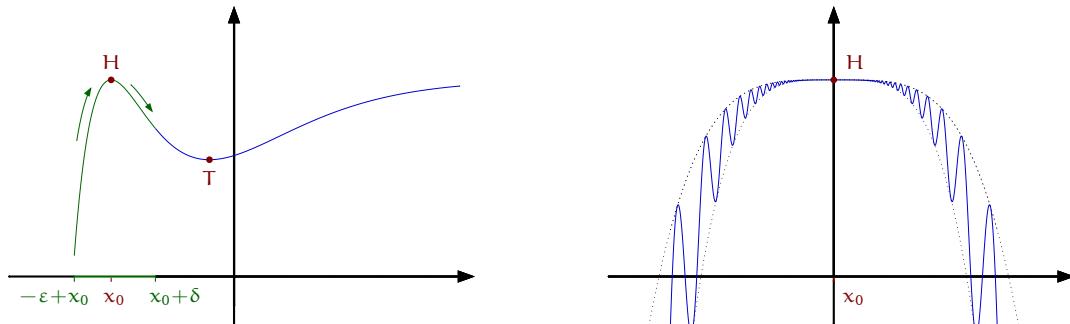


Abb. 11.7 *Lokale Maxima*

Die rechte Abbildung in 11.7 gehört zur Funktion

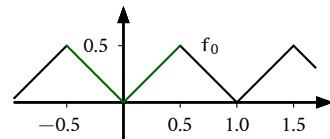
$$f(x) := \begin{cases} 2 - \frac{x^4}{5} \left(2 + \cos\left(\frac{45}{x}\right) \right) & , x \neq 0, \\ 2 & , x = 0. \end{cases} \quad (11.43)$$

Die Faktoren $\frac{1}{5}$ und 45 sind dabei unerheblich, sie dienen lediglich dazu, die unendlich vielen lokalen Maxima und Minima in jeder Umgebung des Maximums $H = [0, 2]$ deutlicher hervortreten zu lassen.

11.2.19 A Zeigen Sie, daß die Funktion (11.43) überall stetig differenzierbar ist und daß sie die obere Einhüllende $x \mapsto 2 - \frac{1}{5} x^4$ und die untere Einhüllende $x \mapsto 2 - \frac{3}{5} x^4$ hat.

11.2.20 Beispiel (*) Das folgende Beispiel einer überall stetigen, aber nirgends differenzierbaren Funktion, geht auf eine Idee des japanischen Mathematikers TAKAGI TEIJI (高木貞治) zurück. Sei

$$\begin{aligned} f_0(x) &:= \{|x - \ell|, \quad x \in [\ell - \frac{1}{2}, \ell + \frac{1}{2}], \ell \in \mathbb{Z}, \\ f_n(x) &:= \frac{1}{4^n} f_0(4^n x). \end{aligned}$$



Damit definieren wir TAKAGIS Funktion t durch

$$t(x) := \sum_{n=0}^{\infty} f_n(x).$$

Die Funktion f_0 ist die periodische Fortsetzung der auf $[-\frac{1}{2}, \frac{1}{2}]$ eingeschränkten Betragsfunktion. f_0 hat die Periodenlänge 1 und Maxima der Höhe $\frac{1}{2}$ an den Stellen $\frac{1}{2} + \ell$, $\ell \in \mathbb{Z}$. f_0 ist offensichtlich stetig. Durch die Konstruktion $f_1(x) := \frac{1}{4} f_0(4x)$ wird f_0 um den Faktor 4 in x- und in y-Richtung gestaucht. f_1 hat also die Periodenlänge $\frac{1}{4}$ und Maxima der Höhe $\frac{1}{8}$ an den Stellen $\frac{1}{8} + \frac{\ell}{4}$, $\ell \in \mathbb{Z}$. Auf diese Weise fortlaufend erhalten wir die periodische, stetige Funktion f_n mit Periodenlänge $\frac{1}{4^n}$ und Maxima der Höhe $\frac{1}{4^n} \cdot \frac{1}{2}$ an den Stellen $\frac{1}{4^n} (\ell + \frac{1}{2})$, $\ell \in \mathbb{Z}$.

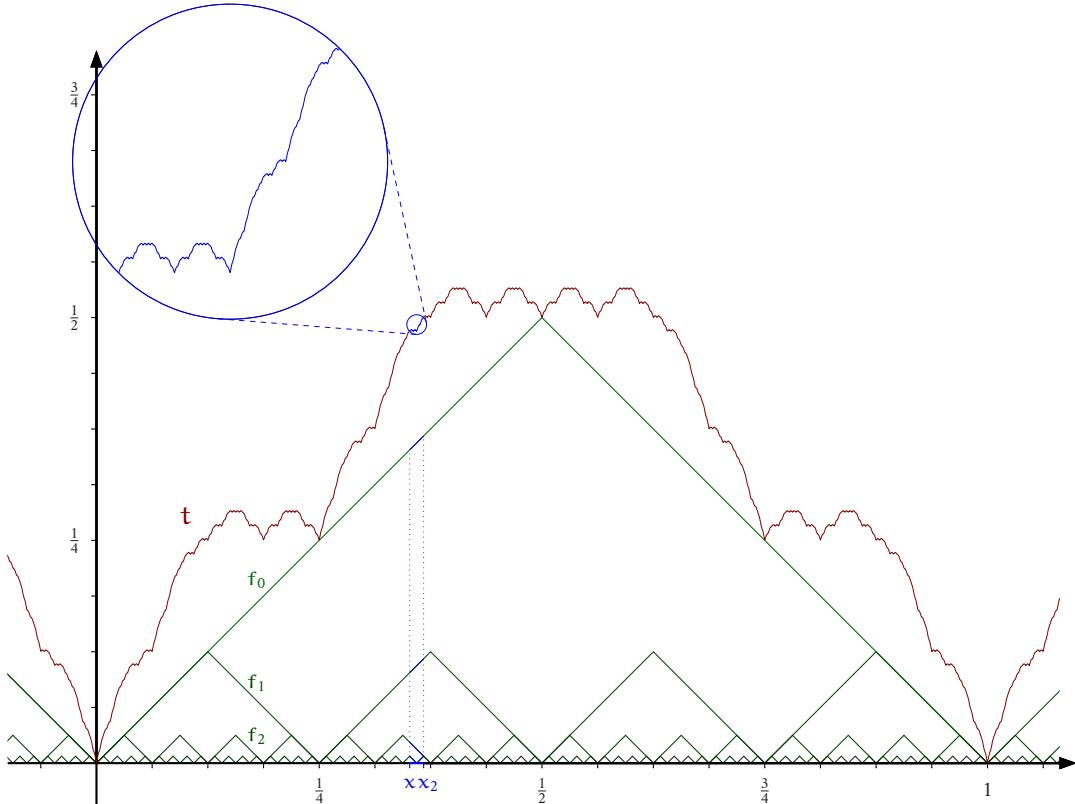


Abb. 11.8 TAKAGI Funktion

Die Reihe, durch die $t(x)$ definiert wird, ist auf \mathbb{R} gleichmäßig konvergent, denn sie hat eine gleichmäßige Majorante:

$$\sum_{n=0}^{\infty} |f_n(x)| = \sum_{n=0}^{\infty} f_n(x) \leq \frac{1}{2} \sum_{n=0}^{\infty} \frac{1}{4^n} < \frac{1}{2} \sum_{n=0}^{\infty} \frac{1}{4^n} = \frac{1}{2} \cdot \frac{4}{3} = \frac{2}{3}.$$

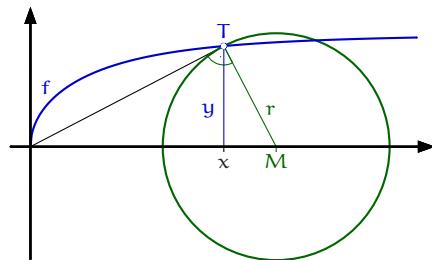
Daher ist t ein gleichmäßiger Grenzwert stetiger Funktionen und somit stetig (Satz 11.1.30). Daß t an keiner Stelle differenzierbar ist, liegt im Wesentlichen daran, daß die Funktionen f_n abwechselnd aus Geradenstücken mit Steigung 1 und -1 zusammengesetzt sind. Für den Differenzenquotienten von t an der Stelle x bilden wir daher eine Folge $(x_n)_{n \in \mathbb{N}}$ durch $x_n := x \pm \frac{1}{4^n} \cdot \frac{1}{4}$, wobei $+$ oder $-$ so gewählt wird, daß x und x_n zu einem gemeinsamen Geradenstück der Funktion f_n gehört. Das ist immer möglich, denn f_n hat die Periodenlänge $\frac{1}{4^n}$, die jeweils zur Hälfte durch ein Geradenstück mit Steigung 1 bzw. -1 belegt ist. Da sich immer zwei benachbarte Dreiecke von f_n in der linken und zwei in der rechten Hälfte eines Dreiecks von

f_{n-1} befinden, gehören x und x_n auch zu einem gemeinsamen Geradenstück von f_{n-1} . Diese Eigenschaft setzt sich auf alle f_r , $r \leq n$ fort. Für f_{n+1} gilt dagegen $f_{n+1}(x_n) = f_{n+1}(x)$, denn x_n befindet sich genau eine Periodenlänge $\frac{1}{4^{n+1}}$ von x entfernt. Ähnliches gilt für die Funktionen f_k , $k > n+1$, denn $\frac{1}{4^{n+1}}$ ist ein ganzzahliges Vielfaches ihrer Periodenlänge. Das bedeutet $\frac{1}{x_n - x} (f_k(x_n) - f_k(x)) = 0$ für alle $k > n$, so daß im Differenzenquotienten von t nur eine endlich Summe zu bilden ist:

$$\frac{t(x_n) - t(x)}{x_n - x} = \sum_{k=0}^n \frac{f_k(x_n) - f_k(x)}{x_n - x} = \sum_{k=0}^n \pm 1.$$

Die letzte Summe ist eine Konsequenz der Konstruktion von $(x_n)_{n \in \mathbb{N}}$, gemäß der x_n und x immer zu einem gemeinsamen Geradenstück der Steigung 1 oder -1 gehören. Eine solche Summe aus 1 und -1 (deren genaue Abfolge wir für den Beweis gar nicht wissen müssen) kann nicht konvergieren, denn die Minimalvoraussetzung für die Konvergenz einer Reihe, nämlich daß die Summanden eine Nullfolge bilden, ist nicht gegeben. Damit hat die Folge der Differenzenquotienten an der Stelle x keinen Grenzwert, so daß t dort nicht differenzierbar ist.

11.2.21 A



Zeigen Sie, daß durch $f(x) := \frac{1}{\sqrt{2}} \sqrt{x\sqrt{x^2 + 4r^2} - x^2}$ die Ordinate y des Punktes T in Abhängigkeit von x beschrieben wird.

Die Skizze legt die Vermutung nahe, daß $\lim_{x \rightarrow \infty} f(x) = r$ gelten sollte. Beweisen Sie das. Dabei können Sie folgendermaßen vorgehen: Wegen der Stetigkeit der Wurzel genügt es zu zeigen, daß der Ausdruck unter der ersten Wurzel gegen $2r^2$ konvergiert. Die Idee dafür könnte sein, diesen Ausdruck in einen Differenzen-

quotienten an der Stelle 0 umzuschreiben, um den Grenzwert als Ableitung einer geeigneten Funktion an dieser Stelle zu erhalten (oder Sie verwenden auf geeignete Weise das dritte Binom).

11.2.22 Für Funktionen, wie etwa $\varphi(x) := e^{-\frac{1}{x}}$ für $x > 0$ und $\varphi(x) := 0$ für $x \leq 0$ stellt sich die Frage nach der Differenzierbarkeit eigentlich nur an der Nahtstelle $x = 0$ (vergl. 11.3.5 viii)). Da wir hier keine Ableitungsregel zur Verfügung haben, müssen wir normalerweise den Weg über den Differenzenquotienten gehen. Bei dem ähnlichen Beispiel 11.2.14 ging daran auch kein Weg vorbei, denn diese Ableitungsfunktion war in $x = 0$ unstetig. Wenn aber zu erwarten ist, daß die Ableitung stetig ist, wäre es wünschenswert, ein hinreichendes Kriterium zur Verfügung zu haben, das mit der Ableitung rechts und links der betreffenden Stelle auskommt. Genauer heißt das: Existieren der linksseitige und der rechtsseitige Grenzwert der Ableitungen an einer Stelle und sind diese gleich, so ist die Funktion dort sogar stetig differenzierbar – vorausgesetzt, die Funktion ist dort überhaupt stetig.

11.2.23 Satz $f : [a, b] \rightarrow \mathbb{R}$ sei stetig und für ein $x_0 \in (a, b)$ auf den Intervallen (a, x_0) und (x_0, b) differenzierbar. Falls die beiden Grenzwerte

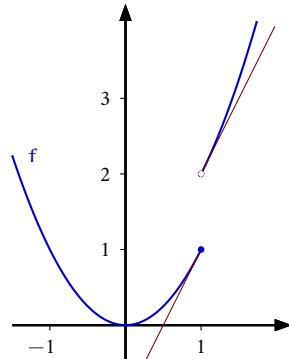
$$\lim_{x \rightarrow x_0^-} f'(x) \text{ und } \lim_{x \rightarrow x_0^+} f'(x)$$

existieren und darüber hinaus gleich sind, ist f an der Stelle x_0 stetig differenzierbar, mit der Ableitung $f'(x_0) = \lim_{x \rightarrow x_0^\pm} f'(x)$.

Beweis. α sei vorläufig der Grenzwert $\lim_{x \rightarrow x_0^\pm} f'(x)$. Wir müssen zunächst die Existenz von $\lim_{x \rightarrow x_0} \frac{f(x) - f(x_0)}{x - x_0}$ nachweisen. Nach Satz 11.1.6 ist das äquivalent dazu, daß der linksseitige Grenzwert $\lim_{x \rightarrow x_0^-} \frac{f(x) - f(x_0)}{x - x_0}$ und der rechtsseitige $\lim_{x \rightarrow x_0^+} \frac{f(x) - f(x_0)}{x - x_0}$ existiert und daß beide Werte gleich sind. Für jedes $x_n \in (a, x_0)$ einer zulässigen Folge $(x_n)_{n \in \mathbb{N}}$, die gegen x_0 konvergiert, gibt es nach dem Mittelwertsatz ein $\xi_n \in (x_n, x_0)$, so daß $\frac{f(x_n) - f(x_0)}{x_n - x_0} = f'(\xi_n)$. Mit $(x_n)_{n \in \mathbb{N}}$ konvergiert auch $(\xi_n)_{n \in \mathbb{N}}$ gegen x_0 , so daß die Voraussetzung des Satzes die Existenz des linksseitigen Grenzwertes $\lim_{x \rightarrow x_0^-} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{n \rightarrow \infty} \frac{f(x_n) - f(x_0)}{x_n - x_0} = \lim_{n \rightarrow \infty} f'(\xi_n) = \alpha$ sicherstellt. Mit derselben Argumentation erhalten wir auch die Existenz des rechtsseitigen Grenzwerts $\lim_{y \rightarrow x_0^+} \frac{f(y) - f(x_0)}{y - x_0} = \lim_{n \rightarrow \infty} \frac{f(y_n) - f(x_0)}{y_n - x_0} = \lim_{n \rightarrow \infty} f'(\eta_n) = \alpha$, mit einem $\eta_n \in (x_0, y_n)$. Dabei ist jetzt $(y_n)_{n \in \mathbb{N}}$ eine zulässige Folge, die von rechts gegen x_0 konvergiert. Das zeigt die Existenz von $f'(x_0) = \alpha = \lim_{x \rightarrow x_0^\pm} f'(x)$ und gleichzeitig, daß f' an der Stelle x_0 stetig ist (s. Satz 11.1.5). \square

Im Verlauf des Beweises ist nicht sofort erkennbar, wo die Stetigkeit von f an der Stelle x_0 genau gebraucht wird. Kann es nicht sein, daß die Voraussetzungen noch stark genug sind, wenn man die Stetigkeit in x_0 nicht fordert, daß sie also automatisch folgt. Die Antwort ist nein, denn man muß sich nur eine stetig differenzierbare Funktion nehmen, sagen wir $x \mapsto x^2$, und sie z. B. an der Stelle $x_0 = 1$ auseinander schneiden:

$$f(x) := \begin{cases} x^2 & x \leq 1 \\ x^2 + 1 & x > 1 \end{cases}.$$



Dann ist $f'(x) = 2x$ für $x < 1$ und für $x > 1$. Offensichtlich ergibt der links- und der rechtsseitige Grenzwert den Wert 2, aber die Funktion ist an der Stelle 1 natürlich nicht differenzierbar. Wo also wurde die Stetigkeit im Beweis verwendet? Die Voraussetzungen des Mittelwertsatzes verlangen sie. f muß auf $[x_n, x_0]$ bzw. $[x_0, y_n]$ stetig und auf (x_n, x_0) bzw. (x_0, y_n) differenzierbar sein.

11.3 Ableitung von Potenzreihen

Eine Funktionen f , die durch eine Potenzreihe $\sum_{k=0}^{\infty} a_k x^k$ mit positivem Konvergenzradius R gegeben ist, kann man formal ableiten, indem man in der Summe jeden Summanden $a_k x^k$ ableitet. Auf diese Weise erhält man eine neue Potenzreihe $\sum_{k=1}^{\infty} k a_k x^{k-1}$, die denselben Konvergenzradius hat:

$$\frac{1}{\limsup_{k \rightarrow \infty} \sqrt[k]{k |a_k|}} = \frac{1}{\limsup_{k \rightarrow \infty} \sqrt[k]{k} \sqrt[k]{|a_k|}} = \frac{1}{\limsup_{k \rightarrow \infty} \sqrt[k]{|a_k|}} = R.$$

Allerdings müssen wir uns erst noch davon überzeugen, daß es sich bei dieser Potenzreihe auch wirklich um die Ableitung von f handelt. Dafür ist etwas Vorbereitung nötig.

Da das Polynom $y \mapsto y^k - x^k$ bei $y = x$ eine Nullstelle hat, muß die Polynomdivision mit dem Linearfaktor $y - x$ ohne Rest aufgehen. Als Ergebnis erhält man

$$\begin{aligned} y^k - x^k &= (y - x) \sum_{\ell=0}^{k-1} y^{k-1-\ell} x^\ell \\ &= (y - x)(y^{k-1} + y^{k-2}x + y^{k-3}x^2 + \dots + y^2x^{k-3} + yx^{k-2} + x^{k-1}). \end{aligned} \quad (11.44)$$

Um zu zeigen, daß f an jeder Stelle $x \in (-R, R)$ differenzierbar ist, arbeiten wir mit der linearisierten Version 11.2.3. Für $x, y \in (-R, R)$ gilt

$$f(y) - f(x) = \sum_{k=1}^{\infty} a_k (y^k - x^k) = \sum_{k=1}^{\infty} a_k \sum_{\ell=0}^{k-1} y^{k-1-\ell} x^\ell (y - x) = \Theta(y)(y - x),$$

mit der Funktion $\Theta(y) := \sum_{k=1}^{\infty} \sum_{n=0}^{k-1} a_k y^{k-1-n} x^n$. Wir müssen zeigen, daß Θ an der Stelle x stetig ist. Das machen wir genauso, wie beim Beweis der Stetigkeit von Potenzreihen, indem wir die gleichmäßige Konvergenz der Reihe nachweisen. Satz 11.1.30 liefert dann das gewünschte Ergebnis. Wir wählen dazu eine Zahl $0 < r < R$, so daß $x, y \in [-r, r]$ gilt. Das Konvergenz-Kriterium für die Reihe $\sum_{k=1}^{\infty} k |a_k| r^{k-1}$ ist erfüllt, da die formale Ableitung der Potenzreihe für $f(r)$ den Konvergenzradius R hat und daher nach Korollar 11.1.31 auf $[-r, r]$ absolut und gleichmäßig konvergiert. Es gibt also ein $n_\varepsilon \in \mathbb{N}$, so daß

$$\begin{aligned} \sum_{k=n+1}^{\infty} \left| \sum_{\ell=0}^{k-1} a_k y^{k-1-\ell} x^\ell \right| &\leq \sum_{k=n+1}^{\infty} |a_k| \sum_{\ell=0}^{k-1} |y|^{k-1-\ell} |x|^\ell \leq \sum_{k=n+1}^{\infty} |a_k| \sum_{\ell=0}^{k-1} r^{k-1-\ell} r^\ell \\ &= \sum_{k=n+1}^{\infty} k |a_k| r^{k-1} < \varepsilon \end{aligned}$$

für alle $n \geq n_\varepsilon$ gilt. Damit konvergiert $\Theta(y)$ absolut und gleichmäßig bzgl. $y \in [-r, r]$ und ist daher nach Satz 11.1.30 stetig. Satz 11.2.3 beweist nun, daß f an der Stelle x differenzierbar ist und dort die Ableitung $f'(x) = \Theta(x) = \sum_{k=1}^{\infty} k a_k x^{k-1}$ hat, die aus der Potenzreihe von f durch Differenzieren der einzelnen Summanden entsteht. Wir haben dieses Ergebnis für Potenzreihen mit Entwicklungspunkt $x_0 = 0$ erhalten. Die allgemeine Situation $f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k$ kann aber durch $g(x) := f(x + x_0) = \sum_{k=0}^{\infty} a_k |x|^k$ auf die spezielle zurückgeführt werden. g ist differenzierbar. Da f aus g durch die lineare Transformation

$f(x) = g(x - x_0)$ hervorgeht, ist nach der Kettenregel auch f differenzierbar, mit der Ableitung $f'(x) = g'(x - x_0) = \sum_{k=1}^{\infty} k a_k (x - x_0)^{k-1}$. Damit ist der folgende Satz bewiesen.

11.3.1 Satz Eine Potenzreihe $f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k$ mit einem Konvergenzradius $R > 0$ ist für $|x - x_0| < R$ beliebig oft differenzierbar. Ihre Ableitung ist durch die Potenzreihe

$$f'(x) = \sum_{k=1}^{\infty} k a_k (x - x_0)^{k-1} \quad (11.45)$$

mit demselben Konvergenzradius R gegeben.

Beweis. Der Beweis ist im Wesentlichen oben schon erbracht worden. Daß f beliebig oft differenzierbar ist, ergibt sich natürlich daraus, daß f' selbst wieder eine Potenzreihe ist, auf die das Ergebnis erneut angewendet werden kann. \square

11.3.2 Beispiel Von der Exponentialfunktion \exp wissen wir bereits $\exp' = \exp$ (Satz 11.2.6). Dieses Ergebnis sollte sich aber auch mittels (11.45) gewinnen lassen:

$$\exp'(x) = \sum_{k=1}^{\infty} k \frac{x^{k-1}}{k!} = \sum_{k=1}^{\infty} \frac{x^{k-1}}{(k-1)!} = \sum_{\ell=0}^{\infty} \frac{x^{\ell}}{\ell!} = \exp(x).$$

11.3.3 Die Regeln von DE L'HOSPITAL Eine direkte Folgerung aus dem verallgemeinerten Mittelwertsatz 11.2.12 sind die Regeln von DE L'HOSPITAL zur Bestimmung von sogenannten *unbestimmten Ausdrücken* der Form $\frac{0}{0}$, oder $\frac{\infty}{\infty}$. Damit ist folgendes gemeint: Zwei stetig differenzierbare Funktionen f und g nehmen an einer Stelle a den Wert $f(a) = g(a) = 0$ an. Dann hat die Funktion $\frac{f}{g}$ an der Stelle a zunächst keinen Funktionswert, denn der wäre eben der unbestimmte Ausdruck $\frac{f(a)}{g(a)} = \frac{0}{0}$. Gesucht ist eine Möglichkeit, für diese Funktion doch einen Funktionswert an der Stelle a so festzulegen, daß sie stetig bleibt. Das kann nur durch Bestimmung des Grenzwerts $\lim_{x \rightarrow a} \frac{f(x)}{g(x)}$ gelingen, vorausgesetzt, dieser Grenzwert existiert überhaupt.

Das Standardbeispiel für diese Situation ist die Funktion $x \mapsto \frac{\sin(x)}{x}$ für $x \neq 0$ (auch wenn wir schon wissen, daß der Grenzwert der Funktionswerte für $x \rightarrow 0$ den Wert 1 hat, vergl. 11.21). An diesem Beispiel kann man die Idee verstehen, die zur Lösung solcher Probleme verwendet wird. Die simple Beobachtung ist, daß $\frac{\sin(x)}{x}$ eigentlich den Differenzenquotient für die Ableitung von \sin an der Stelle $x_0 = 0$ darstellt:

$$\frac{\sin(x)}{x} = \frac{\sin(x) - \sin(0)}{x - 0} \xrightarrow{x \rightarrow 0} \sin'(0) = \cos(0) = 1.$$

Wenn wir das Beispiel etwas abwandeln, etwa $\frac{\sin(x)}{\sin(2x)}$, so handelt es sich nicht mehr um einen Differenzenquotienten, aber um einen Ausdruck, der noch mit dem verallgemeinerten Mittelwertsatz untersucht werden kann:

$$\frac{\sin(x)}{\sin(2x)} = \frac{\sin(x) - \sin(0)}{\sin(2x) - \sin(0)} = \frac{\cos(\xi)}{2\cos(2\xi)},$$

mit einer Stelle $\xi \in (0, x)$. Für $x \rightarrow 0$ strebt ξ nach dem Sandwich-Prinzip ebenfalls gegen 0 und wegen der Stetigkeit der Funktion $x \mapsto \frac{\cos(x)}{2\cos(2x)}$ konvergiert $\frac{\cos(\xi)}{2\cos(2\xi)}$ gegen $\frac{\cos(0)}{2\cos(0)} = \frac{1}{2}$.

Damit ist die Existenz des Grenzwerts $\lim_{x \rightarrow 0} \frac{\sin(x)}{\sin(2x)}$ nachgewiesen und sein Wert $\frac{1}{2}$ ist bestimmt. Diese Beispiel zeigt, was für den allgemeinen Fall zu erwarten ist. Falls der Grenzwert $\lim_{x \rightarrow a} \frac{f'(x)}{g'(x)}$ existiert, dann existiert auch $\lim_{x \rightarrow a} \frac{f(x)}{g(x)}$ und beide Werte stimmen überein. Jedoch gilt die Umkehrung im Allgemeinen nicht, d. h., wir können den Grenzwert $\lim_{x \rightarrow a} \frac{f(x)}{g(x)}$ nicht dafür verwenden, um die Existenz von $\lim_{x \rightarrow a} \frac{f'(x)}{g'(x)}$ nachzuweisen. Als Beispiel wählen wir $f(x) := x^2 \sin(\frac{1}{x})$ für $x \neq 0$, $f(0) := 0$ und $g(x) := \sin(x)$. Die Existenz von $\lim_{x \rightarrow 0} \frac{f(x)}{g(x)}$ können wir direkt nachrechnen:

$$\lim_{x \rightarrow 0} \frac{x^2 \sin(\frac{1}{x})}{\sin(x)} = \lim_{x \rightarrow 0} \frac{x}{\sin(x)} \cdot \lim_{x \rightarrow 0} x \sin(\frac{1}{x}) = 1 \cdot 0 = 0.$$

Der Grenzwert $\lim_{x \rightarrow 0} \frac{f'(x)}{g'(x)}$ existiert jedoch nicht, denn

$$\frac{f'(x)}{g'(x)} = \frac{2x \sin(\frac{1}{x}) - \cos(\frac{1}{x})}{\cos(x)} = \frac{2x \sin(\frac{1}{x})}{\cos(x)} - \frac{\cos(\frac{1}{x})}{\cos(x)}.$$

Der Bruch $\frac{2x \sin(\frac{1}{x})}{\cos(x)}$ konvergiert gegen 0, aber der zweite oszilliert unendlich oft zwischen ≈ -1 und ≈ 1 . Um das zu sehen, wählen wir die Nullfolge $x_n := \frac{1}{n\pi}$. Dann gilt $\frac{\cos(\frac{1}{x_n})}{\cos(x_n)} = \frac{\cos(n\pi)}{\cos(x_n)} = \frac{(-1)^n}{\cos(\frac{1}{n\pi})}$. Während der Nenner gegen 1 konvergiert, oszilliert der Zähler unendlich oft zwischen 1 und -1.

11.3.4 Satz (Die Regeln von L'HOSPITAL) *f und g seien Funktionen, die auf dem Intervall (a, b) differenzierbar sind. Dabei sind auch die Fälle $(-\infty, b)$, (a, ∞) und $(-\infty, \infty) = \mathbb{R}$ zulässig. Außerdem sei $g'(x) \neq 0$ auf (a, b) .*

- i) Es gelte $\lim_{x \rightarrow b^-} f(x) = \lim_{x \rightarrow b^-} g(x) = 0$. Falls dann der Grenzwert $\lim_{x \rightarrow b^-} \frac{f'(x)}{g'(x)}$ existiert, dann existiert auch $\lim_{x \rightarrow b^-} \frac{f(x)}{g(x)}$ und beide Werte stimmen überein. Entsprechende Aussagen gelten für $x \rightarrow a^+$, $x \rightarrow \infty$ und $x \rightarrow -\infty$.
- ii) Gilt $f(x) \rightarrow \infty$ und $g(x) \rightarrow \infty$ für $x \rightarrow b^-$ und existiert $\lim_{x \rightarrow b^-} \frac{f'(x)}{g'(x)}$, dann existiert auch $\lim_{x \rightarrow b^-} \frac{f(x)}{g(x)}$ und beide Grenzwerte stimmen überein. Wieder gelten die entsprechenden Aussagen auch für die Fälle $x \rightarrow a^+$, $x \rightarrow \infty$ und $x \rightarrow -\infty$.

Bevor wir den Satz beweisen, müssen wir noch präzise fassen, was unter $f(x) \rightarrow \infty$ für $x \rightarrow b$ zu verstehen ist. Anschaulich soll es bedeuten, daß $f(x)$ beliebig groß wird, wenn x dem Wert b immer näher kommt. Genauer:

$$\begin{aligned} f(x) \rightarrow \infty \text{ für } x \rightarrow b &\Leftrightarrow \forall r \in \mathbb{R}^+ \exists \delta > 0 \forall x \neq b \wedge |x - b| < \delta f(x) > r, \\ f(x) \rightarrow -\infty \text{ für } x \rightarrow b &\Leftrightarrow \forall r \in \mathbb{R}^+ \exists \delta > 0 \forall x \neq b \wedge |x - b| < \delta f(x) < -r, \\ f(x) \rightarrow \infty \text{ für } x \rightarrow \infty &\Leftrightarrow \forall r \in \mathbb{R}^+ \exists \delta > 0 \forall x > \delta f(x) > r, \\ f(x) \rightarrow -\infty \text{ für } x \rightarrow \infty &\Leftrightarrow \forall r \in \mathbb{R}^+ \exists \delta > 0 \forall x < -\delta f(x) > r, \\ f(x) \rightarrow \infty \text{ für } x \rightarrow -\infty &\Leftrightarrow \forall r \in \mathbb{R}^+ \exists \delta > 0 \forall x > \delta f(x) < -r, \\ f(x) \rightarrow -\infty \text{ für } x \rightarrow -\infty &\Leftrightarrow \forall r \in \mathbb{R}^+ \exists \delta > 0 \forall x < -\delta f(x) < -r, \end{aligned}$$

$$f(x) \rightarrow -\infty \text{ für } x \rightarrow -\infty \Leftrightarrow \forall r \in \mathbb{R}^+ \exists \delta > 0 \forall x < -\delta f(x) < -r.$$

Das heißt etwa im ersten Fall, daß es zu einer (großen) Zahl $r > 0$ immer eine Umgebung $(b - \delta, b + \delta)$ von b gibt, so daß f auf $(b - \delta, b + \delta) \setminus \{b\}$ nur noch Funktionswerte annimmt, die größer als r sind. Ein Beispiel ist eine Funktion, die an der Stelle b einen Pol ohne Vorzeichenwechsel hat, etwa $f(x) := \frac{1}{(x-b)^2}$.

Genau genommen, müssen wir uns auch darüber verständigen, was wir unter $x \rightarrow \infty$ bzw. $x \rightarrow -\infty$ verstehen wollen: Im Geiste unserer Vereinbarung 11.1.2 führen wir das auf sogenannte *bestimmt divergente* Folgen $(x_n)_{n \in \mathbb{N}}$ zurück. Sie sind dadurch gekennzeichnet, daß für jede (große) Zahl $r > 0$ ein Index $n_r \in \mathbb{N}$ existiert, so daß $x_n > r$ bzw. $x_n < -r$ für alle $n \geq n_r$ gilt. Die Folge $(2n)_{n \in \mathbb{N}}$ aller geraden Zahlen ist eine solche Folge, $((-1)^n 2n)_{n \in \mathbb{N}}$ jedoch nicht.

Man kann in den Definitionen noch verlangen, daß $s \rightarrow b^-$ statt $x \rightarrow b$ erfüllt ist, daß sich x also nur von einer Seite dem Wert b nähert. Das würde z. B. für die erste Äquivalenz bedeuten:

$$f(x) \rightarrow \infty \text{ für } x \rightarrow b^- \Leftrightarrow \forall r \in \mathbb{R}^+ \exists \delta > 0 \forall x < b \wedge b - x < \delta f(x) > r.$$

Ein Beispiel ist $f(x) := \frac{1}{1-x} \rightarrow \infty$ für $x \rightarrow 1^-$. f ist eine Funktion, die einem Pol mit Vorzeichenwechsel bei 1 hat. Dagegen ist $f(x) := \left| \frac{\cos(\frac{1}{x})}{\sin(x)} \right| \rightarrow \infty$ für $x \rightarrow 0^+$ mit unserer Definition nicht abgedeckt, denn diese Funktion nimmt auf $(0, 1)$ zwar beliebig große Werte an, wird aber an den unendlich vielen Stellen $\left(\frac{2}{(2n+1)\pi} \right)_{n \in \mathbb{N}}$ immer wieder 0. Daher kann es zu einem $r > 0$ kein $\delta > 0$ geben, so daß $f(x) > r$ für alle $0 < x < \delta$ gilt.

Beweis.

Zu i): Wir gehen von $\lim_{x \rightarrow b^-} f(x) = \lim_{x \rightarrow b^-} g(x) = 0$ und $\lim_{x \rightarrow b^-} \frac{f'(x)}{g'(x)} =: c$ aus (den Fall $x \rightarrow a^+$ beweist man analog). Dann können f und g durch $f(b) := 0$ und $g(b) := 0$ zu stetigen Funktionen auf $(a, b]$ fortgesetzt werden (klar, denn für $x \rightarrow b^-$ gilt $f(x) \rightarrow 0 = f(b)$ und $g(x) \rightarrow 0 = g(b)$). Daher sind die Voraussetzungen des verallgemeinerten Mittelwertsatzes 11.2.12 für jedes Intervall $[x, b]$ erfüllt ($a < x < b$). Für jedes x_n einer Folge $(x_n)_{n \in \mathbb{N}}$ aus (a, b) , die gegen b konvergiert, gibt es demnach ein $\xi_n \in (x_n, b)$, mit der Eigenschaft

$$\frac{f(x_n)}{g(x_n)} = \frac{f(x_n) - f(b)}{g(x_n) - g(b)} = \frac{f'(\xi_n)}{g'(\xi_n)}.$$

Nach dem Sandwich-Prinzip strebt auch die Folge $(\xi_n)_{n \in \mathbb{N}}$ gegen b , weshalb die rechte Seite $\frac{f'(\xi_n)}{g'(\xi_n)}$, den Voraussetzungen gemäß, gegen c konvergiert. Das zeigt $\lim_{n \rightarrow \infty} \frac{f(x_n)}{g(x_n)} = c$ und, da die Folge $(x_n)_{n \in \mathbb{N}}$ beliebig war, auch $\lim_{x \rightarrow b^-} \frac{f(x)}{g(x)} = c$.

Die Fälle $x \rightarrow \pm\infty$ zeigen wir am Ende des Beweises für i) und ii) gemeinsam.

Zu ii): Dieser Fall ist technisch wesentlich aufwendiger zu beweisen, da wir die Funktionen nicht stetig auf den Randpunkt b fortsetzen können. Das macht es schwerer den verallgemeinerten Mittelwertsatz anzuwenden. Wir führen einen Widerspruchsbeweis, d. h., wir gehen von $\lim_{x \rightarrow b^-} \frac{f'(x)}{g'(x)} = c$ aus und nehmen an, daß $\frac{f(x)}{g(x)}$ nicht gegen c konvergiert. Das bedeutet, es gibt ein $\varepsilon > 0$ und eine zulässige Folge $(x_n)_{n \in \mathbb{N}}$ mit Grenzwert b , so daß für unendlich viele $n \in \mathbb{N}$

die Abschätzung $\left| \frac{f(x_n)}{g(x_n)} - c \right| \geq \varepsilon$ gilt. Wenn wir zu der Teilfolge übergehen, die nur aus diesen Folgengliedern besteht, können wir die Gültigkeit dieser Ungleichung für alle $n \in \mathbb{N}$ annehmen. Da $|f(x_n)|$ und $|g(x_n)|$ gegen ∞ divergieren, gibt es für jedes $n \in \mathbb{N}$ ein $m_n \in \mathbb{N}$, so daß $\left| \frac{f(x_n)}{g(x_{m_n})} \right| < \frac{1}{n}$ und $\left| \frac{g(x_n)}{g(x_{m_n})} \right| < \frac{1}{n}$ gilt. Das verwenden wir im verallgemeinerten Mittelwertsatz:

$$\frac{f(x_{m_n}) - f(x_n)}{g(x_{m_n}) - g(x_n)} = \frac{f'(\xi_n)}{g'(\xi_n)},$$

mit einem ξ_n zwischen x_n und x_{m_n} . (ξ_n) ist daher eine zulässige Folge, die gegen b konvergiert. Wir haben somit $\lim_{n \rightarrow \infty} \frac{f'(\xi_n)}{g'(\xi_n)} = c$ und können folgendermaßen abschätzen:

$$\begin{aligned} \varepsilon &\leq \left| \frac{f(x_{m_n})}{g(x_{m_n})} - c \right| = \left| \frac{f(x_{m_n}) - f(x_n)}{g(x_{m_n}) - g(x_n)} \left(1 - \frac{g(x_n)}{g(x_{m_n})} \right) + \frac{f(x_n)}{g(x_{m_n})} - c \right| \\ &= \left| \left(\frac{f'(\xi_n)}{g'(\xi_n)} - c \right) \left(1 - \frac{g(x_n)}{g(x_{m_n})} \right) + \frac{f(x_n)}{g(x_{m_n})} - c \frac{g(x_n)}{g(x_{m_n})} \right| \\ &\leq \left| \frac{f'(\xi_n)}{g'(\xi_n)} - c \right| \left(1 + \frac{|g(x_n)|}{|g(x_{m_n})|} \right) + \frac{|f(x_n)|}{|g(x_{m_n})|} + |c| \frac{|g(x_n)|}{|g(x_{m_n})|} \\ &< \left| \frac{f'(\xi_n)}{g'(\xi_n)} - c \right| \left(1 + \frac{1}{n} \right) + \frac{1}{n} (1 + |c|) \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

Das ist der gesuchte Widerspruch. Also muß doch $\lim_{x \rightarrow b^-} \frac{f(x)}{g(x)} = c$ gelten.

Bleibt der Fall $x \rightarrow \infty$. Wir gehen von $\lim_{x \rightarrow \infty} \frac{f'(x)}{g'(x)} = c$ und $\lim_{x \rightarrow \infty} f(x) = \lim_{x \rightarrow \infty} g(x) = 0$ bzw. $f(x) \xrightarrow{x \rightarrow \infty} \infty$ und $g(x) \xrightarrow{x \rightarrow \infty} \infty$ aus. Um das auf die bereits behandelten Fälle zurückzuführen, definieren wir die Funktionen $F(x) := f(\frac{1}{x})$ und $G(x) := g(\frac{1}{x})$. Für sie gilt $\lim_{x \rightarrow 0^+} F(x) = \lim_{x \rightarrow 0^+} G(x) = 0$, bzw. $F(x) \rightarrow \infty$ und $G(x) \rightarrow \infty$ für $x \rightarrow 0^+$. Außerdem haben wir

$$\frac{F'(x)}{G'(x)} = \frac{-\frac{1}{x^2} f'(\frac{1}{x})}{-\frac{1}{x^2} g'(\frac{1}{x})} = \frac{f'(\frac{1}{x})}{g'(\frac{1}{x})} \xrightarrow{x \rightarrow 0^+} c.$$

Damit wissen wir $c = \lim_{x \rightarrow 0^+} \frac{F(x)}{G(x)}$. Für jede Folge $(x_n)_{n \in \mathbb{N}}$, die bestimmt gegen ∞ divergiert, konvergiert $(\frac{1}{x_n})_{n \in \mathbb{N}}$ gegen 0^+ . Es folgt also $c = \lim_{n \rightarrow \infty} \frac{F(\frac{1}{x_n})}{G(\frac{1}{x_n})} = \lim_{n \rightarrow \infty} \frac{f(x_n)}{g(x_n)}$ und, da die Folge $(x_n)_{n \in \mathbb{N}}$ beliebig war: $c = \lim_{x \rightarrow \infty} \frac{f(x)}{g(x)}$. \square

Man mache sich klar, daß wir mit dem Satz auch die Fälle $f(x) \rightarrow -\infty$, $g(x) \rightarrow \infty$ und $f(x) \rightarrow \infty$, $g(x) \rightarrow -\infty$, sowie $f(x) \rightarrow -\infty$, $g(x) \rightarrow -\infty$ abgedeckt haben.

11.3.5 Beispiel Für die Beispiele verwenden wir auch den natürlichen Logarithmus \ln , der auf Seite 306 eingeführt wird.

i) Wir wollen die Funktion $h(x) := x \ln(x)$ von ihrem Definitionsbereich $(0, \infty)$ stetig auf $[0, \infty)$ fortsetzen. Dafür müssen wir versuchen, den Grenzwert $\lim_{x \rightarrow 0^+} x \ln(x) = \lim_{x \rightarrow 0^+} \frac{\ln(x)}{\frac{1}{x}}$ zu bestimmen. Er ist von der Art „ $\frac{\infty}{\infty}$ “, genau genommen „ $\frac{-\infty}{\infty}$ “. Wenden wir die Regeln von DE L'HOSPITAL an und versuchen $\lim_{x \rightarrow 0^+} \frac{\ln'(x)}{\frac{d}{dx} \frac{1}{x}}$ zu berechnen:

$$\lim_{x \rightarrow 0^+} \frac{\ln'(x)}{\frac{d}{dx} \frac{1}{x}} = \lim_{x \rightarrow 0^+} \frac{\frac{1}{x}}{-\frac{1}{x^2}} = \lim_{x \rightarrow 0^+} -x = 0.$$

Das bedeutet $\lim_{x \rightarrow 0^+} x \ln(x) = 0$. Damit wird $h(x) := \begin{cases} x \ln(x), & x > 0 \\ 0, & x = 0 \end{cases}$ eine stetige Funktion.

ii) $\lim_{x \rightarrow \frac{\pi}{2}^+} \tan(x) - \frac{1}{\cos(x)} = \lim_{x \rightarrow \frac{\pi}{2}^+} \frac{\sin(x) - 1}{\cos(x)} \stackrel{\text{L'H}}{=} \lim_{x \rightarrow \frac{\pi}{2}^+} \frac{\cos(x)}{-\sin(x)} = -\cot\left(\frac{\pi}{2}\right) = 0.$

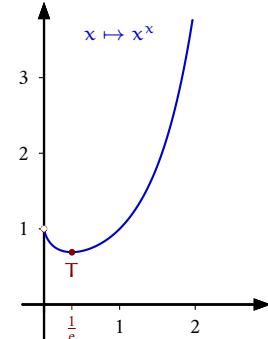
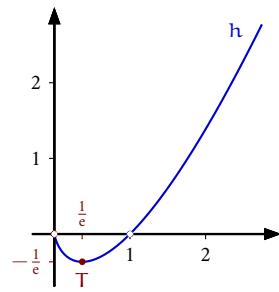
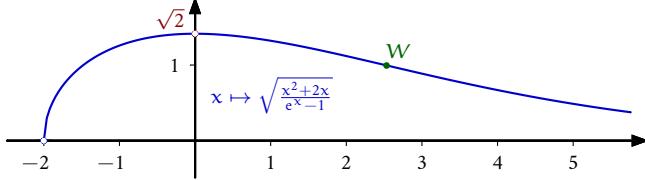
Hier haben wir das Problem „ $\infty - \infty$ “ auf einen unbestimmten Ausdruck der Art „ $\frac{0}{0}$ “ zurückgeführt und die Stetigkeit der Funktion \cot verwendet.

iii) $\lim_{x \rightarrow \infty} x^3 e^{-x} = \lim_{x \rightarrow \infty} \frac{x^3}{e^x} \stackrel{\text{L'H}}{=} \lim_{x \rightarrow \infty} \frac{3x^2}{e^x} \stackrel{\text{L'H}}{=} \lim_{x \rightarrow \infty} \frac{6x}{e^x} \stackrel{\text{L'H}}{=} \lim_{x \rightarrow \infty} \frac{6}{e^x} = 0.$

$$\lim_{x \rightarrow \infty} x^n e^{-x} = \lim_{x \rightarrow \infty} \frac{x^n}{e^x} \stackrel{\text{L'H}}{=} \lim_{x \rightarrow \infty} \frac{nx^{n-1}}{e^x} \stackrel{\text{L'H}}{=} \dots \stackrel{\text{L'H}}{=} \lim_{x \rightarrow \infty} \frac{n!}{e^x} = 0.$$

iv) $\lim_{x \rightarrow 0^+} x^x = \lim_{x \rightarrow 0^+} e^{x \ln(x)} \stackrel{\text{i)}}{=} e^0 = 1.$

v) $\lim_{x \rightarrow 0^\pm} \sqrt{\frac{x^2 + 2x}{e^x - 1}} \stackrel{\text{L'H}}{=} \lim_{x \rightarrow 0^\pm} \sqrt{\frac{2x + 2}{e^x}} = \sqrt{2}.$



Die Stetigkeit der Wurzelfunktion ermöglicht es uns, die Regel von DE L'HOSPITAL unter der Wurzel anzuwenden.

vi) Die Funktion $k(x) := x \ln(|1 - e^{-x}|)$ ist zu untersuchen. Der Definitionsbereich ist $D_k = \mathbb{R} \setminus \{0\}$. Für $x = 0$ erhalten wir einen unbestimmten Ausdruck der Art „ $0 \cdot \infty$ “, der leicht in die Form „ $\frac{\infty}{\infty}$ “ gebracht werden kann:

$$\lim_{x \rightarrow 0^+} x \ln(|1 - e^{-x}|) = \lim_{x \rightarrow 0^+} \frac{\ln(1 - e^{-x})}{\frac{1}{x}} \stackrel{\text{L'H}}{=} \lim_{x \rightarrow 0^+} \frac{-x^2 e^{-x}}{1 - e^{-x}} = \lim_{x \rightarrow 0^+} \frac{-x e^{-x}}{\frac{1 - e^{-x}}{x}} = 0,$$

denn der Zähler konvergiert gegen 0 und der Nenner gegen 1, wie eine erneute Anwendung der Regeln von DE L'HOSPITAL zeigt. Für $x < 0$ ist $k(x) = x \ln(e^{-x} - 1)$.

$$\lim_{x \rightarrow 0^-} x \ln(e^{-x} - 1) = \lim_{x \rightarrow 0^-} \frac{\ln(e^{-x} - 1)}{\frac{1}{x}} \stackrel{\text{L'H}}{=} \lim_{x \rightarrow 0^-} \frac{x^2 e^{-x}}{e^{-x} - 1} = \lim_{x \rightarrow 0^+} \frac{x e^{-x}}{\frac{e^{-x} - 1}{x}} = 0.$$

Durch $k(0) := 0$ lässt sich k also zu einer stetigen Funktion auf ganz \mathbb{R} fortsetzen. Das asymptotische Verhalten von k für $x \rightarrow \infty$ liefert den unbestimmten Ausdruck „ $\infty \cdot 0$ “, den wir in „ $\frac{0}{0}$ “ umwandeln:

$$\lim_{x \rightarrow \infty} x \ln(1 - e^{-x}) = \lim_{x \rightarrow \infty} \frac{\ln(1 - e^{-x})}{\frac{1}{x}} \stackrel{l'H}{=} \lim_{x \rightarrow \infty} \frac{-x^2 e^{-x}}{1 - e^{-x}} \stackrel{iii)}{=} 0.$$

Das Verhalten von k für $x \rightarrow -\infty$: $k(x) = x \ln(e^{-x} - 1) = x \ln(e^{-x}(1 - e^x)) = x \ln(e^{-x}) + x \ln(1 - e^x) = -x^2 + x \ln(1 - e^x)$. Der letzte Ausdruck konvergiert gegen 0:

$$\lim_{x \rightarrow -\infty} x \ln(1 - e^x) = \lim_{x \rightarrow -\infty} \frac{\ln(1 - e^x)}{\frac{1}{x}} \stackrel{l'H}{=} \lim_{x \rightarrow -\infty} \frac{x^2 e^x}{1 - e^x} \stackrel{iii)}{=} 0.$$

Das bedeutet $k(x) + x^2 \xrightarrow{x \rightarrow -\infty} 0$. Also ist $x \mapsto -x^2$ eine Näherungskurve von k für $x \rightarrow -\infty$.

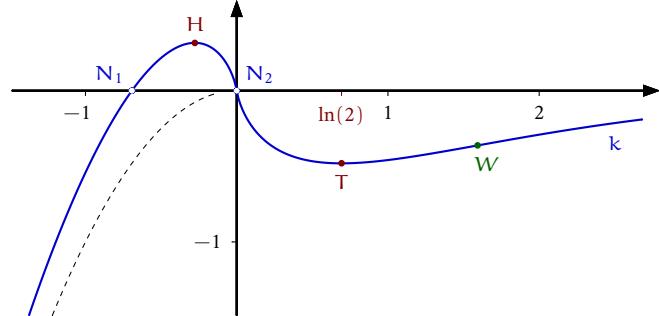
Die Extremstellen sind wohl nicht exakt zu bestimmen: Für $x > 0$ ist

$$k'(x) = \ln(1 - e^{-x}) + \frac{x e^{-x}}{1 - e^{-x}}.$$

$k'(x) = 0$ führt auf kein schönes NEWTON-Verfahren (vergl. Abschnitt 11.7), könnte man meinen. Überraschenderweise ist eine Lösung x der Gleichung $-x = \ln(1 - e^{-x})$ auch eine von $k'(x) = 0$. Denn dann ist $e^{-x} = 1 - e^{-x}$ und daher $k'(x) = \ln(1 - e^{-x}) + x = 0$. Die Gleichung $e^{-x} = 1 - e^{-x}$ lässt sich aber leicht lösen: $x = \ln(2)$. Das führt auf den Tiefpunkt $T := [\ln(2), -\ln(2)^2]$.

11.3.6 A Für die Punkte H und W ergeben geeignete NEWTON-Verfahren:

$$H \approx [-0.2764, 0.3163] \text{ und } W \approx [1.5936, -0.3620].$$



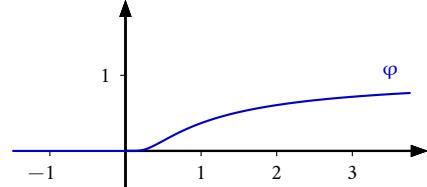
vii) Die Funktion

$$\varphi(x) := \begin{cases} e^{-\frac{1}{x}}, & x > 0 \\ 0, & x \leq 0 \end{cases} \quad (11.46)$$

ist beliebig oft differenzierbar, mit $\varphi^{(n)}(0) = 0$ für alle $n \in \mathbb{N}$.

Die Stetigkeit dieser Funktion ist klar. Für $x > 0$ ist $\varphi'(x) = \frac{1}{x^2} e^{-\frac{1}{x}}$ und für $x < 0$ gilt natürlich $\varphi'(x) = 0$. Mit Hilfe der Regeln von DE L'HOSPITAL zeigen wir $\lim_{x \rightarrow 0^+} \varphi'(x) = 0$:

$$\lim_{x \rightarrow 0^+} \frac{1}{x^2} e^{-\frac{1}{x}} = \lim_{x \rightarrow 0^+} \frac{\frac{1}{x^2}}{e^{\frac{1}{x}}} \stackrel{l'H}{=} \lim_{x \rightarrow 0^+} \frac{-\frac{2}{x^3}}{-\frac{1}{x^2} e^{\frac{1}{x}}} = \lim_{x \rightarrow 0^+} \frac{\frac{2}{x}}{e^{\frac{1}{x}}} \stackrel{l'H}{=} \lim_{x \rightarrow 0^+} \frac{\frac{2}{x^2}}{\frac{1}{x^2} e^{\frac{1}{x}}} = \lim_{x \rightarrow 0^+} \frac{2}{e^{\frac{1}{x}}} = 0.$$



Wegen $\varphi'(x) \xrightarrow{x \rightarrow 0^-} 0$ folgt die Differenzierbarkeit von φ an der Stelle 0 mit $\varphi'(0) = 0$ aus Satz 11.2.23. Für die weiteren Ableitungen wiederholen wir das Verfahren. Für $x > 0$ ist $\varphi''(x) =$

$(\frac{1}{x^4} - \frac{2}{x^3})e^{-\frac{1}{x}}$, $\varphi'''(x) = (\frac{6}{x^4} - \frac{6}{x^5} + \frac{1}{x^6})e^{-\frac{1}{x}}$ und allgemein $\varphi^{(n)}(x) = p_n(\frac{1}{x})e^{-\frac{1}{x}}$, mit einem Polynom p_n vom Grade $2n$. Wir wiederholen obige Rechnung für Ausdrücke der Form $\frac{1}{x^k} e^{-\frac{1}{x}}$ (am besten per Induktion) und erhalten $\lim_{x \rightarrow 0^+} \frac{1}{x^k} e^{-\frac{1}{x}} = 0$ und damit $\lim_{x \rightarrow 0^+} \varphi^{(n)}(x) = 0$. Zusammen mit $\varphi^{(n)}(x) = 0$ für $x < 0$ erhalten wir nach Satz 11.2.23 die Existenz von $\varphi^{(n)}(x)$ für alle $x \leq 0$ und alle $n \in \mathbb{N}$. Insbesondere ist $\varphi^{(n)}(0) = 0$. Die Differenzierbarkeit an allen Stellen $x > 0$ ergibt sich aus der Kettenregel.

11.3.7 A Bestimmen Sie die folgenden Grenzwerte.

- i) $\lim_{x \rightarrow \infty} x \cdot \tan(\frac{1}{x})$, ii)* $\lim_{x \rightarrow 0} \cosh(x)^{\frac{1}{x^2}}$, iii)* $\lim_{x \rightarrow 0} \frac{1}{\sin^2(x)} - \frac{1}{x^2}$,
- iv) $\lim_{x \rightarrow 0} \frac{1 - \cos(x)}{x^2}$, v) $\lim_{x \rightarrow 1} \frac{x^n - 1}{x^m - 1}$.
- vi) Schließen Sie die Definitionslücken der Funktion $f(x) := \frac{2^x - 8}{3x^2 - 27}$.

11.4 Ableitung von Umkehrfunktionen

Die Differentialrechnung stellt ein hinreichendes Kriterium zur Verfügung, das die Existenz einer Umkehrfunktion garantiert.

11.4.1 Satz Sei f auf einem Intervall $[a, b]$ stetig, auf (a, b) differenzierbar, und es sei $f'(x) \geq 0$ (≤ 0) für alle $x \in (a, b)$, wobei $f'(x) = 0$ nur für endlich viele Stellen x möglich sein soll. Dann ist f auf $[a, b]$ streng monoton wachsend (fallend) und hat eine stetige, streng monotone Umkehrfunktion $f^{-1} : [f(a), f(b)] \rightarrow [a, b]$ ($f^{-1} : [f(b), f(a)] \rightarrow [a, b]$).

Beweis. Wir nehmen o. B. d. A. $f'(x) \geq 0$ an und untersuchen zunächst den Fall, daß $f'(x) > 0$ für alle $x \in (a, b)$ gilt. Für $x < y$ und $x, y \in [a, b]$ folgt dann aus dem Mittelwertsatz 11.2.11 die Existenz eines $\xi \in (x, y)$ mit $f(y) - f(x) = f'(\xi)(y - x) > 0$, also $f(x) < f(y)$. Das zeigt bereits, daß f auf $[a, b]$ streng monoton wächst.

Nun der allgemeinere Fall, in dem f' an den endlich vielen Stellen $x_1 < x_2 < \dots < x_n$ Nullstellen haben darf. Dann ist nach unseren bisherigen Überlegungen f auf den Intervallen $[a, x_1]$, $[x_1, x_2], \dots [x_n, b]$ streng monoton wachsend, denn f' ist im Inneren all dieser Intervalle strikt positiv. Da sich die Intervalle genau an den Stellen x_k überschneiden, setzt sich die strenge Monotonie von einem Intervall auf das nächste, schließlich also auf das ganze Intervall $[a, b]$ fort. Aus der strengen Monotonie folgt natürlich leicht die Injektivität von f . Die verbleibenden Behauptungen des Satzes ergeben sich jetzt aus Satz 11.1.24. \square

11.4.2 Satz f sei auf dem Intervall I stetig, invertierbar und an der Stelle $x \in I$ differenzierbar und $f'(x) \neq 0$. Dann ist f^{-1} an der Stelle $y := f(x)$ differenzierbar und hat dort die Ableitung

$$f'^{-1}(y) = \frac{1}{f'(f^{-1}(y))} = \frac{1}{f'(x)}. \quad (11.47)$$

Beweis. Nach Satz 11.2.3 gibt es eine an der Stelle x stetige Funktion Θ mit der Eigenschaft $f(t) - f(x) = \Theta(t)(t - x)$ und $\Theta(x) = f'(x) \neq 0$. Laut Satz 11.1.24 ist f^{-1} stetig, so daß die Funktion $s \mapsto \Theta(f^{-1}(s))$ in y stetig ist und daher in einer Umgebung von y keine Nullstelle hat (Lemma 11.1.13). Für ein s dieser Umgebung sei $t := f^{-1}(s)$. Dann gilt

$$f^{-1}(s) - f^{-1}(y) = t - x = \frac{1}{\Theta(t)}(f(t) - f(x)) = \frac{1}{\Theta(f^{-1}(s))}(s - y).$$

Auch $s \mapsto \frac{1}{\Theta(f^{-1}(s))}$ ist in y stetig. Nach Satz 11.2.3 ist f^{-1} an der Stelle y differenzierbar, mit der Ableitung $f^{-1}'(y) = \frac{1}{\Theta(f^{-1}(y))} = \frac{1}{\Theta(x)} = \frac{1}{f'(x)} = \frac{1}{f'(f^{-1}(y))}$. \square

Man kann sich die Formel (11.47) leicht herleiten, indem man die Gleichung $f(f^{-1}(y)) = y$ auf beiden Seiten nach y ableitet. Auf der linken Seite wird die Kettenregel verwendet: $f'(f^{-1}(y)) \cdot f^{-1}'(y) = 1$. Löst man das nach $f^{-1}'(y)$ auf, so ergibt sich (11.47). Auf diese Weise läßt sich diese Formel allerdings nicht beweisen, denn die Anwendung der Kettenregel setzt die Differenzierbarkeit der beteiligten Funktionen voraus, die für f^{-1} erst zu zeigen ist.

11.4.3 Beispiel Die Funktion $f : x \mapsto x^n$ ist auf jedem Intervall stetig umkehrbar, wenn n ungerade ist, denn $f'(x) = nx^{n-1}$ ist, bis auf die Nullstelle bei $x = 0$, überall positiv. Ihre Umkehrfunktion ist $f^{-1}(y) = \sqrt[n]{y}$. Für gerade n ist f auf $[0, \infty)$ einzuschränken.

11.4.4 Beispiel (Logarithmus) Ein weiteres Beispiel ist natürlich die Exponentialfunktion. Wegen $0 < \exp(x) = \exp'(x)$ sind die notwendigen Voraussetzungen auf jedem Intervall $[a, b]$, also auf ganz \mathbb{R} erfüllt und \exp daher streng monoton wachsend (auch wenn wir das auf andere Weise schon erfahren haben, siehe Lemma 10.2.25). Die Umkehrfunktion \exp^{-1} von \exp ist der *natürliche Logarithmus*. Er wird mit dem Funktionssymbol \ln bezeichnet. Jedem der Potenzrechengesetze entspricht ein Rechengesetz für den Logarithmus:

$$\ln(xy) = \ln(x) + \ln(y) \quad (11.48)$$

$$\ln\left(\frac{x}{y}\right) = \ln(x) - \ln(y) \quad (11.49)$$

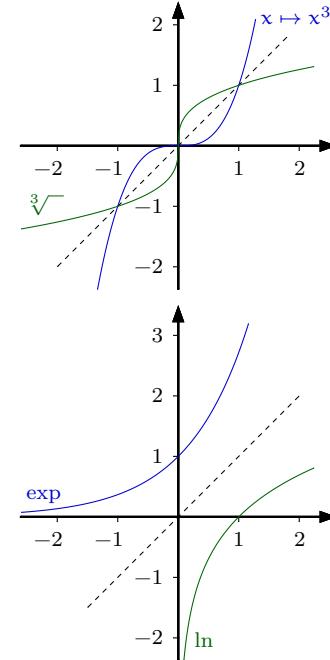
$$\ln(x^y) = y \ln(x) \quad (11.50)$$

Die ersten beiden Gleichungen gelten für alle $x, y \in \mathbb{R}^+$, die letzte aber zunächst nur für $x \in \mathbb{R}^+$ und $y \in \mathbb{Q}$. Wir erhalten sie aus der Beziehung $f \circ f^{-1}(x) = x$ zwischen Funktion und Umkehrfunktion, die hier $e^{\ln(x)} = x$ lautet. Auf xy angewendet bedeutet sie

$$e^{\ln(xy)} = xy = e^{\ln(x)}e^{\ln(y)} = e^{\ln(x)+\ln(y)}.$$

Da die Exponentialfunktion injektiv ist, folgt (11.48) durch Vergleich der Exponenten. Analog zeigt man (11.49). Für (11.50):

$$e^{\ln(x^y)} = x^y = (e^{\ln(x)})^y = e^{y \ln(x)}.$$



Das Problem mit (11.50) ist, daß x^y auf der linken Seite noch gar nicht für alle $y \in \mathbb{R}$ definiert ist, sondern nur für $y \in \mathbb{Q}$ durch die gewöhnlichen Potenzrechengesetze abgedeckt wird. In dieser Beobachtung liegt aber auch schon die Lösung des Problems. Wir verwenden die rechte Seite von (11.50), um durch

$$b^a := e^{a \ln(b)} \quad (11.51)$$

die Potenz mit Basis $b > 0$ für beliebige Exponenten $a \in \mathbb{R}$ zu definieren. Für $a \in \mathbb{Q}$ stimmt sie dann mit b^a aus den Potenzrechengesetzen überein: $e^{a \ln(b)} = e^{\ln(b^a)} = b^a$. Durch diese Definition wird die Funktion $a \mapsto b^a$ stetig differenzierbar, für $b > 1$ streng monoton wachsend und für $0 < b < 1$ streng monoton fallend (Übung 11.4.6). Damit hat sie nach Satz 11.4.1 eine stetige Umkehrfunktion. Diese wird mit \log_b bezeichnet. Wiederholt man die Überlegungen, die zu den Gleichungen (11.48)-(11.50) geführt haben, so erhält man die Rechengesetze für \log_b , in dem man dort \ln einfach durch \log_b ersetzt. Die Definition (11.51) lässt vermuten, daß hinter der Potenz zu einer beliebigen Basis und dem zugehörigen Logarithmus letztlich die Exponentialfunktion und der natürliche Logarithmus stecken. Also sollte \log_b über \ln zu berechnen sein. Das ist auch tatsächlich der Fall, wie folgende Überlegung zeigt. Die Gleichung $x = b^a = e^{a \ln(b)}$ hat per Definition die Lösung $a = \log_b(x)$, aber auch $\ln(x) = a \ln(b)$, oder $a = \frac{\ln(x)}{\ln(b)}$. Also unterscheidet sich \log_b von \ln nur um den Vorfaktor $\frac{1}{\ln(b)}$:

$$\log_b(x) = \frac{1}{\ln(b)} \ln(x). \quad (11.52)$$

Der Logarithmus zur Basis 10 wird üblicherweise einfach durch \log bezeichnet, wobei man allerdings im englischsprachigen Raum mit dieser Konvention vorsichtig sein muß, da hier mit \log oft \ln gemeint ist.

11.4.5 Beispiel Die Helligkeit der Sterne wurden schon in der Antike in sechs Helligkeitsstufen (*Magnituden* m) unterteilt. Dabei ist $m = 1$ den hellsten Sternen und $m = 6$ den Sternen zugeordnet, die bei besten Beobachtungsbedingungen gerade noch mit bloßem Auge zu sehen sind. Man spricht dann von einem Stern der dritten Größenklasse, wenn er die Magnitude $m = 3$ hat. Bei dieser Festlegung war das Auge sozusagen die Messapparatur. Für reproduzierbare Ergebnisse braucht man jedoch eine Definition durch physikalische Größen. Ein natürliches Maß ist der Strahlungsfluß (oder die Strahlungsintensität) E der Strahlung, die auf der Erde ankommt, also die (über alle Frequenzen addierte) Strahlungsleistung je Flächeneinheit. Der Zusammenhang mit der gesamten Strahlungsleistung des Sterns, der sogenannten *Leuchtkraft* L , ist dann

$$E = \frac{L}{4\pi r^2}.$$

Dabei ist r die Entfernung des Sterns von der Erde. $4\pi r^2$ ist der Inhalt der Kugeloberfläche einer Kugel, in deren Mittelpunkt sich der Stern und auf deren Oberfläche sich die Erde befindet.

Das menschliche Auge nimmt unterschiedliche Intensitäten logarithmisch wahr. Man setzte daher fest, daß ein Stern 1-ter Größe den 100-fachen Strahlungsfluß eines Sterns 6-ter Größe aufweist. Der Unterschied um eine Magnitude bedeutet daher den $\sqrt[5]{100}$ -fachen Strahlungsfluß (das ist etwa das 2.511-Fache). Mit fortschreitender Messtechnik erweiterte man die Skala der sechs Magnituden zu größeren und kleineren Werten. Um mit diesen Skalen arbeiten zu

können, braucht man einen Referenzpunkt. Bis ins Jahr 1922 diente dafür der *Polarstern*, für den eine Helligkeit von $m = 2$ festgesetzt wurde (heute verwendet man zu diesem Zweck eine ganze Reihe sehr genau vermessener Sterne unterschiedlicher Größenklassen). Der Zusammenhang zwischen Strahlungsfluß und Helligkeit ist $E \sim \sqrt[5]{100}^{-m}$. Den Proportionalitätsfaktor braucht man dabei nicht zu wissen, da man sich immer auf relative Größen bezieht. Für einen Stern 1 in der Entfernung r_1 , der Leuchtkraft L_1 und einen Stern 2 in der Entfernung r_2 mit der Leuchtkraft L_2 ist

$$\frac{E_1}{E_2} = \frac{L_1}{L_2} \cdot \frac{r_2^2}{r_1^2} = \sqrt[5]{100}^{m_2 - m_1} = 10^{\frac{2}{5}(m_2 - m_1)},$$

also

$$m_2 - m_1 = 2.5 \cdot \log\left(\frac{L_1}{L_2}\right) + 5 \cdot \log\left(\frac{r_2}{r_1}\right).$$

Die Helligkeiten, die Sterne am Nachthimmel zeigen, eignen sich nicht, um sie zu klassifizieren. Es kann sich dabei ja um sehr helle Sterne in großer Entfernung, oder um dunklere in geringer Entfernung handeln. Man nennt m daher *scheinbare Helligkeit*. Um Vergleichbarkeit zu erreichen, stellt man sich den beobachteten Stern in der Standardentfernung von 10 *Parsek* (pc) vor ($1 \text{ pc} \approx 3.3 \text{ Lj} \approx 3.0856 \cdot 10^{13} \text{ km} \approx 2.063 \cdot 10^5 \text{ AE}$). Seine Helligkeit M bezeichnet man als *absolute Helligkeit*. Da dieser fiktive zweite Stern dieselbe Leuchtkraft hat wie der beobachtete Stern, gilt

$$m - M = 5 \cdot \log\left(\frac{r}{10 \text{ pc}}\right).$$

Das ist das sogenannte *Entfernungsmodul*. Man verwendet es, um aus einer bekannten absoluten Helligkeit und der gemessenen scheinbaren auf den Abstand des Sterns zu schließen. Es ist eine eigene Geschichte, wie Astronomen die absolute Helligkeit eines Sterns bestimmen. Der am besten untersuchte Stern ist natürlich unsere Sonne. Für sie ist $m = -26.7$ und $r \approx 1.5 \cdot 10^8 \text{ km}$. Daraus können wir auf die absolute Helligkeit

$$M = -26.7 - 5 \cdot \log\left(\frac{1.5 \cdot 10^8}{3.0856 \cdot 10^{14}}\right) \approx 4.86$$

schließen. Unsere Sonne ist also nur von etwa 5-ter Magnitude.

11.4.6 A Zeigen Sie mit Hilfe von (11.51), daß die Potenzrechengesetze auch für beliebige Exponenten aus \mathbb{R} gültig sind und daß $x \mapsto b^x$ eine stetig differenzierbare, für $b > 1$ streng monoton wachsende und für $0 < b < 1$ streng monoton fallende Funktion ist. Wie sieht ihr Graph aus, wie die Ableitung? Zeigen Sie die Rechengesetze für \log_b , die (11.48)-(11.50) entsprechen. Was ist $\log_b(1)$ und was $\log_b(b)$?

11.4.7 Beispiel

i) Die Exponentialfunktion hat die stetige Umkehrfunktion \ln . Da $\exp'(x) = \exp(x) > 0$ für alle $x \in \mathbb{R}$ gilt, sind die Voraussetzungen von Satz 11.4.2 für alle x erfüllt. Die Anwendung der Regel (11.47) ergibt

$$\ln'(y) = \frac{1}{\exp(\ln(y))} = \frac{1}{y} \quad (11.53)$$

für alle $y > 0$ (denn y muß im Wertebereich \mathbb{R}^+ von \exp liegen).

ii) Die Funktion \sin hat die Ableitung \cos . Diese Funktion ist auf dem Intervall $(-\frac{\pi}{2}, \frac{\pi}{2})$ positiv und nimmt nur auf den Randpunkten $-\frac{\pi}{2}$ und $\frac{\pi}{2}$ den Wert 0 an. Nach Satz 11.4.1 ist \sin auf $[-\frac{\pi}{2}, \frac{\pi}{2}]$ stetig invertierbar. Die Umkehrfunktion bezeichnet man systematisch mit \sin^{-1} , oder mit dem traditionellen Namen *Arcussinus* \arcsin . Auf $(-\frac{\pi}{2}, \frac{\pi}{2})$ sind die Voraussetzungen obigen Satzes erfüllt, so daß \arcsin auf dem Bild $(-1, 1)$ dieses Intervalls ableitbar ist:

$$\arcsin'(y) = \frac{1}{\cos(\arcsin(y))} = \frac{1}{\sqrt{1 - \sin^2(\arcsin(y))}} = \frac{1}{\sqrt{1 - y^2}}. \quad (11.54)$$

Bei dieser Rechnung haben wir berücksichtigt, daß $\sin^2(x) + \cos^2(x) = 1$ gilt und daß $\cos(x)$ auf $(-\frac{\pi}{2}, \frac{\pi}{2})$ positiv ist, so daß in $\cos(x) = \pm \sqrt{1 - \sin^2(x)}$ nur das + in Frage kommt (vgl. Abbildung 12.5).

iii) Die Funktion \cos hat die Ableitung $-\sin$, die auf dem Intervall $(0, \pi)$ negativ ist und auf den Randpunkten 0 und π verschwindet. \cos ist also auf $[0, \pi]$ stetig invertierbar. Die Umkehrfunktion \cos^{-1} wird meist als *Arcuscosinus* \arccos bezeichnet. Sie hat auf $(-1, 1)$ die Ableitung

$$\arccos'(y) = \frac{-1}{\sin(\arccos(y))} = \frac{-1}{\sqrt{1 - \cos^2(\arccos(y))}} = -\frac{1}{\sqrt{1 - y^2}}. \quad (11.55)$$

Die Details dieser Rechnung mache man sich klar, so wie es oben vorgeführt wurde.

Die Ableitungen von \arcsin und \arccos stimmen bis auf das Vorzeichen überein, so daß auch zwischen \arcsin und \arccos eine einfache Beziehung bestehen sollte. Sie läßt sich durch die Gleichung $\sin(x) = \cos(\frac{\pi}{2} - x)$ (siehe (11.15)) gewinnen. Aus $0 \leq y = \sin(x)$ folgt dann einerseits $x = \arcsin(y)$ und andererseits $\frac{\pi}{2} - x = \arccos(y)$, also $\arccos(y) = \frac{\pi}{2} - \arcsin(y)$ (vgl. die Schaubilder auf Seite 401). Leitet man diese Gleichung ab, so erhalten wir $\arccos'(y) = -\arcsin'(y)$ (vgl. Abbildung 12.6).

iv) Die Funktion \tan hat für $x \in (-\frac{\pi}{2}, \frac{\pi}{2})$ nach Übung 11.2.7 die Ableitung $\tan'(x) = \frac{1}{\cos^2(x)} = 1 + \tan^2(x) > 0$. Damit sind die Voraussetzungen von Satz 11.4.1 für jedes Intervall $[a, b] \in (-\frac{\pi}{2}, \frac{\pi}{2})$ erfüllt, so daß \tan auf jedem dieser Intervalle, also auf ganz $(-\frac{\pi}{2}, \frac{\pi}{2})$ stetig invertierbar ist (vgl. Abbildung 12.7). Die Umkehrfunktion auf dem Wertebereich \mathbb{R} von \tan heißt *Arcustangens* und wird durch \arctan bezeichnet. Es gilt

$$\arctan'(y) = \frac{1}{1 + \tan^2(\arctan(y))} = \frac{1}{1 + y^2}. \quad (11.56)$$

v) Die Funktion $\cot = \frac{\cos}{\sin} = \frac{1}{\tan}$ hat die Ableitung $\cot'(x) = -\frac{1}{\sin^2(x)} = -(1 + \cot^2(x))$ auf dem Intervall $(0, \pi)$. Wie für \tan argumentiert man, um die Existenz der Umkehrfunktion *Arcuscotangens* arccot auf dem Intervall $(0, \pi)$ nachzuweisen. Es gilt $\tan(x) = \cot(\frac{\pi}{2} - x)$, so daß in derselben Weise, wie in iii) vorgeführt, $\operatorname{arccot}(y) = \frac{\pi}{2} - \arctan(y)$ folgt (vgl. Abbildung 12.8). Die Ableitung dieser Gleichung ergibt

$$\operatorname{arccot}'(y) = -\frac{1}{1 + y^2}. \quad (11.57)$$

Aus der Beziehung $\cot(x) = \frac{1}{\tan(x)}$ läßt sich ein weiterer Zusammenhang zwischen \arctan und arccot gewinnen. Für $y > 0$ ist $x > 0$. $y = \cot(x)$ hat die Lösung $x = \operatorname{arccot}(y)$.

Andererseits gilt aber auch $\tan(x) = \frac{1}{\cot(x)} = \frac{1}{y}$ und damit $x = \arctan(\frac{1}{y})$. Also gilt $\operatorname{arccot}(y) = \arctan(\frac{1}{y})$. Für $y < 0$ stimmt das aber nicht mehr, denn \tan ist auf $(-\frac{\pi}{2}, \frac{\pi}{2})$ und \cot auf $(0, \pi)$ umkehrbar. Für $y < 0$ muß x in $(\frac{\pi}{2}, \pi)$ liegen. Hier gilt $x = \operatorname{arccot}(y)$. Den zweiten Teil unserer Überlegung für $y > 0$ müssen wir anpassen. Für $x \in (\frac{\pi}{2}, \pi)$ ist \tan nicht mehr umkehrbar, d. h. die Gleichung $\tan(x) = \frac{1}{y}$ kann nicht einfach durch \arctan nach x aufgelöst werden. Dazu müssen wir sie so umschreiben, daß \tan wieder auf $(-\frac{\pi}{2}, \frac{\pi}{2})$ operiert. Das ist leicht, denn es gilt $\tan(x) = \tan(x - \pi)$ (vgl. Abbildung 12.7) und $x - \pi \in (-\frac{\pi}{2}, 0)$. Das bedeutet, wir können die Gleichung $\tan(x - \pi) = \tan(x) = \frac{1}{y}$ mittels \arctan lösen: $x - \pi = \arctan(\frac{1}{y})$. Für $y < 0$ gilt daher $\operatorname{arccot}(y) = \arctan(\frac{1}{y}) + \pi$. Wir erhalten

$$\operatorname{arccot}(y) = \begin{cases} \arctan(\frac{1}{y}) + \pi & , y < 0 \\ \frac{\pi}{2} & , y = 0 \\ \arctan(\frac{1}{y}) & , y > 0 \end{cases} \quad (11.58)$$

vi) Die Funktion *Sinushyperbolicus* \sinh ist durch

$$\sinh(x) := \frac{1}{2}(e^x - e^{-x}) \quad (11.59)$$

für alle $x \in \mathbb{R}$ definiert. Ihre Ableitung ist die Funktion *Kosinushyperbolicus* \cosh :

$$\cosh(x) := \frac{1}{2}(e^x + e^{-x}). \quad (11.60)$$

\sinh ist punktsymmetrisch zum Ursprung ($\sinh(-x) = -\sinh(x)$) und \cosh ist achsensymmetrisch ($\cosh(-x) = \cosh(x)$). Offensichtlich gilt $\cosh(x) > 0$ für alle $x \in \mathbb{R}$, so daß \sinh auf ganz \mathbb{R} streng monoton steigt. Wegen $\sinh(0) = 0$ muß demnach $\sinh(x) > 0$ für $x > 0$ gelten. Da auch $\cosh' = \sinh$ gilt, ist \cosh auf \mathbb{R}_0^+ streng monoton steigend. $\cosh(0) = 1$ zeigt dann $\cosh(x) > 1$ für alle $x > 0$. Als Ergebnis dieser Überlegungen erhalten wir die Invertierbarkeit von \sinh auf ganz \mathbb{R} und die von \cosh auf \mathbb{R}_0^+ (vgl. die Schaubilder 12.9 und 12.10). Die Umkehrfunktionen von \sinh bzw. \cosh heißen *Areasinus* bzw. *Areacosinus*. Sie werden durch Arsinh bzw. Arcosh , meist aber einfach durch die systematischen Ausdrücke \sinh^{-1} bzw. \cosh^{-1} bezeichnet. Wir können sie direkt berechnen. Dafür muß die Gleichung $y = \sinh(x) = \frac{1}{2}(e^x - e^{-x})$ nach x aufgelöst werden. Multipliziert man sie mit e^x , so ergibt sich

$$0 = (e^x)^2 - 2ye^x - 1.$$

Wir setzen $u := e^x$ und erhalten die quadratische Gleichung $u^2 - 2yu - 1 = 0$, mit den Lösungen $u_{1/2} = y \pm \sqrt{y^2 + 1}$. Für positive y ist wegen der Monotonie der Wurzel $\sqrt{y^2 + 1} > \sqrt{y^2} = y$. Damit u positiv wird, muß daher das $+$ in der Lösungsformel gewählt werden. Das bedeutet $e^x = u = y + \sqrt{y^2 + 1}$, also

$$\sinh^{-1}(y) = \ln(y + \sqrt{y^2 + 1}). \quad (11.61)$$

Mit dieser expliziten Formel kann man auch die Ableitung von \sinh^{-1} ausrechnen (das ist eine gute Übung). Einfacher ist es jedoch, wie bei \sin vorzugehen und die zentrale Beziehung zwischen \sinh und \cosh zu verwenden, nämlich

$$\cosh^2(x) - \sinh^2(x) = 1. \quad (11.62)$$

Diese Gleichung lässt sich durch simples Ausmultiplizieren beweisen. Verwenden wir (11.47):

$$\sinh^{-1}'(y) = \frac{1}{\cosh(\sinh^{-1}(y))} = \frac{1}{\sqrt{\sinh^2(\sinh^{-1}(y)) + 1}} = \frac{1}{\sqrt{y^2 + 1}}. \quad (11.63)$$

Diese Überlegungen lassen sich für die Funktion \cosh leicht anpassen und ergeben

$$\cosh^{-1}(y) = \ln(y + \sqrt{y^2 - 1}), \quad y \geq 1, \quad (11.64)$$

$$\cosh^{-1}'(y) = \frac{1}{\sqrt{y^2 - 1}}, \quad y > 1. \quad (11.65)$$

vii) Die Funktion *Tangenshyperbolicus*

$$\tanh(x) := \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (11.66)$$

ist auf ganz \mathbb{R} definiert, nimmt die Werte zwischen -1 und 1 an und hat die Ableitung $\tanh' = \frac{1}{\cosh^2} = 1 - \tanh^2 > 0$ (nachrechnen, vgl. Abbildung 12.11)). \tanh ist daher streng monoton wachsend und lässt sich umkehren. Die Umkehrfunktion \tanh^{-1} gewinnt man wieder durch Lösen einer quadratischen Gleichung (nämlich $(1-y)u^2 = 1+y$):

$$\tanh^{-1}(y) = \frac{1}{2} \ln\left(\frac{1+y}{1-y}\right), \quad |y| < 1, \quad (11.67)$$

$$\tanh^{-1}'(y) = \frac{1}{1-y^2}, \quad |y| < 1. \quad (11.68)$$

Die Berechnung der Ableitung verläuft analog zu der von \arctan .

11.4.8 A Bestimmen Sie für die Funktion $g(x) := \frac{x^2}{1+x^2}$ den maximalen Bereich nicht negativer Zahlen, auf dem sie umkehrbar ist und berechnen Sie dort ihre Umkehrfunktion g^{-1} . Zeichnen Sie g und g^{-1} in ein gemeinsames Koordinatensystem. Bestimmen Sie dann die Tangente t von g^{-1} an der Stelle $x = 0.5$ und zeichnen Sie sie ebenfalls ein.

11.4.9 A Verwenden Sie die EULER-Formel $e^{ix} = \cos(x) + i \sin(x)$, um sich die Beziehungen

$$\begin{aligned} \cosh(ix) &= \cos(x), & \sinh(ix) &= i \sin(x), \\ \tanh(ix) &= i \tan(x), & \coth(ix) &= -i \cot(x) \end{aligned}$$

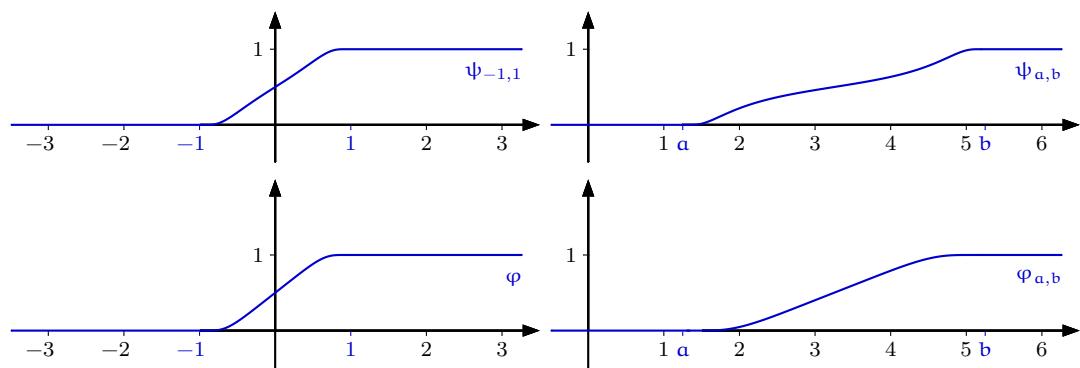
klar zu machen, die die Namensgebungen \sinh , \cosh etc. rechtfertigen.

11.4.10 A Überlegen Sie sich, daß die Funktionen

$$\psi_{a,b}(x) := \begin{cases} 0 & , \quad x \leq a \\ \frac{2}{\pi} \arctan(e^{\frac{1}{a-x}} e^{\frac{1}{b-x}}) & , \quad a < x < b \\ 1 & , \quad b \leq x \end{cases} \quad (11.69)$$

$$\varphi(x) := \begin{cases} 0 & , \quad x \leq -1 \\ \frac{1}{2}(1 + \tanh(\tan(\frac{\pi}{2}x))) & , \quad -1 < x < 1 \\ 1 & , \quad 1 \leq x \end{cases} \quad (11.70)$$

und $\varphi_{a,b}(x) := \varphi\left(\frac{1}{b-a}(2x - a - b)\right)$ beliebig oft differenzierbar sind und daß ihre Kurvenverläufe wie in den folgenden Abbildungen aussehen.



11.5 Stammfunktionen

11.5.1 Definition Sei $f: D_f \rightarrow \mathbb{R}$ eine reelle Funktion. Eine Funktion $F: D_f \rightarrow \mathbb{R}$ heißt Stammfunktion von f , wenn $F'(x) = f(x)$ für alle $x \in D_f$ gilt.

Für viele Funktionen können wir Stammfunktionen angeben. Für \exp ist \exp selbst eine Stammfunktion, für \sin ist $-\cos$ und für \cos ist \sin jeweils eine. Für $\text{id}^n: x \mapsto x^n$ ist $\frac{1}{n+1} \text{id}^{n+1}$ eine Stammfunktion, denn $\frac{1}{n+1} \frac{d}{dx} x^{n+1} = \frac{n+1}{n+1} x^n = x^n$. Das gilt für alle $n \in \mathbb{R} \setminus \{-1\}$. Die Standardhyperbel $x \mapsto \frac{1}{x}$ hat eine eigene Stammfunktion, nämlich $x \mapsto \ln(|x|)$. Für $x > 0$ ist das (11.53), für $x < 0$ folgt es zusammen mit der Kettenregel ebenfalls aus (11.53): $\frac{d}{dx} \ln(|x|) = \frac{d}{dx} \ln(-x) = \frac{-1}{-x} = \frac{1}{x}$. Überhaupt liefern die Funktionen aus 11.4.7 interessante Beispiele für Stammfunktionen:

$f(x)$	$F(x)$	$f(x)$	$F(x)$
$x^n \quad (n \neq -1)$	$\frac{x^{n+1}}{n+1}$	$\frac{1}{x}$	$\ln(x)$
e^x	e^x	$\cos(x)$	$\sin(x)$
$\sin(x)$	$-\cos(x)$	$\tan(x)$	$-\ln(\cos(x))$
$\frac{1}{\sqrt{1-x^2}}$	$\arcsin(x)$	$\cot(x)$	$\ln(\sin(x))$
$\frac{1}{\sqrt{1+x^2}}$	$\sinh^{-1}(x)$	$\frac{1}{1+x^2}$	$\arctan(x)$
$\frac{1}{\sqrt{x^2-1}}$	$\cosh^{-1}(x)$	$\frac{1}{1-x^2}$	$\frac{1}{2} \ln\left(\left \frac{1+x}{1-x}\right \right)$
$f(ax+b)$	$\frac{1}{a} F(ax+b)$	$\frac{g'(x)}{g(x)}$	$\ln(g(x))$

11.5.2 Lemma Zwei Stammfunktionen F und G einer Funktion f auf dem Intervall (a, b) unterscheiden sich nur um eine additive Konstante c : $G(x) = F(x) + c$ für alle $x \in (a, b)$.

Beweis. Die Funktion $G - F$ ist auf (a, b) differenzierbar, mit der Ableitung $(G - F)'(x) = G'(x) - F'(x) = f(x) - f(x) = 0$ für alle $x \in (a, b)$. Nach Korollar 11.2.13 ist $G - F$ konstant, d.h., es gibt ein $c \in \mathbb{R}$, so daß $G(x) - F(x) = c$ für alle $x \in (a, b)$ gilt. \square

Natürlich ist klar, daß mit \sin auch $2 + \sin$ eine Stammfunktion von \cos ist. Das Lemma sagt uns, daß wir keine allgemeineren Stammfunktionen als $c + \sin$ zu erwarten haben.

11.5.3 Satz Für eine Potenzreihe $f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k$ mit einem Konvergenzradius $R > 0$ ist auf $(x_0 - R, x_0 + R)$ eine Stammfunktion durch die Potenzreihe

$$F(x) = \sum_{k=0}^{\infty} \frac{1}{k+1} a_k (x - x_0)^{k+1} = \sum_{k=1}^{\infty} \frac{1}{k} a_{k-1} (x - x_0)^k \quad (11.71)$$

gegeben. F hat denselben Konvergenzradius wie f .

Beweis. Daß F eine Stammfunktion von f ist, folgt einfach aus Satz 11.3.1. Da sich der Konvergenzradius durch Ableitung nicht ändert, haben F und f denselben. \square

Probieren wir das an einer bekannten Funktion aus: Eine Stammfunktion von \exp ist gemäß dem Satz

$$F(x) = \sum_{k=1}^{\infty} \frac{1}{k(k-1)!} x^k = \sum_{k=1}^{\infty} \frac{1}{k!} x^k = \exp(x) - 1.$$

Sie unterscheidet sich von der erwarteten Stammfunktion nur um die additive Konstante -1 .

Wann immer wir eine Potenzreihenentwicklung der Ableitung f' einer Funktion kennen, haben wir über diesen Satz ohne großen Aufwand auch eine für f . Etwa für \arctan . Es gilt $\arctan'(x) = \frac{1}{1+x^2} = \sum_{k=0}^{\infty} (-1)^k x^{2k}$, mit Konvergenzradius 1. Daher gilt für $x \in (-1, 1)$

$$\arctan(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} x^{2k+1} \quad (11.72)$$

Eigentlich müssen wir vorsichtshalber $\arctan(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} x^{2k+1} + c$ ansetzen, denn nach Satz 11.5.3 legt \arctan' die Stammfunktion \arctan nur bis auf eine additive Konstante fest, die durch Auswertung an einer bekannten Stelle bestimmt werden kann. Für $x = 0$ ergibt die Potenzreihe den Wert 0, ebenso wie $\arctan(0)$. Daher ist $c = 0$. Man kann (11.72) dazu verwenden, eine Vorschrift zur Berechnung von π anzugeben. Wegen $\arctan(1) = \frac{\pi}{4}$, sollte

$$\frac{\pi}{4} = \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots \quad (11.73)$$

gelten, eine Formel, die LEIBNIZ angegeben hat. Das ist tatsächlich auch der Fall, aber das ist zunächst überhaupt nicht klar. Das Problem liegt darin, daß $x = 1$ nicht in $(-1, 1)$ liegt. Nur auf diesem offenen Intervall wissen wir nach Satz 11.5.3 die Darstellung (11.72). Wünschenswert ist ein Satz, der sicherstellt, daß die Konvergenz einer Potenzreihe auf dem Rand des Konvergenzbereichs auch zu einem Funktionswert der dargestellten Funktion führt. Der ABELSche Grenzwertsatz stellt das für reelle Potenzreihen sicher.

11.5.4 ABELScher Grenzwertsatz Für eine endliche Summe $\sum_{k=1}^n a_k b_k$ von Produkten führen wir die sogenannte *ABELSche Summation* folgendermaßen ein. Wir bilden die Partialsummen $A_n := \sum_{k=1}^n a_k$ und damit

$$\begin{aligned} \sum_{k=1}^n a_k b_k &= A_1 b_1 + (A_2 - A_1) b_2 + \dots + (A_k - A_{k-1}) b_k + \dots + (A_n - A_{n-1}) b_n \\ &= A_1(b_1 - b_2) + A_2(b_2 - b_3) + \dots + A_{n-1}(b_{n-1} - b_n) + A_n b_n. \end{aligned}$$

Das ergibt

$$\sum_{k=1}^n a_k b_k = \sum_{k=1}^{n-1} A_k(b_k - b_{k+1}) + A_n b_n, \quad A_k = \sum_{\ell=1}^k a_\ell. \quad (11.74)$$

Man kann diese rein algebraische Beziehung als eine diskrete Version der partiellen Integration auffassen (vergl. 12.9). Damit haben wir das wesentliche Hilfsmittel für den Beweis des folgenden Satzes.

11.5.5 Satz (ABELSches Kriterium) $(a_n)_{n \in \mathbb{N}}$ sei eine komplexwertige und $(b_n)_{n \in \mathbb{N}}$ eine reellwertige Folge von Funktionen auf einem gemeinsamen Definitionsbereich D , die den folgenden Bedingungen genügen:

- i) Für alle $x \in D$ ist die Folge $(b_n(x))_{n \in \mathbb{N}}$ monoton fallend.
- ii) $(b_n)_{n \in \mathbb{N}}$ ist gleichmäßig beschränkt: $\exists M > 0 \forall n \in \mathbb{N} \forall x \in D |b_n(x)| \leq M$.
- iii) Die Reihe $\sum_{k=1}^{\infty} a_k(x)$ konvergiert gleichmäßig bezüglich $x \in D$.

Dann konvergiert auch $\sum_{k=1}^{\infty} a_k(x)b_k(x)$ gleichmäßig bezüglich $x \in D$.

Bedingungen i) verlangt nur, daß $b_n(x)$ monoton fallend ist, nicht daß es sich um eine Nullfolge handelt. Die gleichmäßige Konvergenz der Reihe $\sum_{k=1}^{\infty} a_k(x)$ bedeutet, daß es für jedes $\varepsilon > 0$ ein n_ε gibt, so daß für alle $n \geq n_\varepsilon$ und für alle $x \in D$ die Abschätzung $|\sum_{k=n}^{\infty} a_k(x)| < \varepsilon$ gilt (vergl. Definition 11.1.29).

Beweis. Wir bilden die Funktionen $A := \sum_{k=1}^{\infty} a_k$ und $A_n := \sum_{k=1}^n a_k$. Dann gibt es zu jedem $\varepsilon > 0$ ein $n_\varepsilon \in \mathbb{N}$, so daß für alle $n \geq n_\varepsilon$ und für alle $x \in D$ die Abschätzung $|A(x) - A_n(x)| < \varepsilon$ erfüllt ist. Wir versuchen, die gleichmäßige CAUCHY-Bedingung für die Folge $(\sum_{k=1}^n a_k(x)b_k(x))_{n \in \mathbb{N}}$ nachzuweisen. Zunächst verwenden wir die ABELSche Summation:

$$\begin{aligned} \sum_{k=n+1}^m a_k b_k &= \sum_{k=1}^m a_k b_k - \sum_{k=1}^n a_k b_k = \sum_{k=n}^{m-1} A_k(b_k - b_{k+1}) + A_m b_m - A_n b_n \\ &= \sum_{k=n}^{m-1} (A_k - A)(b_k - b_{k+1}) + (A_m - A)b_m - (A_n - A)b_n. \end{aligned}$$

Für alle $m > n \geq n_\varepsilon$ und alle $x \in D$ gilt dann

$$\begin{aligned} \left| \sum_{k=n+1}^m a_k(x)b_k(x) \right| &\leq \sum_{k=n}^{m-1} |A_k(x) - A(x)| (b_k(x) - b_{k+1}(x)) \\ &\quad + |A_m(x) - A(x)| |b_m(x)| + |A_n(x) - A(x)| |b_n(x)| \\ &< \varepsilon \sum_{k=n}^{m-1} (b_k(x) - b_{k+1}(x)) + 2\varepsilon M \\ &= \varepsilon (b_n(x) - b_m(x)) + 2\varepsilon M < 4\varepsilon M. \end{aligned}$$

Damit ist die CAUCHY-Bedingung für die Folge $(\sum_{k=1}^n a_k(x)b_k(x))_{n \in \mathbb{N}}$ gleichmäßig bzgl. $x \in D$ erfüllt. Sie ist nach Lemma 10.1.51 auch gleichmäßig bzgl. $x \in D$ gegen $\sum_{k=1}^{\infty} a_k(x)b_k(x)$ konvergent. \square

Ein ähnliches Ergebnis, das auf der ABELSchen Summation beruht:

11.5.6 Satz (DIRICHLET-Kriterium) $(a_n)_{n \in \mathbb{N}}$ sei eine komplexwertige und $(b_n)_{n \in \mathbb{N}}$ eine Folge positiver Funktionen auf einem gemeinsamen Definitionsbereich D , mit folgenden Eigenschaften:

- i) Für alle $x \in D$ ist die Folge $(b_n(x))_{n \in \mathbb{N}}$ monoton fallend.

- ii) $(b_n(x))_{n \in \mathbb{N}}$ konvergiert gleichmäßig bezüglich $x \in D$ gegen Null.
 iii) Die Folge $(\sum_{k=1}^n a_k(x))_{n \in \mathbb{N}}$ ist gleichmäßig beschränkt bezüglich $x \in D$.
 Dann ist $\sum_{k=1}^{\infty} a_k(x)b_k(x)$ gleichmäßig bezüglich $x \in D$ konvergent.

Bedingung iii) verlangt dabei nicht, daß die Reihe konvergiert, es reicht, wenn $|\sum_{k=1}^n a_k(x)| \leq M$ für ein geeignetes $M > 0$, für alle $x \in D$ und alle $n \in \mathbb{N}$ erfüllt ist.

Beweis. Die ABELSche Summation

$$\sum_{k=n+1}^m a_k b_k = \sum_{k=n}^{m-1} A_k(b_k - b_{k+1}) + A_m b_m - A_n b_n$$

verwenden wir direkt zur Abschätzung

$$\begin{aligned} \left| \sum_{k=n+1}^m a_k(x)b_k(x) \right| &\leq \sum_{k=n}^{m-1} |A_k(x)|(|b_k(x) - b_{k+1}(x)| \\ &\quad + |A_m(x)|b_m(x) + |A_n(x)|b_n(x)) \\ &< M \sum_{k=n}^{m-1} (b_k(x) - b_{k+1}(x)) + M(b_m(x) + b_n(x)) \\ &= M(b_n(x) - b_m(x)) + M(b_m(x) + b_n(x)) = 2Mb_n(x) < \varepsilon. \end{aligned}$$

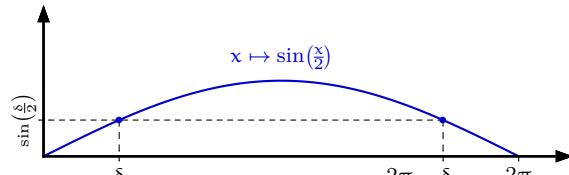
ab einem geeigneten n_ε , denn $(b_n(x))_{n \in \mathbb{N}}$ ist eine gleichmäßige Nullfolge. Also ist $(\sum_{k=1}^n a_k(x)b_k(x))_{n \in \mathbb{N}}$ eine gleichmäßige CAUCHY-Folge und daher gleichmäßig konvergent. \square

11.5.7 Beispiel Eine schöne Anwendung dieses Satzes ist die Reihe

$$\sum_{k=1}^{\infty} \frac{e^{ikx}}{k} \tag{11.75}$$

Sie konvergiert für jedes $\delta > 0$ auf dem Intervall $[\delta, 2\pi - \delta]$ gleichmäßig. Wir wählen dafür $a_k(x) := e^{ikx}$ und $b_k(x) := \frac{1}{k}$. Dann gilt

$$\begin{aligned} \left| \sum_{k=1}^n e^{ikx} \right| &= \left| \sum_{k=1}^n (e^{ix})^k \right| = \left| \frac{e^{inx} - e^{i(n+1)x}}{1 - e^{ix}} \right| \\ &= \left| \frac{1 - e^{inx}}{1 - e^{ix}} \right| = \left| \frac{e^{-in\frac{x}{2}} - e^{in\frac{x}{2}}}{e^{-i\frac{x}{2}} - e^{i\frac{x}{2}}} \right| \\ &= \frac{|\sin(n\frac{x}{2})|}{|\sin(\frac{x}{2})|} \leq \frac{1}{\sin(\frac{x}{2})}. \end{aligned}$$



Jetzt läßt sich Satz 11.5.6 anwenden, und die Behauptung folgt unmittelbar. Die komplexen Zahlen e^{ix} durchlaufen, da $\delta > 0$ beliebig klein gewählt werden kann, alle Punkte z des komplexen Einheitskreises, mit Ausnahme von $z = 1$. Das heißt, die Potenzreihe $\sum_{k=1}^{\infty} \frac{z^k}{k}$, die den Konvergenzradius $R = 1$ hat, konvergiert auch auf allen Randpunkten, von $z = 1$ abgesehen. Damit ist es uns für diese spezielle Potenzreihe gelungen, das Konvergenzverhalten auf dem Rand der Konvergenzbereichs vollständig zu bestimmen.

11.5.8 Satz (ABELScher Grenzwertsatz) Konvergiert die reelle Reihe $\sum_{k=0}^{\infty} c_k x^k$ an der Stelle $x = R > 0$, dann ist sie auf $[0, R]$ gleichmäßig konvergent und definiert eine stetige Funktion $[0, R] \ni x \mapsto \sum_{k=0}^{\infty} c_k x^k$.

Beweis. Wir definieren die monoton fallende Funktionenfolge $(b_k)_{k \in \mathbb{N}}$ durch $b_k(x) := \frac{x^k}{R^k}$, $x \in [0, R]$. Die Funktionen b_k sind offensichtlich durch 1 gleichmäßig beschränkt. Außerdem wählen wir die konstanten Funktionen $a_k(x) := c_k R^k$. Dann ist laut Voraussetzung $\sum_{k=1}^{\infty} a_k(x) = \sum_{k=1}^{\infty} c_k R^k$ gleichmäßig konvergent bzgl. $x \in [0, R]$, da die Reihe gar nicht von x abhängt. Die Voraussetzungen von Satz 11.5.5 sind damit erfüllt und die Reihe $\sum_{k=1}^{\infty} a_k(x)b_k(x) = \sum_{k=1}^{\infty} c_k R^k \frac{x^k}{R^k} = \sum_{k=1}^{\infty} c_k x^k$ gleichmäßig konvergent. Also ist die Funktion $x \mapsto \sum_{k=0}^{\infty} c_k x^k$ der gleichmäßige Grenzwert der stetigen Funktionen $x \mapsto \sum_{k=0}^n c_k x^k$ und daher nach Satz 11.1.30 stetig. \square

11.5.9 Beispiel Mit diesem Ergebnis ist Gleichung (11.73) kein Problem mehr. Die Potenzreihe $\arctan(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} x^{2k+1}$ konvergiert auch an der Stelle $x = 1$. Daher ist sie auf $[0, 1]$ gleichmäßig konvergent und bestimmt dort eine stetige Funktion, die auf $[0, 1]$ mit \arctan übereinstimmt. Da dieser auf ganz $[0, 1]$ stetig ist, muß die Reihe auch für $x = 1$ den Funktionswert des \arctan ergeben, also $\frac{\pi}{4}$.

Übrigens ist Gleichung (11.73) nicht sehr gut geeignet, um brauchbare Näherungen für π zu finden, denn die Reihe konvergiert nur sehr langsam. Man kann die \arctan -Reihe aber trotzdem verwenden, um eine sehr effiziente Methode zur Berechnung von π zu gewinnen. Dafür ist folgende überraschende Gleichung der Ausgangspunkt:

$$(5 + i)^4(239 - i) = 4 \cdot 13^4(1 + i).$$

Die Zahl $5 + i$ hat die Polardarstellung $5 + i = \sqrt{26} e^{i \arctan(\frac{1}{5})}$. Daher gilt $(5 + i)^4 = 4 \cdot 13^2 e^{4i \arctan(\frac{1}{5})}$. $239 - i$ wird durch $13^2 \sqrt{2} e^{-i \arctan(\frac{1}{239})}$ wiedergegeben, und die rechte Seite der Gleichung durch $4 \cdot 13^4 \sqrt{2} e^{i \frac{\pi}{4}}$. Daher haben wir

$$4 \cdot 13^4 \sqrt{2} e^{4i \arctan(\frac{1}{5}) - i \arctan(\frac{1}{239})} = 4 \cdot 13^4 \sqrt{2} e^{i \frac{\pi}{4}}.$$

Vergleicht man die Exponenten, so folgt

$$4 \arctan\left(\frac{1}{5}\right) - \arctan\left(\frac{1}{239}\right) = \frac{\pi}{4}.$$

Genau genommen wissen wir das zunächst nur bis auf ein Vielfaches von 2π . Wegen $0 < 4 \arctan\left(\frac{1}{5}\right) - \arctan\left(\frac{1}{239}\right) \leq \frac{4}{5} \approx \frac{\pi}{4} \approx 0.78$, muß dieses Vielfache einfach 0 sein. Verwenden wir (11.72):

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \left(\frac{16}{5^{2k+1}} - \frac{4}{239^{2k+1}} \right) = \pi. \quad (11.76)$$

Die ersten vier Summanden ergeben den Wert 3.14159177, der bereits auf 5 Stellen mit $\pi = 3.1415926\dots$ übereinstimmt.

11.6 Die Taylor-Entwicklung

Aus dem verallgemeinerten Mittelwertsatz 11.2.12 lässt sich auf überraschende Weise die TAYLOR-Formel gewinnen, die eine der wichtigsten Methoden bereitstellt, um Funktionswerte solcher Funktionen, wie sin, cos, exp, ln, arcsin, etc. berechenbar zu machen. Das bedeutet, daß die Bestimmung etwa von $\sin(x)$ auf die vier Grundrechenarten zurückgeführt werden muß, denn bei Licht betrachtet, können wir nichts anderes. Allerdings ist zu erwarten, daß im Allgemeinen unendlich viele solcher Rechenvorgänge durchzuführen sind, d. h., daß ein Grenzwert zu bilden ist, um $\sin(x)$ exakt festzulegen. Diese Anforderungen werden bestens von Potenzreihen erfüllt. Da wir über die bereits gut Bescheid wissen (vergl. 10.2.13 und 11.3), ist es nicht schwer die allgemeine Form einer Potenzreihe zu finden, die eine Funktion f in einer Umgebung eines Punktes x_0 – dem sogenannten *Entwickelpunkt* – beschreibt: Wir wählen der Einfachheit halber $x_0 = 0$. Dann folgt aus

$$f(x) = \sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + \cdots + a_n x^n + \cdots$$

bereits $a_0 = f(0)$. Durch Ableitung erhalten wir

$$f'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1} = a_1 + 2a_2 x + 3a_3 x^2 + 4a_4 x^3 + \cdots + n a_n x^{n-1} + \cdots,$$

woraus $a_1 = f'(0)$ abzulesen ist. Wiederholen wir das, bis das Bildungsgesetz der Koeffizienten a_k erkennbar wird:

$$\begin{aligned} f''(x) &= \sum_{n=2}^{\infty} (n-1) n a_n x^{n-2} = 2a_2 + 2 \cdot 3a_3 x + 3 \cdot 4a_4 x^2 + \cdots + (n-1) n a_n x^{n-2} \cdots, \\ f'''(x) &= \sum_{n=3}^{\infty} (n-2)(n-1) n a_n x^{n-3} = 3! a_3 + 4! a_4 x + \cdots + \frac{n!}{(n-3)!} a_n x^{n-3} + \cdots, \\ f^{(k)}(x) &= \sum_{n=k}^{\infty} \frac{n!}{(n-k)!} a_n x^{n-k} = k! a_k + (k+1)! a_{k+1} x + \frac{(k+2)!}{2} a_{k+2} x^2 + \cdots. \end{aligned}$$

Das zeigt, daß $a_k = \frac{f^{(k)}(0)}{k!}$ für $x_0 = 0$, bzw., als Ergebnis einer analogen Rechnung, $a_k = \frac{f^{(k)}(x_0)}{k!}$ gelten muß.

Jetzt ist der Weg vorgezeichnet – wie man meinen könnte: Für jede beliebig oft differenzierbare Funktion f bilden wir die zugehörige *TAYLOR-Reihe* T_f ,

$$T_f(x, x_0) := \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k. \quad (11.77)$$

Dann müssen wir nur noch zeigen, daß die *TAYLOR-Polynome* $T_{f,n}$,

$$T_{f,n}(x, x_0) := \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k, \quad (11.78)$$

in einer Umgebung von x_0 gegen f konvergieren. Das Problem dabei ist aber, daß uns das so ohne Weiteres nicht gelingen wird. Um die Schwierigkeiten zu verdeutlichen: Wann immer wir eine Vermutung für den Wert c einer Reihe $\sum_{k=0}^{\infty} c_k$ haben, nützt uns das normalerweise ohne zusätzliche Informationen nichts. Dem Hörensagen nach hat etwa die Reihe $\sum_{k=1}^{\infty} \frac{1}{k^2}$ den Wert $\frac{\pi^2}{6}$. Das können wir aber nicht verwerten, denn wir haben keinerlei Anhaltspunkte, wie wir $|\sum_{k=1}^n \frac{1}{k^2} - \frac{\pi^2}{6}| \xrightarrow{n \rightarrow \infty} 0$ zeigen sollen. Vor dieser Situation stehen wir auch bei der TAYLOR-Entwicklung einer Funktion f . Wir haben zwar die Vermutung, daß $f(x)$ der Wert der TAYLOR-Reihe $T_f(x, x_0)$ ist, aber im Augenblick keine Methode, um das im konkreten Fall zu verifizieren. Es wird sich zeigen, daß wir noch nicht einmal darauf vertrauen können, daß $T_f(x, x_0) = f(x)$ gilt, wenn wir mit unseren Techniken zur Untersuchung von Reihen nachgewiesen haben, daß $(T_{f,n}(x, x_0))_{n \in \mathbb{N}}$ konvergiert (11.6.8, vi), vii).

Hier setzt der Satz von TAYLOR an, der in einer Formel einen verwertbaren Zusammenhang zwischen $f(x)$ und $T_{f,n}(x, x_0)$ herstellt. Die grundlegende Idee zu ihrer Gewinnung besteht in der Anwendung des verallgemeinerten Mittelwertsatzes 11.2.12 auf den Fehler

$$R_{f,n}(x, x_0) := f(x) - T_{f,n}(x, x_0).$$

der Approximation von $f(x)$ durch das TAYLOR-Polynom $T_{f,n}(x, x_0)$. Dabei ist entscheidend, diesen Fehler als Funktion des Entwicklungspunktes x_0 und nicht als Funktion von x aufzufassen. Wegen $R_{f,n}(x, x) = 0$ gibt es nach dem verallgemeinerten Mittelwertsatz ein ξ_n zwischen x und x_0 , so daß, mit einer Funktion $\delta_n(x, x_0)$ mit $\delta_n(x, x) = 0$ und $\delta'_n(x, \xi) \neq 0$ für alle ξ zwischen x und x_0 ,

$$\frac{R_{f,n}(x, x_0)}{\delta_n(x, x_0)} = \frac{R_{f,n}(x, x_0) - R_{f,n}(x, x)}{\delta_n(x, x_0) - \delta_n(x, x)} = \frac{R'_{f,n}(x, \xi_n)}{\delta'_n(x, \xi_n)}$$

gilt. Die Funktion δ_n legen wir später genauer fest. Die Idee dahinter wird deutlich, wenn wir $R'_{f,n}(x, x_0)$ ausrechnen und dabei auf eine *Teleskopsumme* stoßen:

$$\begin{aligned} R'_{f,n}(x, x_0) &= \frac{d}{dx_0} \left(f(x) - \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k \right) \\ &= \sum_{k=1}^n \frac{f^{(k)}(x_0)}{(k-1)!} (x - x_0)^{k-1} - \sum_{k=0}^n \frac{f^{(k+1)}(x_0)}{k!} (x - x_0)^k \\ &= \sum_{\ell=0}^{n-1} \frac{f^{(\ell+1)}(x_0)}{\ell!} (x - x_0)^\ell - \sum_{k=0}^n \frac{f^{(k+1)}(x_0)}{k!} (x - x_0)^k = -\frac{(x - x_0)^n}{n!} f^{n+1}(x_0). \end{aligned}$$

Auf diese Weise haben wir einen sehr allgemeinen Ausdruck für das sogenannte *Restglied* $R_{f,n}(x, x_0)$ der TAYLOR-Entwicklung erhalten:

$$R_{f,n}(x, x_0) = -\frac{f^{n+1}(\xi_n)}{n!} (x - \xi_n)^n \cdot \frac{\delta_n(x, x_0)}{\delta'_n(x, \xi_n)}. \quad (11.79)$$

In dieser Form wird es normalerweise nicht verwendet. Wir geben die gebräuchlichen Versionen für δ_n an. Dazu wählen wir $\delta_n(x, x_0) := (x - x_0)^p$, mit einem beliebigen $p \in \{1, \dots, n+1\}$ und erhalten das Restglied nach SCHLÖMILCH

$$R_{f,n}(x, x_0) = \frac{1}{p \cdot n!} f^{(n+1)}(\xi_n) (x - \xi_n)^{n+1-p} (x - x_0)^p. \quad (11.80)$$

Die Wahl $p := n+1$ liefert die meist verwendete Version, nämlich das Restglied nach LAGRANGE

$$R_{f,n}(x, x_0) = \frac{f^{(n+1)}(\xi_n)}{(n+1)!} (x - x_0)^{n+1} \quad (11.81)$$

und $p := 1$ ergibt das Restglied nach CAUCHY

$$R_{f,n}(x, x_0) = \frac{1}{n!} f^{(n+1)}(\xi_n) (x - \xi_n)^n (x - x_0) \quad (11.82)$$

$$= \frac{1}{n!} f^{(n+1)}(x_0 + \Theta_n(x - x_0)) (1 - \Theta_n)^n (x - x_0)^{n+1}. \quad (11.83)$$

Hier haben wir eine zweite Version angegeben, die man erhält, wenn man verwendet, daß eine Stelle ξ_n zwischen x_0 und x immer in der Form $\xi_n = x_0 + \Theta_n(x - x_0)$ mit einem $\Theta_n \in (0, 1)$ darstellbar ist.

Der Vollständigkeit halber geben wir noch die Integralversion von $R_{f,n}(x, x_0)$ an (vergl. Seite 392)

$$R_{f,n}(x, x_0) = \frac{1}{n!} \int_{x_0}^x (x - t)^n f^{(n+1)}(t) dt. \quad (11.84)$$

Welche Form man auch immer bevorzugt, wir haben durch das Restglied $R_{f,n}(x, x_0)$ ein Maß für den Fehler der Approximation von $f(x)$ durch die zugehörigen TAYLOR-Polynome (11.78)

$$f(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k + R_{f,n}(x, x_0).$$

Das ist die *TAYLOR-Formel*, die für jede Funktion f gilt, die $n+1$ mal differenzierbar ist. ξ_n bzw. Θ_n ist normalerweise nicht bekannt. Man könnte daher zu dem Schluß kommen, daß die TAYLOR-Formel zur Berechnung von $f(x)$ nicht verwendet werden kann. Das stimmt auch, zur exakten Berechnung von $f(x)$ ist sie normalerweise nicht geeignet, aber um Näherungswerte zu finden, die für ausreichend großes n meistens beliebig gut gewählt werden können, sehr wohl. Darüber hinaus gestattet sie es, das Problem der Konvergenz der TAYLOR-Reihe gegen $f(x)$ anzugehen, denn wir haben jetzt eine Formel für den Approximationsfehler $R_{f,n}(x, x_0)$.

11.6.1 Satz (Satz von TAYLOR) Eine reellwertige Funktion $f : (a, b) \rightarrow \mathbb{R}$ sei $n+1$ mal differenzierbar. Dann gibt es für alle $x, x_0 \in (a, b)$ jeweils ein ξ_n zwischen x und x_0 , also $\xi_n \in (x, x_0)$, oder $\xi_n \in (x_0, x)$, bzw. ein $\Theta_n \in (0, 1)$, so daß

$$f(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k + R_{f,n}(x, x_0) \quad (11.85)$$

mit

$$\begin{aligned} R_{f,n}(x, x_0) &= \frac{1}{p \cdot n!} f^{(n+1)}(\xi_n) (x - \xi_n)^{n+1-p} (x - x_0)^p \\ &= \frac{1}{p \cdot n!} f^{(n+1)}(x_0 + \Theta_n(x - x_0)) (1 - \Theta_n)^{n+1-p} (x - x_0)^{n+1} \end{aligned}$$

gilt ($1 \leq p \leq n+1$). Ist f unendlich oft differenzierbar und gilt $\lim_{n \rightarrow \infty} R_{f,n}(x, x_0) = 0$, so ist die TAYLOR-Reihe konvergent und hat den Funktionswert $f(x)$:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k. \quad (11.86)$$

Wir sagen dann, daß f an der Stelle x_0 in eine TAYLOR-Reihe entwickelt werden kann. Eine hinreichende Bedingung dafür ist die Existenz von Zahlen $r \in \mathbb{N}_0$ und $c > 0$, so daß für fast alle $n \in \mathbb{N}$ und alle x aus einem Intervall $[x_0 - \delta, x_0 + \varepsilon]$ die Abschätzung

$$|f^{(n)}(x)| \leq c n^r \quad (11.87)$$

erfüllt ist ($\delta \geq 0, \varepsilon \geq 0$). In diesem Fall konvergiert die TAYLOR-Reihe auf dem Intervall gleichmäßig gegen f , d. h., f läßt sich auf dem Intervall gleichmäßig in eine TAYLOR-Reihe entwickeln.

Beweis. Die TAYLOR-Formel (11.85) haben wir schon bewiesen. $R_{f,n}(x, x_0) \xrightarrow{n \rightarrow \infty} 0$ für ein $x \in (a, b)$ bedeutet, daß gemäß (11.85) $T_{f,n}(x) = f(x) - R_{f,n}(x, x_0)$ gegen $f(x)$ konvergiert. Das heißt also, die TAYLOR-Reihe an der Stelle x ist konvergent und hat $f(x)$ als Wert.

Die hinreichende Konvergenzbedingung: Für ein $x \in [x_0 - \delta, x_0 + \varepsilon]$ liegt auch die Zwischenstelle ξ_n in diesem Intervall. Mit (11.81) und $\rho := \max\{\delta, \varepsilon\}$ haben wir eine gleichmäßige Abschätzung

$$|R_{f,n}(x, x_0)| \leq c \frac{(n+1)^k}{(n+1)!} \rho^{n+1} \xrightarrow{n \rightarrow \infty} 0$$

des Fehlers $R_{f,n}(x, x_0)$ und damit die gleichmäßige Konvergenz der TAYLOR-Reihe auf $[x_0 - \delta, x_0 + \varepsilon]$. Das liegt daran, daß die Reihe $\sum_{k=0}^{\infty} n^k \frac{\rho^n}{n!}$ konvergiert, wie leicht aus den Quotientenkriterium folgt: $\frac{(n+1)^k \rho^{n+1} n!}{(n+1)! n^k \rho^n} = \left(\frac{n+1}{n}\right)^k \cdot \frac{\rho}{n+1} \xrightarrow{n \rightarrow \infty} 0$. \square

11.6.2 Bemerkung Ist die Entwicklungsstelle $x_0 = 0$, so schreiben wir $T_{f,n}(x)$, $T_f(x)$ und $R_{f,n}(x)$ statt $T_{f,n}(x, 0)$, $T_f(x, 0)$ und $R_{f,n}(x, 0)$.

11.6.3 Beispiel Untersuchen wir das Restglied für die bekannten Funktionen \exp , \sin , \cos und \ln .

i) Zuerst die Exponentialfunktion \exp . Für ein beliebiges $a > 0$ gilt für $x \in [-a, a]$ die Abschätzung $|R_{\exp,n}(x)| = \frac{|x|^{n+1}}{(n+1)!} e^{\xi_n} \leq \frac{|a|^{n+1}}{(n+1)!} e^a \xrightarrow{n \rightarrow \infty} 0$, (vergl. (10.13)). Das zeigt insbesondere, daß die TAYLOR-Polynome $T_{\exp,n}(x) = \sum_{k=0}^n \frac{x^k}{k!}$ auf jedem der Intervalle $[-a, a]$ gleichmäßig bzgl. x gegen e^x konvergieren. Wir haben demnach, wie wir bereits wissen, für alle $x \in \mathbb{R}$ die Entwicklung $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$.

ii) Die trigonometrischen Funktionen \sin und \cos . Da die n -te Ableitung von \sin und von \cos bis auf ein Vorzeichen wieder \sin oder \cos ist, haben wir $|\sin^{(n+1)}(\xi_n)| \leq 1$ und $|\cos^{(n+1)}(\xi_n)| \leq 1$. (11.87) ist also auf jedem Intervall $[-a, a]$ erfüllt, d. h., für jedes $x \in \mathbb{R}$ sind $\sin(x)$ und $\cos(x)$ durch ihre TAYLOR-Reihen gegeben. Für \sin haben wir $\sin'(x) = \cos(x)$, $\sin''(x) = -\sin(x)$, $\sin^{(3)}(x) = -\cos(x)$ und $\sin^{(4)}(x) = \sin(x)$. Ab hier wiederholen sich die Ableitungen in genau derselben Reihenfolge, wie die ersten vier. Wir erhalten $\sin(0) = 0$, $\sin'(0) = 1$, $\sin''(0) = 0$, $\sin^{(3)}(0) = -1$, $\sin^{(4)}(0) = 0$, $\sin^{(5)}(0) = 1$, $\sin^{(6)}(0) = 0$, ... Versuchen wir das Bildungsgesetz aufzustellen: Offensichtlich verschwinden alle geraden Ableitungen an der Stelle 0. Die ungeraden Ableitungen sind abwechselnd 1 und -1 . Setzen wir in die TAYLOR-Formel ein, so erhalten wir

$$\sin(x) = x - \frac{1}{3!} x^3 + \frac{1}{5!} x^5 \dots + \frac{(-1)^k}{(2k+1)!} x^{2k+1} \dots = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} x^{2k+1}. \quad (11.88)$$

Die Approximation von \sin durch die TAYLOR-Polynome $T_{\sin,n} =: T_n$ bis zur Ordnung $n = 21$ ist auf Seite 395 zu sehen.

Die Überlegungen für \cos verlaufen analog zu denen für \sin und ergeben

$$\cos(x) = 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \dots + \frac{(-1)^k}{(2k)!}x^{2k} \dots = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!}x^{2k}. \quad (11.89)$$

iii) Der Logarithmus. Die Funktion \ln lässt sich nicht um $x_0 = 0$ herum entwickeln, da die Funktion dort ja gar nicht definiert ist. Wir könnten das zum Anlaß nehmen, einmal $x_0 \neq 0$ zu wählen. Allerdings versucht man das bei den wichtigen Grundfunktionen zu vermeiden und nimmt dafür lieber in Kauf, die Funktion geeignet zu verschieben. Daher entwickelt man üblicherweise $\ln(x+1)$ in eine TAYLOR-Reihe. Die nötigen Ableitungen sind schnell gefunden:

$$\begin{aligned} \frac{d}{dx} \ln(x+1) &= \frac{1}{1+x} = (1+x)^{-1}, & \frac{d^2}{dx^2} \ln(x+1) &= -(1+x)^{-2}, \\ \frac{d^3}{dx^3} \ln(x+1) &= 2(1+x)^{-3}, & \frac{d^4}{dx^4} \ln(x+1) &= -2 \cdot 3(1+x)^{-4}, \\ \frac{d^5}{dx^5} \ln(x+1) &= 4!(1+x)^{-5}, & \frac{d^n}{dx^n} \ln(x+1) &= (-1)^{n-1}(n-1)!(1+x)^{-n}. \end{aligned}$$

Das Restglied nach LAGRANGE ist daher

$$R_{\ln,n}(x) = \frac{(-1)^n n! x^{n+1}}{(n+1)!(1+\xi_n)^{n+1}} = \frac{(-1)^n}{n+1} \left(\frac{x}{1+\xi_n} \right)^{n+1}.$$

Für $x \in [0, 1]$, also $\xi_n \in (0, x)$, können wir das leicht abschätzen:

$$|R_{\ln,n}(x)| = \frac{1}{n+1} \left(\frac{x}{1+\xi_n} \right)^{n+1} \leq \frac{x^{n+1}}{n+1} \leq \frac{1}{n+1} \xrightarrow{n \rightarrow \infty} 0.$$

Für $x \in (-1, 0)$ gelingt uns das nur für $x \geq -\frac{1}{2}$. Wir wissen dann $1 + \xi_n > \frac{1}{2}$ und daher $\frac{|x|}{1+\xi_n} < 2 \cdot \frac{1}{2} = 1$. Damit erhalten wir

$$|R_{\ln,n}(x)| = \frac{1}{n+1} \left(\frac{|x|}{1+\xi_n} \right)^{n+1} < \frac{1}{n+1} \xrightarrow{n \rightarrow \infty} 0.$$

Auf dem Intervall $[-\frac{1}{2}, 1]$ liegt Konvergenz der TAYLOR-Reihe gegen $\ln(x+1)$ vor:

$$\ln(x+1) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k \quad (11.90)$$

Natürlich vermuten wir auch auf dem verbleibenden Intervall $(-1, -\frac{1}{2})$ die Gültigkeit dieser Formel. Immerhin ist die Reihe auf diesem Bereich noch absolut konvergent. Allerdings ist die Konvergenz der TAYLOR-Reihe noch keine Garantie dafür, daß ihr Wert mit dem Funktionswert übereinstimmt (11.6.8, vi), vii), viii)). Die Vermutung ist in diesem Falle jedoch wahr, aber das müssen wir auf andere Weise bestätigen. Mit dem Restglied nach LAGRANGE kommen wir dabei nicht weiter: Da wir die genaue Lage von ξ_n , z. B. für $x = -\frac{2}{3}$, nicht kennen, haben wir außer $-\frac{2}{3} < \xi_n < 0$ keine weiteren Informationen. Daher wissen wir nur $1 + \xi_n > \frac{1}{3}$, woraus sich

bestenfalls auf $|R_{\ln,n}(x)| = \frac{1}{n+1} \left(\frac{|x|}{1+\xi_n} \right)^{n+1} < \frac{1}{n+1} (\frac{2}{3} \cdot 3)^{n+1} = \frac{2^{n+1}}{n+1}$ schließen lässt. Das nützt offensichtlich nichts.

Versuchen wir es mit dem Restglied (11.83) nach CAUCHY:

$$|R_{\ln,n}(x)| = \frac{(1-\Theta)^n}{(1+\Theta x)^{n+1}} |x|^{n+1},$$

mit einem $\Theta \in (0, 1)$, das von n und x abhängt (wir unterdrücken den Index n). Aus $-1 < x \leq 1$ folgt $1 - \Theta < 1 + \Theta x \leq 1 + \Theta$ und daraus $\frac{1}{1+\Theta x} < \frac{1}{1-\Theta}$. Damit können wir bequem abschätzen:

$$\begin{aligned} |R_{\ln,n}(x)| &< \frac{1}{1+\Theta x} \frac{(1-\Theta)^n}{(1-\Theta)^n} |x|^{n+1} \\ &= \frac{|x|^{n+1}}{1+\Theta x} \leq M_x |x|^{n+1} \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

$M_x := \max\{1, \frac{1}{1+x}\}$ ist das Maximum der stetigen Funktion $[0, 1] \ni \Theta \mapsto \frac{1}{1+\Theta x}$. Damit haben wir (11.90) auf dem gesamten Intervall $(-1, 1]$ gezeigt. Immerhin liegt 1 noch in dem Bereich, auf dem die TAYLOR-Reihe gegen den Funktionswert konvergiert. Daher erhalten wir, als ein Nebenprodukt unserer Überlegungen, den Wert der LEIBNIZ-Reihe (10.38)

$$\ln(2) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}. \quad (11.91)$$

Übrigens ist das keine gute Methode, um $\ln(2)$ zu berechnen, denn diese Reihe konvergiert außerordentlich langsam. Die ersten 12 Summanden ergeben gerade mal ≈ 0.6532107 , gegenüber $\ln(2) = 0.6931471805599453 \dots$. Für $x = -\frac{1}{2}$ folgt aus (11.90) eine wesentlich bessere:

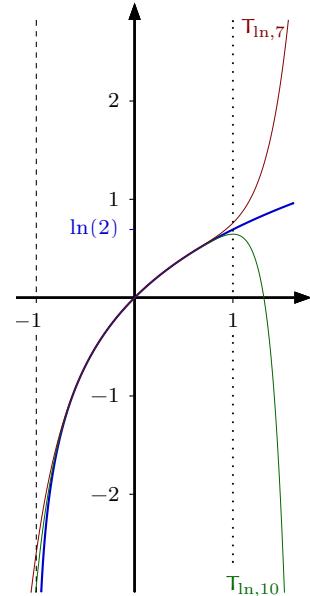
$$\ln(2) = \sum_{n=1}^{\infty} \frac{1}{n2^n}.$$

Hier hat man nach 12 Additionen bereits ≈ 0.69313 . Das kann durch

$$\ln\left(\frac{1+x}{1-x}\right) = \sum_{n=0}^{\infty} \frac{2x^{2n+1}}{2n+1} \quad (11.92)$$

noch einmal deutlich verbessert werden. Für $x = -\frac{1}{3}$ haben wir dann $\ln\left(\frac{1}{2}\right) = -\ln(2) = -\frac{2}{3} \sum_{n=0}^{\infty} \frac{1}{9^n(2n+1)}$. Nach 12 Additionen ergibt das den Wert 0.6931471805598 als Approximation für $\ln(2)$.

iv) Für die Funktion $f(x) := \frac{1}{1-x}$ brauchen wir keine TAYLOR-Entwicklung durchzuführen, denn wir kennen sie schon – oder? Trotzdem ist es sehr instruktiv, wenn wir es dennoch tun, da wir hier alle Bestimmungsstücke der Theorie konkret ausrechnen können. Die nötigen Ableitungen sind $f^{(n)}(x) = \frac{n!}{(1-x)^{n+1}}$. Wir entwickeln natürlich um $x_0 = 0$. Dann ist $f^{(n)}(0) = n!$ und daher $T_{f,n}(x) = \sum_{k=0}^n x^k$, wie erwartet.



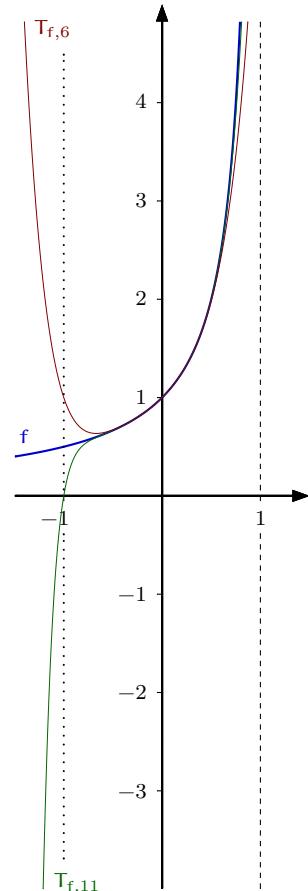
Das Restglied nach LAGRANGE hat die Form $R_{f,n}(x) = \frac{x^{n+1}}{(1-\xi_n)^{n+1}}$. Es macht ähnliche Probleme, wie das von $\ln(x+1)$. Wir untersuchen daher gleich das Restglied nach CAUCHY:

$$\begin{aligned} |R_{f,n}(x)| &= \frac{(n+1)!(1-\Theta)^n|x|^{n+1}}{n!(1-\Theta x)^{n+2}} \\ &< \frac{(n+1)|x|^{n+1}}{(1-\Theta x)^2} \leq N_x(n+1)|x|^{n+1} \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

$N_x := \max\{1, \frac{1}{(1-x)^2}\}$ ist das Maximum der stetigen Funktion $[0, 1] \ni \Theta \mapsto \frac{1}{(1-\Theta x)^2}$. Damit haben wir die Konvergenz der TAYLOR-Reihe gegen $f(x)$ auf $(-1, 1)$ explizit nachgerechnet – natürlich nur so zur Übung, denn eigentlich wissen wir ja alles über die geometrische Reihe:

$$\begin{aligned} \frac{1}{1-x} &= \sum_{k=0}^{\infty} x^k = \sum_{k=0}^n x^k + \sum_{k=n+1}^{\infty} x^k \\ &= T_{f,n}(x) + x^{n+1} \sum_{\ell=0}^{\infty} x^{\ell} = T_{f,n}(x) + \frac{x^{n+1}}{1-x}. \end{aligned}$$

Daher ist $R_{f,n}(x) = \frac{x^{n+1}}{1-x}$, was offensichtlich für $|x| < 1$ gegen Null konvergiert. Berechnen wir ξ_n , auch wenn das jetzt eigentlich nicht mehr nötig ist. Dazu haben wir $(1-\xi_n)^{n+1} = 1-x$ aufzulösen, mit dem Ergebnis $1-\xi_n = \sqrt[n+1]{1-x} \xrightarrow{n \rightarrow \infty} 1$. Mit dieser Information über ξ_n von Anfang an wäre die Konvergenz des Restglieds kein Problem gewesen.



v) Die bisherigen Beispiele zeigen, daß die direkte Abschätzung des Restglieds mitunter nicht ganz einfach ist, selbst bei so elementaren Funktionen wie $x \mapsto \frac{1}{1-x}$, oder $x \mapsto \ln(1+x)$. Wir werden daher, wann immer es möglich ist, versuchen, die TAYLOR-Entwicklung auf andere Weise zu finden. Die Funktion $g(x) := e^{-x^2}$ ist dafür ein erstes Beispiel. Hier fangen wir gar nicht erst mit den Ableitungen an, sondern verwenden die bekannte TAYLOR-Reihe der Exponentialfunktion: $e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}$. Für $z = -x^2$ erhalten wir daraus

$$e^{-x^2} = \sum_{k=0}^{\infty} \frac{(-x^2)^k}{k!} = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k}}{k!}.$$

Da die Darstellung einer Funktion durch eine Potenzreihe bei gegebenem Entwicklungspunkt eindeutig ist (vergl. Lemma 11.6.5), haben wir die TAYLOR-Entwicklung von g gefunden.

11.6.4 Definition Eine reelle Funktion $f: D_f \rightarrow \mathbb{R}$ auf einem offenen Intervall D_f heißt an einer Stelle $x_0 \in D_f$ analytisch, wenn es eine offene Umgebung von x_0 gibt, auf der f eine Potenzreihenentwicklung

$$f(x) = \sum_{k=0}^{\infty} a_k (x - x_0)^k$$

hat. Ist f an jeder Stelle $x_0 \in D_f$ analytisch, so heißt f analytisch.

Die Funktion f heißt an der Stelle $x_0 \in D_f$ glatt, wenn sie in einer Umgebung von x_0 beliebig oft differenzierbar ist. f heißt glatt, wenn sie an jeder Stelle $x_0 \in D_f$ glatt ist.

Eine in x_0 analytische Funktion ist nach Satz 11.3.1 in einer Umgebung von x_0 beliebig oft differenzierbar, also glatt in x_0 . Die Umkehrung gilt nicht, wie wir in Beispiel 11.6.8, vii) sehen werden.

11.6.5 Lemma Die Potenzreihendarstellung einer in x_0 analytischen Funktion ist eindeutig.

Beweis. Aus $f(x) = \sum_{k=0}^{\infty} a_k(x - x_0)^k = \sum_{k=0}^{\infty} b_k(x - x_0)^k$ folgt $f(x_0) = a_0 = b_0$. Da man Potenzreihen gliedweise ableiten darf, haben wir auch $f'(x_0) = a_1 = b_1$, $f''(x_0) = 2a_2 = 2b_2$, $f'''(x_0) = 3!a_3 = 3!b_3$ und allgemein $f^{(n)}(x_0) = n!a_n = n!b_n$. Aus diesen Gleichungen folgt unmittelbar $a_k = b_k$ für alle $k \in \mathbb{N}_0$. \square

11.6.6 Lemma Eine in x_0 analytische Funktion f läßt sich in eine TAYLOR-Reihe entwickeln, die mit der Potenzreihendarstellung von f übereinstimmt.

Beweis. Sei $f(x) = \sum_{k=0}^{\infty} a_k(x - x_0)^k$. Dann folgt, wie oben: $a_k = \frac{f^{(k)}(x_0)}{k!}$. Damit ist $f(x)$ in einer Umgebung von x_0 durch seine TAYLOR-Reihe gegeben. Daher ist $T_{f,n}(x, x_0) = \sum_{k=0}^n a_k(x - x_0)^k$ und $R_{f,n}(x, x_0) = \sum_{k=n+1}^{\infty} a_k(x - x_0)^k$. Da die Potenzreihe laut Voraussetzung auf einer Umgebung von x_0 konvergent sein soll, muß der Reihenrest $\sum_{k=n+1}^{\infty} a_k(x - x_0)^k$ für $n \rightarrow \infty$ gegen Null konvergieren. Nach dem Satz von TAYLOR ist f in einer Umgebung von x_0 in eine TAYLOR-Reihe entwickelbar, die durch die Potenzreihendarstellung von f selbst gegeben ist. \square

Dieses Lemma gehört zu denen, die eigentlich selbstverständlich sind. Seine Aufgabe besteht nur darin, sicherzustellen, daß die Eigenschaft einer Funktion analytisch zu sein und die Darstellung durch eine TAYLOR-Reihe, wie erwartet, zusammenpassen: *Eine analytische Funktion ist ihre eigene TAYLOR-Entwicklung*.

11.6.7 Satz Sind f und g in x_0 analytisch, dann gilt das auch für fg und, falls $g(x_0) \neq 0$ erfüllt ist, auch für $\frac{f}{g}$.

Beweis. Wir gehen von $f(x) = \sum_{\ell=0}^{\infty} a_{\ell}(x - x_0)^{\ell}$ und $g(x) = \sum_{k=0}^{\infty} b_k(x - x_0)^k$ aus. Beide Reihen haben einen nicht verschwindenden Konvergenzradius. R sei der kleinere der beiden. Dann sind beide Reihen in $(x_0 - R, x_0 + R)$ absolut konvergent. Wir haben dort also den Doppelreihensatz 10.2.21 für die Doppelreihe zur Verfügung, die zu $f(x)g(x)$ gehört:

$$\begin{aligned} f(x)g(x) &= \sum_{\ell=0}^{\infty} g(x)a_{\ell}(x - x_0)^{\ell} = \sum_{\ell=0}^{\infty} \sum_{k=0}^{\infty} a_{\ell}b_k(x - x_0)^{\ell+k} \\ &= \sum_{\mu=0}^{\infty} \sum_{\ell=0}^{\mu} a_{\ell}b_{\mu-\ell}(x - x_0)^{\mu} = \sum_{k=0}^{\infty} c_k(x - x_0)^k, \end{aligned}$$

mit $c_k := \sum_{\ell=0}^k a_{\ell}b_{k-\ell}$... muß fortgesetzt werden. \square

11.6.8 Beispiel Die beiden letzten Lemmata eröffnen die Möglichkeit, die TAYLOR-Entwicklung wichtiger Funktionen auf andere Weise zu finden, als dadurch, alle Ableitungen zu berechnen. Wir beziehen uns dabei immer wieder auf die Möglichkeit, für eine Potenzreihe die Ableitung bzw. eine Stammfunktion gliedweise bilden zu dürfen (Satz 11.3.1 und 11.5.3). Für die Logarithmusfunktion $x \mapsto \ln(1+x)$ etwa ist die Ableitung durch $x \mapsto \frac{1}{1+x}$ gegeben, eine Funktion deren TAYLOR-Entwicklung eine geometrische Reihe mit Konvergenzradius 1 ist: $\frac{1}{1+x} = \sum_{k=0}^{\infty} (-1)^k x^k$. Eine geeignete Stammfunktion ist die Ausgangsfunktion: $\ln(1+x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{k+1}}{k+1} + c$. Die Konstante c ist Null, wie durch Einsetzen von $x = 0$ leicht bestätigt werden kann. Die Indextransformation $\ell = k + 1$ führt auf (11.90). Der Vorteil dieser Vorgehensweise besteht darin, daß wir ohne großen Aufwand die Reihenentwicklung bekommen und darüber hinaus auch keinen Ärger mit Abschätzungen des TAYLOR-Restglieds haben. Wir wissen aus Satz 11.5.3 und 11.5.8, daß die Potenzreihe auf ihrem Konvergenzbereich $(-1, 1]$ tatsächlich die Funktion $x \mapsto \ln(x+1)$ wiedergibt.

Ein anderes Beispiel ist die Funktion $x \mapsto \frac{x}{(1-x)^2}$. Sie ist, bis auf den Faktor x , die Ableitung von $x \mapsto \frac{1}{1-x} = \sum_{k=0}^{\infty} x^k$. Also ist ihre TAYLOR-Entwicklung $\frac{x}{(1-x)^2} = x \cdot \sum_{k=1}^{\infty} kx^{k-1} = \sum_{k=1}^{\infty} kx^k$. Der Konvergenzbereich ist offensichtlich $(-1, 1)$.

Wie werden immer wieder auf diese Methoden zurückgreifen, um die TAYLOR-Entwicklung wichtiger Funktionen zu bestimmen. Aber zunächst stellen wir das Standardbeispiel für eine Funktion vor, deren TAYLOR-Reihe zwar überall konvergiert, aber, bis auf uninteressante Ausnahmen, nicht gegen den Funktionswert. Damit wird noch einmal unterstrichen, daß die TAYLOR-Reihe alleine nicht ausreicht, um zu entscheiden, ob die TAYLOR-Entwicklung einer Funktion möglich ist, oder nicht. Man braucht immer noch zusätzliche Informationen, etwa wie die oben vorgestellten, oder, wenn nichts anderes hilft, eben das Verhalten des Restglieds.

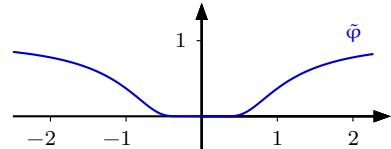
vi) Die Funktion (11.46),

$$\varphi(x) = \begin{cases} e^{-\frac{1}{x}} & , x > 0 \\ 0 & , x \leq 0 \end{cases}$$

die wir in den Beispielen 11.3.5 zu den Regeln von DE L'HOSPITAL untersucht haben, ist glatt. Ihre Ableitungen an der Stelle $x_0 = 0$ sind $\varphi^{(n)}(0) = 0$. Damit ist die zugehörige TAYLOR-Reihe schnell hingeschrieben: $T_{\varphi}(x) = 0$. Außer auf dem wenig spannenden Intervall $(-\infty, 0)$, stimmt diese Reihe nicht mit dem Funktionswert $\varphi(x)$ überein.

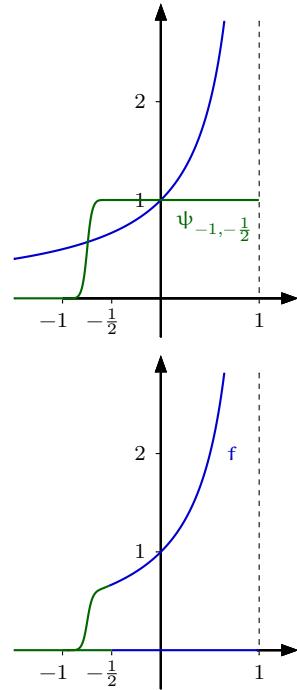
vii) Wir können φ etwas abändern: $\tilde{\varphi}(x) := \varphi(x^2)$, also

$$\tilde{\varphi}(x) = \begin{cases} e^{-\frac{1}{x^2}} & , x \neq 0 \\ 0 & , x = 0 \end{cases} .$$



Als Verkettung glatter Funktionen ist $\tilde{\varphi}$ glatt. Außerdem erbt sie über die Kettenregel die Eigenschaft $\tilde{\varphi}^{(n)}(0) = 0$ von φ . Damit ist auch ihre TAYLOR-Reihe einfach anzugeben: $T_{\tilde{\varphi}}(x) = 0$, mit unendlichem Konvergenzradius. Sie stimmt nur noch an der Stelle 0 mit dem Funktionswert von $\tilde{\varphi}$ überein.

viii) Die beiden vorangehenden Beispiele erscheinen etwas extrem, in dem Sinne, daß die TAYLOR-Entwicklung keine brauchbare Information über die entwickelte Funktion enthält. Wir können aber auch Funktionen angeben, für die die Entwicklung auf einem Teil des Konvergenzbereichs die Funktion wiedergibt und auf einem anderen nicht. Dafür müssen wir nur eine analytische Funktion, sagen wir $x \mapsto \frac{1}{1-x}$, mit einer anderen zu einer glatten Funktion geeignet fortsetzen, und zwar so, daß die ursprüngliche Funktion auf einem Teilintervall, das den Entwicklungspunkt x_0 enthält, nicht verändert wird. Dazu eignet sich die Funktion $\psi_{a,b}$ ((11.69) aus Übung 11.4.10). Wir definieren f durch die Vorschrift $f(x) := \psi_{-1,-\frac{1}{2}}(x) \frac{1}{1-x}$. Da $\psi_{-1,-\frac{1}{2}}(x) = 1$ oberhalb von $x = -\frac{1}{2}$ gilt, ist $f(x)$ in diesem Bereich einfach durch $\frac{1}{1-x}$ gegeben. Bilden wir also die TAYLOR-Reihe von f , mit Entwicklungspunkt $x_0 = 0$, so entsteht dabei die geometrische Reihe, die auf $(-1, 1)$ konvergiert. Allerdings stimmt sie mit $f(x)$ nur auf $[-\frac{1}{2}, 1)$ überein, denn auf $(-1, -\frac{1}{2})$ liefert sie ebenfalls $\frac{1}{1-x}$, aber das ist hier nicht $f(x)$.



ix) Eine wichtige Klasse von Funktionen, deren TAYLOR-Entwicklung wir kennen sollten, ist durch $x \mapsto (1+x)^\alpha$, bei beliebigem $\alpha \in \mathbb{R}$, definiert. Die Entwicklung ist eine Verallgemeinerung des binomischen Lehrsatzes 1.2.16. Wir haben keine Funktion zur Hand, auf die wir sie zurückführen könnten, um so den Weg über die Ableitungen zu vermeiden. Daher:

$$\begin{aligned} \frac{d}{dx}(1+x)^\alpha &= \alpha(1+x)^{\alpha-1}, & \frac{d^2}{dx^2}(1+x)^\alpha &= \alpha(\alpha-1)(1+x)^{\alpha-2}, & \dots \\ \frac{d^k}{dx^k}(1+x)^\alpha &= \alpha(\alpha-1)\cdots(\alpha-k+1)(1+x)^{\alpha-k}. \end{aligned}$$

Die Koeffizienten der TAYLOR-Reihe sind dann durch $\frac{1}{k!} \alpha(\alpha-1)\cdots(\alpha-k+1)$ gegeben, wenn wir um $x_0 = 0$ entwickeln. Sie sehen auf den ersten Blick etwas umständlich aus. Nehmen wir für den Moment einmal $\alpha \in \mathbb{N}$ an. Dann verschwinden die Koeffizienten ab dem Index $\alpha + 1$ und sind für $0 \leq k \leq \alpha$ die gewöhnlichen Binomialkoeffizienten

$$\frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!} = \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)(\alpha-k)\cdots 2}{k!(\alpha-k)!} = \frac{\alpha!}{k!(\alpha-k)!} = \binom{\alpha}{k}.$$

Wir definieren daher für alle $\alpha \in \mathbb{R}$ die *verallgemeinerten Binomialkoeffizienten*

$$\binom{\alpha}{k} := \begin{cases} \frac{1}{k!} \alpha(\alpha-1)\cdots(\alpha-k+1) & , k \in \mathbb{N} \\ 1 & , k = 0 \end{cases} \quad (11.93)$$

Dann wird $\binom{0}{0} = 1$ und $\binom{0}{k} = 0$ für $k \in \mathbb{N}$. Außerdem bricht die Folge der Binomialkoeffizienten automatisch nach $k = \alpha$ ab, sollte α eine natürliche Zahl sein. Die vermutete TAYLOR-Entwicklung ist die *binomische Reihe*

$$(1+x)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k. \quad (11.94)$$

Testen wir sie für $\alpha = -1$. Dann ist $\binom{-1}{k} = \frac{1}{k!}(-1)(-2)\cdots(-k) = (-1)^k$ und daher $(1+x)^{-1} = \sum_{k=0}^{\infty} (-1)^k x^k$, wie es sein soll. Mit dem Quotientenkriterium finden wir den Konvergenzradius $R = 1$:

$$\frac{\left| \binom{\alpha}{k+1} \right|}{\left| \binom{\alpha}{k} \right|} = \frac{|\alpha(\alpha-1)\cdots(\alpha-k+1)(\alpha-k)|k!}{|\alpha(\alpha-1)\cdots(\alpha-k+1)|(k+1)!} = \frac{|\alpha-k|}{k+1} \xrightarrow{k \rightarrow \infty} 1.$$

Da gilt natürlich nur für $\alpha \notin \mathbb{N}$, denn für $\alpha \in \mathbb{N}$ bricht die Reihe ja ab und ergibt die bekannte binomische Formel (1.32). Es verhält sich also alles so, wie man es von einer Erweiterung des binomischen Lehrsatzes erwarten sollte. Das einzige verbleibende Problem ist, daß (11.94) bisher nur eine Vermutung ist. Um sie zu bestätigen, müssen wir die TAYLOR-Restglieder abschätzen. Wir verwenden die Version von CAUCHY. Wegen $\frac{1}{n!} \frac{d^{n+1}}{dx^{n+1}} (1+x)^\alpha = (n+1) \binom{\alpha}{n+1} (1+x)^{\alpha-n-1}$ haben wir für $x \in (-1, 1)$, wie schon mehrfach vorgeführt, den folgenden Ausdruck abzuschätzen:

$$(n+1) \left| \binom{\alpha}{n+1} \right| (1+\Theta x)^{\alpha-1} \frac{(1-\Theta)^n}{(1+\Theta x)^n} |x|^{n+1} < K_x (n+1) \left| \binom{\alpha}{n+1} \right| |x|^{n+1} \xrightarrow{n \rightarrow \infty} 0.$$

Dabei ist K_x das Maximum der stetigen Funktion $[0, 1] \ni \Theta \mapsto (1+\Theta x)^{\alpha-1}$. Die Konvergenz gegen 0 folgt aus der absoluten Konvergenz der abgeleiteten binomischen Reihe.

Wir berechnen (11.94) für den wichtigen Fall $\alpha = -\frac{1}{2}$. Für $n \geq 1$ erhalten wir

$$\binom{-\frac{1}{2}}{n} = \frac{(-\frac{1}{2})(-\frac{3}{2})(-\frac{5}{2})\cdots(-\frac{2n-1}{2})}{n!} = (-1)^n \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2 \cdot 4 \cdot 6 \cdots 2n} = (-1)^n \frac{(2n-1)!!}{(2n)!!}.$$

Dabei sind die sogenannten *Doppelfakultäten* $(2n)!!$ bzw $(2n-1)!!$ folgendermaßen definiert:

$$(2n)!! := 2 \cdot 4 \cdot 6 \cdots 2n = 2^n n!, \quad (2n-1)!! := 1 \cdot 3 \cdot 5 \cdots (2n-1) = \frac{(2n)!}{2^n n!}. \quad (11.95)$$

Mit $\frac{(2k-1)!!}{(2k)!!} = \frac{1}{4^k} \frac{(2k)!}{k! k!} = \frac{1}{4^k} \binom{2k}{k}$ ergibt das

$$\frac{1}{\sqrt{1+x}} = \sum_{k=0}^{\infty} \frac{(-1)^k}{4^k} \binom{2k}{k} x^k = 1 - \frac{1}{2}x + \frac{3}{8}x^2 - \frac{5}{16}x^3 + \frac{35}{128}x^4 - \cdots. \quad (11.96)$$

x) Wir verwenden (11.96), um die TAYLOR-Entwicklung von \arcsin zu finden. Wegen $\arcsin'(z) = \frac{1}{\sqrt{1-z^2}}$ haben wir zunächst in (11.96) $-x^2$ einzusetzen und davon eine Stammfunktion zu bilden (die additive Konstante ist hier 0):

$$\arcsin(x) = \sum_{k=0}^{\infty} \frac{1}{(2k+1)4^k} \binom{2k}{k} x^{2k+1} = x + \frac{1}{6}x^3 + \frac{3}{40}x^5 + \frac{5}{112}x^7 + \cdots. \quad (11.97)$$

Wegen $\arccos(x) = \frac{\pi}{2} - \arcsin(x)$, (vergl. Seite 309) haben wir damit auch

$$\arccos(x) = \frac{\pi}{2} - \sum_{k=0}^{\infty} \frac{1}{(2k+1)4^k} \binom{2k}{k} x^{2k+1}. \quad (11.98)$$

xi) Die Reihenentwicklung für \arctan haben wir uns schon in (11.72) überlegt:

$$\arctan(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} x^{2k+1}.$$

Wegen $\operatorname{arccot}'(x) = -\frac{1}{1+x^2}$, hat $\operatorname{arccot}(x)$ im Wesentlichen dieselbe Reihenentwicklung, wie $\arctan(x)$, nämlich $c - \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} x^{2k+1}$. Aus $\operatorname{arccot}(0) = \frac{\pi}{2}$ folgt

$$\operatorname{arccot}(x) = \frac{\pi}{2} - \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} x^{2k+1}. \quad (11.99)$$

Da für $x > 0$ die Beziehung $\operatorname{arccot}(x) = \arctan(\frac{1}{x})$ besteht, hat man über (11.99) die Möglichkeit, die Berechnung von $\arctan(x)$ für $x > 1$ auf die \arctan -Reihe für $0 < x < 1$ zurückzuführen.

Wir kennen die Reihenentwicklung für die Umkehrfunktionen aller trigonometrischen Funktionen und für \sin und \cos . Es fehlen die Reihen für \tan und \cot . Überraschenderweise sind diese alles andere als einfach zu finden. Der Versuch, alle Ableitungen von \tan oder \cot zu bilden, führt auf keine erkennbare Struktur. Wir werden sehen, daß ohne die Kenntnis der sogenannten BERNOULLI-Zahlen, keine Chance besteht, die Reihenentwicklung von \tan und \cot zu verstehen.

11.6.9 Die BERNOULLI-Zahlen

Die BERNOULLI-Polynome $(B_k(t))_{k \in \mathbb{N}_0}$, für $t \in \mathbb{R}$, entstehen als Koeffizienten der Reihenentwicklung der Funktionenfamilie $\{b_t | t \in \mathbb{R}\}$:

$$b_t(x) := \frac{xe^{tx}}{e^x - 1} =: \sum_{k=0}^{\infty} B_k(t) \frac{x^k}{k!}, \quad x \neq 0, \quad b_t(0) := 1. \quad (11.100)$$

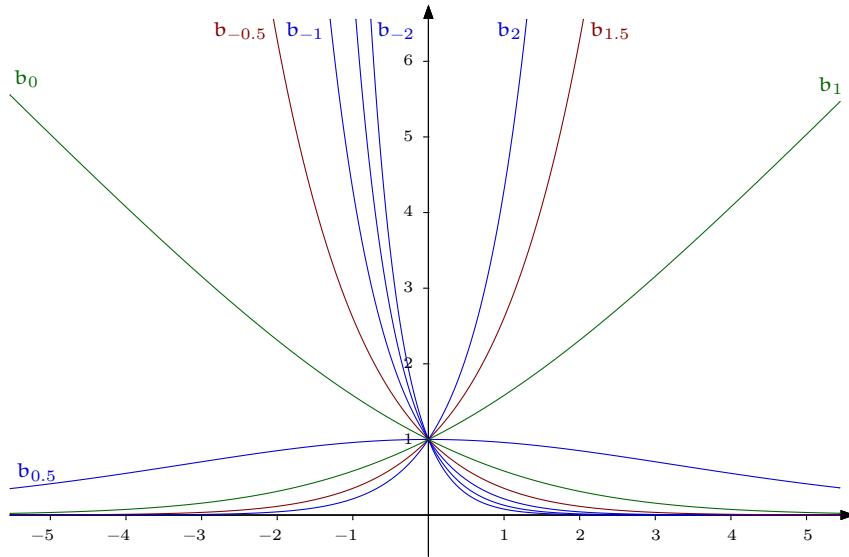
Die eigentlichen BERNOULLI-Zahlen sind durch $B_k := B_k(0)$ definiert, d.h., sie sind durch

$$\frac{x}{e^x - 1} = \sum_{k=0}^{\infty} B_k \frac{x^k}{k!} \quad (11.101)$$

bestimmt ($x \neq 0$). Es gilt $b_t(-x) = b_{1-t}(x)$, d.h., durch Spiegeln an der y -Achse geht b_t in b_{1-t} über.

Wir verwenden die Regeln von DE L'HOSPITAL, um zu zeigen, daß die durch (11.100) definierten Funktionen b_t an der Stelle $x = 0$ stetig differenzierbar sind (an allen anderen Stellen ist b_t natürlich beliebig oft differenzierbar):

$$\begin{aligned} \lim_{x \rightarrow 0^{\pm}} b_t(x) &= \lim_{x \rightarrow 0^{\pm}} \frac{xe^{tx}}{e^x - 1} \stackrel{L'H}{=} \lim_{x \rightarrow 0^{\pm}} \frac{e^{tx} + txe^{tx}}{e^x} = 1 = b_t(0). \\ \lim_{x \rightarrow 0^{\pm}} b'_t(x) &= \lim_{x \rightarrow 0^{\pm}} \frac{(1-x)e^x - 1}{(e^x - 1)^2} \stackrel{L'H}{=} \lim_{x \rightarrow 0^{\pm}} \frac{-xe^x}{2(e^{2x} - e^x)} \\ &\stackrel{L'H}{=} \lim_{x \rightarrow 0^{\pm}} \frac{-xe^x - e^x}{2(2e^{2x} - e^x)} = -\frac{1}{2} = b'_t(0). \end{aligned}$$



Wegen $b_t(x) = e^{tx} b_0(x)$ ist $b'_t(x) = e^{tx} (tb_0(x) + b'_0(x))$. Daher ist $b'_t(0) = t - \frac{1}{2}$ und b'_t stetig.

Bevor wir die Eigenschaften der BERNOULLI-Zahlen untersuchen, geben wir erst einmal einen Hinweis, warum sie etwas mit \tan und \cot zu tun haben:

$$b_0(2x) = \frac{2x}{e^{2x} - 1} = x \frac{e^{2x} + 1}{e^{2x} - 1} - x = x \frac{e^x + e^{-x}}{e^x - e^{-x}} - x = x \coth(x) - x.$$

Das ist noch nicht das gewünschte Ergebnis, aber immerhin schon einmal $\coth(x) = 1 + \frac{1}{x} b_0(2x)$. Von hier zu \cot und dann zu \tan ist es nicht mehr weit: $\cot(x) = i \coth(ix)$ ist der erste und $\cot(x) - 2 \cot(2x) = \cot(x) - \frac{\cot^2(x)-1}{\cot(x)} = \frac{1}{\cot(x)} = \tan(x)$ der zweite Zusammenhang (vergl. (11.20)).

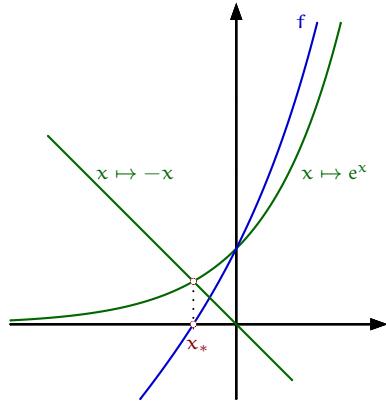
11.6.10 A Zeigen Sie (11.92).

11.6.11 Beispiel (EULER-Formel) Wir wissen, daß die Exponentialreihe, ebenso wie die Sinus- und die Kosinusreihe, jeweils einen unendlichen Konvergenzradius hat. Wir können jedes $z \in \mathbb{C}$ in die Exponentialfunktion einsetzen, insbesondere $z = ix$, und haben es dann mit einer absolut konvergenten Reihe zu tun. Diese können wir nach Satz 10.2.4 in zwei Reihen aufspalten, nämlich in die mit geraden und die mit ungeraden Potenzen. Beachten wir dabei noch $i^{2k} = (i^2)^k = (-1)^k$, so erhalten wir ohne Mühe die berühmte EULER-Formel:

$$\begin{aligned} e^{ix} &= \lim_{n \rightarrow \infty} \sum_{k=0}^{2n+1} \frac{i^k}{k!} x^k = \lim_{n \rightarrow \infty} \left[\sum_{k=0}^n \frac{i^{2k}}{(2k)!} x^{2k} + \sum_{k=0}^n \frac{i i^{2k}}{(2k+1)!} x^{2k+1} \right] \\ &= \lim_{n \rightarrow \infty} \sum_{k=0}^n \frac{(-1)^k}{(2k)!} x^{2k} + i \lim_{n \rightarrow \infty} \sum_{k=0}^n \frac{(-1)^k}{(2k+1)!} x^{2k+1} = \cos(x) + i \sin(x). \end{aligned}$$

11.7 Das Newton-Verfahren

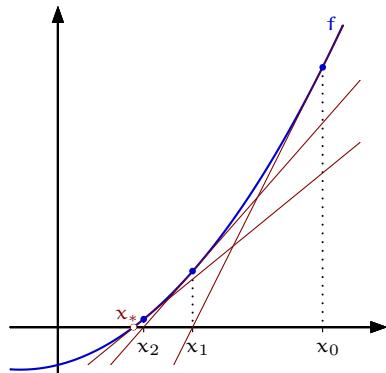
Viele Gleichungen, wie etwa $e^x = -x$, sind nicht geschlossen lösbar. Eine Skizze der Situation zeigt aber, daß es eine Lösung x_* gibt. Normalerweise lassen sich solche Gleichungen auf die Nullstellenbestimmung geeigneter Funktionen zurückführen. Im Beispiel ist es die Funktion $f(x) := e^x + x$. Wir brauchen daher ein Verfahren, mit dem wir die Nullstellen von Funktionen beliebig genau bestimmen können. Sind die Funktionen ausreichend glatt, also wenigstens stetig differenzierbar, dann kann man mit Hilfe des *Newton-Verfahren* recht schnell gute Näherungen gewinnen. Die Idee zum Verfahren ist sehr anschaulich: Man startet mit einem Schätzwert x_0 der Nullstelle und bildet die Tangente t_0 von f an dieser Stelle. Deren Schnittpunkt x_1 mit der x -Achse ist der erste Näherungswert. Man verwendet ihn als neuen Schätzwert, wiederholt mit ihm das Verfahren und erhält die nächste Näherung x_2 , usw.



Für die Näherung x_n läßt sich leicht eine Rekursion gewinnen. Dazu bestimmen wir den Schnittpunkt der Tangente $t_0(x) = f(x_0) + f'(x_0)(x - x_0)$ mit der x -Achse. $t_0(x_1) = 0$ führt auf $f'(x_0)(x_1 - x_0) = -f(x_0)$ und das einfach auf $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$. Wenn im nächsten Schritt x_1 zum neuen Schätzwert wird, erhalten wir die nächste Näherung x_2 durch $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$. Es ist jetzt klar, daß die Rekursion durch

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (11.102)$$

gegeben ist, mit einem Startwert x_0 , von dem offensichtlich $f'(x_0) \neq 0$ zu fordern ist. Tatsächlich sollte in einer kleinen Umgebung der Nullstelle $f'(x) \neq 0$ gelten (von der Nullstelle selbst abgesehen).



Für unser Beispiel wählen wir den Startwert $x_0 = 0$. Als Rekursion erhalten wir

$$x_{n+1} = x_n - \frac{e^{x_n} + x_n}{e^{x_n} + 1} = e^{x_n} \frac{x_n - 1}{e^{x_n} + 1}.$$

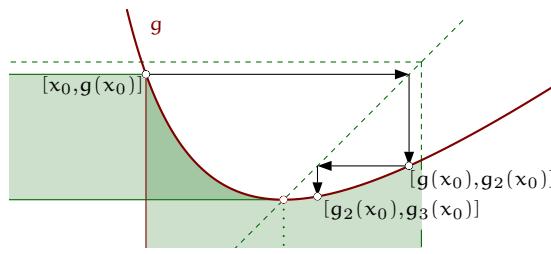
Die Schätzwerte sind dann der Reihe nach $x_0 = 0$, $x_1 = -0.5$, $x_2 \approx -0.5663110031972182$, $x_3 \approx -0.5671431650348622$, $x_4 \approx -0.5671432904097811$, $x_5 \approx -0.5671432904097838$, $x_6 \approx -0.5671432904097838$. Bei einer Genauigkeit von 16 Stellen nach dem Komma bleibt das Ergebnis jetzt konstant (was natürlich im Allgemeinen nicht heißt, daß es sich um das genaue Ergebnis handelt). Wir haben in diesem Beispiel eine rasante Zunahme der Genauigkeit: Pro Iterationsschritt hat sich die Anzahl der Stellen, die sich nicht mehr verändern und daher wohl als korrekt anzusehen sind, etwa verdoppelt. Das ist kein Zufall, wie der folgende Satz zeigt.

11.7.1 Satz Sei f eine auf dem Intervall $[a, b]$ zweimal stetig differenzierbare Funktion, mit den Eigenschaften $f(a)f(b) < 0$ und $f'(x) \neq 0$ f. a. $x \in [a, b]$. Dann gibt es genau eine Nullstelle

$x_* \in (a, b)$ von f und eine Zahl $0 < \delta < 1$, so daß das NEWTON-Verfahren für einen Startwert x_0 in der Umgebung $[x_* - \delta, x_* + \delta]$ von x_* in quadratischer Ordnung gegen x_* konvergiert. D.h., es gibt eine Zahl $M > 0$, so daß $|x_{n+1} - x_*| \leq M|x_n - x_*|^2$ für alle $n \in \mathbb{N}$ gilt.

Beweis. Aus $f(a)f(b) < 0$ folgt, daß die Funktionswerte auf dem Rand des Intervalls verschiedene Vorzeichen haben. Nach dem Zwischenwertsatz 11.1.10 gibt es mindestens eine Nullstelle $x_* \in (a, b)$ von f . Da f' auf $[a, b]$ keine Nullstelle hat, muß dort entweder $f'(x) > 0$ oder $f'(x) < 0$ gelten. Daher ist f entweder streng monoton wachsend oder fallend. Damit ist klar, daß x_* die einzige Nullstelle von f in $[a, b]$ ist. Nun zum NEWTON-Verfahren. Dabei wird die Funktion $g(x) := x - \frac{f(x)}{f'(x)}$ mit einem Startwert $x_0 \in [a, b]$ iteriert: Die Folge $(x_n)_{n \in \mathbb{N}}$ der Schätzwerte entsteht durch $x_n := g_n(x_0)$ mit $g_n(x) := \underbrace{g \circ g \circ \cdots \circ g}_{n-\text{mal}}(x) = g(g(g(\cdots g(x) \cdots)))$ und $g_0(x) := x$. Um sicherzustellen, daß sich die Bedingungen für g während der Iteration nicht ändern, muß x_0 ausreichend nah bei x_* gewählt werden. Daß das möglich ist, überlegen wir uns mit Hilfe des Mittelwertsatzes 11.2.11 für g . Zunächst ist klar, daß $g(x) = x$ äquivalent zu $f(x) = 0$ ist. Auf $[a, b]$ ist das nur für x_* möglich. Mit $g'(x) = \frac{f(x)f''(x)}{f'(x)^2}$ erhalten wir durch zweimalige Anwendung des Mittelwertsatzes

$$\begin{aligned}|g(x) - x_*| &= |g(x) - g(x_*)| = \left| \frac{f(\xi)f''(\xi)}{f'(\xi)^2} \right| |x - x_*| = \frac{|f(\xi) - f(x_*)||f''(\xi)|}{f'(\xi)^2} |x - x_*| \\ &= \frac{|f'(\eta)||f''(\xi)|}{f'(\xi)^2} |\xi - x_*| |x - x_*| \leq \frac{|f'(\eta)||f''(\xi)|}{f'(\xi)^2} |x - x_*|^2.\end{aligned}$$



Die erste Anwendung bezieht sich auf g und liefert eine Stelle ξ zwischen x und x_* . Die zweite betrifft f und ergibt die Zwischenstelle η zwischen ξ und x_* . Nun ersetzen wir $|f'(\eta)|$ und $\frac{|f''(\xi)|}{|f'(\xi)|^2}$ jeweils durch den maximalen Funktionswert M_1 bzw. M_2 , den die stetigen Funktionen $|f'|$ bzw. $\frac{|f''|}{|f'|^2}$ auf $[a, b]$ annehmen. Wir erhalten die

Abschätzung $|g(x) - x_*| \leq M|x - x_*|^2$, mit $M := M_1 M_2$. Jetzt sorgen wir dafür, daß x so nah bei x_* liegt, daß $M|x - x_*| < 1$ gilt. Wir wählen also $\delta < \min\{M^{-1}, 1\}$ und haben für alle $x \in U_\delta := [x_* - \delta, x_* + \delta]$ die Abschätzung $|g(x) - x_*| \leq M|x - x_*|^2 \leq M\delta^2 \leq \delta$. Das bedeutet insbesondere, daß $g(x)$ wieder in U_δ liegt. Wir haben diese Abschätzung also auch für $g(g(x))$ und für alle weiteren Iterationen von g zur Verfügung. Wir wählen x_0 in U_δ und erhalten

$$\begin{aligned}|x_n - x_*| &= |g_n(x_0) - x_*| = |g(g_{n-1}(x_0)) - x_*| \leq M|g_{n-1}(x_0) - x_*|^2 \\ &\leq M \cdot M^2 |g_{n-2}(x_0) - x_*|^4 \leq M^{\sum_{k=0}^{n-2} 2^k} |g_{n-3}(x_0) - x_*|^{2^3}\end{aligned}$$

$$\begin{aligned} &\leq M^{\sum_{k=0}^{\ell-1} 2^k} |g_{n-\ell}(x_0) - x_*|^{2^\ell} \leq M^{\sum_{k=0}^{n-1} 2^k} |x_0 - x_*|^{2^n} = M^{2^n - 1} |x_0 - x_*|^{2^n} \\ &\leq (M\delta)^{2^n - 1} \delta \leq \lambda^{2^n - 1}, \end{aligned}$$

mit $\lambda := M\delta < 1$. Das zeigt die Konvergenz in quadratischer Ordnung gegen x_* für die Folge $(x_n)_{n \in \mathbb{N}}$ des NEWTON-Verfahrens. \square

Jetzt lässt sich auch die Stellenverdopplung beim NEWTON-Verfahren verstehen. Die Genauigkeit der Schätzung x_n drückt sich durch die Anzahl s_n gültiger Stellen aus, oder anders gesehen, durch die erste Stelle $s_n + 1$, an der die Dezimalentwicklungen von x_n und x_* verschieden sind. Die Dezimalentwicklung von $|x_n - x_*|$ hat daher die Form $\sum_{k=s_n+1}^{\infty} a_k 10^{-k}$, mit $9 \geq a_{s_n+1} \geq 1$. Das lässt sich abschätzen:

$$10^{-s_n-1} \leq \sum_{k=s_n+1}^{\infty} a_k 10^{-k} \leq 9 \cdot 10^{-s_n} \sum_{k=s_n+1}^{\infty} 10^{-k-s_n} = 9 \cdot 10^{-s_n} \sum_{\ell=1}^{\infty} 10^{-\ell} = 10^{-s_n}.$$

Wir wenden den Logarithmus \log zur Basis 10 auf diese Ungleichungen an. Wegen $\log = \frac{1}{\ln(10)} \ln$ ist er streng monoton wachsend und respektiert daher die Ungleichungen (vergl. Satz 11.4.1 und Beispiel 11.4.3). Wir erhalten $-s_n - 1 \leq \log(|x_n - x_*|) \leq -s_n$, oder $s_n + 1 \geq |\log(|x_n - x_*|)| = -\log(|x_n - x_*|) \geq s_n$. Die Genauigkeit s_n von x_n ist demnach durch den ganzzahligen Anteil von $|\log(|x_n - x_*|)|$ gegeben. Wir machen also nur einen kleinen Fehler, wenn wir diesen Ausdruck zur Abschätzung der gültigen Stellen verwenden. Wegen $\log(|x_n - x_*|) \leq (2^n - 1) \log(\lambda)$ gilt $s_n \approx |\log(|x_n - x_*|)| \geq (2^n - 1) |\log(\lambda)|$. x_n liefert daher etwa $s_n \approx \lceil (2^n - 1) |\log(\lambda)| \rceil$ gültige Stellen und x_{n+1} dann $s_{n+1} \approx \lceil (2 \cdot (2^n - 1) + 1) |\log(\lambda)| \rceil \geq \lceil 2 \cdot (2^n - 1) |\log(\lambda)| \rceil \geq 2 \lceil (2^n - 1) |\log(\lambda)| \rceil \approx 2s_n$.

Testen wir das an dem auf Seite 331 vorgestellten Beispiel $f(x) = e^x + x$ auf dem Intervall $[-1.2, 0]$, mit $\delta = 0.5$. Hier sind $x \mapsto \frac{f''(x)}{f'(x)^2} = \frac{e^x}{(e^x + 1)^2}$ und $x \mapsto f'(x) = e^x + 1$ monoton wachsend, so dass $M = f'(0) \frac{f''(0)}{f'(0)^2} = 0.25$ und damit $\lambda = 0.125$ gewählt werden kann. Damit ergeben sich $s_1 \approx \lceil |\log(0.125)| \rceil = 0$, $s_2 \approx 2$, $s_3 \approx 6$, $s_4 \approx 13$ gültige Nachkommastellen, was mit dem Beispiel ganz gut übereinstimmt. Diese Zahlenreihe darf man aber nur als Demonstration des qualitativen Verhaltens der Iteration verstehen. In der Praxis wird man das Intervall $U_\delta = [x_* - \delta, x_* + \delta]$, auf dem alle Abschätzungen basieren, nur ungefähr kennen. Wenn δ klein sein muß, lässt es sich nicht angeben, da x_* ja nur in einer ersten Schätzung x_0 bekannt ist. Wenn man also keine weitere Informationen hat, wird man versuchen, die erste Schätzung so gut wie eben möglich zu machen und hoffen, daß man sich damit bereits im Einzugsbereich von U_δ befindet.

Unter zusätzlichen Voraussetzungen an f'' , die in der Praxis oft erfüllt sind, erhält man monoton fallende, oder monoton wachsende Approximationsfolgen $(x_n)_{n \in \mathbb{N}}$.

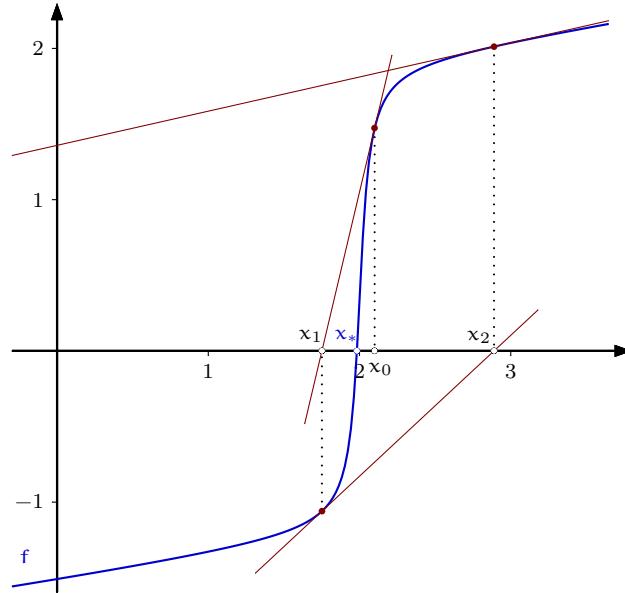
11.7.2 Beispiel Wir suchen die Nullstelle x_* von $f(x) := \arctan(x^3 + 2x^2 - 16) + \frac{x}{6}$.

Aus $f(1.8)f(2.1) \approx -1.481696$ folgt $x_* \in (1.8, 2.1)$. Starten wir mit $x_0 = 2.1$, dann verraten schon die ersten Schritte, daß wir nicht nah genug bei x_* begonnen haben:

$$\begin{aligned}x_1 &= 1.75134961199951, \\x_2 &= 2.88939505910958, \\x_3 &= -6.01406091623807.\end{aligned}$$

Für $x_0 = 2$ dagegen erhalten wir

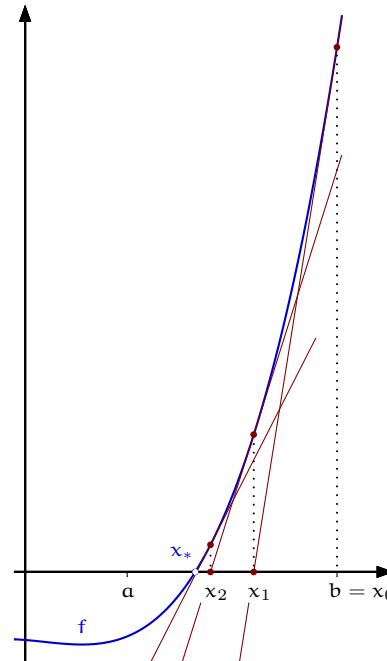
$$\begin{aligned}x_1 &= 1.983471074380165, \\x_2 &= 1.982732722205831, \\x_3 &= 1.982729286832358, \\x_4 &= 1.982729286756874, \\x_5 &= 1.982729286756874, \\f(x_5) &\approx -7.21644 \cdot 10^{-15}.\end{aligned}$$



Die Situation wird besser, wenn zusätzlich zu den Bedingungen aus Satz 11.7.1 noch bekannt ist, daß f'' auf dem Intervall $[a, b]$ keinen Vorzeichenwechsel aufweist (im Beispiel ist $f''(x) > 0$ für $x < 2.0009988$). Untersuchen wir zunächst den Fall $f'(x) > 0$ und $f''(x) > 0$ für alle x aus $[a, b]$. Dann ist f eine streng monoton wachsende Linkskurve. Daher gilt $f(b) > 0$. Wir verwenden die Notation aus dem Beweis zu Satz 11.7.1 und wählen $x_0 = b$. Dann gilt für die erste Näherung $x_1: x_1 = g(x_0) = x_0 - \frac{f(x_0)}{f'(x_0)} < x_0$. Nach dem Mittelwertsatz gibt es ein $\xi \in (x_1, x_0)$ mit der Eigenschaft

$$\begin{aligned}x_1 - x_* &= g(x_0) - g(x_*) = g'(\xi)(x_0 - x_*) \\&= \frac{f'(\xi)f''(\xi)}{f'(\xi)^2}(x_0 - x_*) \geq 0.\end{aligned}$$

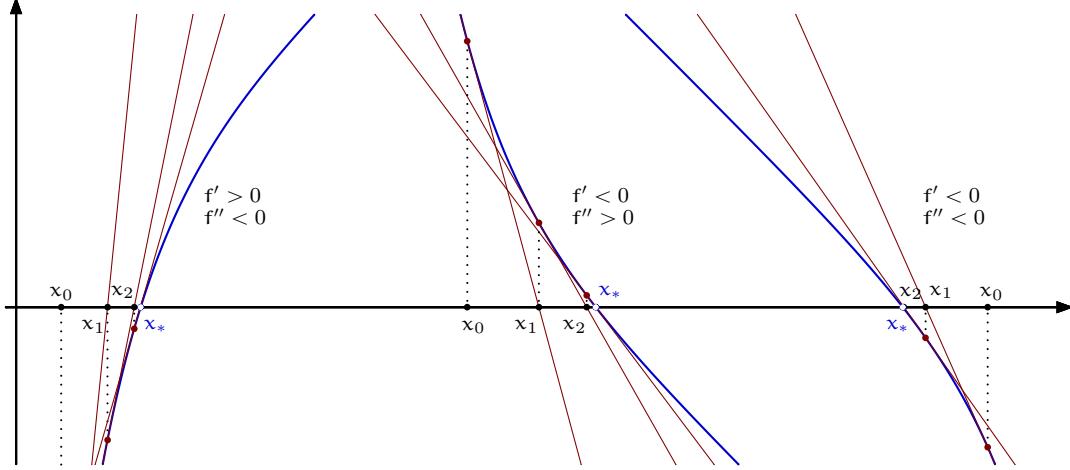
Damit wissen wir $x_0 > x_1 \geq x_*$. Falls zufällig $x_1 = x_*$ gelten sollte, sind wir fertig. Andernfalls haben wir für x_1 und x_* dieselbe Ausgangslage, wie für x_0 und x_* , so daß wir für $x_2 = g(x_1)$ die Abschätzung $x_0 > x_1 > x_2 \geq x_*$



erhalten. Es ist jetzt klar, daß das NEWTON-Verfahren eine monoton fallende Folge $(x_n)_{n \in \mathbb{N}}$ von Schätzwerten mit x_* als einer unteren Schranke liefert. Nach dem Satz von der monotonen Konvergenz 10.1.28 ist die Folge konvergent und hat einen Grenzwert $\bar{x} \geq x_*$. Da g stetig ist, folgt $g(\bar{x}) = \lim_{n \rightarrow \infty} g(x_n) = \lim_{n \rightarrow \infty} x_{n+1} = \bar{x}$. Das bedeutet $f(\bar{x}) = 0$, und da auf $[a, b]$ nur eine Nullstelle von f existiert, muß $x_* = \bar{x}$ gelten. Das Ergebnis unserer Analyse ist eine monoton fallende Folge $(x_n)_{n \in \mathbb{N}}$ von Schätzwerten x_n , die gegen die Nullstelle x_* von f konvergiert,

solange x_0 oberhalb von x_* in dem Bereich gewählt wird, in dem $f'(x) > 0$ und $f''(x) > 0$ gilt. Ist dieser Bereich groß, was durchaus oft vorkommt, dann muß der erste Schätzwert nicht besonders gut gewählt werden.

Drei weitere Ausgangslagen sind günstig: Für alle $x \in [a, b]$ gilt entweder $f'(x) > 0, f''(x) < 0$, oder $f'(x) < 0, f''(x) > 0$, oder $f'(x) < 0, f''(x) < 0$. Im ersten und zweiten Fall liefert $x_0 = a$ eine monoton wachsende Folge $(x_n)_{n \in \mathbb{N}}$ und im dritten ergibt $x_0 = b$ eine monoton fallende Folge von Schätzwerten, die gegen die jeweilige Nullstelle konvergiert.



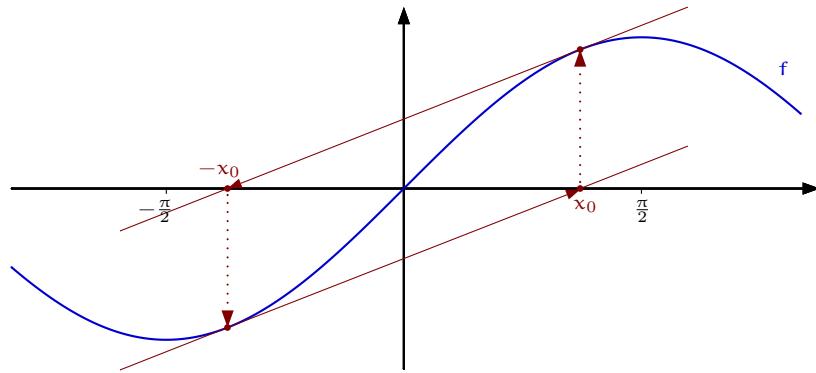
Wir überlegen uns exemplarisch noch den ersten dieser Fälle, denn die anderen sind völlig analog zu behandeln: Für $x_0 = a < x_*$ ist $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} > x_0$, denn $f(x_0) < 0$ und $f'(x_0) > 0$. Außerdem folgt aus dem Mittelwertsatz, daß x_1 nicht jenseits der Nullstelle liegt: $x_1 - x_* = g(x_0) - g(x_*) = \frac{f(\xi)f''(\xi)}{f'(\xi)^2} (\xi)(x_0 - x_*) < 0$, denn $\frac{f(\xi)f''(\xi)}{f'(\xi)^2} \geq 0$ und $x_0 - x_* < 0$. Daher haben wir die Abschätzung $x_0 < x_1 \leq x_*$, die als Ausgangssituation der nächsten Iteration dienen kann: Ist $x_1 = x_*$, so sind wir fertig, andernfalls wiederholen wir die Überlegung mit $x_1 < x_*$ und erhalten $x_0 < x_1 < x_2 \leq x_*$ etc. Für die monoton wachsende Folge $(x_n)_{n \in \mathbb{N}}$ ergibt sich, wie oben vorgeführt, $x_* = \lim_{n \rightarrow \infty} x_n$.

11.7.3 Satz f sei auf dem Intervall $[a, b]$ zweimal stetig differenzierbar und habe folgende weiteren Eigenschaften: $f(a)f(b) < 0$ und f' , sowie f'' haben auf $[a, b]$ keine Nullstellen. Dann gibt es genau eine Nullstelle $x_* \in (a, b)$ von f . Darüber hinaus lässt sich über das NEWTON-Verfahren eine monotone Folge $(x_n)_{n \in \mathbb{N}}$ von Schätzwerten für x_* gemäß folgender Regeln bilden:

- i) $f'(x) > 0, f''(x) > 0$ für alle $x \in [a, b]$: $x_0 = b$ und $x_n \searrow x_*$.
- ii) $f'(x) > 0, f''(x) < 0$ für alle $x \in [a, b]$: $x_0 = a$ und $x_n \nearrow x_*$.
- iii) $f'(x) < 0, f''(x) > 0$ für alle $x \in [a, b]$: $x_0 = a$ und $x_n \nearrow x_*$.
- iv) $f'(x) < 0, f''(x) < 0$ für alle $x \in [a, b]$: $x_0 = b$ und $x_n \searrow x_*$.

11.7.4 A Wenn $f''(x_*) = 0$ gilt, lässt sich die Monotonie des NEWTON-Verfahrens nicht mehr beweisen (man würde allerdings in einer solchen Situation darüber nachdenken, ob man das Verfahren nicht auf f'' anwendet). Die folgende Skizze zeigt, was passieren könnte, wenn x_0

unglücklich gewählt wird. Bestimmen Sie dieses x_0 für $f(x) := \sin(x)$. Was passiert, wenn x_0 ein wenig größer oder kleiner ausfällt?



11.7.5 A Wenden Sie das NEWTON-Verfahren auf die Funktion $f(x) := x^2 - a$ mit positivem a an. Welche Iteration erhalten Sie?

11.8 Polynome

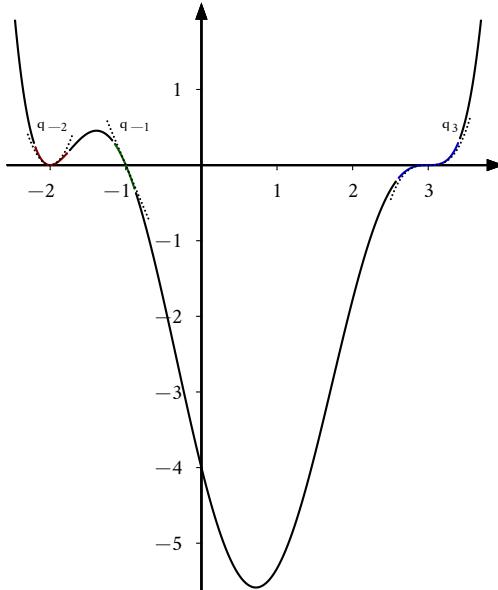
11.8.1 Definition Ein Polynom ist eine Funktion $p: \mathbb{C} \rightarrow \mathbb{C}$ oder $p: \mathbb{R} \rightarrow \mathbb{R}$ mit der Rechenvorschrift

$$p(z) := \sum_{k=0}^n a_k z^k. \quad (11.103)$$

Im ersten Fall sprechen wir von einem komplexen Polynom (oder nur von einem Polynom), im zweiten von einem reellen Polynom. Die Zahlen $a_k \in \mathbb{C}$ bzw. $a_k \in \mathbb{R}$ sind die Koeffizienten von p . a_n wird als Leitkoeffizient bezeichnet und a_0 als Absolutglied. n ist der Grad von p . Wir bezeichnen ihn durch $\text{grad}(p)$. Für das Nullpolynom (also das Polynom mit der Funktionsgleichung $p(z) = 0$ für alle $z \in \mathbb{C}$, oder $z \in \mathbb{R}$) setzt man den Grad durch $-\infty$ fest. p heißt normiert, falls $a_n = 1$ gilt. Wir nennen p gerade (ungerade), falls in $p(z)$ nur gerade (ungerade) Potenzen auftreten, falls also die Koeffizienten mit ungeradem (geradem) Index verschwinden. Die Bausteine $z \mapsto a_k z^k$ von p werden Monome genannt.

Das Polynom $p(z) := 4.1z^6 - 2z^4 + iz^2 + e$ ist gerade, hat den Grad 6, den Leitkoeffizient 4.1 und das Absolutglied e .

Von besonderem Interesse sind die Nullstellen eines Polynoms. Diese sind einfach abzulesen, sollte p z. B. in der Form $p(z) = 2(z+3)^2(z-1)(z-\pi)(z-2i)^3$ vorliegen. Die Nullstellenmenge von p ist $\{-3, 1, \pi, 2i\}$. Dabei sind die Nullstellen $z = 1$ und $z = \pi$ einfach, die Nullstelle $z = -3$ zweifach und $z = 2i$ dreifach. Die Ausdrücke $(z+3)$, $(z-1)$, $(z-2i)$ etc. heißen Linearfaktoren von p . Jeder Linearfaktor repräsentiert eine Nullstelle. Die Potenz, mit denen sie als Faktor in $p(z)$ auftreten, heißt Vielfachheit der Nullstelle. Im Beispiel hat -3 die Vielfachheit 2, $2i$ die Vielfachheit 3, und die restlichen Nullstellen haben die Vielfachheit 1. Was zweifache, dreifache, etc. Nullstellen vor einfachen auszeichnet, macht man sich am besten an einem reellen Polynom klar, etwa an dem Polynom sechsten Grades $q(x) := \frac{1}{27}(x+2)^2(x+1)(x-3)^3$ (auch wenn q nicht in der Form (11.103) vorliegt, kann man doch durch Addieren der Vielfachheiten den Grad erhalten – durch Ausmultiplizieren ergibt sich nämlich dadurch die höchste auftretende Potenz). -2 ist eine doppelte Nullstelle. Da in der nächsten Umgebung von -2 keine weiter Nullstelle liegt, ändert der Teil $\frac{1}{27}(x+1)(x-3)^3$ sein Vorzeichen dort nicht. Er ändert auch seinen Wert $\frac{125}{27}$, den er bei -2 annimmt nur wenig (denn Polynome sind stetige Funktionen). Daher verhält sich q in einer kleinen Umgebung von -2 wie die Funktion $q_{-2}(x) := \frac{125}{27}(x+2)^2$, d. h. wie eine um zwei Einheiten nach links verschobene (leicht gestreckte) Normalparabel $x \mapsto x^2$. Insbesondere findet an einer doppelten Nullstelle kein Vorzeichenwechsel statt. q hat hier augenscheinlich eine



Extremstelle. Dieselben Betrachtungen stellen wir für die einfache Nullstelle -1 an. Hier verhält sich q lokal wie die Funktion $q_{-1}(x) := -\frac{64}{27}(x+1)$, also im Wesentlichen wie die um eine Einheit nach links geschobene zweite Winkelhalbierende $x \mapsto -x$. Diese hat einen Vorzeichenwechsel von $+$ zu $-$. Schließlich ist das Verhalten von q in der Nähe von 3 wie das von $q_3(x) := \frac{100}{27}(x-3)^3$, also im Wesentlichen wie das einer um drei Einheiten nach rechts verschobenen kubischen Parabel $x \mapsto x^3$. Dort hat q offensichtlich einen Sattelpunkt. Es ist nun klar, wie vierfache, fünffache, etc. Nullstellen zu interpretieren sind.

Wie man sieht erfährt man viel über ein Polynom, wenn es in der Form wie q vorliegt. Wir sagen für ein solches Polynom, daß es *vollständig in Linearfaktoren zerfällt*, oder *vollständig faktorisiert*. Natürlich stellen sich sofort zwei Fragen, nämlich erstens, wie eine solche Faktorisierung zu bestimmen ist und zweitens, ob es immer eine gibt. Für die zweite gibt es zwei Antworten: Für ein reelles Polynom mit reellen Nullstellen ist die Faktorisierung nicht immer möglich – das einfachste Beispiel dafür stellt $r(x) := x^2 + 1$ dar – für ein Polynom mit möglicherweise komplexen Nullstellen aber immer. Das ist eine Folge des *Fundamentalsatzes der Algebra* 11.8.23, der besagt, daß jedes Polynom wenigstens *eine* Nullstelle hat (möglicherweise in \mathbb{C} wahlgemerkt, bei r wäre das z. B. i). Wieso das eine vollständige Faktorisierung zur Folge hat, klären wir, nachdem wir uns mit der *Polynomdivision* vertraut gemacht haben. Die erste Frage, nach der Bestimmung einer Faktorisierung, d. h. nach dem Auffinden von Nullstellen, hat mehrere Antworten. Für $\text{grad}(p) = 1$ handelt es sich bei $p(x) = a_1x + a_0$ um eine Gleichung. Hier sollte die Nullstelle leicht zu finden sein. Für $\text{grad}(p) = 2$, also für $p(x) = ax^2 + bx + c$ (in der vertrauten Schulnotation) bedeutet es, eine *quadratische* Gleichung zu lösen. Mit der sogenannten *Mitternachtsformel* $x_{1/2} = \frac{1}{2a}(-b \pm \sqrt{b^2 - 4ac})$ haben wir dafür eine Lösungsformel (die im Falle $b^2 - 4ac < 0$ richtig interpretiert werden muß). Auch für $\text{grad}(p) = 3$ und $\text{grad}(p) = 4$ gibt es Lösungsformeln. Es ist aber zu erwarten, daß sie nicht so einfach wie die Mitternachtsformel sind. Es hat viele Bemühungen gebraucht, bis endlich bewiesen werden konnte, daß es ab $\text{grad}(p) = 5$ keine Lösungsformeln mehr gibt. Das ist nicht damit zu verwechseln, daß in diesen Fällen keine Lösungen, also Nullstellen mehr vorhanden sind. Es bedeutet, daß keine noch so komplizierte Formel existiert, in die einfach die Koeffizienten von p einzusetzen wären, um alle Lösungen auszurechnen. Ab $\text{grad}(p) = 5$ kann es also ein ernsthaftes Problem darstellen, die Nullstellen eines Polynoms zu finden. Wenn man keine raten kann (wie und wann das gehen könnte behandeln wir gleich und ja, das ist durchaus eine zulässige Methode), muß man sich mit Näherungsmethoden behelfen. Dazu wäre es wünschenswert, wenn man ungefähr weiß, wo die Nullstellen liegen. Da ein Polynom vom Grad n höchstens n verschiedene Nullstellen haben kann – das werden wir in Kürze zeigen – müssen diese in einem begrenzten Gebiet liegen. Für ein reelles Polynom würde man darüber hinaus noch gerne wissen, wie viele reelle Nullstellen es hat.

Bevor wir uns den skizzierten Themen zuwenden können, brauchen wir einen Satz, der die Wohldefiniertheit einiger der bisher verwendeten Begriffe sicherstellt.

11.8.2 Satz Für jedes Polynom p und jedes fest gewählte $z_0 \in \mathbb{K}$ ($= \mathbb{R}$, oder $= \mathbb{C}$) gibt es eine Darstellung $p(z) = \sum_{k=0}^n b_k(z - z_0)^k$, $z \in \mathbb{K}$. Wir sagen, p ist um z_0 entwickelt. Dabei sind die Entwicklungskoeffizienten b_k eindeutig und durch $b_k = \frac{1}{k!} p^{(k)}(z_0)$ gegeben. $p^{(k)}$ ist dabei die (formale) k -te Ableitung von p .

Beweis. Wir gehen von $p(z) = \sum_{k=0}^n b_k(z - z_0)^k$ aus und nehmen an, p ließe sich auch mit einem anderen Satz b'_0, b'_1, \dots, b'_m berechnen: $p(z) = \sum_{k=0}^m b'_k(z - z_0)^k$. Ist $m \neq n$, dann ergänzen wir in der Summe mit der kleineren oberen Grenze die fehlenden Koeffizienten durch 0. Wir können daher o. B. d. A. von $m = n$ ausgehen. Die ersten beiden Koeffizienten müssen wegen $p(z_0) = b_0 = b'_0$ schon mal gleich sein. $k_0 \geq 1$ sei also der erste Index, für den $b_{k_0} \neq b'_{k_0}$ gilt. Dann folgt $0 = \sum_{k=k_0}^n (b_k - b'_k)(z - z_0)^k = (z - z_0)^{k_0} \sum_{k=k_0}^n (b_k - b'_k)(z - z_0)^{k-k_0}$ für alle $z \in \mathbb{K}$, insbesondere für alle $z \neq z_0$. Für diese z muß $q(z) := \sum_{k=k_0}^n (b_k - b'_k)(z - z_0)^{k-k_0} = 0$ gelten. q ist ein Polynom und daher stetig. Das bedeutet $q(0) = \lim_{z \rightarrow z_0} q(z) = 0 = b_{k_0} - b'_{k_0}$, im Widerspruch dazu, daß $b_{k_0} \neq b'_{k_0}$ gelten sollte. Also gibt es keinen Index, für den die Koeffizienten verschieden sind.

Die Entwicklung um z_0 : Für ein Polynom $p(z) = \sum_{k=0}^n a_k z^k$ sei die Ableitung p' durch die gewohnten Rechenregeln definiert (die wir für reelle Polynome kennen): $p'(z) := \sum_{k=1}^n k a_k z^{k-1}, \dots, p^{(\ell)}(z) = \sum_{k=\ell}^n k(k-1)\cdots(k+1-\ell) a_k z^{k-\ell} = \sum_{k=\ell}^n \frac{k!}{(k-\ell)!} a_k z^{k-\ell}$. Um die Koeffizienten b_k der Entwicklung $p(z) = \sum_{k=0}^n b_k(z - z_0)^k$ zu bestimmen, bilden wir nach denselben Rechenregeln $p^{(\ell)}(z) = \sum_{k=\ell}^n \frac{k!}{(k-\ell)!} b_k (z - z_0)^{k-\ell}$. Das bedeutet $p^{(\ell)}(z_0) = \frac{\ell!}{0!} b_\ell = \ell! b_\ell$. Das zeigt, daß in einer Entwicklung um z_0 die Koeffizienten durch (11.8.2) festgelegt sind. Bleibt zu zeigen, daß eine Entwicklung um z_0 immer möglich ist. Dafür gibt es mehrere Möglichkeiten. Naheliegend ist es, die Koeffizienten b_ℓ gemäß (11.8.2) in $\sum_{\ell=0}^n b_\ell(z - z_0)^\ell$ einzusetzen und nachzurechnen, daß dieser Ausdruck mit $p(z) = \sum_{k=0}^n a_k z^k$ übereinstimmt. Der Ansatz $b_\ell = \frac{1}{\ell!} p^{(\ell)}(z_0) = \sum_{k=\ell}^n \frac{k!}{\ell!(k-\ell)!} a_k z_0^{k-\ell} = \sum_{k=\ell}^n a_k \binom{k}{\ell} z_0^{k-\ell}$ sollte der Richtige sein. Die folgende Rechnung zeigt das:

$$\begin{aligned} \sum_{\ell=0}^n b_\ell(z - z_0)^\ell &= \sum_{\ell=0}^n \sum_{k=\ell}^n a_k \binom{k}{\ell} z_0^{k-\ell} (z - z_0)^\ell \stackrel{*}{=} \sum_{k=0}^n a_k \sum_{\ell=0}^k \binom{k}{\ell} z_0^{k-\ell} (z - z_0)^\ell \\ &\stackrel{(1.32)}{=} \sum_{k=0}^n a_k (z_0 + z - z_0)^k = \sum_{k=0}^n a_k z^k = p(z) \end{aligned}$$

Eine andere Möglichkeit, das einzusehen, kommt aus der linearen Algebra: Die Menge $\{q_0, q_1, \dots, q_n\}$, mit den Polynomen $q_k(z) := (z - z_0)^k$, ist eine Basis für den $n + 1$ -dimensionalen Vektorraum \mathcal{P}_n aller Polynome von höchstens n -tem Grad. Die Entwicklung von p um z_0 ist also einfach die Basisdarstellung von p in dieser Basis. \square

Wir haben im Schritt *) die Regel

$$\sum_{\ell=0}^n \sum_{k=\ell}^n a_{\ell k} = \sum_{k=0}^n \sum_{\ell=0}^k a_{\ell k} \quad (11.104)$$

verwendet, um die Reihenfolge der Summen zu vertauschen. Was dahintersteckt, kann man sich wie folgt veranschaulichen. Zunächst ist klar, daß die beiden Summen nicht so ohne Weiteres ihre Reihenfolge ändern können, da die zweite den Summationsindex ℓ der ersten in ihrer unteren Grenze enthält. Es handelt sich um die Summierung einer oberen Dreiecksmatrix. Dabei kann man folgendermaßen vorgehen: Man addiert zunächst die Zeileneinträge und anschließend die Ergebnisse, oder aber zuerst die Spalteneinträge und dann deren Resultate. Beides

muß zum gleichen Ergebnis führen. Der Ausdruck $\sum_{\ell=0}^n \sum_{k=\ell}^n a_{\ell k}$ bedeutet, daß die Summe $\sum_{\ell=0}^n$ der Zeilensummen $\sum_{k=\ell}^n a_{\ell k}$ gebildet wird, der Ausdruck $\sum_{k=0}^n \sum_{\ell=0}^k a_{\ell k}$ dagegen ist die Summe $\sum_{k=0}^n$ der Spaltensummen $\sum_{\ell=0}^k a_{\ell k}$:

	0	1	2	3	...	k	...	n	
0	a_{00}	a_{01}	a_{02}	a_{03}	...	a_{0k}	...	a_{0n}	$\sum_{k=0}^n a_{0k}$
1		a_{11}	a_{12}	a_{13}	...	a_{1k}	...	a_{1n}	$\sum_{k=1}^n a_{1k}$
\vdots		\ddots	\vdots	\vdots		\vdots		\vdots	\vdots
ℓ			$a_{\ell\ell}$...	$a_{\ell k}$...	$a_{\ell n}$		$\sum_{k=\ell}^n a_{\ell k}$
\vdots			\ddots	\vdots	\vdots		\vdots	\vdots	\vdots
\vdots				a_{kk}			\vdots	\vdots	\vdots
\vdots					\ddots	\vdots	\vdots	\vdots	\vdots
n							a_{nn}		$\sum_{k=n}^n a_{nk}$
	$\sum_{\ell=0}^0 a_{\ell 0}$	$\sum_{\ell=0}^1 a_{\ell 1}$	$\sum_{\ell=0}^2 a_{\ell 2}$	$\sum_{\ell=0}^k a_{\ell k}$...	$\sum_{\ell=0}^n a_{\ell n}$	$\sum_{k=0}^n \sum_{\ell=0}^k a_{\ell k} = \sum_{\ell=0}^n \sum_{k=\ell}^n a_{\ell k}$

Eigentlich zeigt sich erst jetzt, daß der Grad eines Polynoms p wohldefiniert ist, denn es hat genau einen Koeffizienten mit größtem Index, durch den $\text{grad}(p)$ festgelegt wird. Dieser Satz hat aber auch noch eine praktische Konsequenz. Mitunter läßt sich ein Polynom $p(z) = \sum_{k=0}^n a_k z^k$ auf eine Weise bestimmen, die zu einer scheinbar unterschiedlichen Darstellung $p(z) = \sum_{k=0}^n \tilde{a}_k z^k$ führt. Der Satz garantiert, daß ein sogenannter *Koeffizientenvergleich* durchgeführt werden darf, d. h., daß $a_k = \tilde{a}_k$ für $k = 0, \dots, n$ gilt. Auf diese Weise erfährt man manchmal etwas Neues über die Koeffizienten, mitunter lernt man sie so überhaupt erst kennen. Als Beispiel verschaffen wie uns eine Formel für die Koeffizienten des Produkts zweier Polynome.

11.8.3 Satz (CAUCHY-Produkt) Das Produkt zweier Summen $\sum_{k=0}^n s_k$ und $\sum_{\ell=0}^m t_\ell$ ist durch

$$\left(\sum_{k=0}^n s_k \right) \left(\sum_{\ell=0}^m t_\ell \right) = \sum_{k=0}^n \sum_{\ell=0}^m s_k t_\ell = \sum_{r=0}^{n+m} \sum_{k=\max\{0, m-r\}}^{\min\{n, r\}} s_k t_{r-k} \quad (11.105)$$

gegeben. Für das Produkt zweier Polynome $p(z) = \sum_{k=0}^n a_k z^k$ und $q(z) = \sum_{k=0}^m b_k z^k$ bedeutet das

$$p(z)q(z) = \sum_{k=0}^n \sum_{\ell=0}^m a_k b_\ell z^{k+\ell} = \sum_{r=0}^{n+m} \left(\sum_{k=\max\{0, m-r\}}^{\min\{n, r\}} a_k b_{r-k} \right) z^r, \quad (11.106)$$

d. h., die Koeffizienten von pq sind für $r = 0, 1, \dots, m+n$ durch

$$\sum_{k=\max\{0, m-r\}}^{\min\{n, r\}} a_k b_{r-k} \quad (11.107)$$

bestimmt.

Meistens findet man (11.107) in der Form

$$p(z)q(z) = \sum_{r=0}^{n+m} \sum_{k=0}^r a_k b_{r-k} z^r. \quad (11.108)$$

Dabei wird vereinbart, daß die Koeffizienten a_k oder b_{r-k} in dieser Formel, deren Indizes nicht in $\{0, \dots, n\}$ bzw. $\{0, \dots, m\}$ liegen, durch 0 zu ersetzen sind.

Beweis. Wir machen uns die Produktformel an dem Beispiel $n = 4$ und $m = 5$, also an

$$(s_0 + s_1 + s_2 + s_3 + s_4)(t_0 + t_1 + t_2 + t_3 + t_4 + t_5)$$

klar. Die Produkte, die beim Ausmultiplizieren auftreten, lassen sich systematisch in einer 4×5 -Matrix anordnen. Die Idee besteht jetzt darin, die Diagonalen zu summieren. Hier stehen Produkte, deren Indizes eine konstante Summe r haben. Für $r = 3$ ergibt das $s_0 t_3 + s_1 t_2 + s_2 t_1 + s_3 t_0 = \sum_{k=0}^3 s_k t_{3-k}$ (die vierte Diagonale von rechts oben nach links unten summiert) und für $r = 4$ ebenfalls $\sum_{k=0}^4 s_k t_{4-k}$. Man könnte also $\sum_{k=0}^r s_k t_{r-k}$ für eine beliebige Diagonale mit Indexsumme r vermuten. Allerdings stimmt das schon für $r = 5$ nicht mehr, denn hier würde der Summand $s_5 t_0$ auftreten, den es gar nicht gibt. Der letzte gültige ist $s_4 t_0 = s_n t_0$. Wir müssen in unserer Summenformel also dafür sorgen, daß k bis r wächst, falls $r \leq n$ gilt und nur bis n , falls $r > n$ eintritt. Dazu ersetzen wir die obere Grenze in der Summe einfach durch das Minimum der beiden Zahlen n und r : $\sum_{k=0}^{\min\{n,r\}} s_k t_{r-k}$. Testen wir diese Formel für $r = 7$. Wir erhalten $s_0 t_7 + s_1 t_6 + s_2 t_5 + \dots + s_4 t_3$. Hier stimmen die ersten beiden Summanden nicht, da es t_7 und t_6 nicht gibt. Wir müssen dafür sorgen, daß $r - k \leq m$, also $k \geq r - m$ gilt. Außerdem soll k bei 0 starten, falls $r - m$ negativ sein sollte (wie für $r = 4$ und $m = 5$). Das erreichen wir durch die untere Grenze $\max\{0, m - r\}$ in der Summenformel. Damit sind wir bei (11.105) angekommen.

$r = 0$	1	2	3	4	5	6	7	8	9
$s_0 t_0$	$s_0 t_1$	$s_0 t_2$	$s_0 t_3$	$s_0 t_4$	$s_0 t_5$	$s_0 t_6$	$s_0 t_7$	$s_0 t_8$	$s_0 t_9$
$s_1 t_0$	$s_1 t_1$	$s_1 t_2$	$s_1 t_3$	$s_1 t_4$	$s_1 t_5$	$s_1 t_6$	$s_1 t_7$	$s_1 t_8$	$s_2 t_9$
$s_2 t_0$	$s_2 t_1$	$s_2 t_2$	$s_2 t_3$	$s_2 t_4$	$s_2 t_5$	$s_2 t_6$	$s_2 t_7$	$s_2 t_8$	$s_2 t_9$
$s_3 t_0$	$s_3 t_1$	$s_3 t_2$	$s_3 t_3$	$s_3 t_4$	$s_3 t_5$	$s_3 t_6$	$s_3 t_7$	$s_3 t_8$	$s_3 t_9$
$s_4 t_0$	$s_4 t_1$	$s_4 t_2$	$s_4 t_3$	$s_4 t_4$	$s_4 t_5$	$s_4 t_6$	$s_4 t_7$	$s_4 t_8$	$s_4 t_9$
$s_5 t_0$	$s_5 t_1$	$s_5 t_2$	$s_5 t_3$	$s_5 t_4$	$s_5 t_5$	$s_5 t_6$	$s_5 t_7$	$s_5 t_8$	$s_5 t_9$
$s_6 t_0$	$s_6 t_1$	$s_6 t_2$	$s_6 t_3$	$s_6 t_4$	$s_6 t_5$	$s_6 t_6$	$s_6 t_7$	$s_6 t_8$	$s_6 t_9$

(11.106) ist jetzt eine simple Folgerung aus diesem Ergebnis: Wenn man $s_k := a_k z^k$ und $t_\ell := b_\ell z^\ell$ wählt, summiert man über die Produkte $a_k b_{r-k} z^k z^{r-k} = a_k b_{r-k} z^r$, aus denen man z^r ausklammern kann. Die verbleibende Summe (11.106) ist der Koeffizient von z^r . \square

11.8.4 Beispiel $p(z) := 6z^4 + 2z^3 - 4z^2 + z - 2$, $q(z) := 3z^5 - 2z^4 + 0z^3 + 3z^2 - z + 1$. Dann sind die Koeffizienten c_k von pq durch $c_9 = 18$, $c_8 = -12 + 6 = -6$, $c_7 = -4 - 12 = -16$, $c_6 = 18 + 8 + 3 = 29$, $c_5 = -6 + 6 - 2 - 6 = -8$, $c_4 = 6 - 2 - 12 + 4 = -4$, $c_3 = 2 + 4 + 3 = 9$, $c_2 = -4 - 1 - 6 = -11$, $c_1 = 1 + 2 = 3$ und $c_0 = -2$ gegeben: $p(z)q(z) = 18z^9 - 6z^8 - 16z^7 + 29z^6 - 8z^5 - 4z^4 - 11z^2 + 3z - 2$.

Natürlich haben wir dabei nicht in die Formel (11.107) eingesetzt. Das wäre mühsam. Sie kann allenfalls beim Programmieren von Nutzen sein. Aber sie liefert die Idee, gemäß der man solche Polynome wie p und q systematisch multipliziert. Nämlich indem man, von der höchsten Potenz von pq beginnend, diejenigen Koeffizienten a_k von p und b_{r-k} von q miteinander multipliziert, deren zugehörigen Potenzen k und $r-k$ die konstante Summe r haben und als Summe den Koeffizienten von z^r für pq ergeben. In diesem Beispiel ist 9 der höchste Koeffizient von pq , also $c_9 = 6 \cdot 3$. Der zweite entsteht, indem wir die Koeffizienten von p in der Reihenfolge absteigender Potenzen (dabei vergisst man keine) mit denen von q multiplizieren, die als Summe der zugehörigen Potenzen 8 ergeben: $c_8 = 6 \cdot (-2) + 2 \cdot 3 = -6$. Die Potenz 7 kann als $4+3, 3+4$ und $2+5$ entstehen. Daher ist $c_7 = 6 \cdot 0 + 2 \cdot (-2) + (-4) \cdot 3 = -16$. Auf diese Weise fährt man fort, bis man bei $c_0 = -2 \cdot 1 = -2$ angekommen ist.

Eine weitere Folgerung aus Satz 11.8.2 betrifft mehrfache Nullstellen eines Polynoms.

11.8.5 Lemma Ein Polynom p hat genau dann eine mehrfache Nullstelle z_0 , wenn $p(z_0) = 0$ und $p'(z_0) = 0$ gilt.

Beweis. Falls p in z_0 wenigstens eine doppelte Nullstelle hat, gilt $p(z) = q(z)(z - z_0)^2$. Also ist $p'(z) = q'(z)(z - z_0)^2 + 2q(z)(z - z_0)$. Das zeigt $p'(z_0) = 0$. Für die Umkehrung entwickeln wir p um z_0 : $p(z) = \sum_{k=0}^n b_k(z - z_0)^k$. Aus $p(z_0) = 0$ folgt $b_0 = 0$ und $p'(z_0) = 0$ hat $b_1 = 0$ zur Folge. Das heißt $p(z) = (z - z_0)^2 \sum_{k=2}^n b_k(z - z_0)^{k-2}$, also mindestens eine doppelte Nullstelle für p .

Eine andere Möglichkeit das zu beweisen, macht keinen Gebrauch von Satz 11.8.2, verlangt aber einen kleinen Vorgriff auf Korollar 11.8.11: Aus $p(z_0) = 0$ folgt, daß die Polynomdivision von $p(z)$ mit $(z - z_0)$ ohne Rest aufgeht. Es gibt daher ein Polynom p_1 mit der Eigenschaft $p(z) = p_1(z)(z - z_0)$. Dann ist $p'(z) = p'_1(z)(z - z_0) + p_1(z)$. Aus $0 = p'(z_0) = p_1(z_0)$ folgt, nach demselben Korollar, die Existenz eines weiteren Polynoms p_2 , so daß $p_1(z) = p_2(z)(z - z_0)$ gilt. Daraus folgt unmittelbar $p(z) = p_2(z)(z - z_0)^2$. \square

11.8.6 Polynome mit rationalen Koeffizienten Für ein solches Polynom $r(x) = \sum_{k=0}^n a_k x^k$, $a_k \in \mathbb{Q}$, gibt es eine Methode, um zu entscheiden, ob es rationale Nullstellen hat oder nicht und um diese gegebenenfalls auch alle auszurechnen. Da es sich um Nullstellen von r handelt, also um Lösungen der Gleichung $\sum_{k=0}^n a_k x^k = 0$, können wir o. B. d. A. davon ausgehen, daß alle Koeffizienten natürliche Zahlen sind. Ist das nicht der Fall, so kann man r mit dem Hauptnenner aller a_k multiplizieren (was die Nullstellen nicht beeinflußt) und so diese Situation herstellen.

11.8.7 Satz Wenn die Gleichung $a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 = 0$ mit ganzzahligen Koeffizienten a_k eine rationale Lösung $\frac{p}{q}$ hat (natürlich mit $\text{ggT}(p, q) = 1$), dann muß p ein Teiler von a_0 und q ein Teiler von a_n sein. Ist $a_n = 1$, so sind rationale Lösungen ganzzahlig.

Beweis. Wir können $a_0 \neq 0$ annehmen, denn andernfalls ist die Aussage $p \mid a_0$ trivial. Auch $a_n \neq 0$ schränkt die Allgemeinheit nicht ein.

Aus $a_n \frac{p^n}{q^n} + a_{n-1} \frac{p^{n-1}}{q^{n-1}} + \cdots + a_2 \frac{p^2}{q^2} + a_1 \frac{p}{q} + a_0 = 0$ folgt, nachdem man diese Gleichung mit q^n multipliziert und nach $a_0 q^n$ umgestellt hat

$$\begin{aligned}-a_0 q^n &= a_n p^n + a_{n-1} p^{n-1} q + a_{n-2} p^{n-2} q^2 + \cdots + a_2 p^2 q^{n-2} + a_1 p q^{n-1} \\ &= (a_n p^{n-1} + a_{n-1} p^{n-2} q + a_{n-2} p^{n-3} q^2 + \cdots + a_2 p q^{n-2} + a_1 q^{n-1}) \cdot p.\end{aligned}$$

Da der Ausdruck in der Klammer eine natürliche Zahl darstellt, folgt $p \mid a_0 q^n$. Das hat nach Korollar 3.1.9 bereits $p \mid a_0$ zur Folge, denn p und q sind teilerfremd.

Wir hätten die Gleichung auch nach $-a_n p^n$ auflösen und dann, genau wie oben, auf $q \mid a_n$ schließen können. \square

11.8.8 Beispiel $r(x) := \frac{2}{3}x^3 - \frac{79}{27}x^2 + 3x + \frac{10}{27}$ hat offensichtlich rationale Koeffizienten. Die Gleichung $r(x) = 0$ multiplizieren wir mit 27 und erhalten $2 \cdot 3^2 x^3 - 79 x^2 + 81 x + 2 \cdot 5 = 0$. Für eine rationale Nullstelle $\frac{p}{q}$ muß p in $\{-1, 1, -2, 2, -5, 5\}$ und q in $\{-1, 1, -2, 2, -3, 3, -9, 9\}$ liegen. Als rationale Lösungen kommen also nur die Elemente aus $\{\pm 1, \pm 2, \pm 3, \pm \frac{1}{2}, \pm \frac{5}{2}, \pm \frac{1}{3}, \pm \frac{2}{3}, \pm \frac{5}{3}, \pm \frac{1}{9}, \pm \frac{2}{9}, \pm \frac{5}{9}\}$ in Frage. Sollte durch Einsetzen in $27 \cdot r$ keine dieser 22 Zahlen das Ergebnis 0 liefern, dann wissen wir, daß keine rationale Lösung existiert. Gibt es dagegen solche Lösungen, dann müssen sie in dieser Menge zu finden sein. Nach höchstens 22 Berechnungen wissen wir Bescheid. Natürlich kann man die Tests beenden, sobald drei Lösungen gefunden sind. Fangen wir mit den ganzzahligen Kandidaten an und probieren $x = 1$. Das führt auf $18 - 79 + 81 + 10 = 30$. Für $x = -1$ erhalten wir $-18 - 79 - 81 + 10 = -168$. Versuchen wir $x = \pm 2$: $18 \cdot 8 - 79 \cdot 4 + 81 \cdot 2 + 10 = 0$ und $-18 \cdot 8 - 79 \cdot 4 - 81 \cdot 2 + 10 = -612$. $x = 2$ ist eine erste Lösung. Jetzt müssen wir die restlichen Kandidaten prüfen. Das kann ein erheblicher Rechenaufwand sein, aber nach endlich vielen Rechenschritten herrscht Sicherheit über alle rationalen Lösungen. Wenn das Ergebnis nur wichtig genug ist, wird man diese Arbeit in Kauf nehmen (zumal man sie ja einem Computeralgebra-System überlassen kann). In diesem Beispiel gibt es tatsächlich noch zwei weitere rationale Lösungen.

Ist der Leitkoeffizient 1, so reduziert sich das Verfahren darauf, die Teiler von a_0 zu testen. Alle rationalen Lösungen sind dann sogar ganzzahlig, denn ein Nenner q müßte die Zahl 1 teilen. Nehmen wir $s(x) := x^5 + x^4 - 18x^3 + 14x^2 + 21x + 5$. Mögliche ganzzahlige Lösungen können nur ± 1 oder ± 5 sein. Man rechnet schnell $s(-1) = 18 + 14 - 21 + 5 = 16$ und $s(1) = 2 - 18 + 14 + 21 + 5 = 24$ nach. Bleiben -5 und 5 . Bei der Verteilung der Koeffizienten ist klar, daß nur -5 möglich ist, $+5$ liefert bei den positiven Koeffizienten zu große Zahlen, die durch $-18 \cdot 5^3$ nicht mehr kompensiert werden können. $s(-5) = -4 \cdot 5^2 \cdot 5^2 + 90 \cdot 25 + 14 \cdot 25 - 20 \cdot 5 = 4 \cdot 25 - 20 \cdot 5 = 0$. s hat daher nur eine rationale Nullstelle, nämlich -5 . Ob es weitere, dann irrationale Nullstellen gibt, können wir im Augenblick nicht entscheiden.

11.8.9 Polynomdivision Für eine ganze Zahl p und eine natürliche Zahl q ist immer eine Zerlegung $p = tq + r$, $0 \leq r < q$, mit eindeutig bestimmten Zahlen $t \in \mathbb{Z}$ und $r \in \mathbb{N}_0$ möglich. Wir haben das *Teilen mit Rest* genannt (vergl. 3.1.2). Eine solche Zerlegung ist auch für Polynome p und q möglich, wenn $0 \leq r < q$ durch $\text{grad}(r) < \text{grad}(q)$ ersetzt wird. Dafür müssen wir zunächst klären, wie sich der Grad von Polynomen bei Addition und Multiplikation verhält.

Für zwei Polynome p und q gilt

$$\text{grad}(p \pm q) \leq \max\{\text{grad}(p), \text{grad}(q)\}, \quad (11.109)$$

$$\text{grad}(p \cdot q) = \text{grad}(p) + \text{grad}(q). \quad (11.110)$$

Diese beiden Eigenschaften sind fast offensichtlich. Durch Addition der Polynome $p(z) = a_n z^n + \dots + a_0$ und $q(z) = b_m z^m + \dots + b_0$ kann der Grad des Polynoms $p + q$ nicht größer als der maximale Grad der Ausgangspolynome werden, aber durchaus kleiner. Dafür braucht nur $n = m$ und $a_n = -b_n$ zu gelten. Das Produkt $p \cdot q(z) = p(z)q(z) = a_n b_m z^{n+m} + a_n b_{m-1} z^{n+m-1} + \dots + a_0 b_0$ hat den Grad $n+m = \text{grad}(p)+\text{grad}(q)$. Übrigens ist die uneingeschränkte Gültigkeit der Gleichung (11.110) der Grund dafür, daß man den Grad des Nullpolynoms 0 durch $-\infty$ festgelegt hat. Hätte man dafür irgendeine andere Zahl gewählt, so erhielte man für ein beliebiges Polynom p die Gleichung $\text{grad}(0) = \text{grad}(0 \cdot p) = \text{grad}(0) + \text{grad}(p)$, also $\text{grad}(p) = 0$. Dagegen ergibt $-\infty = \text{grad}(0) = \text{grad}(0 \cdot p) = \text{grad}(0) + \text{grad}(p) = -\infty + \text{grad}(p) = -\infty$ keinen Widerspruch. Durch diese Wahl wird auch richtig wiedergegeben, daß die Menge der Polynome nullteilerfrei ist, daß also aus $pq = 0$ immer $p = 0$, oder $q = 0$ folgt (das ist klar, wie man durch kurzes Nachdenken feststellt): $pq = 0$ bedeutet $\text{grad}(pq) = -\infty = \text{grad}(p)+\text{grad}(q)$. Das geht nur für $\text{grad}(p) = -\infty$, oder $\text{grad}(q) = -\infty$.

11.8.10 Satz Für ein Polynom p und ein Polynom $q \neq 0$ gibt es eindeutig bestimmte Polynome t und r mit den Eigenschaften $\text{grad}(r) < \text{grad}(q)$ und

$$p = tq + r. \quad (11.111)$$

Beweis. Wir zeigen zunächst die Eindeutigkeit. Aus $p = tq + r = t'q + r'$ folgt $(t - t')q = r' - r$. Wäre dieses Polynom nicht das Nullpolynom, dann wäre nach (11.109) $\text{grad}(r' - r) < \text{grad}(q)$, denn das gilt sowohl für r als auch für r' . Auf der linken Seite kann dann $t - t'$ nicht verschwinden, so daß nach (11.110) $\text{grad}(r - r') = \text{grad}((t - t')q) \geq \text{grad}(q)$ folgt. Das ist ein Widerspruch. Damit ist $r = r'$ und $(t - t')q = 0$. Wegen $q \neq 0$ muß $t - t' = 0$ gelten.

Die Existenz der Darstellung (11.111) ließe sich mittels vollständiger Induktion nach $n = \text{grad}(p)$ beweisen. Da der Beweis aber konstruktiv ist, führen wir nur den entscheidenden Schritt vor, den man dann leicht zur vollständigen Induktion ausbauen kann. Zunächst behandeln wir den Sonderfall $\text{grad}(p) < \text{grad}(q)$. Hier ist $t = 0$ und $p = r$.

Jetzt sei $p(z) := a_n z^n + a_{n-1} z^{n-1} + a_{n-2} z^{n-2} + \dots + a_1 z + a_0$ und $q(z) := b_m z^m + b_{m-1} z^{m-1} + b_{m-2} z^{m-2} + \dots + b_1 z + b_0$, sowie $m \leq n$. Die Idee besteht nun einfach darin, $q(z)$ mit einem so gewählten Monom $t_1(z) := c_k z^k$ zu multiplizieren, daß sich in der Differenz $p(z) - q(z) \cdot t_1(z)$ der erster Summand $a_n z^n$ von $p(z)$ heraushebt. Es ist klar, daß dafür $k = n - m$ und $c_k = \frac{a_n}{b_m}$ gewählt werden muß:

$$\begin{aligned} p(z) - q(z) \frac{a_n}{b_m} z^{n-m} &= a_n z^n + a_{n-1} z^{n-1} + a_{n-2} z^{n-2} + \dots + a_1 z + a_0 \\ &\quad - a_n z^n - a_n \frac{b_{m-1}}{b_m} z^{n-1} - a_n \frac{b_{m-2}}{b_m} z^{n-2} - \dots - a_n \frac{b_0}{b_m} z^{n-m} \\ &= a'_{n-1} z^{n-1} + a'_{n-2} z^{n-2} + \dots + a'_{n-m} z^{n-m} \\ &\quad + a_{n-1-m} z^{n-1-m} \dots a_1 z + a_0 =: p_1(z), \end{aligned}$$

mit den neuen Koeffizienten a'_k für das Polynom p_1 . Die genaue Form

$$a'_k = a_k - a_n \frac{b_k}{b_m} \quad (11.112)$$

ist für die folgenden Überlegungen nicht weiter wichtig. Entscheidend ist, daß wir jetzt eine Darstellung $p = t_1 q + p_1$ erhalten haben, mit einem Polynom p_1 , dessen Grad höchstens $\text{grad}(p) - 1$ ist. Sollte bereits $\text{grad}(p_1) < \text{grad}(q)$ gelten, so sind wir fertig. Andernfalls wiederholen wir das Verfahren mit p_1 und einem neuen Monom t_2 , um $p_1 = t_2 q + p_2$ zu erhalten. Das Polynom p_2 erfüllt jetzt $\text{grad}(p_2) < \text{grad}(p_1)$, wobei auch der Fall $\text{grad}(p_2) = -\infty$ eintreten kann, falls $p_2 = 0$ sein sollte. Auf diese Weise fahren wir fort, bis das erste mal ein Polynom p_k mit $\text{grad}(p_k) < \text{grad}(q)$ erscheint. Es ist klar, daß dieser Fall schließlich eintreten muß, weil der Grad sich in jedem Schritt um wenigstens eins vermindert. Dieses Polynom ist r . Wir haben damit die gewünschte Form erhalten:

$$\begin{aligned} p &= t_1 q + p_1 = t_1 q + t_2 q + p_2 = \dots = t_1 q + t_2 q + \dots + t_k q + r \\ &= (t_1 + t_2 + \dots + t_k)q + r = tq + r, \end{aligned}$$

mit $t := t_1 + t_2 + \dots + t_k$. □

Wie für ganze Zahlen nennen wir q einen *Teiler* von p , falls in (11.111) $r = 0$ gilt. Wir schreiben dafür $q \mid p$. Ein Polynom t heißt *gemeinsamer Teiler* von p und q , falls $t \mid p$ und $t \mid q$ erfüllt ist. t heißt *größter gemeinsamer Teiler* von p und q , falls t ein gemeinsamer Teiler von p und q ist, und falls der Grad jedes anderen gemeinsamen Teilers den von t nicht übertreffen kann. Natürlich schreiben wir dafür wieder $\text{ggT}(p, q)$. Man überlegt sich schnell, daß der größte gemeinsame Teiler bis auf einen Vorfaktor aus \mathbb{K} eindeutig bestimmt ist. Ein Verfahren, mit dem man sich einen solchen Teiler verschaffen kann, stellt der euklidische Algorithmus dar, sinngemäß auf Polynome übertragen. Die Überlegungen dazu kann jeder leicht selbst anstellen, so daß wir sie hier nicht noch einmal vorführen. Das Folgende gilt, außer wenn es explizit anders bestimmt wird, für Polynome über \mathbb{R} oder \mathbb{C} .

11.8.11 Korollar Für ein Polynom p mit der Nullstelle z_0 geht die Polynomdivision mit dem Linearfaktor $q(z) := (z - z_0)$ ohne Rest auf, d. h., es gibt ein Polynom t vom Grad $\text{grad}(p) - 1$ mit der Eigenschaft $p(z) = t(z)(z - z_0)$. Insbesondere kann, außer für das Nullpolynom, die Anzahl der Nullstellen nicht den Grad des Polynoms übertreffen. Dabei sind die Vielfachheiten mitzurechnen.

Beweis. Nach Satz 11.8.10 existiert die Darstellung $p(z) = t(z)(z - z_0) + r(z)$, mit einem Restpolynom r von kleinerem Grad als $\text{grad}(q) = 1$. Daher muß $\text{grad}(r) = 0$, oder $\text{grad}(r) = -\infty$ gelten, d. h., r ist konstant. Aus $0 = p(z_0) = t(z_0)(z_0 - z_0) + r$ folgt $r = 0$. Da für jede Nullstelle eines nicht konstanten Polynoms p die Division mit dem entsprechenden Linearfaktor den Grad um eins erniedrigt, kann es nicht mehr als $\text{grad}(p)$ Nullstellen geben. Für ein konstantes Polynom $p \neq 0$ ist das natürlich ebenfalls wahr. □

11.8.12 Satz (Identitätssatz für Polynome) Stimmen zwei Polynome p und q mit Graden nicht größer als $n \geq 0$ an mehr als n verschiedenen Stellen überein, so muß $p = q$ gelten.

Beweis. Wir nehmen an, daß $p(z_i) = q(z_i)$ für wenigsten $n + 1$ verschiedene Stellen z_i gilt. Dann hat das Polynom $p - q$ nach (11.109) einen Grad, der höchsten n ist, aber $n + 1$ verschiedene Nullstellen. Nach Korollar 11.8.11 ist das nur für das Nullpolynom möglich. Das bedeutet $p - q = 0$, oder $p = q$. \square

11.8.13 Beispiel Für $p(z) := 4z^4 - 16z^3 + 17z^2 - 4z + 4$ ist $z_0 = 2$ eine Nullstelle. Die Polynomdivision mit $z - 2$ sollte also ohne Rest aufgehen. Der Beweis von Satz 11.8.10 zeigt, wie wir dabei vorgehen: Wir müssen $z - 2$ mit $4z^3$ multiplizieren und das Ergebnis $4z^4 - 8z^3$ von $p(z)$ abziehen, damit der führende Summand $4z^4$ verschwindet. Das machen wir nach folgendem Rechenschema: In der ersten Zeile schreiben wir die Aufgabe $(4z^4 - 16z^3 + 17z^2 - 4z + 4) : (z - 2) =$. Auf der rechten Seite tragen wir der Reihe nach die Monome $t_1(z) = 4z^3$, $t_2(z) = -8z^2$, usw. als Summe $4z^3 - 8z^2 + \dots$ ein. Beim ersten Schritt ziehen wir $4z^4 - 8z^3$ von $p(z)$ ab, indem wir diesen Ausdruck unter die passenden Potenzen von $p(z)$ schreiben und, wie in den Anfängen der Schulzeit, die Koeffizienten voneinander abziehen. Da beim Addieren weniger Rechenfehler auftreten, hat es sich bewährt, nicht $-(4z^4 - 8z^3)$ zu verwenden, sondern den ausmultiplizierten Ausdruck $-4z^4 + 8z^3$. Dieser kann jetzt zu $p(z)$ addiert werden und liefert das Polynom $p_1(z) = -8z^3 + 17z^2 - 4z + 4$, mit dem wir auf dieselbe Weise verfahren. Nach höchstens vier Durchgängen ist die Polynomdivision beendet.

$$\begin{array}{r}
 (4z^4 - 16z^3 + 17z^2 - 4z + 4) : (z - 2) = 4z^3 - 8z^2 + z - 2 \\
 -4z^4 + 8z^3 \\
 \hline
 -8z^3 + 17z^2 - 4z + 4 \\
 8z^3 - 16z^2 \\
 \hline
 z^2 - 4z + 4 \\
 -z^2 + 2z \\
 \hline
 -2z + 4 \\
 2z - 4 \\
 \hline
 0
 \end{array}$$

Wir erhalten $p(z) = 4z^4 - 16z^3 + 17z^2 - 4z + 4 = (4z^3 - 8z^2 + z - 2)(z - 2)$. Wenn wir jetzt nach weiteren Nullstellen suchen, müssen wir uns mit dem Polynom $t(z) = 4z^3 - 8z^2 + z - 2$ befassen, das nur noch von drittem Grade ist. $t(2) = 4 \cdot 8 - 8 \cdot 4 + 2 - 2 = 0$ zeigt, daß 2 sogar eine doppelte Nullstelle darstellt. Eine erneute Polynomdivision liefert $t(z) = (4z^2 + 1)(z - 2)$, also $p(z) = (4z^2 + 1)(z - 2)^2$ (nachrechnen). Die verbleibenden Nullstellen $\frac{i}{2}$ und $-\frac{i}{2}$ kann man jetzt einfach ablesen.

11.8.14 Beispiel Wir suchen von der Funktion $f(x) := \frac{x^4 - 2x^3 + x^2 - 18}{2x^2 - 5x - 3}$ für den Bereich $|x| \rightarrow \infty$ eine Näherungskurve n . Wenn wir dafür den Bruch in der etwas ungewohnten Weise $(x^4 - 2x^3 + x^2 - 18) : (2x^2 - 5x - 3)$ schreiben, wird klar, was zu tun ist. Wir führen mit $x^4 - 2x^3 + x^2 - 18$ und $2x^2 - 5x - 3$ eine Polynomdivision durch und erhalten $x^4 - 2x^3 + x^2 - 18 = t(x)(2x^2 - 5x - 3) + r(x)$, mit einem Polynom t vom Grad 2 und einem Restpolynom r mit $\text{grad}(r) < 2$. Wenn wir diese Gleichung durch $2x^2 - 5x - 3$ teilen, erhalten wir auf der linken

Seite $f(x)$ und auf der rechten $t(x) + \frac{r(x)}{2x^2 - 5x - 3}$. Der letzte Bruch geht für $|x| \rightarrow \infty$ gegen Null, denn der Zählergrad ist nach Satz 11.8.10 kleiner als der Nennergrad.

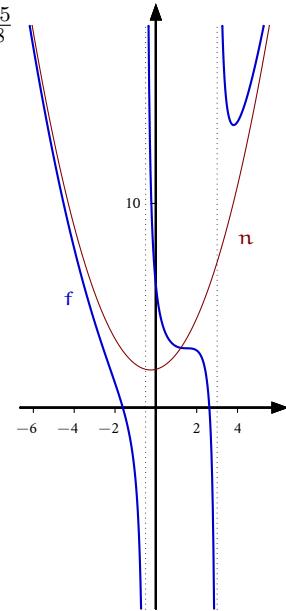
$$\begin{array}{r} (x^4 - 2x^3 + x^2 - 18) : (2x^2 - 5x - 3) = \frac{1}{2}x^2 + \frac{1}{4}x + \frac{15}{8} \\ -x^4 + \frac{5}{2}x^3 + \frac{3}{2}x^2 \\ \hline \frac{1}{2}x^3 + \frac{5}{2}x^2 - 18 \\ -\frac{1}{2}x^3 + \frac{5}{4}x^2 + \frac{3}{4}x \\ \hline \frac{15}{4}x^2 + \frac{3}{4}x - 18 \\ -\frac{15}{4}x^2 + \frac{75}{8}x + \frac{45}{8} \\ \hline \frac{81}{8}x - \frac{99}{8} \end{array}$$

Das bedeutet

$$x^4 - 2x^3 + x^2 - 18 = \left(\frac{1}{2}x^2 + \frac{1}{4}x + \frac{15}{8}\right)(2x^2 - 5x - 3) + \frac{81x - 99}{8},$$

oder

$$\begin{aligned} f(x) &= \frac{\left(\frac{1}{2}x^2 + \frac{1}{4}x + \frac{15}{8}\right)(2x^2 - 5x - 3) + \frac{81x - 99}{8}}{2x^2 - 5x - 3} \\ &= \frac{1}{2}x^2 + \frac{1}{4}x + \frac{15}{8} + \frac{81x - 99}{8(2x^2 - 5x - 3)}. \end{aligned}$$



Der polynomiale Anteil $\frac{1}{8}(4x^2 + 2x + 15)$ von $f(x)$ liefert die Rechenvorschrift der *Näherungskurve* n . Hier ist es eine Parabel. Sollte bei dem Verfahren eine Gerade als Näherungskurve herauskommen, dann nennt man sie *Asymptote*. Diese ist also nur eine besonders einfache Näherungskurve.

11.8.15 Satz (Interpolationspolynom) Für $n+1$ Punkte $P_i := (z_i, w_i) \in \mathbb{K}^2$, $i = 0, 1, \dots, n$, mit paarweise verschiedenen Koordinaten z_i und wenigstens einem $w_i \neq 0$, gibt es genau ein Polynom p vom Grad n mit der Eigenschaft $p(z_i) = w_i$ für $i = 0, \dots, n$. p heißt Interpolationspolynom für die Punkte P_i und kann folgendermaßen berechnet werden:

$$p(z) = \sum_{k=0}^n w_k \prod_{i \neq k}^n \frac{z - z_i}{z_k - z_i}. \quad (11.113)$$

Beweis. Zunächst die Eindeutigkeit: Gäbe es zwei verschiedene Polynome p und q , jeweils vom Grade n , mit der Eigenschaft $p(z_i) = w_i = q(z_i)$ für $i = 1, \dots, n+1$, dann würden diese an $n+1$ Stellen übereinstimmen und müßten nach Satz 11.8.12 gleich sein.

Die Existenz des Interpolationspolynoms ist ein Ergebnis der linearen Algebra. Sei \mathcal{P}_n der Vektorraum der Polynome vom Grad kleiner oder gleich n . Die Menge der Monome $\mathcal{B} := \{p_k \mid p_k(z) := z^k, 0 \leq k \leq n\}$ ist linear unabhängig (vergl. Übung 6.4.5). Jedes Polynom ist offensichtlich eine Linearkombination dieser Monome. Damit ist \mathcal{B} auch erzeugend für \mathcal{P}_n und daher eine Basis. Wir erhalten $\dim \mathcal{P}_n = n+1$. Nun führen wir die lineare Abbildung $A : \mathcal{P}_n \rightarrow \mathbb{K}^n$ ein, die jedem Polynom p den Vektor $A_p := [p(z_0), p(z_1), \dots, p(z_n)]^t \in \mathbb{K}^n$ seiner Funktionswerte an den Stellen z_0, z_1, \dots, z_n zuordnet. Man mache sich klar, daß aus der Eigenschaft $(s \cdot p + t \cdot q)(z) = s \cdot p(z) + t \cdot q(z)$ die Linearität von A unmittelbar folgt. Für zwei

verschiedene Polynome p und q können die Vektoren ihrer Funktionswerte nicht übereinstimmen, wie wir oben bereits gezeigt haben (kurzes Nachdenken). Das bedeutet aber, daß A injektiv und wegen $\dim \mathcal{P}_n = \dim \mathbb{K}^{n+1}$ auch surjektiv ist (vergl. Satz ...) Unsere Aufgabe besteht jetzt nur noch darin, die Umkehrabbildung A^{-1} zu bestimmen, denn diese ordnet einem Vektor $[w_0, w_1, \dots, w_n]^t \in \mathbb{K}^n$ das Polynom p mit der Eigenschaft $p(z_i) = w_i$, also das Interpolationspolynom zu (man mache sich das klar, das ist die zentrale Idee des Beweises). Wir wissen, daß die lineare Abbildung A^{-1} bereits durch die Bilder $A^{-1}e_k$ der kanonischen Basisvektoren festgelegt ist. Wir müssen daher ein Polynom $q_k := A^{-1}e_k$ finden, das an der Stelle z_k den Wert 1 und an allen anderen Stellen z_i den Wert 0 annimmt. Jedes z_i für $i \neq k$ ist eine Nullstelle von q_k . Bis auf einen Vorfaktor bedeutet das $q_k(z) \sim (z-z_0) \cdots (z-z_{k-1})(z-z_{k+1}) \cdots (z-z_n) = \prod_{i \neq k}^n (z - z_i)$. Wir teilen die rechte Seite durch den Wert $\prod_{i \neq k}^n (z_k - z_i)$, den sie an der Stelle z_k annimmt und haben q_k gefunden:

$$q_k(z) = \prod_{i \neq k}^n \frac{z - z_i}{z_k - z_i}.$$

Der Rest wird durch die Linearität von A^{-1} erledigt:

$$p = A^{-1}[w_0, w_1, \dots, w_n]^t = A^{-1} \sum_{k=0}^n w_k e_k = \sum_{k=0}^n w_k A^{-1} e_k = \sum_{k=0}^n w_k q_k.$$

Wenn wir diesen Ausdruck an einer Stelle z auswerten, ergibt sich (11.113). \square

Die Idee für diesen Beweis stammt von Dr. H. Fischer und wurde dem Autor dankenswerterweise auf einer der vielen gemeinsamen Wanderungen verraten.

Die Formel liefert auch für den Grenzfall, daß alle w_i verschwinden, das richtige Polynom, nämlich das Nullpolynom. Das ist das einzige, das $n + 1$ verschiedene Nullstellen haben kann, bei einem Grad, der nicht größer als n ist. Da das Nullpolynom aber nicht den Grad n hat, wie im Satz verlangt, wurde dieser wenig wichtige Spezialfall ausgeschlossen.

11.8.16 Das HORNER-Schema Soll das Polynom $p(z) := 4z^4 - 16z^3 + 17z^2 - 4z + 4$ an der Stelle 3 ausgewertet werden, dann rechnet man üblicherweise $p(3) = 4 \cdot 3^4 - 16 \cdot 3^3 + 17 \cdot 3^2 - 4 \cdot 3 + 4 = 4 \cdot 81 - 16 \cdot 27 + 17 \cdot 9 - 8 = 324 - 432 + 153 - 8 = 37$. Wenn man das ökonomisch macht, indem man zunächst $3^2 = 9$ berechnet, dann $3^3 = 9 \cdot 3 = 27$ und schließlich $3^4 = 27 \cdot 3 = 81$, bevor man die restlichen Produkte bildet, dann sind dafür insgesamt 3 + 4 Multiplikationen nötig. Wie man sieht, können dabei vorübergehend recht große Zahlen auftreten, obwohl das für das Ergebnis nicht zutrifft. Das HORNER-Schema reduziert die Anzahl der Multiplikationen und sorgt dafür, daß die Zwischenergebnisse nicht übermäßig wachsen. Die Idee dafür ist, $p(z)$ folgendermaßen umzuformen:

$$\begin{aligned} p(z) &= (4z - 16)z^3 + 17z^2 - 4z + 4 = ((4z - 16)z + 17)z^2 - 4z + 4 \\ &= (((4z - 16)z + 17)z - 4)z + 4. \end{aligned}$$

Wenn man diese Klammerausdrücke von innen nach außen berechnet, dann sind bei jedem Rechenschritt jeweils nur eine Multiplikation mit 3 und eine Addition nötig. Insgesamt sind

das nur noch 4 Multiplikationen:

$$\begin{aligned} p(3) &= (((12 - 16) \cdot 3 + 17) \cdot 3 - 4) \cdot 3 + 4 = ((-12 + 17) \cdot 3 - 4) \cdot 3 + 4 \\ &= (15 - 4) \cdot 3 + 4 = 33 + 4 = 37. \end{aligned}$$

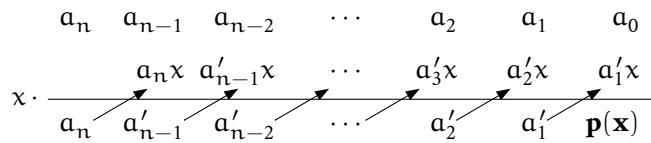
Es bietet sich natürlich an, dieses Verfahren zu schematisieren, um den Arbeitsaufwand weiter zu reduzieren. Dafür schreibt man in der ersten Zeile des sogenannten HORNER-Schemas die Koeffizienten von p , sortiert in absteigender Reihenfolge der Potenzen.

$$\begin{array}{r} 4 & -16 & 17 & -4 & 4 \\ 3 \cdot & \frac{4 \cdot 3}{4 & -4} \\ \hline & & & & \end{array}$$

Der erste Koeffizient wird in die dritte Zeile übertragen, mit 3 multipliziert und das Ergebnis unter den zweiten Koeffizienten -16 geschrieben. Den Wert 3, an dem p ausgewertet wird, schreibt man, nur so zu Erinnerung, zwischen den Anfang der zweiten und der dritten Zeile. Anschließend werden die ersten beiden Einträge der zweiten Spalte addiert und das Ergebnis in der dritten Zeile eingetragen. Das entspricht der Berechnung der innersten Klammer. Mit dem Ergebnis -4 wiederholt sich der Vorgang, bis in der letzten Spalte in der dritten Zeile das Ergebnis steht:

$$\begin{array}{r} 4 & -16 & 17 & -4 & 4 \\ 3 \cdot & \frac{12 & -12 & 15 & 33}{4 & -4 & 5 & 11 & \mathbf{37}} \\ & & & & \end{array}$$

Für ein allgemeines Polynom $p(z) = a_n z^n + a_{n-1} z^{n-1} + a_{n-2} z^{n-2} + \dots + a_1 z + a_0$ sieht das HORNER-Schema zur Berechnung von $p(x)$ folgendermaßen aus:



Das HORNER-Schema kann auch für die Polynomdivision von $p(z)$ mit dem Linearfaktor $z - z_0$ einer Nullstelle z_0 von p verwendet werden. Das beruht auf folgender Beobachtung: Wenn wir $p(z)$ durch $z - z_0$ teilen, erhalten wir im ersten Schritt

$$\begin{array}{r} (a_n z^n + a_{n-1} z^{n-1} + a_{n-2} z^{n-2} + \dots + a_0) : (z - z_0) = a_n z^{n-1} + a'_{n-1} z^{n-2} \dots \\ -a_n z^n + a_n z_0 z^{n-1} \\ \hline (a_{n-1} + a_n z_0) z^{n-1} + a_{n-2} z^{n-2} + \dots + a_0 \\ -a'_{n-1} z^{n-1} + a'_{n-1} z_0 z^{n-2} \end{array}$$

Dabei ist der erste Koeffizient a_n und der zweite $a'_{n-1} := a_{n-1} + a_n z_0$ wie im ersten bzw. zweiten Schritt des HORNER-Schemas zur Berechnung von $p(z_0)$ entstanden. Man kann sich leicht davon überzeugen, daß das auch für die weiteren Koeffizienten des Ergebnispolynoms

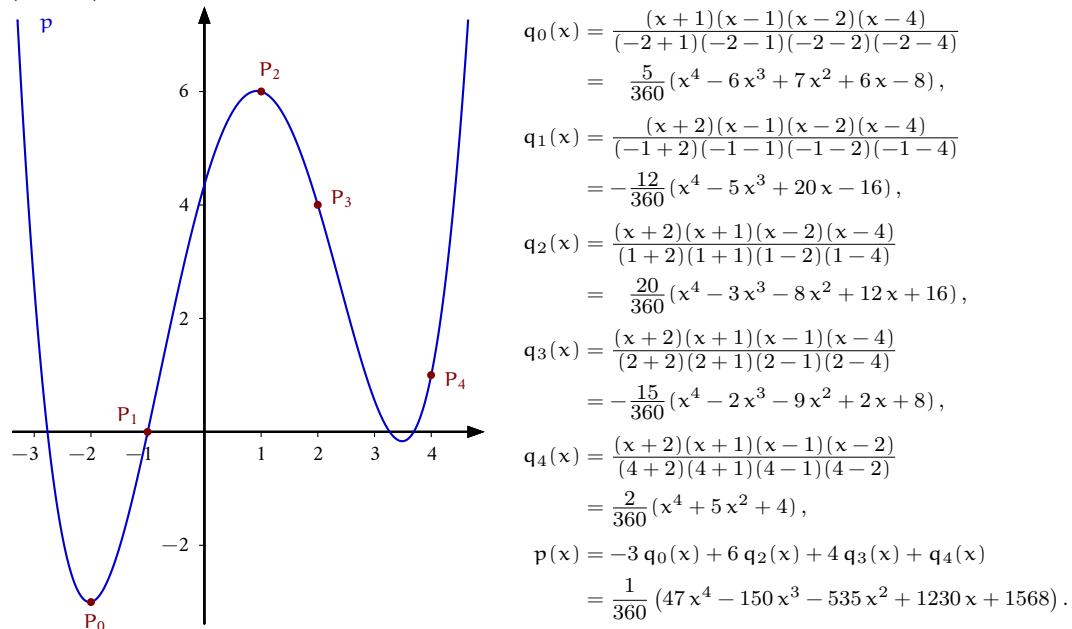
richtig bleibt. Im letzten Schritt des Schemas erhalten wir $p(z_0) = 0$. Als Resultat bekommt man $a_n z^{n-1} + a'_{n-1} z^{n-2} + \dots + a'_0$, mit den Koeffizienten $a_n, a'_{n-1}, \dots, a'_0$, die aus dem HORNER-Schema zur Berechnung von $p(z_0)$ abgelesen werden können.

Für das Beispiel $s(z) = z^5 + z^4 - 18z^3 + 14z^2 + 21z + 5$ und $z_0 = -5$ sieht das folgendermaßen aus:

$$\begin{array}{r} 1 & 1 & -18 & 14 & 21 & 5 \\ -5 \cdot & \hline & -5 & 20 & -10 & -20 & -5 \\ & 1 & -4 & 2 & 4 & 1 & 0 \end{array}$$

Das bedeutet $(z^5 + z^4 - 18z^3 + 14z^2 + 21z + 5) : (z + 5) = z^4 - 4z^3 + 2z^2 + 4z + 1$.

11.8.17 Beispiel Wir suchen das Interpolationspolynom p vierten Grades zu den fünf Punkten $P_0 := [-2, -3], P_1 := [-1, 0], P_2 := [1, 6], P_3 := [2, 4]$ und $P_4 := [4, 1]$. Wir erhalten nach (11.113)



11.8.18 Abschätzung der Nullstellenmenge Um gezielt Näherungsverfahren zur Bestimmung von Nullstellen einsetzen zu können, braucht man ungefähre Bereiche, in denen sich alle Nullstellen befinden, oder in denen man mit Sicherheit keine vorfinden wird. Eine erste grobe Abschätzung liefert die sogenannte CAUCHY-Regel für reelle Polynome.

11.8.19 Satz (CAUCHY-Regel) Für ein reelles, normiertes Polynom $p(x) = \sum_{k=0}^n a_k x^k$ sei N die Anzahl der negativen Koeffizienten a_k . Dann können oberhalb von

$$\max \left\{ \sqrt[n-k]{N|a_k|} \mid a_k < 0 \right\} \quad (11.114)$$

keine reellen Nullstellen mehr liegen.

Beweis. Sei $p(x) = \sum_{a_k > 0} a_k x^k - \sum_{a_k < 0} |a_k| x^k$ und x größer als die Zahl (11.114). Dann können wir folgendermaßen abschätzen:

$$\sum_{a_k > 0} a_k x^k \geq x^n = \frac{1}{N} \sum_{a_k < 0} x^{n-k} x^k > \frac{1}{N} \sum_{a_k < 0} \sqrt[n-k]{N|a_k|}^{n-k} x^k = \sum_{a_k < 0} |a_k| x^k,$$

woraus sofort $p(x) = \sum_{a_k > 0} a_k x^k - \sum_{a_k < 0} |a_k| x^k > 0$ folgt. \square

Testen wir das Verfahren an einem Polynom, bei dem wir die Nullstellen kennen, sagen wir $p(x) := (x+5)(x+2)(x-1)(x-4) = x^4 + 2x^3 - 21x^2 - 22x + 40$. Wir bestimmen $\max\{\sqrt{42}, \sqrt[3]{44}\} = \sqrt{42} \approx 6.49$. Das ist nicht sonderlich genau, verglichen mit der größten Nullstelle 4, aber das war auch nicht zu erwarten, wenn man sich den Beweis genau ansieht. Um auch eine Abschätzung für die kleinste Nullstelle zu erhalten, wenden wir das Verfahren auf das gespiegelte Polynom $q(x) := p(-x) = x^4 - 2x^3 - 21x^2 + 22x + 40$ an (wäre der Grad von p ungerade, so müssten wir natürlich $q(x) = -p(-x)$ verwenden). Die obere Grenze für q ist $\max\{4, \sqrt{42}\} = \sqrt{42}$. Also liegen die reellen Nullstellen im Intervall $[-\sqrt{42}, \sqrt{42}]$. Jetzt wandeln wir das Beispiel leicht ab: $r(x) := (x+5)(x-1)(x-2)(x-4)(x^2 + 2x + 1300) = x^6 + 1275x^4 - 2580x^3 - 27216x^2 + 80520x - 52000$. Als obere Grenze ergibt sich $\max\{\sqrt[3]{3 \cdot 2580}, \sqrt[4]{3 \cdot 27216}, \sqrt[6]{3 \cdot 52000}\} \approx 19.8$, bei der reellen Nullstellenmenge $\{-5, 1, 2, 4\}$. Für die untere Grenze ergibt sich -16.9 . Zieht man die großen Koeffizienten in Betracht, dann ist das Ergebnis gar nicht so schlecht, verglichen mit den wirklichen Nullstellen dient es aber gerade noch zur groben Orientierung.

11.8.20 Satz Für ein normiertes Polynom $p(x) = \sum_{k=0}^n a_k x^k$ liegen alle Nullstellen in der abgeschlossenen Kreisscheibe um 0 mit dem Radius

$$R_1 := \max \left\{ 1, \sum_{k=0}^{n-1} |a_k| \right\}. \quad (11.115)$$

Beweis. Wir definieren $q(z) := \sum_{k=0}^{n-1} a_k z^k$. Wir können von $q \neq 0$ ausgehen, da andernfalls 0 die einzige Nullstelle ist. Es sei $|z| > R$, also insbesondere $|z| > 1$. Dann gilt

$$|q(z)| \leq \sum_{k=0}^{n-1} |a_k| |z|^k \leq |z|^{n-1} \sum_{k=0}^{n-1} |a_k| \leq |z|^{n-1} R_1.$$

Mit Hilfe der umgekehrten Dreiecksungleichung erhalten wir daraus

$$|p(z)| = |z^n + q(z)| \geq |z|^n - |q(z)| \geq |z|^n - |z|^{n-1} R_1 = |z|^{n-1} (|z| - R_1) > 0.$$

Daher müssen Nullstellen z von p die Bedingung $|z| \leq R_1$ erfüllen. \square

Das Beispiel 11.8.13 $p(z) = 4z^4 - 16z^3 + 17z^2 - 4z + 4 = 4(z^4 - 4z^3 + \frac{17}{4}z^2 - z + 1)$ ergibt als Radius $R_1 = 4 + \frac{17}{4} + 1 + 1 = 10.25$, bei der Nullstellenmenge $\{2, -\frac{1}{2}, \frac{1}{2}\}$. Das ist nicht gerade eine überwältigende Genauigkeit. Wenn man sich vergegenwärtigt, daß die Nullstellenmenge

von p in der eines Produkts pq mit einem weiteren Polynom q enthalten ist, könnte man durch geschickte Wahl von q die Genauigkeit verbessern. Sind die Koeffizienten von p alle etwa von der gleichen Größenordnung, so ist $q(z) := z \pm 1$ manchmal eine gute Wahl, denn

$$\begin{aligned}(z \pm 1)p(z) &= (z \pm 1) \sum_{k=0}^n a_k z^k = \sum_{k=0}^n a_k z^{k+1} \pm \sum_{k=0}^n a_k z^k = \sum_{\ell=1}^{n+1} a_{\ell-1} z^\ell \pm \sum_{k=0}^n a_k z^k \\ &= z^{n+1} + \sum_{k=1}^n (a_{k-1} \pm a_k) z^k \pm a_0.\end{aligned}$$

In unserem Beispiel ergibt sich mit $(z + 1)p(z) = 4(z^5 - 3z^4 + \frac{1}{4}z^3 - \frac{13}{4}z^2 + 1)$ die leichte Verbesserung $R_1 = 7.5$.

Es gibt weitere Abschätzungen für die Nullstellen eines normierten Polynoms. Sie beruhen auf der folgenden notwendigen Nullstellenbedingung.

11.8.21 Satz (CAUCHY) Durch $p(z) = \sum_{k=0}^n a_k z^k$ sei ein normiertes Polynom gegeben und $z \neq 0$ sei eine Nullstelle. Dann gilt folgende Ungleichung:

$$|z| \leq |a_{n-1}| + \frac{|a_{n-2}|}{|z|} + \frac{|a_{n-3}|}{|z|^2} + \cdots + \frac{|a_0|}{|z|^{n-1}}. \quad (11.116)$$

Beweis. $p(z) = 0$ bedeutet $-z^n = a_{n-1}z^{n-1} + \cdots + a_1z + a_0$. Die Dreiecksungleichung ergibt $|z|^n \leq |a_{n-1}| |z|^{n-1} + \cdots + |a_1| |z| + |a_0|$ und eine Division mit $|z|^{n-1}$ schließlich (11.116). \square

11.8.22 Korollar Die Nullstellen eines normierten Polynoms $p(z) = \sum_{k=0}^n a_k z^k$ müssen in abgeschlossenen Kreisscheiben um 0 mit den Radien

$$R_2 = \max \{ |a_0|, |a_1| + 1, \dots, |a_{n-1}| + 1 \}, \quad (11.117)$$

$$R_3 = |a_{n-1}| + \frac{|a_{n-2}|}{R_2} + \frac{|a_{n-3}|}{R_2^2} + \cdots + \frac{|a_0|}{R_2^{n-1}}, \quad (11.118)$$

$$R_4 = 2 \cdot \max \left\{ |a_{n-1}|, \frac{|a_{n-2}|}{|a_{n-1}|}, \frac{|a_{n-3}|}{|a_{n-2}|}, \dots, \frac{1}{2} \frac{|a_0|}{|a_1|} \right\}. \quad (11.119)$$

liegen. Dabei gilt R_4 nur für den Fall, daß kein Koeffizient von p verschwindet.

Beweis. Wir nehmen an, z sei eine Nullstelle mit $|z| > R_2$. z muß die Ungleichung (11.116) erfüllen:

$$\begin{aligned}|z| &\leq |a_{n-1}| + \frac{|a_{n-2}|}{|z|} + \frac{|a_{n-3}|}{|z|^2} + \cdots + \frac{|a_0|}{|z|^{n-1}} < |a_{n-1}| + \frac{|a_{n-2}|}{R_2} + \frac{|a_{n-3}|}{R_2^2} + \cdots + \frac{|a_0|}{R_2^{n-1}} \\ &\leq \max \{ |a_k| \mid k = 1, \dots, n-1 \} \left(1 + \frac{1}{R_2} + \frac{1}{R_2^2} + \cdots + \frac{1}{R_2^{n-2}} \right) + \frac{|a_0|}{R_2^{n-1}} \\ &\leq (R_2 - 1) \frac{1 - \frac{1}{R_2^{n-1}}}{1 - \frac{1}{R_2}} + \frac{R_2}{R_2^{n-1}} = R_2,\end{aligned}$$

im Widerspruch zu $|z| > R_2$. Dabei haben wir den Trick $\max\{|a_k| \mid k = 1, \dots, n-1\} = \max\{|a_k| \mid k = 1, \dots, n-1\} + 1 - 1 = \max\{|a_k| + 1 \mid k = 1, \dots, n-1\} - 1 \leq R_2 - 1$ eingesetzt.

R_3 ist der Schnittpunkt der ersten Winkelhalbierenden mit der Funktion $f(x) := |a_{n-1}| + \frac{|a_{n-2}|}{x} + \dots + \frac{|a_1|}{x^{n-2}} + \frac{|a_0|}{x^{n-1}}$. Da f monoton fallend ist, bei $x = 0$ eine senkrechte Asymptote hat und für $x \rightarrow \infty$ eine waagrechte in der Höhe $|a_{n-1}|$, gibt es genau einen solchen Schnittpunkt R_3 . Gäbe es eine Nullstelle z von p mit $|z| > R_3$, so würde das den Widerspruch

$$|z| \leq |a_{n-1}| + \frac{|a_{n-2}|}{|z|} + \dots + \frac{|a_0|}{|z|^{n-1}} < |a_{n-1}| + \frac{|a_{n-2}|}{R_3} + \dots + \frac{|a_0|}{R_3^{n-1}} = f(R_3) = R_3$$

ergeben. Also müssen alle Nullstellen z die Ungleichung $|z| \leq R_3$ erfüllen.

Für eine Nullstelle z mit $|z| > R_4$ schätzen wir einen typischen Summanden in (11.116) ab. Für $k = 2, \dots, n-1$ gilt

$$\frac{|a_{n-k}|}{|z|^{k-1}} < \frac{|a_{n-k}|}{2 \frac{|a_{n-2}|}{|a_{n-1}|} \cdot 2 \frac{|a_{n-3}|}{|a_{n-2}|} \cdots 2 \frac{|a_{n-k}|}{|a_{n-(k-1)}|}} = \frac{|a_{n-1}|}{2^{k-1}}.$$

Wegen der Sonderrolle von a_0 in (11.119) gilt für $k = n$:

$$\frac{|a_0|}{|z|^{n-1}} < \frac{|a_{n-1}|}{2^{n-2}}.$$

Damit ergibt sich aus (11.116) der Widerspruch

$$\begin{aligned} |z| &< |a_{n-1}| \left(1 + \frac{1}{2} + \frac{1}{4} + \dots + \frac{1}{2^{n-2}} + \frac{1}{2^{n-2}} \right) = |a_{n-1}| \left(\sum_{k=0}^{n-1} \frac{1}{2^k} + \frac{1}{2^{n-2}} \sum_{k=1}^{\infty} \frac{1}{2^k} \right) \\ &= |a_{n-1}| \sum_{k=0}^{\infty} \frac{1}{2^k} = 2 |a_{n-1}| \leq R_4. \end{aligned} \quad \square$$

Im Beispiel 11.8.13 erhalten wir durch (11.115) mit $R_2 = 5$ gegenüber $R_1 = 7.5$ eine weitere Verbesserung. Wenn wir die Bedingungen auf $(z+1)p(z) = 4(z^5 - 3z^4 + \frac{1}{4}z^3 - \frac{13}{4}z^2 + 1)$ anwenden, erhalten wir $R_2 = 4.25$ und $R_3 = 3.37$, als genäherte Lösung von $x = 3 + \frac{0.25}{x} + \frac{3.25}{x^2} + \frac{1}{x^4}$. R_4 liefert für p keine Verbesserung.

11.8.23 Satz (Fundamentalsatz der Algebra) *Jedes Polynom, das nicht konstant ist, hat in \mathbb{C} wenigstens eine Nullstelle.*

Beweis (H. Leinfelder 1981). Die Idee des Beweises besteht darin, zunächst für den Betrag $|p|$ des Polynoms p ein Minimum nachzuweisen, um anschließend dessen Wert als Null zu identifizieren.

Wir gehen von einem normierten Polynom p aus: $p(z) = \sum_{k=0}^n a_k z^k$. Durch $R := 1 + \sum_{k=0}^{n-1} |a_k|$ ist eine Nullstellenschranke von p mit der Eigenschaft $|p(z)| > R$ für $|z| > R$ gegeben. Für ein solches z können wir nämlich folgendermaßen abschätzen:

$$\left| \sum_{k=0}^{n-1} a_k z^k \right| \leq \sum_{k=0}^{n-1} |a_k| |z|^k \leq |z|^{n-1} \sum_{k=0}^{n-1} |a_k| = |z|^{n-1} (R - 1) < |z|^{n-1} (|z| - 1).$$

Es folgt $|p(z)| \geq |z|^n - |\sum_{k=0}^{n-1} a_k z^k| > |z|^n - |z|^{n-1}(|z| - 1) = |z|^{n-1} > R^{n-1} > R$. D.h., mögliche Nullstellen von p müssen in der abgeschlossenen Kreisscheibe K um 0 mit dem Radius R liegen.

Auf K hat die stetige Funktion $|p|$ ein Minimum, das nicht größer als $|p|(0) = |a_0|$ sein kann. Wegen $|p|(z) > R > |a_0|$ für $|z| > R$, liefert $|p|$ außerhalb von K keine kleineren Werte. Daher befindet sich in K sogar ein globales Minimum von $|p|$. Nun, da die Existenz eines Minimums von $|p|$ gesichert ist, können wir annehmen, daß es sich an der Stelle $z = 0$ befindet (andernfalls verschieben wir p geeignet). Wir wollen $a_0 = 0$ zeigen.

Für eine beliebige komplexe Zahl z vom Betrage 1 definieren wir die stetige Funktion $f : \mathbb{R} \rightarrow \mathbb{R}$ durch $f(t) := |p|^2(tz) = |p(tz)|^2$. k sei der erste Index ≥ 1 , für den $a_k \neq 0$ gilt. Wir erhalten

$$\begin{aligned} f(t) - f(0) &= (t^n z^n + \dots + t^k z^k a_k + a_0)(t^n \bar{z}^n + \dots + t^k \bar{z}^k \bar{a}_k + \bar{a}_0) - |a_0|^2 \\ &= (\bar{z}^k \bar{a}_k a_0 + z^k a_k \bar{a}_0)t^k + t^{k+1}r(t), \end{aligned}$$

mit einem geeigneten reellen Polynom r , dessen genaue Form für das Folgende nicht weiter wichtig ist. Da $f(0)$ ein Minimum von f darstellt, ist für hinreichend kleines $t > 0$ die Ungleichung $\frac{f(t)-f(0)}{t^k} = 2\operatorname{Re}(z^k a_k \bar{a}_0) + t \cdot r(t) \geq 0$ erfüllt. Im Grenzwert $t \searrow 0$ folgt daraus $2\operatorname{Re}(z^k a_k \bar{a}_0) \geq 0$. Falls $a_0 \neq 0$ seien sollte, gilt für eine geeignete Zahl w aus dem Einheitskreis $a_k \bar{a}_0 = w|a_k||a_0| \neq 0$. Jetzt verfügen wir über z und wählen dafür eine k -te Wurzel aus $-\bar{w}$ (vergl. 5.2.9). Das ergibt den Widerspruch $0 \leq \operatorname{Re}(-\bar{w}w|a_k||a_0|) = -|a_k||a_0| < 0$. Daher bleibt nur $a_0 = 0$, also $p(0) = 0$. \square

11.8.24 Korollar Jedes Polynom p vom Grade $n \geq 1$ lässt sich vollständig in Linearfaktoren zerlegen, d.h., für $p(z) = \sum_{k=0}^n a_k z^k$ gibt es Zahlen $z_i \in \mathbb{C}$, $i = 1, \dots, n$, so daß für alle $z \in \mathbb{C}$ gilt:

$$p(z) = a_n(z - z_1)(z - z_2) \cdots (z - z_n).$$

Beweis. Nach dem Fundamentalsatz hat p wenigstens eine Nullstelle $z_1 \in \mathbb{C}$. Nach Korollar 11.8.11 geht die Polynomdivision mit dem Linearfaktor $z - z_1$ ohne Rest auf: $p(z) = p_1(z)(z - z_1)$, mit einem Polynom p_1 vom Grade $n - 1$. Auf p_1 wenden wir den Fundamentalsatz erneut an und erhalten $p_1(z) = p_2(z)(z - z_2)$, für ein geeignetes $z_2 \in \mathbb{C}$ und ein Polynom p_2 vom Grade $n - 2$. Nach n Schritten sind wir bei einem Polynom p_n vom Grade 0 angekommen, also bei einer Konstanten, bei der es sich nur um a_n handeln kann (warum?)

$$p(z) = p_1(z)(z - z_1) = p_2(z)(z - z_1)(z - z_2) = a_n(z - z_1)(z - z_2) \cdots (z - z_n). \quad \square$$

11.8.25 STURMSche Ketten Wir haben Lokalisierungssätze für die Nullstellen eines Polynoms kennengelernt. Im Folgenden interessieren wir uns für reelle Polynome und für ihre reellen Nullstellen (ohne das immer wieder zu betonen). Für ein Polynom p ohne mehrfache Nullstellen läßt sich mit Hilfe einer *STURMSchen Kette* ihre Anzahl bestimmen. Eine solche Kette entsteht aus $p_0 := p$ und $p_1 := p'$ nach folgendem Verfahren:

$$p_0 = q_1 p_1 - p_2, \quad p_1 = q_2 p_2 - p_3,$$

$$\begin{aligned}
 p_2 &= q_3 p_3 - p_4, & p_3 &= q_4 p_4 - p_5, \\
 &\vdots & &\vdots \\
 p_{k-1} &= q_k p_k - p_{k+1}, & p_k &= q_{k+1} p_{k+1} - p_{k+2}, \\
 &\vdots & &\vdots \\
 p_{n-2} &= q_{n-1} p_{n-1} - p_n & p_{n-1} &= q_n p_n.
 \end{aligned}$$

Das Polynom $-p_{k+1}$ ist der Rest der Division von p_{k-1} durch p_k . Bei jeder Division reduziert sich der Grad des Restes mindestens um 1, so daß das Verfahren bei einem gemeinsamen Teiler p_n von p_0 und p_1 endet. Die Voraussetzung, daß p keine mehrfache Nullstellen hat, oder, was dazu äquivalent ist, daß $p_0 = p$ und $p_1 = p'$ keine gemeinsamen Nullstellen aufweisen, ist entscheidend für die folgenden Überlegungen: $v(x)$ sei die Anzahl der Vorzeichenwechsel benachbarter Elemente des Vektors $s(x) := [p_0(x), p_1(x), \dots, p_n(x)]$. Wir interessieren uns für das Verhalten von $v(x)$ in der Nähe einer Nullstelle x_0 eines Elements der Kette, wenn x von einem Wert kleiner als x_0 zu einem Wert anwächst, der größer als, oder gleich x_0 ist. Dabei vereinbaren wir, daß für den Fall $1 \leq k \leq n-1$ und $p_k(x) = 0$, die Vorzeichen von $p_{k-1}(x)$ und $p_{k+1}(x)$ für einen möglichen Vorzeichenwechsel verwendet werden.

- i) $p_k(x_0) = 0$ und $p_{k+1}(x_0) = 0$ ist nicht möglich.
- ii) p_n hat keine reellen Nullstellen.
- iii) Aus $p_k(x_0) = 0$ folgt $p_{k-1}(x_0) = -p_{k+1}(x_0) \neq 0$.
- iv) Beim Übergang von $x < x_0$ zu $x \geq x_0$ verringert sich $v(x)$ um eins, wenn $p_0(x_0) = 0$ gilt. Der Wert ändert sich nicht, wenn es sich um die Nullstelle eines anderen Elements der STURMSchen Kette handelt.

Zu i) : Aus $p_k(x_0) = 0$ und $p_{k+1}(x_0) = 0$ folgt $p_{k-1}(x_0) = q_k(x_0)p_k(x_0) - p_{k+1}(x_0) = 0$ und daraus genauso $p_{k-2}(x_0) = \dots = p_1(x_0) = p_0(x_0) = 0$, entgegen der Voraussetzung, daß p_0 und p_1 keine gemeinsame Nullstellen haben.

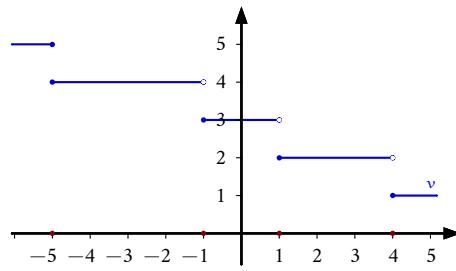
Zu ii) : p_n ist ein gemeinsamer Teiler aller p_k . $p_n(x_0) = 0$ hätte also $p_k(x_0) = 0$ für alle Elemente der Kette zur Folge, was nach i) nicht möglich ist.

Zu iii) : Da $p_{k-1}(x_0) \neq 0$ und $p_{k+1}(x_0) \neq 0$ aus $p_k(x_0) = 0$ folgt ($k = n$ ist nach ii) nicht möglich), haben wir $p_{k-1}(x_0) = q_k(x_0)p_k(x_0) - p_{k+1}(x_0) = -p_{k+1}(x_0)$.

Zu iv) : Zunächst sei $p_0(x_0) = 0$. Dann ist $p_1(x_0) \neq 0$. Wegen der Stetigkeit von p_1 gilt $p_1(x) \neq 0$ sogar in einer kleinen Umgebung $U(x_0)$ von x_0 , auf der p_1 daher keinen Vorzeichenwechsel erfahren kann. Für $p_1(x_0) > 0$ ist p_0 auf $U(x_0)$ streng monoton wachsend und hat bei x_0 einen Vorzeichenwechsel von $-$ nach $+$. Für $U(x_0) \ni x < x_0$ gilt daher $p_0(x) < 0$ und $p_1(x) > 0$, während für $U(x_0) \ni x \geq x_0$ $p_0(x) \geq 0$ und $p_1(x) > 0$ gilt. Beim Übergang von $x < x_0$ zu $x \geq x_0$ ist also der Vorzeichenwechsel zwischen $p_0(x)$ und $p_1(x)$ verschwunden. Dasselbe Ergebnis erhält man für den Fall $p_1(x_0) < 0$.

Jetzt sei $p_k(x_0) = 0$ und $1 \leq k < n-1$ (falls auch $p_0(x_0) = 0$ gelten sollte, muß sogar $k > 1$ sein). Nach ii) gilt $p_{k-1}(x_0) \neq 0$ und $p_{k+1}(x_0) \neq 0$. Wegen der Stetigkeit der Polynome gibt es wieder eine Umgebung $U(x_0)$, in der keine Nullstelle von p_{k-1} oder p_{k+1} liegt. Wegen iii)

haben sie dort entgegengesetztes Vorzeichen. Egal, welches Vorzeichen $p_k(x)$ für $x < x_0$ bzw. $x > x_0$ hat, es muß mit einem der verschiedenen Vorzeichen von $p_{k-1}(x)$ und $p_{k+1}(x)$ übereinstimmen. Daher gibt es zwischen $p_{k-1}(x)$, $p_k(x)$ und $p_{k+1}(x)$ für $x < x_0$ und für $x > x_0$ jeweils nur einen Vorzeichenwechsel. Das stimmt auch noch für $x = x_0$, denn wegen $p_k(x_0) = 0$ zählt hier der Vorzeichenwechsel von $p_{k-1}(x_0)$ und $p_{k+1}(x_0)$. Beim Übergang von $x < x_0$ zu $x \geq x_0$ liefert die Nullstelle von p_k daher keinen Beitrag zur Änderung von $v(x)$. Solch eine Änderung kann nur durch eine Nullstelle von p_0 verursacht werden.



Die Vorzeichenfunktion v können wir zur Zählung reeller Nullstellen eines Polynoms p in einem Intervall $(a, b]$ verwenden: $v(a) - v(b)$ ist genau die Anzahl der reellen Nullstellen von p in $(a, b]$. v ist eine fast überall stetige Funktion, mit den reellen Nullstellen von p als Sprungstellen der Höhe -1 . Dabei ist v rechtsseitig stetig. Damit ist gemeint: Nähern wir uns mit x von rechts einer Nullstelle x_0 von p_0 , so strebt $v(x)$ gegen $v(x_0)$. Nähern wir uns aber von links, so strebt $v(x)$ gegen $v(x_0) + 1$. Damit ist klar, daß v eine monoton fallende Treppenfunktion ist. Jetzt beobachten wir das Verhalten von $v(x)$, während x von a nach b wandert. $v(x)$ bleibt bei dem Wert $v(a)$, solange x nicht die erste Nullstelle oberhalb von a erreicht hat. Ist das eingetreten, dann nimmt v den Wert $v(a) - 1$ an, bei der nächsten den Wert $v(a) - 2$ usw., bis $x = b$ erreicht ist. Sollte b selbst auch eine Nullstelle sein, so springt $v(x)$ hier ein letztes mal. Daher unterscheidet sich $v(a)$ von $v(b)$ um die Anzahl der Nullstellen in $(a, b]$. In der Skizze ist $v(-5) - v(5) = 4 - 1 = 3$, d. h. im Intervall $(-5, 5]$ gibt es 3 reelle Nullstellen von p_0 . Um alle Nullstellen zu erfassen, sollten a und b mindestens so groß sein, daß die Nullstellenschranken von p_0 in $(a, b]$ liegen. Im Beispiel liefert $v(-6) - v(5) = 4$ die Anzahl aller Nullstellen.

11.8.26 Beispiel Wir testen das Verfahren an einem Beispiel, bei dem wir alle Nullstellen kennen: $p(x) := \frac{1}{100}(x-4)(x-1)(x+1)(x+5)(x^2+2)$. Wir wählen $p_0(x) := x^6 + x^5 - 19x^4 + x^3 - 22x^2 - 2x + 40$. Da es nur um die Nullstellen geht, können wir unbequeme Vorfaktoren weglassen. Dann ist $p_1(x) = p'_0(x) = 6x^5 + 5x^4 - 76x^3 + 3x^2 - 44x - 2$. Die Polynomdivision von p_0 mit p_1 ergibt

$$p_0(x) = \left(\frac{1}{6}x + \frac{1}{36}\right)p_1(x) - \frac{233}{36}x^4 + \frac{47}{18}x^3 - \frac{59}{4}x^2 - \frac{4}{9}x + \frac{721}{18}$$

Daher ist $p_2(x) = \frac{233}{36}x^4 - \frac{47}{18}x^3 + \frac{59}{4}x^2 + \frac{4}{9}x - \frac{721}{18}$. Mit diesem Ausdruck will man nicht die nächste Polynomdivision durchführen, wenn man es vermeiden kann. Da wir uns nur für Nullstellen und Vorzeichen interessieren, haben wir die Freiheit, das Ergebnis für $p_2(x)$ mit dem Hauptnenner der Koeffizienten zu multiplizieren, oder gemeinsame positive Vielfache herauszudividieren. Für p_2 ist der Hauptnenner offensichtlich 36. Wir rechnen also mit $p_2(x) = 233x^4 - 94x^3 + 531x^2 + 16x - 1442$ weiter und erhalten:

$$p_1(x) = \left(\frac{6}{233}x + \frac{1729}{233^2}\right)p_2(x) - \frac{1}{233^2}(4705776x^3 + 777600x^2 + 400464x - 2384640)$$

Der größte gemeinsame Teiler aller Koeffizienten des Kandidaten für p_3 ist 1296. Kürzen wir diesen heraus (und lassen den gemeinsamen Vorfaktor weg), dann haben wir $p_3(x) = 3631x^3 + 600x^2 + 309x - 1840$. Auf diese Weise fahren wir fort:

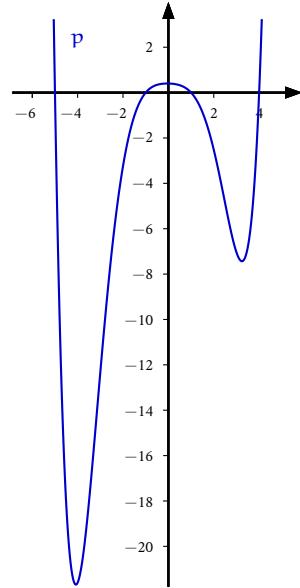
$$\begin{aligned} p_0(x) &= x^6 + x^5 - 19x^4 + x^3 - 22x^2 - 2x + 40, \\ p_1(x) &= 6x^5 + 5x^4 - 76x^3 + 3x^2 - 44x - 2, \\ p_2(x) &= 233x^4 - 94x^3 + 531x^2 + 16x - 1442 \\ p_3(x) &= 3631x^3 + 600x^2 + 309x - 1840, \\ p_4(x) &= -7192x^2 - 1961x + 20361, \\ p_5(x) &= -41959x + 11551, \\ p_6(x) &= -1. \end{aligned}$$

Für die Nullstellenschranke nehmen wir $R_3 = 5.1$, als genäherte Lösung von $x = 1 + \frac{19}{x} + \frac{1}{x^2} + \frac{22}{x^3} + \frac{2}{x^4} + \frac{40}{x^5}$. Alle reellen Nullstellen müssen demnach im Intervall $[-5.1, 5.1]$ liegen. Wir wählen $a = -6$, $b = 6$ und berechnen

$$\begin{aligned} s(-6) &= [13300, -23390, 339850, -766390, -226785, 263305, -1] \\ s(6) &= [29260, 36562, 299434, 805910, -250317, -240203, -1] \end{aligned}$$

Das heißt $v(-6) = 5$ und $v(6) = 1$. Das Verfahren der STURMSchen Ketten gibt also die 4 reellen Nullstellen richtig wieder, mit denen wir das Beispiel konstruiert haben.

Man sieht an dem Beispiel, daß man die Anzahl reeller Nullstellen eines Polynoms herausbekommt, auch wenn das Verfahren rechenaufwendig ist. Tatsächlich lassen sich die nötigen Polynomdivisionen gut durch ein Computeralgebra-Programm erledigen. Das Verfahren hat trotzdem einen gravierenden Schönheitsfehler, nämlich die Bedingung, daß p keine mehrfachen Nullstellen haben darf. Das sieht man dem Polynom üblicherweise nicht an. Schließlich dient das Verfahren ja gerade dazu, etwas über die Nullstellen zu erfahren. Glücklicherweise gibt es für dieses Problem eine befriedigende Lösung. Dafür müssen wir uns nur ins Gedächtnis rufen, daß die STURMSche Kette für p eigentlich der euklidische Algorithmus zum Auffinden eines größten gemeinsamen Teilers von p und p' ist. Und der ist p_n . Das Verfahren liefert uns damit einen Hinweis auf mögliche mehrfache Nullstellen von p : Ist x_0 eine solche Nullstelle, dann ist das Polynom, das durch den Linearfaktor $(x - x_0)$ gebildet wird, ein gemeinsamer Teiler von p und p' und daher auch ein Teiler von p_n . Daran, daß p_n nicht einfach das konstante Polynom ist, wie im Beispiel 11.8.26, läßt sich also erkennen, daß p mehrfache Nullstellen hat. Diese Argumentation macht keinen Gebrauch davon, daß es sich bei p um ein reelles Polynom handelt. Eine gemeinsame Nullstelle könnte also durchaus auch komplex sein. Aber weil p reell ist, muß mit jeder komplexen Nullstelle z auch \bar{z} eine sein: $p(\bar{z}) = \sum_{k=0}^n a_k z^k = \sum_{k=0}^n a_k \bar{z}^k = \bar{p}(z) = 0$. Damit ist mit $x - z$ auch $x - \bar{z}$ und folglich $(x - z)(x - \bar{z}) = x^2 - 2\operatorname{Re}(z)x + |z|^2$ ein gemeinsamer Teiler von $p(x)$, $p'(x)$ und von $p_n(x)$. Wenn also p_n nicht konstant ist, aber keine reellen Nullstellen hat, dann haben p und p'



gemeinsame komplexe Nullstellen, die unser Verfahren nicht beeinträchtigen. Hat p_n jedoch reelle Nullstellen, dann sind es mehrfache von p . Ist eine solche Nullstelle für p k -fach, dann ist sie für p' nur noch $k - 1$ -fach, ebenso wie für p_n . Wenn wir also p durch p_n teilen, dann sind im Ergebnis \tilde{p} alle Nullstellen nur noch einfach und keine ist verloren gegangen. Anschließend muß die STURMSche Kette für \tilde{p} gebildet werden, um die Anzahl der reellen Nullstellen von \tilde{p} zu bestimmen, die auch die Anzahl der reellen Nullstellen von p darstellt. Der Preis dafür, daß jetzt auch mehrfache Nullstellen zulässig sind, ist also, daß das Verfahren der STURMSchen Kette möglicherweise zweimal durchgeführt werden muß, wenn sich herausstellen sollte, daß p_n nicht das konstante Polynom ist.

11.8.27 Beispiel Wir untersuchen $p(x) := \frac{1}{100}(x^8 - 2x^7 - 6x^6 + 14x^5 - 24x^4 + 34x^3 - 26x^2 + 18x - 9)$. Dann ist $p_1(x) = \frac{1}{100}(8x^7 - 14x^6 - 36x^5 + 70x^4 - 96x^3 + 102x^2 - 52x + 18)$, oder besser in gekürzter Form $p_1(x) = 4x^7 - 7x^6 - 18x^5 + 35x^4 - x48x^3 + 51x^2 - 26x + 9$.

Die STURMSche Kette:

$$\begin{aligned} p_0(x) &= x^8 - 2x^7 - 6x^6 + 14x^5 - 24x^4 + 34x^3 - 26x^2 + 18x - 9, \\ p_1(x) &= 4x^7 - 7x^6 - 18x^5 + 35x^4 - x48x^3 + 51x^2 - 26x + 9 \\ p_2(x) &= 31x^6 - 66x^5 + 157x^4 - 292x^3 + 261x^2 - 226x + 135, \\ p_3(x) &= 263x^5 - 488x^4 + 506x^3 - 506x^2 + 243x - 18, \\ p_4(x) &= -235x^4 + 622x^3 - 622x^2 + 622x - 387, \\ p_5(x) &= -x^3 + x^2 - x + 1, \\ p_6(x) &= 0. \end{aligned}$$

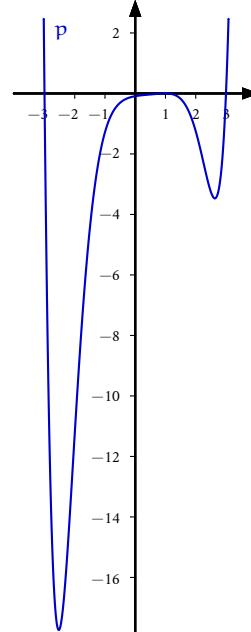
Dabei entstehen die letzten beiden Ergebnisse aus

$$\begin{aligned} p_3(x) &= -\left(\frac{263}{235}x - \frac{48906}{55225}\right)p_4(x) + \frac{19920672}{55225}(x^3 - x^2 - 1), \\ p_4(x) &= (235x - 387)(-x^3 + x^2 - x + 1). \end{aligned}$$

Der größte gemeinsame Teiler von p_0 und p_1 ist das Polynom p_5 , von dem die reelle Nullstelle 1 leicht zu sehen ist. Damit ist $p_5(x) = -(x - 1)(x^2 + 1)$ schnell gefunden. Die Stelle $x = 1$ ist eine doppelte Nullstelle von p .

Die beiden komplexen Nullstellen $\pm i$ stören das Verfahren eigentlich nicht, aber wenn wir schon mal dabei sind, dividieren wir durch $-p_5$ und sind die Vielfachheiten aller Nullstellen los. Das Verfahren startet jetzt mit einem Polynom, das nur noch den Grad 5 hat: $p_0(x) = (x^5 - x^4 - 8x^3 + 8x^2 - 9x + 9)(x^3 - x^2 + x - 1)$.

$$\begin{aligned} \tilde{p}_0(x) &= x^5 - x^4 - 8x^3 + 8x^2 - 9x + 9 \\ \tilde{p}_1(x) &= 5x^4 - 4x^3 - 24x^2 + 16x - 9 \\ \tilde{p}_2(x) &= 21x^3 - 24x^2 + 41x - 54, \\ \tilde{p}_3(x) &= 187x^2 - 150x + 27, \\ \tilde{p}_4(x) &= -115x + 189, \\ \tilde{p}_5(x) &= -1. \end{aligned}$$



Zur Abschätzung der Nullstellen von \tilde{p}_0 verwenden wir die ungefähre Lösung $R_3 = 3.8$ der Gleichung $x = 1 + \frac{8}{x} + \frac{8}{x^2} + \frac{9}{x^3} + \frac{9}{x^4}$. Die reellen Nullstellen suchen wir in $(-4, 4]$:

$$\begin{aligned}s(-4) &= [-595, \quad 1079, \quad -1946, \quad 3619, \quad 649, \quad -1], \\ s(4) &= [\quad 357, \quad 695, \quad 1070, \quad 2419, \quad -271, \quad -1].\end{aligned}$$

Es gibt demnach $v(-4) - v(4) = 4 - 1 = 3$ reelle Nullstellen von \tilde{p}_0 , von denen eine 1 ist. Das bedeutet $\tilde{p}_0(x) = (x^4 - 8x^2 - 9)(x - 1)$. Wir müssen also noch die Nullstellen des Polynoms $r_1(x) := x^4 - 8x^2 - 9$ finden. Wir wissen aber, daß darin noch $x^2 + 1$ als Teiler vorhanden sein muß: $r_1(x) = (x^2 + 1)(x^2 - 9)$. Die fehlenden Nullstellen sind daher -3 und 3 . Fassen wir zusammen, was wir herausgefunden haben: $p(x) = \frac{1}{100} (x^2 + 1)^2(x - 1)^2(x + 3)(x - 3)$.

11.8.28 Die Regel von DESCARTES Für reelle Polynome $p(x) = \sum_{k=0}^n a_k x^k$ kann man an der Anzahl der Vorzeichenwechsel benachbarter Koeffizienten a_k und a_{k+1} einen Hinweis auf die Anzahl der positiven Nullstellen erhalten. Die Regel besagt, daß die Anzahl der Vorzeichenwechsel, möglicherweise um eine gerade Zahl vermindert, die Zahl der positiven Nullstellen wiedergibt. Die Methode kann also recht ungenau sein, wie das Beispiel $p(x) = x^9 - x^8 + 4x^7 - 4x^6 + 6x^5 - 6x^4 + 4x^3 - 4x^2 + x - 1$ zeigt. Hier haben wir 9 Vorzeichenwechsel, so daß es 9, 7, 5, 3 oder nur eine positive Nullstelle geben könnte. Wenn wir das gespiegelte Polynom $\tilde{p}(x) := p(-x) = -x^9 - x^8 - 4x^7 - 4x^6 - 6x^5 - 6x^4 - 4x^3 - 4x^2 - x - 1$ untersuchen, erfahren wir etwas über die negativen Nullstellen von p . Wie man sieht, gibt es hier gar keine Vorzeichenwechsel, so daß p keine negativen Nullstellen haben kann – und diese Aussage ist nicht ungenau. Tatsächlich gibt es nur eine reelle Nullstelle, denn $p(x) = (x^2 + 1)^4(x - 1)$. Das Polynom q , das durch $q(x) := x^3 + 4x^2 - 2x - 20$ gegeben ist, hat nur einen Vorzeichenwechsel, so daß es genau eine positive Nullstelle gibt. $\tilde{q}(x) = -x^3 + 4x^2 + 2x - 20$ hat zwei Vorzeichenwechsel, d. h., q kann noch zwei negative Nullstellen haben, oder aber keine. Diese Beispiele zeigen, daß mitunter auch genaue Aussagen zu erhalten sind.

11.8.29 Satz (Regel von DESCARTES) Ein reelles Polynom $p(x) = \sum_{k=0}^n a_k x^k$, mit der Eigenschaft $p(0) \neq 0$, hat so viele positive Nullstellen wie die Anzahl der Vorzeichenwechsel benachbarter Koeffizienten von p , möglicherweise um eine gerade Zahl vermindert. Dabei wird die Vielfachheit einer Nullstelle mitgezählt. Sollten Koeffizienten a_k verschwinden, dann werden sie beim Vorzeichenwechsel ignoriert.

Bevor wir den Beweis führen, wollen wir an einigen Beispielen sehen, woher die Vorzeichenwechsel stammen können und was es mit der Verminderung der Nullstellenzahl auf sich hat. Dreh- und Angelpunkt aller Überlegungen ist das Korollar 11.8.24, nach dem jedes normierte Polynom die Form $p(x) = (x - x_1)(x - x_2) \cdots (x - x_n)$ hat, mit Nullstellen $x_i \in \mathbb{C}$. Da p reell ist, muß mit jeder komplexen Nullstelle x_k auch $\overline{x_k}$ eine Nullstelle sein. Wir multiplizieren die Linearfaktoren von $p(x)$ aus:

$$\begin{aligned}p(x) &= (x - x_1)(x - x_2)(x - x_3) \cdots (x - x_n) \\ &= x^n \\ &\quad - (x_1 + x_2 + x_3 + \cdots + x_n) x^{n-1}\end{aligned}$$

$$\begin{aligned}
& + (x_1x_2 + x_1x_2 + \dots + x_1x_k + x_2x_3 + \dots + x_2x_n + \dots + x_{n-1}x_n) x^{n-2} \\
& - (x_1x_2x_3 + x_1x_2x_4 + \dots + x_{n-2}x_{n-1}x_n) x^{n-3} \\
& \dots \\
& + (-1)^n x_1x_2 \dots x_n \\
& = \sum_{j=0}^n \left[(-1)^j \sum_{i_1 < i_2 < \dots < i_j} x_{i_1}x_{i_2} \dots x_{i_j} \right] x^{n-j} = \sum_{j=0}^n a_{n-j} x^{n-j}, \\
a_k & = (-1)^{n-k} \sum_{i_1 < i_2 < \dots < i_{n-k}} x_{i_1}x_{i_2} \dots x_{i_{n-k}}, \quad 0 \leq k < n. \tag{11.120}
\end{aligned}$$

Der Ausdruck (11.120), der die Koeffizienten von p aus den Nullstellen berechnet, ist der *Viertasche Wurzelsatz*, den die meisten (wenn überhaupt) nur für $n = 2$ kennen, wo er für eine quadratische Gleichung $x^2 + a_1x + a_0 = 0$ die möglichen Lösungen x_1 und x_2 über $a_1 = -(x_1 + x_2)$ und $a_0 = x_1x_2$ mit den beiden Koeffizienten verknüpft. Nehmen wir für den Moment an, daß alle Nullstellen positiv sind. Dann enthalten die Summen in (11.120) jeweils nur positive Einträge, so daß die a_k abwechselnd positiv und negativ sind. Bei den $n + 1$ Koeffizienten gibt es demnach genau n Vorzeichenwechsel. Diese Zahl stimmt mit der Anzahl positiver Nullstellen übereinstimmt. Sollten alle Nullstellen negativ sein, dann haben die Summanden in (11.120) die Form $|x_{i_1}| |x_{i_2}| \dots |x_{i_{n-k}}|$. Die Koeffizienten weisen jetzt keine Vorzeichenwechsel auf, im Einklang mit der Anzahl positiver Nullstellen.

Künftig bezeichnen wir mit $v(p)$ die Anzahl der Vorzeichenwechsel benachbarter Koeffizienten von p . Die Multiplikation eines normierten Polynoms p mit $x - \lambda$ erhöht die Anzahl der Vorzeichenwechsel mindestens um eine ungerade Zahl ≥ 1 , wenn λ positiv ist. Falls p keine Vorzeichenwechsel hat, folgt das, weil das absolute Glied $\tilde{a}_0 = -\lambda a_0$ in $\tilde{p}(x) := p(x)(x - \lambda) = \sum_{k=0}^{n+1} \tilde{a}_k x^k$ negativ und der Leitkoeffizient positiv ist. Es muß also mindestens einen Vorzeichenwechsel geben. Ist es wirklich nur einer, dann gibt es eine Position $k_1 < n + 1$, so daß oberhalb alle Koeffizienten nicht negativ sind, \tilde{a}_{k_1} negativ ist und alle folgenden Koeffizienten nicht mehr positiv werden. Soll eine Situation eintreten, die von der eben geschilderten verschieden ist, dann kann das nur dadurch geschehen, daß zwischen 0 und $n + 1$ eine ungerade Anzahl > 1 an Vorzeichenwechsel stattfinden, denn nur so kann man von dem positiven Vorzeichen des Leitkoeffizienten zum negativen des absoluten Gliedes gelangen. In der Situation eines beliebigen normierten Polynoms p behalten wir diese Überlegung als Muster bei. Wir zeigen, daß für jeden Vorzeichenwechsel von p einer von \tilde{p} oberhalb dieser Stelle zu finden sein wird und daß ein zusätzlicher durch \tilde{a}_0 entsteht. Damit folgt $v(\tilde{p}) \geq v(p) + 1$. Zwischen diesen Positionen, an denen mit Sicherheit ein Vorzeichenwechsel stattfinden muß, könnten aber noch weitere zu finden sein. Unsere obige Überlegung zeigt, daß es sich dabei immer um eine gerade Anzahl handeln muß, so daß sich $v(\tilde{p})$ von $v(p)$ um eine ungerade Zahl unterscheidet.

In der allgemeinen Situation eines beliebigen normierten Polynoms p müssen wir in $p(x) = \sum_{k=0}^n a_k x^k$ die Positionen der Vorzeichenwechsel markieren. Wir können davon ausgehen, daß p Vorzeichenwechsel hat. Die Positionen, an denen sie stattfinden, sind durch

$$\begin{aligned}
k_1 &:= \max \{ k < n \mid a_k < 0 \}, & k_2 &:= \max \{ k < k_1 \mid a_k > 0 \}, \\
k_3 &:= \max \{ k < k_2 \mid a_k < 0 \}, & \dots
\end{aligned}$$

gegeben. Das bedeutet $a_n = 1 > 0$, $a_k \geq 0$ für $n > k > k_1$ und $a_{k_1} < 0$. Weiter gilt $a_k \leq 0$ für $k_1 > k > k_2$ und $a_{k_2} > 0$, usw. Im Beispiel $p(x) := x^7 - 2x^6 - x^5 + 3x^4 - x^2 - x + 1$ etwa ist $k_1 = 6$, $k_2 = 4$, $k_3 = 2$ und $k_4 = 0$.

Berechnen wir jetzt die Koeffizienten \tilde{a}_k von \tilde{p} :

$$\begin{aligned}\tilde{p}(x) &= (x - \lambda) \sum_{k=0}^n a_k x^k = \sum_{k=0}^n a_k x^{k+1} - \sum_{k=0}^n \lambda a_k x^k = \sum_{k=1}^{n+1} a_{k-1} x^k - \sum_{k=0}^n \lambda a_k x^k \\ &= x^{n+1} + \sum_{k=1}^n (a_{k-1} - \lambda a_k) x^k - \lambda a_0.\end{aligned}$$

Wir erhalten $\tilde{a}_{n+1} = 1$, $\tilde{a}_k = a_{k-1} - \lambda a_k$ und $\tilde{a}_0 = -\lambda a_0$. Es gilt $\tilde{a}_{k_1+1} = a_{k_1} - \lambda a_{k_1+1} \leq a_{k_1} < 0$, denn $\lambda a_{k_1+1} \leq 0$. Also müssen zwischen $n+1$ und k_1+1 eine ungerade Anzahl an Vorzeichenwechseln auftreten, aber wenigstens der eine, der bei p zwischen n und k_1 vorhanden ist. $\tilde{a}_{k_2+1} = a_{k_2} - \lambda a_{k_2+1} \geq a_{k_2} > 0$, denn $-\lambda a_{k_2+1} \geq 0$. Das zeigt, daß auch für den Vorzeichenwechsel von p zwischen k_1 und k_2 eine ungerade Anzahl bei \tilde{p} zwischen k_2+1 und k_1+1 vorhanden sein müssen. Der letzte Vorzeichenwechsel von p findet an einer Stelle $k_r \geq 0$ statt. Er produziert eine ungerade Anzahl an Vorzeichenwechseln von \tilde{p} zwischen k_r+1 und $k_{r-1}+1$. Da a_{k_r} dasselbe Vorzeichen wie a_0 hat, gilt das auch für $\tilde{a}_{k_r+1} = a_{k_r} - \lambda a_{k_r+1}$, denn falls a_{k_r+1} nicht Null ist, hat dieser Koeffizient das entgegengesetzte Vorzeichen von a_{k_r} . Das bedeutet, daß gegenüber p wenigstens ein zusätzlicher Vorzeichenwechsel bei \tilde{p} vorhanden sein muß, nämlich einer zwischen \tilde{a}_{k_r+1} und $\tilde{a}_0 = -\lambda a_0$. Wir fassen unsere Überlegungen zu einem Lemma zusammen:

11.8.30 Lemma Sei p ein reelles normiertes Polynom, mit der Eigenschaft $p(0) \neq 0$. $v(p)$ sei die Anzahl der Vorzeichenwechsel benachbarter Koeffizienten von p . Dann gilt für das Polynom \tilde{p} , das durch $\tilde{p}(x) := p(x)(x - \lambda)$, $\lambda > 0$, definiert ist

$$v(\tilde{p}) = v(p) + 2r - 1,$$

für ein geeignetes $r \in \mathbb{N}$.

Beweis von Satz 11.8.29. Zunächst überlegen wir uns, daß für eine ungerade Zahl $v(p)$ wenigstens eine positive Nullstelle vorhanden sein muß. In diesem Fall ist nämlich $p(0) = a_0 < 0$. In $p(x) = x^n \left(1 + \frac{a_{n-1}}{x} + \frac{a_{n-2}}{x^2} + \dots + \frac{a_1}{x^{n-1}} + \frac{a_0}{x^n}\right)$ strebt die Klammer gegen 1, wenn x nur groß genug wird. Ab einem ausreichend großen x ist sie $> \frac{1}{2}$ und $p(x) > \frac{x^n}{2} > 0$. Nach dem Zwischenwertsatz 11.1.10 muß es eine positive Nullstelle von p geben. Anders sieht es aus, wenn $v(p)$ gerade ist. Hier ist nämlich $a_0 > 0$. Sollte p positive Nullstellen haben, dann gibt es eine größte x_m . Für $x > x_m$ gilt dann $p(x) > 0$. Auf dem Intervall $[0, x_m]$ nimmt p sein Minimum an. Verschieben wir das Polynom p nur weit genug nach oben, indem wir a_0 ausreichend vergrößern, dann hat das Polynom in $[0, x_m]$ und oberhalb keine Nullstellen mehr, weil das Minimum positiv geworden ist. Bei geradem $v(p)$ ist es demnach möglich, daß keine positiven Nullstellen vorhanden sind, im Einklang mit der Behauptung des Satzes. Natürlich ist das mit Sicherheit für $v(p) = 0$ der Fall.

Ab jetzt sei $v(p) > 0$ und erst einmal ungerade. Wir bezeichnen mit $n(p)$ die Anzahl positiver Nullstellen von p . Dann gilt $n(p) \geq 1$, mit einer positiven Nullstelle x_1 . Wir haben also $p(x) = p_1(x)(x - x_1)$ mit einem Polynom p_1 , für das $v(p_1)$ nach Lemma 11.8.30 um eine ungerade Zahl kleiner als $v(p)$ ist. Sollte $v(p_1) = 0$ sein, dann gibt es keine weitere positive Nullstelle. Andernfalls ist $v(p_1) > 0$ gerade, so daß weiterhin $n(p_1) = 0$ möglich ist. In beiden Fällen sind wir fertig, denn $n(p) = 1 \leq v(p)$ und $v(p) - n(p)$ ist gerade. Bleibt die Situation $v(p_1) > 0$ mit einer weiteren Nullstelle x_2 zu untersuchen (x_2 darf mit x_1 übereinstimmen). Wir haben $p_1(x) = p_2(x)(x - x_2)$, jetzt mit ungeradem $v(p_2)$. Damit hat p_2 wenigstens eine positive Nullstelle x_3 , so daß $p_1(x) = p_3(x)(x - x_2)(x - x_3)$ (diese Argumentationskette liefert die Begründung dafür, daß sich die Anzahl der Nullstellen nur um eine gerade Zahl ändern kann). $v(p_3)$ ist wieder gerade. Ist $v(p_3) = 0$, oder $v(p_3) > 0$ und $n(p_3) = 0$, dann sind wir fertig, denn wir haben $v(p) \geq 3$ und $n(p) = 3$. Andernfalls haben wir wieder dieselbe Ausgangssituation wie oben: $v(p_3) > 0$ mit einer weiteren Nullstelle x_4 . Ab hier wiederholt sich das Verfahren, jedes mal mit der Möglichkeit zweier weiterer Nullstellen. Da sich die Anzahl der Vorzeichenwechsel bei jeder Nullstelle um eine ungerade Zahl reduziert, bricht das Verfahren nach endlich vielen Schritten ab. Um es k mal zu durchlaufen muß $v(p) \geq 2k + 1$ sein und $n(p) = 2k + 1$, im Einklang mit der Behauptung des Satzes.

Ist $v(p) > 0$ und gerade, dann sind wir für $n(p) = 0$ fertig. Andernfalls gibt es, wie wir oben vorgeführt haben, eine gerade Anzahl positiver Nullstellen $n(p) \leq v(p)$. \square

11.9 Kurvendiskussion

Übersicht über das Verfahren: Wir gehen von einer Funktion f aus, die mindestens drei mal stetig differenzierbar ist. Dann befinden sich lokale Maxima und Minima von f im Inneren eines Intervalls nach Lemma 11.2.9 an Stellen x , für die $f'(x) = 0$ gilt. Die Lösungen $\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \dots$ dieser Gleichung sind die Positionen möglicher lokaler Extrema von f . Wir sprechen hier von *Extremstellen*. Die zugehörigen Punkte auf der Kurve sind die Extrema von f . Allerdings muß zunächst geklärt werden, ob überhaupt ein Extremum vorliegt und wenn ja, ob es ein lokales Maximum oder Minimum ist. Ein notwendiges und hinreichendes Kriterium dafür ist die Untersuchung von f' auf einen Vorzeichenwechsel (VZW) in einer Umgebung der betreffenden Stelle. Ein (lokales) Maximum an der Stelle \tilde{x}_2 wird zweifelsfrei dadurch identifiziert, daß die Funktion links von ihr steigt, d.h. positive Werte von $f'(x)$ aufweist und rechts von ihr fällt, was durch negative Werte von $f'(x)$ angezeigt wird. Da f' nur an seinen Nullstellen $\tilde{x}_1, \tilde{x}_2, \tilde{x}_3$ das Vorzeichen wechselt kann, hat $f'(x)$ zwischen \tilde{x}_1 und \tilde{x}_2 immer dasselbe. Daher ist es ausreichend, einen beliebigen Wert $x_{12} \in (\tilde{x}_1, \tilde{x}_2)$ und, aus demselben Grund, einen weiteren Wert $x_{23} \in (\tilde{x}_2, \tilde{x}_3)$ in f' einzusetzen, um einen eventuellen VZW von f' festzustellen. In der Skizze zeigt der Wechsel von $f'(x_{12}) > 0$ (\oplus) zu $f'(x_{23}) < 0$ (\ominus) das lokale Maximum von f an der Stelle \tilde{x}_2 an. Auf dieselbe Weise kann \tilde{x}_3 getestet werden. Das Vorzeichen von f' links von \tilde{x}_3 muß man dabei nicht neu berechnen, da es durch $f'(\tilde{x}_3) < 0$ bereits bekannt ist. In der Skizze ist $f'(\tilde{x}_{34}) > 0$, so daß der VZW von \ominus nach \oplus einen Tiefpunkt von f

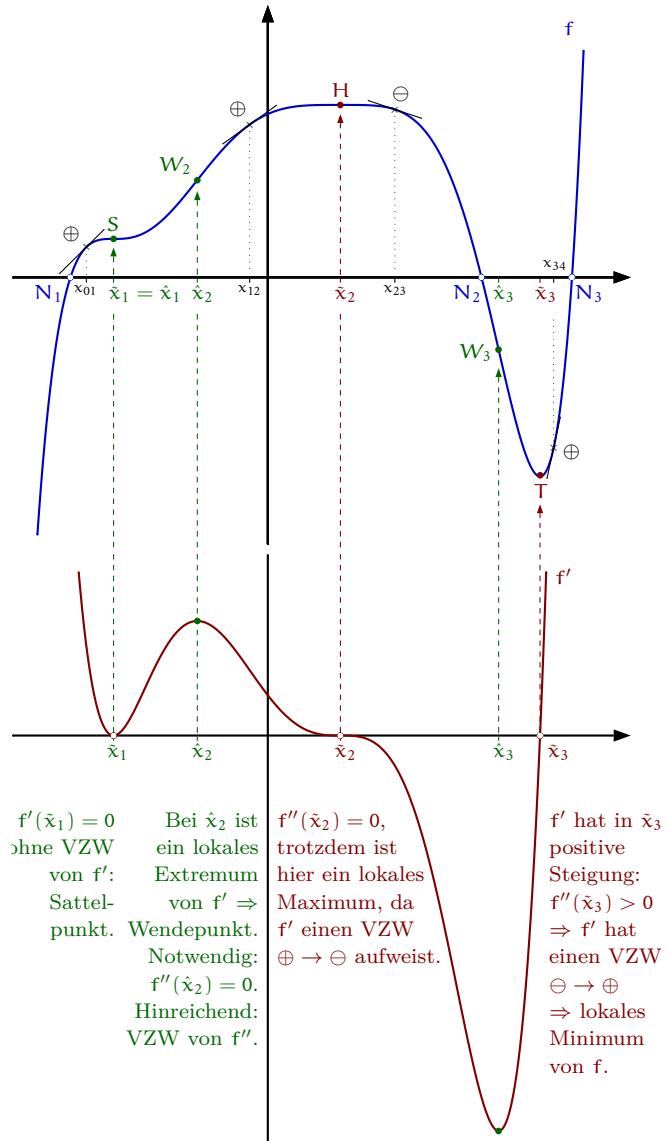


Abb. 11.9 Eine Funktion mit allen wichtigen Punkten

an der Stelle \tilde{x}_3 identifiziert. Die Werte x_{01} und x_{12} links und rechts von \tilde{x}_1 ergeben für f' dasselbe Vorzeichen \oplus , d. h. es findet kein VZW von f' statt. Trotzdem wird die Stelle \tilde{x}_1 zweifelsfrei identifiziert, und zwar als ein Wendepunkt (s. u.) mit waagrechter Tangente, also als ein *Sattelpunkt*. Als Ergebnis unserer Überlegungen können wir festhalten, daß die eben vorgestellte *Vorzeichenmethode* zur Identifikation von Extremstellen nicht versagen kann. Liegt ein VZW für f' an der betreffenden Stelle vor, dann handelt es sich um ein lokales Minimum (beim VZW $\ominus \rightarrow \oplus$) oder um ein lokales Maximum (beim VZW $\oplus \rightarrow \ominus$). Findet kein VZW statt, dann ist die Stelle trotzdem identifiziert, denn dann handelt es sich um einen Sattelpunkt (bei dem Übergang $\oplus \rightarrow \oplus$ um einen steigenden, wie bei \tilde{x}_1 , und bei $\ominus \rightarrow \ominus$ um einen fallenden).

Es gibt eine weitere Methode Extremstellen zu identifizieren. Um sicherzustellen, daß ein Vorzeichenwechsel von f' an einer Lösung von $f'(x) = 0$ vorliegt, sagen wir, bei \tilde{x}_3 , kann man prüfen, ob f' dort eine positive oder negative Steigung hat, d. h. ob $f''(\tilde{x}_3) > 0$ oder $f''(\tilde{x}_3) < 0$ gilt. \tilde{x}_3 z. B. ist eine Nullstelle von f' mit positiver Steigung (von f' wohlgemerkt). Damit muß f' hier einen VZW von \ominus nach \oplus erfahren, so daß hier ein lokales Minimum liegt. Diese sogenannte *Einsetzmethode* liefert nicht immer eine Antwort. Im Gegensatz zur Vorzeichenmethode kann sie versagen. So ist $f''(\tilde{x}_2) = 0$, denn \tilde{x}_2 ist zwar eine Nullstelle von f' , aber die Tangente von f' an dieser Stelle hat offensichtlich die Steigung Null. Trotzdem liegt hier ein lokales Maximum von f (mit der Vorzeichenmethode leicht zu identifizieren), das durch die Einsetzmethode jedoch nicht erkannt wird. Halten wir fest: Durch Einsetzen in f'' kann eine Lösung \tilde{x} von $f'(x) = 0$ als Position eines lokalen Maximums bzw. Minimums identifiziert werden, wenn $f''(\tilde{x}) < 0$ bzw. $f''(\tilde{x}) > 0$ gilt. Sollte das Ergebnis aber $f''(\tilde{x}) = 0$ sein, so ist wieder alles offen, d. h. es kann sich um einen Sattelpunkt (wie bei \tilde{x}_1), aber auch um ein lokales Minimum oder Maximum handeln (wie bei \tilde{x}_2). Es empfiehlt sich in diesem Fall mit der Vorzeichenmethode Klarheit zu schaffen.

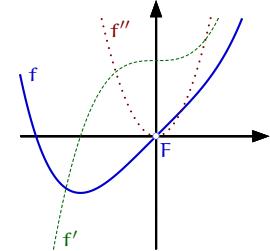
Eine Stelle \hat{x} , an der eine Kurve von einer Rechts- in eine Linkskurve oder von einer Links- in eine Rechtskurve übergeht, nennen wir *Wendestelle* (dabei denkt man sich die Funktion von links nach rechts, also entlang wachsender x -Werte durchlaufen). Der zugehörige Punkt auf der Kurve heißt *Wendepunkt*.

Untersuchen wir zunächst einmal den Übergang von einer Rechts- in eine Linkskurve: Entlang einer Rechtskurve nimmt die Steigung dauernd ab, d. h. die Ableitung f' wird entlang einer Rechtskurve permanent kleiner – solange, bis sie die Stelle erreicht hat, wo sie zur Linkskurve ansetzt. Ab da nimmt die Steigung wieder zu, denn im Verlauf einer Linkskurve wächst die Steigung. An der Wendestelle weist die Funktion demnach lokal die kleinste Steigung auf. Die Ableitung f' muß an dieser Stelle also ein lokales Minimum annehmen. Entsprechendes gilt für den Übergang von einer Linkskurve in eine Rechtskurve.

Halten wir fest: Wendepunkte sind Punkte, die lokal die größte oder kleinste Steigung aufweisen, für die f' also ein lokales Maximum oder Minimum annimmt. Das Problem, Wendepunkte zu finden und zu identifizieren ist damit auf das schon gelöste Problem zurückgeführt, Extremstellen zu finden und zu identifizieren, diesmal allerdings für die Funktion f' . Die Kandidaten für mögliche Positionen sind die Lösungen $\hat{x}_1, \hat{x}_2, \hat{x}_3, \dots$ der Gleichung $f''(x) = 0$, denn

als Extremstelle der Funktion f' muß ihre Ableitung verschwinden. Um sicherzugehen, daß es sich tatsächlich um Wendepunkte handelt, muß nachgewiesen werden, daß auch wirklich lokale Maxima oder Minima von f' vorliegen. d. h. die Ableitung von f' , also f'' , muß in einer Umgebung der betreffenden Punkte einen Vorzeichenwechsel aufweisen. Dabei ist es meist nicht so wichtig, welcher Art dieser genau ist, denn er unterscheidet, ob es sich um einen Übergang von einer Rechts- in eine Linkskurve oder umgekehrt handelt.

In der Skizze 11.9 ist $\tilde{x}_1 = \tilde{x}_1$ bereits bei der Identifikation möglicher Extremstellen als Position eines Sattelpunktes erkannt worden. \tilde{x}_2 und \tilde{x}_3 lassen sich mit der beschriebenen Methode am VZW von f'' als Positionen von Wendepunkten identifizieren. Für \tilde{x}_2 , das auch eine Lösung von $f''(x) = 0$ ist, würde f'' keinen VZW ergeben, so daß hier keine Extremstelle von f' und daher auch kein Wendepunkt von f vorliegt. Allerdings wurde \tilde{x}_2 in diesem Beispiel bereits als x -Koordinate einer Extremstelle identifiziert. Es kann aber vorkommen, daß eine Lösung von $f''(x) = 0$ keinen VZW bei f'' verursacht, ohne daß sie eine Extremstelle ist. Dann handelt es sich um einen sogenannten *Flachpunkt* von f . Ein einfaches Beispiel ist etwa $f(x) := \frac{1}{4}x^4 + x$, mit $f'(x) = x^3 + 1$ und $f''(x) = 3x^2$. Die Stelle $F = (0, 0)$ ist ein Flachpunkt von f .



Die Identifikation eines Wendepunktes, also die Identifikation einer Extremstelle von f' , kann auch mit der oben vorgestellten Einsetzmethode erfolgen. Man setzt die betreffende Stelle \tilde{x} in die zweite Ableitung von f' , also in f''' ein und hofft, daß $f'''(\tilde{x}) \neq 0$ ist. Ist das der Fall, so liegt sicher ein Wendepunkt vor. Wenn nicht, sollte man sich auf nichts einlassen und zur sicheren Vorzeichenmethode zurückkehren.

11.9.1 Kurvendiskussion -- die einzelnen Schritte

1. Nullstellen: Löse die Gleichung $f(x) = 0$.

Zu deren Lösungen x_1, x_2, \dots gehören die Nullstellen $N_1 = [x_1, 0], N_2 = [x_2, 0], \dots$

2. Extremstellen: Löse die Gleichung $f'(x) = 0$.

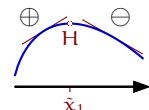
Bestimme zu den Lösungen $\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \dots$ (den möglichen Stellen der lokalen Maxima, Minima oder Sattelpunkte) die zugehörigen y -Werte $\tilde{y}_1 = f(\tilde{x}_1), \tilde{y}_2 = f(\tilde{x}_2), \tilde{y}_3 = f(\tilde{x}_3), \dots$

Identifikation:

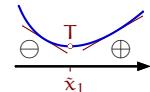
a) Vorzeichenmethode: (notwendig und hinreichend)

Setze einen x -Wert links von \tilde{x}_1 und einen rechts von \tilde{x}_1 in f' ein (die Punkte x sind beliebig wählbar, solange sie sich nicht jenseits benachbarter möglicher Extremstellen befinden) und stelle jeweils das Vorzeichen fest.

VZW für f' $\oplus \rightarrow \ominus$: lokales Max. $H = [\tilde{x}_1, \tilde{y}_1]$

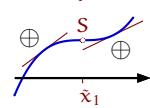


VZW für $f q' \ominus \rightarrow \oplus$: lokales Min. $T = [\tilde{x}_1, \tilde{y}_1]$

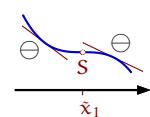


kein

VZW für $f' \oplus \rightarrow \oplus$: Sattelpunkt $S = [\tilde{x}_1, \tilde{y}_1]$



$\ominus \rightarrow \ominus$: Sattelpunkt $S = [\tilde{x}_1, \tilde{y}_1]$



b) alternativ: (nur hinreichend)

Setze \tilde{x}_1 in f'' ein.

$f''(\tilde{x}_1) < 0$: lokales Maximum $H = [\tilde{x}_1, \tilde{y}_1]$,

$f''(\tilde{x}_1) > 0$: lokales Minimum $T = [\tilde{x}_1, \tilde{y}_1]$,

$f''(\tilde{x}_1) = 0$: Teste auf Wendepunkt (s.u.). Falls erfolgreich,
dann liegt ein Sattelpunkt $S = [\tilde{x}_1, \tilde{y}_1]$ vor.

Verfahren genauso mit den restlichen Lösungen $\tilde{x}_2, \tilde{x}_3, \dots$

3. Wendestellen: Löse die Gleichung $f''(x) = 0$.

Bestimme zu den Lösungen $\hat{x}_1, \hat{x}_2, \hat{x}_3, \dots$ (den möglichen Stellen der Wendepunkte) die zugehörigen y -Werte $\hat{y}_1 = f(\hat{x}_1), \hat{y}_2 = f(\hat{x}_2), \hat{y}_3 = f(\hat{x}_3), \dots$

Identifikation:

a) Vorzeichenmethode: (notwendig und hinreichend)

Setze einen x -Wert links von \hat{x}_1 und einen rechts von \hat{x}_1 in f'' ein (die Punkte x sind beliebig wählbar, solange sie sich nicht jenseits benachbarter möglicher Wendestellen befinden) und stelle jeweils das Vorzeichen fest.

VZW für f'' : Wendepunkt $W = [\hat{x}_1, \hat{y}_1]$.

Kein

VZW für f'' : Flachpunkt $F = [\hat{x}_1, \hat{y}_1]$.

b) alternativ: (nur hinreichend)

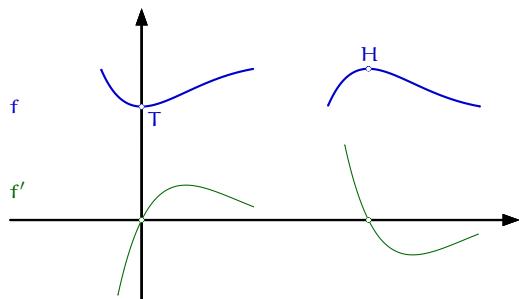
Setze \hat{x}_1 in f''' ein.

$f'''(\hat{x}_1) \neq 0$: Wendepunkt $W = [\hat{x}_1, \hat{y}_1]$.

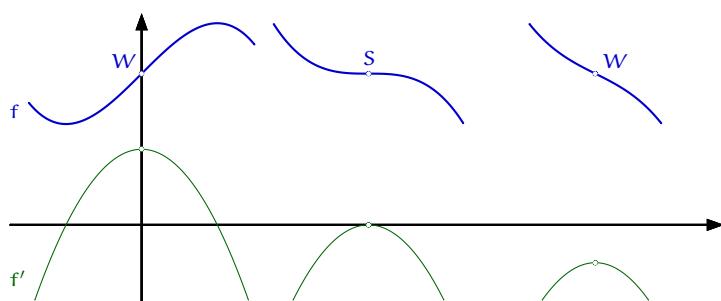
Verfahren genauso mit den restlichen Lösungen $\hat{x}_2, \hat{x}_3, \dots$

Kurzfassung:

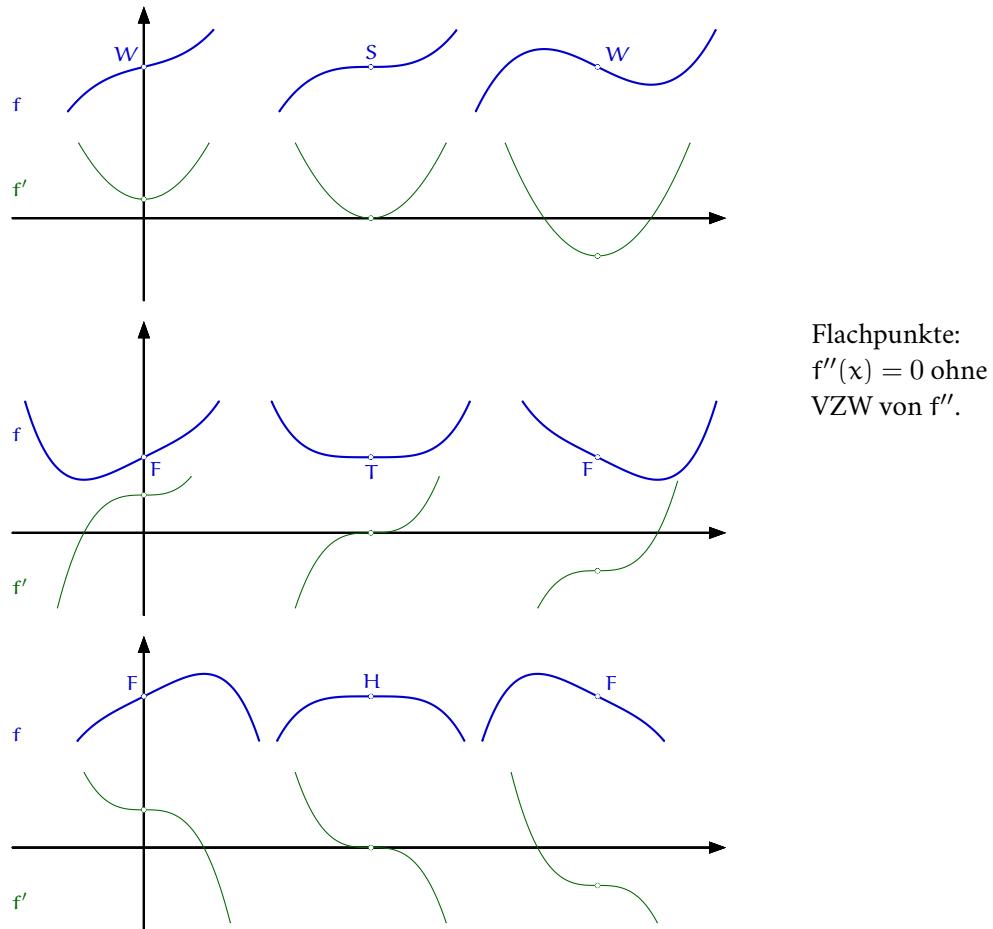
Markante Punkte	mögliche Stellen x	Art	Identifikation	
			notwendig und hinreichend	hinreichend
1. Nullstellen	$f(x) = 0$			
2. Extrempunkte	$f'(\tilde{x}) = 0$	lokales Maximum	VZW von f' bei \tilde{x} ⊕ → ⊖ <i>bergauf - bergab</i>	$f''(\tilde{x}) < 0$
		lokales Minimum	VZW von f' bei \tilde{x} ⊖ → ⊕ <i>bergab - bergauf</i>	$f''(\tilde{x}) > 0$
		Sattelpunkt	kein VZW von f' bei \tilde{x} ⊕ → ⊕ ⊖ → ⊖	$f''(\tilde{x}) = 0$ und $f'''(\tilde{x}) \neq 0$
3. Wendepunkte	$f''(\hat{x}) = 0$	Wendepunkt	VZW von f'' bei \hat{x}	$f'''(\hat{x}) \neq 0$
4. Flachpunkte	$f''(\hat{x}) = 0$	Flachpunkt	kein VZW von f'' bei \hat{x}	

11.9.2 Übersicht ausgezeichneter Kurvenpunkte

Extrempunkte:
 $f'(x) = 0$ und
VZW von f' .



Wendepunkte:
 $f''(x) = 0$ und
VZW von f'' .



11.9.3 A Führen Sie für die folgenden Funktionen eine vollständige Kurvendiskussion durch.

i) $f(x) := \frac{1}{40} (x^4 - 26x^2 + 48x - 23)$

ii) $f(x) := \frac{e^{-x} + x - 1}{e^{-x} - 2}$

iii) $f(x) := \frac{x^3}{x^2 + 1}$
Alle Tangenten durch $[0, -0.5]$.

iv) $f(x) := \frac{3x^3}{3x^2 - 4}$
Alle Tangenten durch $[0.5, 0]$.

v) $f(x) := \arccos(\cos(x))$

vi) $f(x) := 2 + \frac{1}{2} \sin(x) - \cos\left(\frac{x}{2}\right)$

vii) $f(x) := e^{\alpha x} \sin(x), \alpha > 0$

viii) $f(x) := x \arctan(x)$

ix) $f(x) := \begin{cases} \frac{\sin(x)}{x}, & x \neq 0 \\ 1, & x = 0 \end{cases}$

x) $f(x) := \frac{1}{1 + e^{2x}}$

xi) $f(x) := e^x (5 + 2 \cos(2x) - \sin(2x))$

xii) $f(x) := \frac{e^{-x}}{\sin(x) + \cos(x)}$

xiii) $f(x) := \frac{(x+8)^2}{\sqrt{x^2+8}}$
Alle Tangenten durch $[-8, 0]$.

xiv) $f(x) := \frac{1}{24} (x^4 - 32x + 48)$
Alle Tangenten durch $[0, 0]$.

xv) $f(x) := \begin{cases} \frac{3}{2}x + e^{-\frac{6}{x^2} + \frac{3}{2}}, & x \neq 0 \\ 0, & x = 0 \end{cases}$

xvi) $f(x) := \frac{4e^{-x}}{4 - x^3}$

punkt \mathbf{m} , sowie die Asymptoten).

xvii) $f(x) := \sqrt{2x^2 - \sqrt{6}x} - \sqrt{2}x.$

Zeigen Sie, daß f Teil einer Hyperbel ist.
Bestimmen Sie deren Quadrik und finden Sie alle Bestimmungsstücke (a, b , die Brennpunkte $f_{1/2}$ und den Mittel-

xviii) $f(x) := \sin(x) \ln(\sin^2(x))$

Schließen Sie die Definitionslücken.

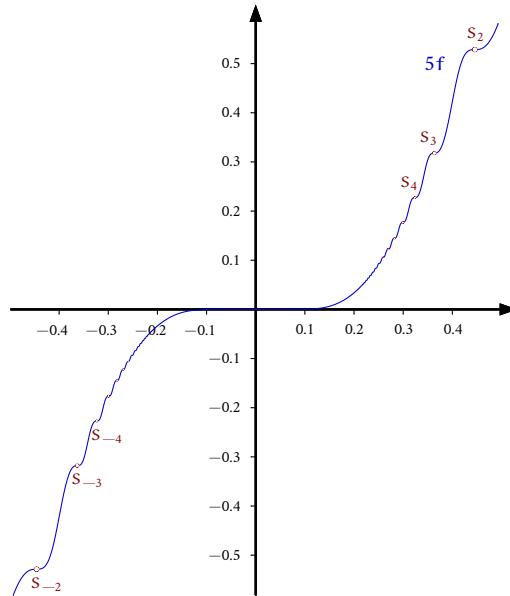
xix) $f(x) := e^{\tan(x)}$

Schließen Sie alle Definitionslücken.

11.9.4 A Man unterteile ein gleichschenkliges, rechtwinkliges Dreieck durch einen geraden Schnitt minimaler Länge in zwei inhaltsgleiche Bereiche.

11.9.5 Bemerkung Das vorgestellte Verfahren zur Kurvendiskussion ist unter der Voraussetzung korrekt, daß sich die Nullstellen, die Extrema und die Wendepunkte nicht häufen. Das heißt, in jedem beschränkten Intervall haben die Gleichungen $f(x) = 0$, $f'(\tilde{x}) = 0$ und $f''(\hat{x}) = 0$ jeweils nur endlich viele Lösungen. Das ist bei gängigen Funktionen der Normalfall. Andernfalls kann es lokale Maxima (oder Minima) geben, bei denen die Funktion vor der betreffenden Stelle nicht monoton wachsend und nach ihr nicht monoton fallend ist. Beispiel 11.2.18 zeigt eine solche Situation. Die Methode des Vorzeichenwechsels von f' zur Identifikation versagt hier. Es ist eben nicht möglich, einen Wert links der betreffenden Stelle einzusetzen, der nicht jenseits weiterer Kandidaten für lokale Maxima oder Minima liegt, da es in jeder Umgebung unendlich viele davon gibt. Entsprechendes gilt für Wendepunkte. Ein Beispiel ist

$$f(x) := \begin{cases} \arccot\left(e^{\frac{1}{|x|}} + \sin\left(e^{\frac{1}{|x|}}\right)\right) \cdot \operatorname{sgn}(x) & , x \neq 0 \\ 0 & , x = 0 \end{cases} \quad (11.121)$$



In jeder ε -Umgebung $(-\varepsilon, \varepsilon)$ von $x = 0$ befinden sich unendlich viele Sattelpunkte. Den Punkt $[0, 0]$ kann man nicht mehr als Sattelpunkt bezeichnen. Ein Sattelpunkt ist insbesondere ein Wendepunkt, d. h., der Kurvenverlauf hat von einer Rechts- in eine Linkskurve zu wechseln, oder umgekehrt. Diese Eigenschaft kann man dem Kurvenpunkt $[0, 0]$ nicht mehr sinnvoll zuordnen, da in jeder ε -Umgebung unendlich viele Änderungen des Kurvenverlaufs stattfinden. Es gibt keine letzte Position links von 0, ab der man von einer Rechtskurve, oder von einer Linkskurve reden könnte.

11.9.6 A Zeigen Sie, daß f stetig differenzierbar ist und daß die Sattelpunkte durch

$$S_n := \left[\frac{1}{\ln((2n-1)\pi)}, \arccot((2n-1)\pi) \right] \text{ und } S_{-n} := \left[\frac{-1}{\ln((2n-1)\pi)}, -\arccot((2n-1)\pi) \right]$$

gegeben sind ($n \in \mathbb{N}$).

11.9.7 Konvexe Funktionen*

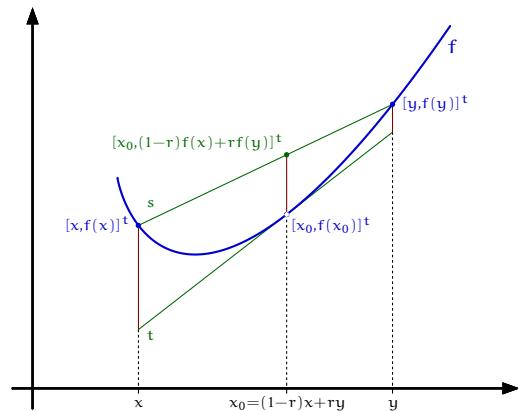
11.9.8 Definition Eine Funktion $f: [a, b] \rightarrow \mathbb{R}$ heißt konvex (konkav), falls für alle $x \neq y \in [a, b]$ und alle $r \in (0, 1)$ die folgende Ungleichung erfüllt ist

$$f((1-r)x + ry) \stackrel{(\geq)}{\leq} (1-r)f(x) + rf(y). \quad (11.122)$$

Sie heißt strikt konvex (konkav), falls für alle $r \in (0, 1)$ das $<$ -Zeichen ($>$ -Zeichen) gilt.

Anschaulich bedeutet die Konvexität einer Funktion, daß sie nach links gekrümmmt ist, daß sie also im Durchlaufsinn von links nach rechts eine *Linkskurve* beschreibt. Das bedeutet, daß die Funktion immer unterhalb jeder ihrer Sehnenabschnitte verläuft. Genauer: Für je zwei Punkte $x < y, x, y \in [a, b]$, ist die zugehörige *Sehne* s von f die Gerade, die die beiden Kurvenpunkte $[x, f(x)]^t$ und $[y, f(y)]^t$ verbindet (vergl. die Skizze). Ihre Gleichung lautet für $x_0 \in [x, y]$:

$$\begin{aligned} x_0 \mapsto s(x_0) &= (1 - \frac{x_0 - x}{y - x})f(x) + \frac{x_0 - x}{y - x}f(y) \\ &= (1 - r)f(x) + rf(y), \end{aligned}$$



mit $r := \frac{x_0 - x}{y - x} \in [0, 1]$. Verläuft f also unterhalb von s , so muß $f(x_0) \leq s(x_0)$ gelten. Das ist aber bereits (11.122), wenn man sich noch $(1-r)x + ry = x + r(y-x) = x_0$ vergegenwärtigt. Andererseits ist (11.122) nichts anderes als $f(x_0) \leq s(x_0)$. Die Eigenschaft unterhalb ihrer Sehnenabschnitte zu verlaufen, ist für eine Funktion demnach äquivalent zur Konvexität.

Es ist leicht einzusehen, daß eine Funktion f genau dann konkav ist, wenn $-f$ konvex ist. Eine konkave Funktion verläuft daher immer oberhalb ihrer Sehnenabschnitte und beschreibt eine Rechtskurve.

11.9.9 Satz Die Funktion f sei auf dem Intervall (a, b) stetig differenzierbar. Dann ist folgendes äquivalent:

- i) f' ist auf (a, b) streng monoton wachsend (fallend).
- ii) f verläuft auf (a, b) , von den Berührpunkten abgesehen, immer strikt oberhalb (unterhalb) jeder ihrer Tangenten.
- iii) f ist auf (a, b) strikt konvex (konkav).

Beweis. Wie beweisen nur die Aussagen über konvexe Funktionen. Ist f konkav, dann erhält man die Behauptungen aus denen für die konvexe Funktion $-f$.

i) \Rightarrow ii): Wir wählen ein beliebiges $x_0 \in (a, b)$ und wollen zeigen, daß für alle $x \neq x_0$ immer $f(x) > t(x)$ gilt, wenn $x \mapsto t(x) := f(x_0) + f'(x_0)(x - x_0)$ die Tangente von f an der Stelle x_0 ist. Nach dem Mittelwertsatz 11.2.11 gilt

$$f(x) - t(x) = f(x) - f(x_0) - f'(x_0)(x - x_0) = (f'(\xi) - f'(x_0))(x - x_0),$$

mit einem $\xi \in (x, x_0)$ bzw. $\xi \in (x_0, x)$, je nach Lage von x relativ zu x_0 . Im ersten Fall ist $\xi < x_0$ und daher $f'(\xi) < f'(x_0)$. Damit ist $f'(\xi) - f'(x_0) < 0$ und $x - x_0 < 0$, so daß $f(x) - t(x) > 0$ folgt. Genauso argumentiert man im zweiten Fall $x > x_0$ und erhält dasselbe Ergebnis.

ii) \Rightarrow iii): Für $x, y \in (a, b)$, $x < y$ und $s \in (0, 1)$ sei $x_0 := (1 - s)x + sy$. Dann gilt $x = (1 - s)x + sx < (1 - s)x + sy = x_0 < (1 - s)y + sy = y$. Bestimmen wir zunächst

$$\begin{aligned} (1 - s)t(x) + st(y) &= (1 - s)(f(x_0) + f'(x_0)(x - x_0)) + s(f(x_0) + f'(x_0)(y - x_0)) \\ &= f(x_0) - f'(x_0)x_0 + f'(x_0)((1 - s)x + sy) = f(x_0). \end{aligned}$$

Mit $t(x) < f(x)$ und $t(y) < f(y)$ folgt daraus bereits die strikte Konvexität von f :

$$f((1 - s)x + sy) = f(x_0) = (1 - s)t(x) + st(y) < (1 - s)f(x) + sf(y).$$

iii) \Rightarrow i): f sei strikt konvex und $x < y$. Wir müssen $f'(x) < f'(y)$ zeigen. Wir wählen dafür zunächst ein $x_0 \in (x, y)$. Dann gilt $x_0 = (1 - t)x + ty$, mit $t = \frac{x_0 - x}{y - x} \in (0, 1)$. Wir erhalten

$$\begin{aligned} f(x_0) &< (1 - t)f(x) + tf(y) = \frac{y - x_0}{y - x}f(x) + \frac{x_0 - x}{y - x}f(y) \Rightarrow \\ 0 &< \frac{y - x_0}{y - x}(f(x) - f(x_0)) + \frac{x_0 - x}{y - x}(f(y) - f(x_0)) \Rightarrow \\ (x_0 - y)(f(x) - f(x_0)) &< (x_0 - x)(f(y) - f(x_0)) \Rightarrow \\ \frac{f(x) - f(x_0)}{x_0 - x} &> \frac{f(y) - f(x_0)}{x_0 - y} \Rightarrow \\ \frac{f(x) - f(x_0)}{x - x_0} &< \frac{f(x_0) - f(y)}{x_0 - y}. \end{aligned}$$

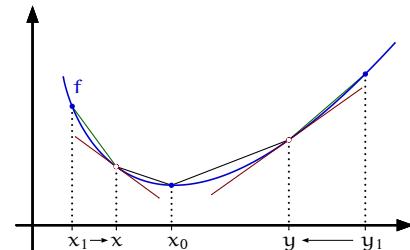
Jetzt wählen wir noch ein $x_1 < x$ und ein $y_1 > y$ und wenden die eben gefundene Abschätzung auf die Ausgangslagen $x_1 < x < x_0$, $x < x_0 < y$ und $x_0 < y < y_1$ an:

$$\frac{f(x_1) - f(x)}{x_1 - x} < \frac{f(x) - f(x_0)}{x - x_0} < \frac{f(x_0) - f(y)}{x_0 - y} < \frac{f(y) - f(y_1)}{y - y_1}.$$

Für $x_1 \rightarrow x$ und $y_1 \rightarrow y$ konvergiert der erste und der letzte Ausdruck gegen $f'(x)$ bzw. $f'(y)$. Das bedeutet

$$f'(x) \leq \frac{f(x) - f(x_0)}{x - x_0} < \frac{f(x_0) - f(y)}{x_0 - y} \leq f'(y),$$

also $f'(x) < f'(y)$.



□

11.9.10 Korollar f sei auf (a, b) zweimal differenzierbar. Es gelte dort $f''(x) \geq 0$ (≤ 0) und $f''(x) = 0$ nur für endlich viele Stellen x . Dann ist f auf (a, b) strikt konvex (konkav).

Beweis. Nach Satz 11.4.1 ist f' auf (a, b) streng monoton wachsend. \square

Im Lichte unserer Überlegungen sind bei einer zweimal stetig differenzierbaren Funktion f die Wendepunkte dort zu finden, wo sie von einem strikt konvexen (mit $f''(x) > 0$) in einen strikt konkaven Abschnitt (mit $f''(x) < 0$), oder von einem strikt konkaven in einen strikt konvexen übergeht. Der Übergang von $f''(x) > 0$ zu $f''(x) < 0$, oder umgekehrt ist natürlich nur an Stellen mit $f''(x) = 0$ möglich. Damit haben wir noch einmal von einem anderen Standpunkt aus unser Verfahren zum Auffinden von Wendepunkten gerechtfertigt: Löse $f''(x) = 0$ und identifizierte die Lösungen, bei denen f'' einen Vorzeichenwechsel erfährt.

11.9.11 Beispiel

Die Exponentialfunktion \exp ist auf ganz \mathbb{R} strikt konvex, denn es gilt $\exp''(x) = \exp(x) > 0$ für alle $x \in \mathbb{R}$. Also verläuft \exp oberhalb ihrer Tangenten, also etwa oberhalb der Tangente t_1 an der Stelle 0, $t_1(x) = x + 1$. Das ergibt eine Abschätzung für die Exponentialfunktion:

$$\exp(x) > 1 + x, \quad x \neq 0. \quad (11.123)$$

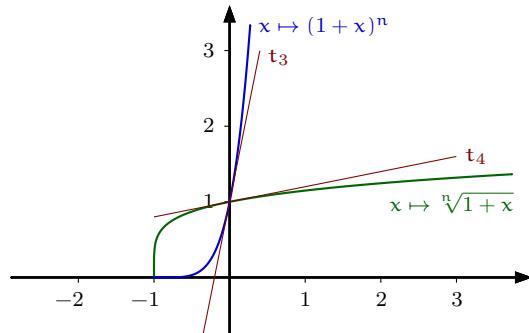
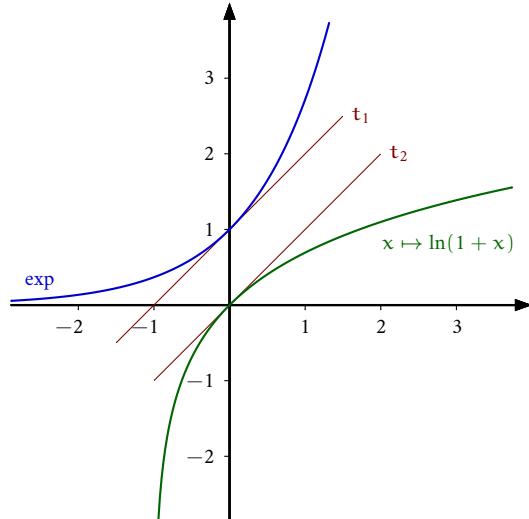
Die Funktion $f(x) := \ln(1 + x)$ ist für alle $x > -1$ strikt konkav, denn $f''(x) = -\frac{1}{(1+x)^2} < 0$. Ihre Tangente an der Stelle 0 hat die Gleichung $t_2(x) = x$. Es gilt demnach die Abschätzung

$$\ln(1 + x) < x, \quad -1 < x \neq 0. \quad (11.124)$$

Die Funktion $g(x) := (1 + x)^n$, $n \geq 2$, ist für $x > -1$ strikt konvex, denn $g''(x) = n(n-1)(1+x)^{n-2} > 0$. Die Abschätzung für diese Funktion ist die *BERNOULLI-Ungleichung* $(1 + x)^n > 1 + nx$, $-1 < x \neq 0$. Die Tangente t_3 an der Stelle 0 hat die Gleichung $t_3(x) = 1 + nx$. Betrachten wir auch die Umkehrfunktion: $h(x) := \sqrt[n]{1+x} = (1+x)^{\frac{1}{n}}$. Ihre zweite Ableitung ist $-\frac{n-1}{n^2}(1+x)^{\frac{1}{n}-2} < 0$, für $x > -1$. h ist für $x > -1$ also strikt konkav.

Die Tangente t_4 an der Stelle 0 lautet $t_4(x) = 1 + \frac{1}{n}x$. Die zugehörige Abschätzung ist die sogenannte *umgekehrte BERNOULLI-Ungleichung*

$$(1 + x)^{\frac{1}{n}} < 1 + \frac{1}{n}x, \quad -1 < x \neq 0. \quad (11.125)$$



11.9.12 A Untersuchen Sie die Funktion $f(x) := \cos(x) - 1 + \frac{x^2}{2}$ auf Konvexität. Finden Sie eine Abschätzung für $\cos(x)$, indem Sie eine geeignete Tangente wählen. Beantworten Sie damit die Frage: Wieviele Lösungen hat die Gleichung $2 - 2\cos(x) = x^2$?

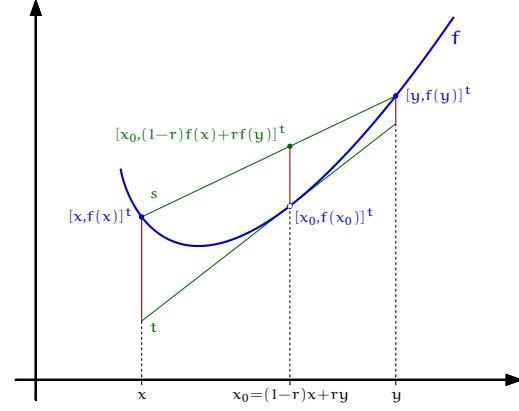
11.9.13 JENSENSche Ungleichung

\mathbf{m} sei eine Konvexitätskombination der Kurvenpunkte $\mathbf{p}_i := [a_i, f(a_i)]^t$ von f . Das bedeutet, für Zahlen $\lambda_i \in (0, 1)$, $i = 1, \dots, n$, mit $\sum_{i=1}^n \lambda_i = 1$ ist

$$\mathbf{m} := \sum_{i=1}^n \lambda_i \mathbf{p}_i = \left[\sum_{i=1}^n \lambda_i a_i, \sum_{i=1}^n \lambda_i f(a_i) \right]^t.$$

Verwendet man $\lambda_1 = 1 - \sum_{i=2}^n \lambda_i$, dann kann man das auch in der Form

$$\mathbf{m} = \mathbf{p}_1 + \sum_{i=2}^n \lambda_i (\mathbf{p}_i - \mathbf{p}_1)$$



schreiben. Das zeigt, daß \mathbf{m} entsteht, indem zu \mathbf{p}_1 die λ_i -Anteile der Verbindungsvektoren $\mathbf{p}_i - \mathbf{p}_1$ addiert werden. Ist f konvex, so erhält man auf diese Weise immer einen Punkt aus dem Bereich, der von dem Graphen zwischen \mathbf{p}_1 und \mathbf{p}_n und der Geraden zwischen \mathbf{p}_1 und \mathbf{p}_n berandet wird (vorausgesetzt \mathbf{p}_1 ist der am weitesten links und \mathbf{p}_n der am weitesten rechts liegende Punkt). Bei $\mathbf{p} := [\sum_{i=1}^n \lambda_i a_i, f(\sum_{i=1}^n \lambda_i a_i)]^t$ handelt es sich dagegen um den Kurvenpunkt, der zur Konvexitätskombination $\mathbf{m} := \sum_{i=1}^n \lambda_i \mathbf{a}_i$ der x -Werte a_1, \dots, a_n gehört. Er muß auf dem Rande besagten Bereichs liegen. Da \mathbf{p} und \mathbf{m} denselben x -Wert m haben, kann der y -Wert von \mathbf{p} nicht größer als der von \mathbf{m} sein. Das ist der Inhalt der JENSENSchen Ungleichung.

11.9.14 Satz (JENSENSche Ungleichung) Sei $f: [a, b] \rightarrow \mathbb{R}$ konvex (konkav), $a_i \in [a, b]$, $\lambda_i \in (0, 1)$ für $i = 1, \dots, n$, mit der Eigenschaft $\sum_{i=1}^n \lambda_i = 1$. Dann gilt

$$f\left(\sum_{i=1}^n \lambda_i a_i\right) \stackrel{(\geq)}{\leq} \sum_{i=1}^n \lambda_i f(a_i). \quad (11.126)$$

Ist f strikt konvex (konkav), dann gilt in (11.126) genau dann das Gleichheitszeichen, wenn $a_1 = a_2 = \dots = a_n$ erfüllt ist.

Beweis. Der Beweis für die Ungleichung erfolgt mit vollständiger Induktion nach $n \geq 2$ und nur für den Fall einer konvexen Funktion f . Für $n = 2$ ist Ungleichung (11.126) einfach die Konvexität von f . Der Schritt von n nach $n + 1$: Dafür wird $1 - \lambda_1 = \sum_{i=2}^{n+1} \lambda_i$ verwendet:

$$f\left(\sum_{i=1}^{n+1} \lambda_i a_i\right) = f\left(\lambda_1 a_1 + \sum_{i=2}^{n+1} \lambda_i a_i\right) = f\left(\lambda_1 a_1 + (1 - \lambda_1) \sum_{i=2}^{n+1} \frac{\lambda_i}{1 - \lambda_1} a_i\right)$$

$$\begin{aligned} &\leq \lambda_1 f(a_1) + (1 - \lambda_1) f\left(\sum_{i=2}^{n+1} \frac{\lambda_i}{1-\lambda_1} a_i\right) \leq \lambda_1 f(a_1) + (1 - \lambda_1) \sum_{i=2}^{n+1} \frac{\lambda_i}{1-\lambda_1} f(a_i) \\ &= \lambda_1 f(a_1) + \sum_{i=2}^{n+1} \lambda_i f(a_i) = \sum_{i=1}^{n+1} \lambda_i f(a_i). \end{aligned}$$

Dabei stammt die erste Abschätzung aus der Konvexität von f und die zweite aus der Induktionsvoraussetzung, da die betreffende Summe n Summanden hat und $\sum_{i=2}^{n+1} \frac{\lambda_i}{1-\lambda_1} = 1$ gilt.

Nun sei f strikt konvex. In (11.126) gilt sicher das Gleichheitszeichen, wenn $a_1 = a_2 = \dots = a_n$ erfüllt ist. Wir müssen die Umkehrung zeigen. Es gelte also $f\left(\sum_{i=1}^n \lambda_i a_i\right) = \sum_{i=1}^n \lambda_i f(a_i)$. Für $n = 2$ und $a_1 \neq a_2$ ergibt sich sofort ein Widerspruch zur strikten Konvexität von f :

$$\lambda_1 f(a_1) + (1 - \lambda_1) f(a_2) = f(\lambda_1 a_1 + (1 - \lambda_1) a_2) < \lambda_1 f(a_1) + (1 - \lambda_1) f(a_2).$$

Für $n > 2$: Sollten nicht alle a_i gleich sein, dann können wir o. B. d. A. $a_1 \neq a_2$ annehmen und erhalten folgenden Widerspruch:

$$\begin{aligned} \sum_{i=1}^{n+1} \lambda_i f(a_i) &= f\left((\lambda_1 + \lambda_2)\left(\frac{\lambda_1}{\lambda_1 + \lambda_2} a_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} a_2\right) + (1 - \lambda_1 - \lambda_2) \sum_{i=3}^n \frac{\lambda_i}{1 - \lambda_1 - \lambda_2} a_i\right) \\ &\leq (\lambda_1 + \lambda_2) f\left(\frac{\lambda_1}{\lambda_1 + \lambda_2} a_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} a_2\right) + (1 - \lambda_1 - \lambda_2) f\left(\sum_{i=3}^n \frac{\lambda_i}{1 - \lambda_1 - \lambda_2} a_i\right) \\ &\leq (\lambda_1 + \lambda_2) f\left(\frac{\lambda_1}{\lambda_1 + \lambda_2} a_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} a_2\right) + \sum_{i=3}^n \lambda_i f(a_i) \\ &< (\lambda_1 + \lambda_2) \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} f(a_1) + \frac{\lambda_2}{\lambda_1 + \lambda_2} f(a_2)\right) + \sum_{i=3}^n \lambda_i f(a_i) = \sum_{i=1}^n \lambda_i f(a_i). \end{aligned}$$

In der letzten Abschätzung wurde die strikte Konvexität von f und $a_1 \neq a_2$ verwendet. \square

11.9.15 Korollar Es seien $a_1, \dots, a_n, \lambda_1, \dots, \lambda_n$ positive Zahlen mit der Eigenschaft $\sum_{i=1}^n \lambda_i = 1$. Dann haben wir folgende Abschätzung:

$$a_1^{\lambda_1} a_2^{\lambda_2} \cdots a_n^{\lambda_n} \leq \sum_{i=1}^n \lambda_i a_i. \quad (11.127)$$

Gleichheit gilt genau für $a_1 = a_2 = \dots = a_n$.

Im Spezialfall $\lambda_1 = \lambda_2 = \dots = \lambda_n = \frac{1}{n}$ lautet die Ungleichung

$$\sqrt[n]{a_1 a_2 \cdots a_n} \leq \frac{1}{n} \sum_{i=1}^n a_i \quad (11.128)$$

und besagt, daß das geometrische Mittel $\sqrt[n]{a_1 a_2 \cdots a_n}$ immer unterhalb des arithmetischen Mittels $\frac{1}{n} \sum_{i=1}^n a_i$ liegt, falls nicht alle a_i gleich sind.

Beweis. Die Funktion \ln ist strikt konkav, denn $\ln''(x) = -\frac{1}{x^2} < 0$. Also muß

$$\ln(a_1^{\lambda_1} a_2^{\lambda_2} \cdots a_n^{\lambda_n}) = \sum_{i=1}^n \lambda_i \ln(a_i) \leq \ln\left(\sum_{i=1}^n \lambda_i a_i\right)$$

gelten. Da die Exponentialfunktion $x \mapsto e^x$ streng monoton wachsend ist, folgt die behauptete Ungleichung (11.127). Sollte hier das Gleichheitszeichen stehen, dann haben wir durch Anwendung von \ln auch $\sum_{i=1}^n \lambda_i \ln(a_i) = \ln(\sum_{i=1}^n \lambda_i a_i)$, woraus sofort $a_1 = a_2 = \dots = a_n$ gefolgert werden kann, denn \ln ist ja strikt konkav. \square

11.9.16 Satz (Youngsche Ungleichung) Für reelle Zahlen $p, q > 1$ mit der Eigenschaft $\frac{1}{p} + \frac{1}{q} = 1$ und für $a, b \geq 0$ gilt

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q} \quad (11.129)$$

Beweis. Eine Beweismöglichkeit ist die Anwendung von 11.9.15: Für $n = 2$, $\lambda_1 := \frac{1}{p}$ und $\lambda_2 := \frac{1}{q}$, sowie $a_1 := a^p$ und $a_2 := b^q$ ergibt sich die Behauptung sofort. Allerdings ist dafür nicht der ganze Aufwand der JENSENSchen Ungleichung nötig.

Es genügt zu wissen, daß der Logarithmus eine konkave Funktion ist. Das heißt, wir haben $\ln(\lambda x + (1 - \lambda)y) \geq \lambda \ln(x) + (1 - \lambda) \ln(y) = \ln(x^\lambda y^{(1-\lambda)})$. Wegen der strengen Monotonie der Exponentialfunktion folgt daraus $\lambda x + (1 - \lambda)y \geq x^\lambda y^{(1-\lambda)}$. Jetzt muß nur noch $\lambda = \frac{1}{p}$, $1 - \lambda = \frac{1}{q}$, $x := a^p$ und $y := b^q$ eingesetzt werden, um die Youngsche Ungleichung zu gewinnen. \square

12 Integralrechnung

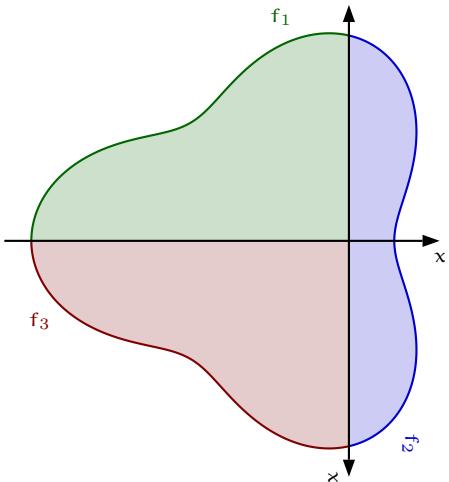
12.1 Das Flächenproblem

Wir formulieren zunächst das allgemeine Flächenproblem:

Man finde ein Verfahren zur Bestimmung des Flächeninhaltes von Flächen, die durch Kurven berandet sind.

Für die meisten Fälle lässt sich das durch eine geeignete Zerlegung der Figur auf das folgende Problem reduzieren:

Finde ein Verfahren zur Bestimmung des Flächeninhalts einer Fläche, die durch den Ausschnitt des Graphen einer Funktion f und der x -Achse begrenzt wird.



12.1.1 Flächeninhalte geometrischer Figuren Welche Flächeninhalte können wir denn bisher berechnen?

Rechteck: $F = a \cdot b.$ Dreieck: $F = \frac{1}{2} a \cdot h.$

Trapez: $F = \frac{1}{2} (a + b) \cdot h,$
denn $c = a - (x + y) = b + x + y$, also $x + y = c - b$ und damit $c = a - c + b$,
oder $c = \frac{1}{2} (a + b).$

Kreis: $F = \pi \cdot r^2.$ (Wieso eigentlich?)

Reduktion auf eine Funktion:

Den Kreis ausgenommen, können wir bisher eigentlich nur den Flächeninhalt von Figuren berechnen, die sich auf Rechtecke reduzieren lassen.

12.1.2 Die Fläche unter einer Kurve (Methode der RIEMANN-Summen)

Eingedenk unserer ernüchternden Beobachtung, daß wir bisher eigentlich nur in der Lage sind, den Flächeninhalt von Figuren auszurechnen, die sich (was ihren Flächeninhalt angeht) auf Rechtecke reduzieren lassen, machen wir aus der Not eine Tugend und bestimmen zunächst nur einen Näherungswert für die Fläche unter einer Kurve, indem wir sie

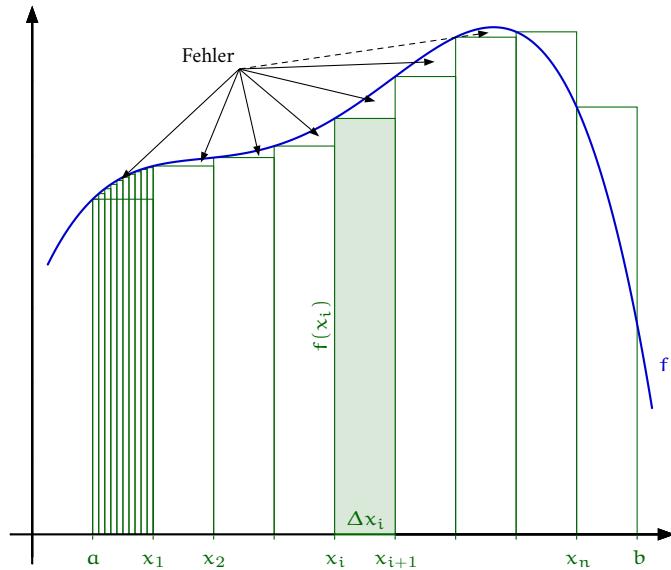
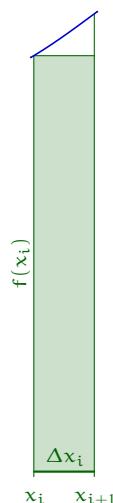


Abb. 12.1 Die Methode der RIEMANN-Summen

durch Rechtecke annähern. Wie schon beim Tangentenproblem nehmen wir dabei einen Fehler in Kauf, in der Hoffnung, daß wir ihn im Nachhinein beliebig klein machen können und daß er nach einem Grenzübergang vollständig verschwindet.

Dazu unterteilen wir das betrachtete Flächenstück unter der Funktion f in möglichst viele kleine Rechtecke der Basisbreite $\Delta x_i = x_{i+1} - x_i$ und der Höhe $f(x_i)$. Dabei bilden die Punkte $a = x_0 < x_1 < x_2 < \dots < x_i < x_{i+1} < \dots < x_n < x_{n+1} = b$ eine Zerlegung \mathcal{Z} für das Basisintervall $[a, b]$ des Flächenstücks.

Die Feinheit von \mathcal{Z} ist die Länge der größten Basisbreite Δx_i . Wir bezeichnen sie mit $|\mathcal{Z}|$.



Wenn wir die Feinheit $|\mathcal{Z}|$ kleiner machen, werden alle Längen Δx_i kleiner. Gleichzeitig wird aber die Anzahl n der Unterteilungspunkte größer, denn bei kleineren Basisbreiten brauchen wir mehr von ihnen, um die Strecke von a bis b zu überdecken. Der Fehler, den wir bei der Überdeckung mit Rechtecken gegenüber dem vermuteten wirklichen Flächeninhalt machen wird normalerweise kleiner, wenn wir die Zerlegung feiner wählen. Das ist für das Intervall $[x_0, x_1]$ angedeutet. Es wird also ein Grenzübergang durchzuführen sein, und zwar der Grenzübergang $|\mathcal{Z}| \rightarrow 0$, um den wirklichen Flächeninhalt zu erhalten. Diesen wollen wir vorläufig mit $F_{a,b}$ bezeichnen. Zunächst die Näherung: Dazu müssen wir nur die Flächeninhalte der Rechtecke mit der Basisbreite Δx_i und der Höhe $f(x_i)$, also $f(x_i)\Delta x_i$, von $i = 0$ bis $i = n$ addieren.

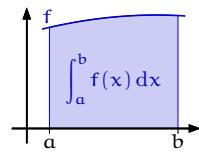
$$F_{a,b} \approx \sum_{i=0}^n f(x_i)\Delta x_i .$$

Nun können wir die Idee zur Berechnung der Fläche $F_{a,b}$ unter dem Graphen von f genauer fassen: $F_{a,b}$ sollte der Grenzwert von $\sum_{i=0}^n f(x_i)\Delta x_i$ für $|\mathcal{Z}| \rightarrow 0$ sein:

$$F_{a,b} = \lim_{|\mathcal{Z}| \rightarrow 0} \sum_{i=0}^n f(x_i)\Delta x_i .$$

Schreibweise: Statt $F_{a,b}$ schreiben wir

$$\int_{x=a}^{x=b} f(x) dx, \quad \text{oder meist} \quad \int_a^b f(x) dx.$$



Dabei ist das Integralzeichen \int als stilisiertes *S* (für *Summe*) anzusehen und $f(x) dx$ soll an $f(x_i)\Delta x_i$ in der RIEMANN-Summe $\sum_{i=0}^n f(x_i)\Delta x_i$ erinnern:

$$\int_a^b f(x) dx := \lim_{|\tilde{x}| \rightarrow 0} \sum_{i=0}^n f(x_i) \Delta x_i. \quad (12.1)$$

$\int_a^b f(x) dx$ heißt *bestimmtes (RIEMANN-)Integral von f in den Grenzen von x = a bis x = b* und gibt (für positive Funktionen f) die Fläche zwischen dem Graphen von f und der x-Achse, in den Grenzen a und b wieder. f heißt *Integrand*. Funktionen, für die das Integral in den Grenzen a und b existiert, nennen wir *über [a, b] (RIEMANN-)integrierbar*. Wenn eine Funktion über jedes endliche Intervall $[a, b]$ integrierbar ist, nennen wir sie *(RIEMANN-)integrierbar*. Mitunter verwenden wir statt x für die Integrationsvariable auch andere Zeichen, wie t, s, etc. Das ist natürlich ohne weiteres möglich, denn ihrer Funktion nach ist die Integrationsvariable eine Summationsvariable, die nach Ausführung des Integrals verschwunden ist:

$$\int_a^b f(x) dx = \int_a^b f(t) dt.$$

Für welche Funktionen f lässt sich dieser Grenzwert denn nun prinzipiell bestimmen?

Satz $\int_a^b f(x) dx$ existiert für alle stetigen Funktionen und alle Funktionen, die sich stückweise aus stetigen Funktionen zusammensetzen.

Diesen Satz nehmen wir zur Kenntnis, wir werden ihn aber nicht beweisen. Für unsere Zwecke ist das auch gar nicht nötig. Seine Aufgabe besteht für uns einfach nur darin, sicherzustellen, daß es genügend viele Funktionen gibt, für die das Flächenproblem lösbar ist. Als Methode, um konkrete Flächeninhalte auszurechnen, eignet sie sich allenfalls für numerische Flächenbestimmungen mit Hilfe eines Computers. Selbst für einfachste Funktionen führt die Methode der RIEMANN-Summen nur sehr schwer zu Ergebnissen, da die auftretenden Summen ohne Computerunterstützung im Allgemeinen keine auswertbaren Ausdrücken ergeben. Wir wollen uns daher möglichst schnell nach einer alternativen Methode umsehen, die es uns in vielen Fällen gestatten wird, die Fläche unter einer Kurve formelmäßig zu bestimmen. Immerhin können wir bei unserer Suche nach dieser Methode nun das Objekt *Fläche unter einer Kurve* einsetzen, denn obiger Satz stellt uns dieses Objekt zur Verfügung. Bevor wir das tun, stellen wir noch die elementaren Eigenschaften des RIEMANN-Integrals zusammen:

- i) $\int_a^b f(x) dx + \int_b^c f(x) dx = \int_a^c f(x) dx. \quad (\text{Additionsregel})$
- ii) $\int_a^b k \cdot f(x) dx = k \cdot \int_a^b f(x) dx. \quad (\text{Faktorregel})$

$$\text{iii)} \quad \int_a^b f(x) dx + \int_a^b g(x) dx = \int_a^b (f(x) + g(x)) dx. \quad (\text{Summenregel})$$

$$\text{iv)} \quad f \geq 0 \Rightarrow \int_a^b f(x) dx \geq 0. \quad (\text{Positivität})$$

12.1.3 Stammfunktionen

Eine Funktion F heißt Stammfunktion von f , falls sie differenzierbar und ihre Ableitung durch f gegeben ist: $F' = f$.

12.1.4 Satz Zwei Stammfunktionen unterscheiden sich nur um eine Konstante.

Beweis. F und G seien zwei Stammfunktionen von f . Es gilt also $F'(x) = f(x) = G'(x)$, oder $F'(x) - G'(x) = (F - G)'(x) = 0$. Also ist $F - G$ eine Funktion, deren Ableitung überall verschwindet. Sie hat daher überall die Steigung Null. Da nur eine konstante Funktion diese Eigenschaft hat, gilt $F(x) - G(x) = c = \text{konst}$, d. h., $F(x) = G(x) + c$. \square

$f(x)$	$F(x)$
x^n	$\frac{1}{n+1} x^{n+1}, (n \neq -1)$
$\sin(x)$	$- \cos(x)$
$\frac{1}{1+x^2}$	$\arctan(x)$

$f(x)$	$F(x)$
$\frac{1}{x}$	$\ln(x)$
$\cos(x)$	$\sin(x)$
e^x	e^x

$f(x)$	$F(x)$
$\tan(x)$	$-\ln(\cos(x))$
$\frac{1}{\sqrt{1-x^2}}$	$\arcsin(x)$
$\frac{1}{1-x^2}$	$\frac{1}{2} \ln\left(\frac{ 1+x }{ 1-x }\right)$

Tabelle 12.1 Einige Stammfunktionen

Um uns davon zu überzeugen, daß diese Tabelle stimmt, müssen wir jeweils nur die Einträge in der $F(x)$ -Spalte ableiten und nachrechnen, daß die Einträge der $f(x)$ -Spalte entstehen:

$\frac{d}{dx} \frac{1}{n+1} x^{n+1} = \frac{n+1}{n+1} x^{n+1-1} = x^n$. $F(x) = \frac{1}{n+1} x^{n+1}$ ist natürlich nur für diejenigen $n \in \mathbb{R}$ eine Stammfunktion von $f(x) = x^n$, für die sich diese Formel überhaupt hinschreiben läßt. Für $n = -1$ ist das offensichtlich nicht mehr der Fall. Die Funktion $f(x) = x^{-1} = \frac{1}{x}$ hat daher auch ihre eigene Regel zur Bildung der Stammfunktion. Man beachte, daß sie nicht durch $\ln(x)$ gegeben ist, wie $\ln'(x) = \frac{1}{x}$ nahelegen könnte, denn in die \ln -Funktion lassen sich nur positive Werte einsetzen, in $\frac{1}{x}$ aber auch negative. Um nachzuweisen, daß $F(x) = \ln(|x|)$ die gesuchte Stammfunktion ist, müssen wir nur $F' = f$ nachrechnen. Nun könnte man denken, daß diese Funktion gar nicht differenzierbar ist, denn $|x|$ ist nicht differenzierbar. Aber $|x|$ ist nur an der Stelle $x = 0$ nicht ableitbar, sonst jedoch überall. 0 ist auch der x -Wert, der die Definitionslücke von \ln markiert – es besteht mithin gar keine Veranlassung, nach der Differenzierbarkeit von F an dieser Stelle zu fragen. Verwenden wir die Definition der *Betragsfunktion*

$$|x| := \begin{cases} x & \text{für } x \geq 0, \\ -x & \text{für } x < 0, \end{cases} \quad (12.2)$$

so erhalten wir

$$\ln(|x|) = \begin{cases} \ln(x) & \text{für } x > 0, \\ \ln(-x) & \text{für } x < 0. \end{cases}$$

Für $x > 0$ wissen wir $\frac{d}{dx} \ln(|x|) = \frac{1}{x}$ bereits. Für $x < 0$ müssen wir sie noch zeigen. Dazu benutzen wir die Kettenregel:

$$\frac{d}{dx} \ln(|x|) = \frac{d}{dx} \ln(-x) = \frac{1}{-x} \cdot (-1) = \frac{1}{x}.$$

Damit kennen wir für einige wenige Funktionen eine Stammfunktion. Wir werden später Techniken kennenlernen, mit deren Hilfe wir auch für kompliziertere Funktionen Stammfunktionen finden können. Vorerst aber wollen wir den zentralen Satz formulieren und beweisen, der den Begriff Stammfunktion mit dem Flächenproblem in Beziehung setzt.

12.1.5 Satz (Hauptsatz der Differential- und Integralrechnung)

Für eine stetige Funktion f ist

$$F(x) := \int_a^x f(t) dt$$

eine Stammfunktion von f , d. h., es gilt $F' = f$.

Beweis. Den Beweis können wir in zwei Schritten führen. Im ersten Schritt stellen wir die Idee vor und kümmern uns nicht um die Details. Im zweiten Schritt gehen wir etwas mehr in die Tiefe.

Erster Schritt: Wir müssen zeigen, daß F differenzierbar ist und $F' = f$ gilt. Da wir vorerst noch keine Ableitungsregel für die Funktion F zur Verfügung haben, müssen wir uns auf die Definition der Ableitung zurückziehen:

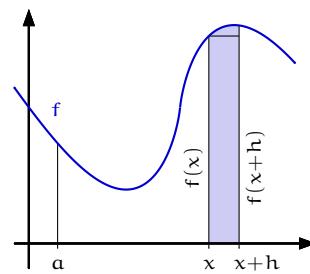
$$F'(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h}$$

Das machen wir zunächst etwas informell, um die Idee zu verstehen: $F(x)$ ist der Flächeninhalt von a bis zu der variablen oberen Grenze x und $F(x+h)$ der bis zu der nah benachbarten Stelle $x+h$. Also ist $F(x+h) - F(x)$ der Flächeninhalt in dem schmalen Streifen von x bis $x+h$. Für sehr kleines h sind $f(x)$ und $f(x+h)$ kaum noch verschieden, da wir f als stetig angenommen haben. Damit machen wir keinen großen Fehler, wenn wir den tatsächlichen Flächeninhalt $F(x+h) - F(x)$ durch den des Rechtecks mit der Höhe $f(x)$ und der Breite h ersetzen. Dieser Fehler wird tatsächlich beliebig klein, wenn wir h gegen Null streben lassen.

Also gilt

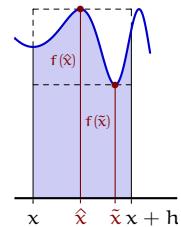
$$\frac{F(x+h) - F(x)}{h} \approx \frac{f(x) \cdot h}{h} = f(x).$$

Jetzt können wir verstehen, wieso man auf die Idee kommen kann, in der Flächenfunktion $F(x)$ die Stammfunktion von f zu vermuten.



2. Schritt: Nachdem wir die Idee verstanden haben, können wir uns jetzt ihrer formalen Absicherung zuwenden. Wir benutzen dafür den Satz vom Maximum (11.1.9), nach dem eine stetige Funktion auf einem abgeschlossenen Intervall $[a, b]$ ein Maximum und ein Minimum hat.

Für uns ist natürlich das Intervall $[x, x+h]$ von Interesse. \tilde{x} sei die Stelle, an der f ein Minimum annimmt und \hat{x} die Stelle des Maximums. Dann verkleinern wir den tatsächlichen Flächeninhalt über dem Intervall $[x, x+h]$, wenn wir ihn durch den Flächeninhalt $f(\tilde{x}) \cdot h$ des einbeschriebenen Rechtecks mit der Höhe des minimalen Funktionswertes $f(\tilde{x})$ ersetzen, und wir vergrößern ihn, wenn wir $f(\hat{x}) \cdot h$, den Flächeninhalt des umfassenden Rechtecks mit der Höhe des Maximums $f(\hat{x})$ nehmen. Es gilt demnach:



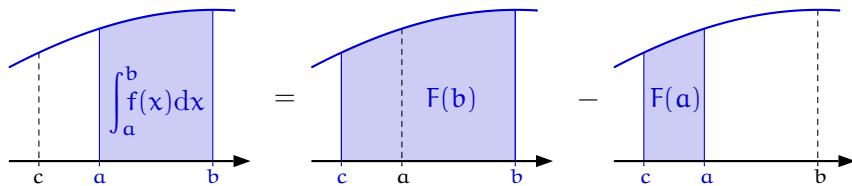
$$f(\tilde{x}) = \frac{f(\tilde{x}) \cdot h}{h} \leq \frac{F(x+h) - F(x)}{h} \leq \frac{f(\hat{x}) \cdot h}{h} = f(\hat{x}).$$

Lassen wir nun h gegen Null streben, so wandert die rechte Seite $x+h$ des Intervalls $[x, x+h]$ gegen x und alle Punkte, die sich in diesem Intervall befinden, ebenfalls. Also konvergieren \tilde{x} und \hat{x} gegen x , und da f stetig ist, die Funktionswerte $f(\tilde{x})$ und $f(\hat{x})$ gegen $f(x)$. Nach dem Sandwich-Prinzip konvergiert dann auch der Differenzenquotient gegen $f'(x)$:

$$F'(x) = \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h} = f(x).$$

Das wollten wir zeigen. □

12.1.6 Flächenberechnung mit Hilfe von Stammfunktionen Die Methode der RIEMANN-Summen ist eine ziemlich schwerfällige Methode, um konkrete Flächeninhalte auszurechnen. Das trifft natürlich um so mehr auf die Flächenfunktion $F(x) = \int_c^x f(t) dt$ zu. Um sie nämlich zu kennen, müßten wir für jedes x die Methode der RIEMANN-Summen anwenden. Allerdings würde sie es uns ermöglichen, den Flächeninhalt, sagen wir von $x=a$ bis $x=b$, einfach als Differenz der Funktionswerte $F(b)$ und $F(a)$ zu erhalten:



Es lohnt sich also, nach einer alternativen Berechnungsmethode für F (genauer: für $F(b) - F(a)$) zu suchen. Inzwischen haben wir den Hauptsatz der Differential- und Integralrechnung zu unserer Verfügung. Wir wissen also, daß die Flächenfunktion F eine Stammfunktion von f ist. Darüber hinaus wissen wir auch, daß sich zwei Stammfunktionen allenfalls um eine Konstante unterscheiden. Für eine weitere Stammfunktion G von f gilt also $F(x) = G(x) + k$, mit einer Konstanten k . Diese Konstante kennen wir nicht, denn wir kennen ja die Flächenfunktion nicht. Die entscheidende Beobachtung ist aber, daß wir sie auch gar nicht kennen müssen, denn in der Differenz $F(b) - F(a)$ fällt sie heraus:

$$F(b) - F(a) = (G(b) + k) - (G(a) + k) = G(b) + k - G(a) - k = G(b) - G(a).$$

In der Formel

$$\int_a^b f(t) dt = F(b) - F(a) \quad (12.3)$$

können wir daher *jede* Stammfunktion F von f nehmen.

Damit haben wir die Berechnung von Flächen auf die Bestimmung von Stammfunktionen zurückgeführt.

Wir müssen zur Berechnung des Flächeninhalts zwischen dem Graphen einer Funktion f und der x -Achse in den Grenzen von a bis b zunächst eine Stammfunktion F von f finden, dann lediglich die beiden Grenzen einsetzen, und schließlich die Differenz $F(b) - F(a)$ bilden. Das schreiben wir in folgender Weise:

$$\int_a^b f(t) dt = [F(t)]_a^b := F(b) - F(a). \quad (12.4)$$

D. h., wir setzen in (12.4) die Integrationsgrenzen nicht sofort ein, denn auf diese Weise hätten wir bei konkreten Funktionen nur die wenig anschauliche Zahl $F(b) - F(a)$ vorzuweisen, der man meist wenig Informationen über die Stammfunktion entnehmen kann. Durch die Schreibweise $[F(t)]_a^b$ haben wir die volle Information, nämlich die über die Stammfunktion und die Grenzen.

Der enge Zusammenhang zwischen Integral und Stammfunktion drückt sich auch in der oft bequem einsetzbaren Schreibweise

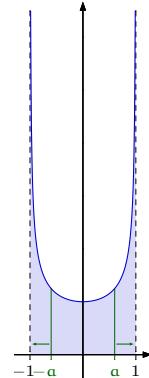
$$\int f(x) dx = F(x) \quad (12.5)$$

für die Stammfunktion F von f aus. Diesen Ausdruck nennt man *unbestimmtes Integral* von f . Offensichtlich gilt

$$\int f'(x) dx = f(x). \quad (12.6)$$

Wir wenden unsere Erkenntnisse auf ein paar Funktionen der Tabelle 12.1 an:

- i) $\int_0^2 x^3 dx = \left[\frac{1}{4}x^4 \right]_0^2 = \frac{1}{4}2^4 = 4.$
- ii) $\int_{-2}^2 e^x dx = \left[e^x \right]_{-2}^2 = e^2 - \frac{1}{e^2}.$
- iii) $\int_0^{2\pi} \sin(t) dt = \left[-\cos(t) \right]_0^{2\pi} = -\cos(2\pi) + \cos(0) = 0.$
- iv) $\int_{-a}^a \frac{1}{\sqrt{1-x^2}} dx = \left[\arcsin(x) \right]_{-a}^a = \arcsin(a) - \arcsin(-a)$
 $= 2 \arcsin(a) \xrightarrow{a \rightarrow 1} 2 \arcsin(1) = \pi.$



Beim letzten Integral haben wir die Grenzen -1 und 1 , an denen wir eigentlich interessiert sind, nicht direkt einsetzen können, weil der Integrand $\frac{1}{\sqrt{1-x^2}}$ an diesen Stellen nicht definiert ist. Daher haben wir uns mit den Grenzen $0 < a < 1$ und $-1 < -a < 0$ einen respektvollen Abstand zu dieser Definitionslücke verschafft und versucht, diesen nach der Integration durch den

Grenzübergang $a \rightarrow 1$ wieder zum Verschwinden zu bringen. Dabei haben wir einen endlichen Flächeninhalt für die unendlich ausgedehnte Fläche erhalten, die von der Funktion $\frac{1}{\sqrt{1-x^2}}$, der x -Achse und den senkrechten Asymptoten bei -1 und 1 begrenzt wird.

Wie wir sehen, liefert Beispiel iii) nicht gerade das gewünschte Ergebnis. Die Fläche, die die Sinus-Funktion mit der x -Achse in den Grenzen von 0 bis 2π einschließt, hat sicher nicht den Flächeninhalt 0 . Was ist falsch gelaufen?

Wir lassen uns von der Methode der RIEMANN-Summen $\sum_{i=0}^n f(x_i) \Delta x_i$ leiten. Hier tragen die Summanden $f(x_i) \Delta x_i$ mehr Information, als nur den Flächeninhalt des zugehörigen Flächenstücks. Mit dem Vorzeichen von $f(x_i)$ wird auch noch angezeigt, ob sie sich oberhalb oder unterhalb der x -Achse befinden. Deshalb wird eine Fläche, die unterhalb der x -Achse liegt mit einem negativen Vorzeichen auftreten. Das müssen wir berücksichtigen, wenn wir an einer Funktion interessiert sind, die positive und negative Funktionswerte aufweist. Am einfachsten geschieht das, indem wir nicht über Nullstellen von Funktionen hinwegintegrieren, sondern uns sozusagen von Nullstelle zu Nullstelle hängeln und die Beträge der dabei entstehenden positiv und negativ gerechneten Flächeninhalte addieren. Für das Sinus-Beispiel könnte das folgendermaßen aussehen:

$$\begin{aligned} \left| \int_0^\pi \sin(t) dt \right| + \left| \int_\pi^{2\pi} \sin(t) dt \right| &= \left| [-\cos(t)]_0^\pi \right| + \left| [-\cos(t)]_\pi^{2\pi} \right| \\ &= |-cos(\pi) + cos(0)| + |-cos(2\pi) + cos(\pi)| \\ &= |2| + |-2| = 4. \end{aligned}$$

Nachdem wir die Flächenberechnung im Wesentlichen auf das Finden von Stammfunktionen zurückgeführt haben, benötigen wir jetzt leistungsfähige Integrationstechniken, um unseren Bestand an bekannten Stammfunktionen aufzustocken.

12.2 Integrationstechniken

Zu der Produkt- und der Kettenregel gibt es eine korrespondierende Integrationsregeln, nämlich die *Produktintegration* und die *Substitution*.

12.2.1 Die Produktintegration

Die Produktregel der Ableitung lautet

$$(u \cdot v)' = u' \cdot v + u \cdot v'. \quad (12.7)$$

Da uv die Stammfunktion von $(uv)'$ ist, ergibt sich aus (12.7) durch Integration auf beiden Seiten:

$$u(x) \cdot v(x) = \int u'(x) \cdot v(x) dx + \int u(x) \cdot v'(x) dx.$$

Wir stellen das nach einem der Integrale um und erhalten bereits die Formel für die *Produktintegration* (auch als *partielle Integration* bezeichnet)

$$\int u'(x) \cdot v(x) dx = u(x) \cdot v(x) - \int u(x) \cdot v'(x) dx, \quad (12.8)$$

oder, in etwas übersichtlicherer, symbolischer Schreibweise:

$$\int u' \cdot v = u \cdot v - \int u \cdot v'. \quad (12.9)$$

Was fangen wir nun mit dieser Formel an? Am besten erlernt man diese Integrationstechnik an konkreten Anwendungsbeispielen.

Wir beginnen mit einem Standardbeispiel:

$$\int x \sin(x) dx.$$

Wenn wir (12.9) anwenden wollen, müssen wir eine der beiden Faktoren, x oder $\sin(x)$ als u' deklarieren. Da (12.9) auf der rechten Seite die Angabe von $u \cdot v$ verlangt, sollten wir die Stammfunktion u von u' angeben können. Das müssen wir bei unserer Wahl berücksichtigen. Im vorliegenden Beispiel ist das aber kein Problem, denn wir können beide Faktoren problemlos integrieren. Wir entscheiden uns für $u'(x) = \sin(x)$. Dann erhalten wir

$$\begin{aligned} \int x \sin(x) dx &= x(-\cos(x)) - \int 1 \cdot (-\cos(x)) dx = -x \cos(x) + \int \cos(x) dx. \\ \int v \ u' &= v \ u - \int v' \ u \end{aligned}$$

Erfolg haben wir mit unserer Wahl, wenn wir das Integral auf der rechten Seite lösen können. Das ist hier der Fall, denn $\int \cos(x) dx = \sin(x)$. Damit haben wir die Stammfunktion von $x \mapsto x \sin(x)$ gefunden:

$$\int x \sin(x) dx = \sin(x) - x \cos(x).$$

Wir können uns immer davon überzeugen, daß wir richtig gerechnet haben:

$$\frac{d}{dx} (\sin(x) - x \cos(x)) = \cos(x) + x \sin(x) - \cos(x) = x \sin(x).$$

Die Funktion $x \mapsto \sin(x) - x \cos(x)$ ist also tatsächlich eine Stammfunktion von $x \mapsto x \sin(x)$. Wir wollen an diesem Beispiel auch demonstrieren, wie eine schlechte Wahl von u und v ausgesehen hätte und woran man das erkennt: Für $u'(x) = x$ und $v(x) = \sin(x)$ erhalten wir

$$\int x \sin(x) dx = \frac{1}{2}x^2 \sin(x) - \frac{1}{2} \int x^2 \cos(x) dx.$$

Jetzt ist das zweite Integral durch den Faktor x^2 noch schwerer zu lösen, als das Ausgangsintegral.

$$i) \int x e^x dx = x e^x - \int e^x dx = x e^x - e^x = (x-1)e^x.$$

Dabei haben wir $u'(x) = e^x$ und $v(x) = x$ gewählt.

$$ii) \int x^2 e^x dx = x^2 e^x - 2 \int x e^x dx = x^2 e^x - 2(x-1)e^x = (x^2 - 2x + 2)e^x.$$

($u'(x) = e^x, v(x) = x^2$)

$$iii) \int e^x \sin(x) dx = e^x \sin(x) - \int e^x \cos(x) dx \\ = e^x \sin(x) - e^x \cos(x) - \int e^x \sin(x) dx.$$

Wir haben die Produktintegration zweimal angewendet. Das erste mal mit $u'(x) = e^x$ und $v(x) = \sin(x)$ und das zweite mal mit $u'(x) = e^x$ und $v(x) = \cos(x)$. Damit haben wir scheinbar nichts gewonnen, weil sich unser gesuchtes Integral reproduziert hat. Entscheidend ist aber, daß es mit einem *negativen* Vorzeichen entstanden ist, denn so können wir die Gleichung einfach nach dem gewünschten Integral auflösen: $2 \int e^x \sin(x) dx = (\sin(x) - \cos(x))e^x$, also $\int e^x \sin(x) dx = \frac{1}{2}(\sin(x) - \cos(x))e^x$.

$$iv) \int \sin^2(x) dx = \int \sin(x) \sin(x) dx \\ = -\cos(x) \sin(x) + \int \cos^2(x) dx \\ = -\cos(x) \sin(x) + \int (1 - \sin^2(x)) dx \\ = -\cos(x) \sin(x) + x - \int \sin^2(x) dx.$$

Wir haben $u'(x) = \sin(x)$ und $v(x) = \cos(x)$ gesetzt. In der zweiten Gleichung machen wir dann von der zentralen Beziehung $\sin^2(x) + \cos^2(x) = 1$, Gebrauch. Dabei reproduziert sich das Ausgangsintegral. Auflösen ergibt

$$\int \sin^2(x) dx = \frac{1}{2}(x - \sin(x) \cos(x)). \quad (12.10)$$

Daraus folgt nun leicht

$$\int \cos^2(x) dx = \frac{1}{2}(x + \sin(x) \cos(x)). \quad (12.11)$$

$$\text{v) } \int x \ln(x) dx = \frac{1}{2}x^2 \ln(x) - \frac{1}{2} \int \frac{x^2}{x} dx = \frac{1}{2}x^2 \ln(x) - \frac{1}{2} \int x dx = \frac{1}{2}x^2 \ln(x) - \frac{1}{4}x^2.$$

Hier haben wir, anders als bisher, $u'(x) = x$ und $v(x) = \ln(x)$ gesetzt. Dann taucht im zweiten Integral natürlich $u(x) = \frac{1}{2}x^2$ auf, was wir bisher immer vermeiden wollten. Da die Ableitung $\ln'(x)$ aber einfach $\frac{1}{x}$ ist, entsteht auf diese Weise trotzdem ein leicht zu berechnendes Integral.

$$\text{vi) } \int \ln(x) dx = \int 1 \cdot \ln(x) dx = x \ln(x) - \int x \frac{1}{x} dx = x \ln(x) - x.$$

Wir haben den Trick angewandt, $u'(x) = 1$ und $v(x) = \ln(x)$ zu setzen.

$$\text{vii) } \int \ln^2(x) dx = x \ln^2(x) - 2 \int x \cdot \ln(x) \cdot \frac{1}{x} dx = x \ln^2(x) - 2 \int \ln(x) dx \\ = x \ln^2(x) - 2x \ln(x) + 2x.$$

- viii) Überzeugen Sie sich durch Ableiten, daß in den Beispielen tatsächlich die Stammfunktionen gefunden wurden.

Die Produktintegration läßt sich natürlich auch für bestimmte Integrale formulieren:

$$\int_a^b u'(x) \cdot v(x) dx = [u(x) \cdot v(x)]_a^b - \int_a^b u(x) \cdot v'(x) dx. \quad (12.12)$$

Etwa

$$\int_0^1 xe^x dx = [xe^x]_0^1 - \int_0^1 e^x dx = [xe^x]_0^1 - [e^x]_0^1 = e - (e - 1) = 1.$$

Normalerweise sollte man aber so vorgehen, daß man zunächst das unbestimmte Integral löst, also die Stammfunktion angibt und dann die Grenzen einsetzt. Auf diese Weise hat man mehr Kontrolle über Integrationsfehler. Denn eine vermeintliche Stammfunktion kann einfach durch Ableiten auf ihre Richtigkeit hin überprüft werden.

In unseren Beispielen haben wir gesehen, daß man mitunter etwas probieren muß, bis man den richtigen Ansatz zur Lösung des Integrals gefunden hat. Trotzdem kann es passieren, daß ein Integral, dessen Integrand als Produkt auftritt, nicht mit der Produktintegration gelöst werden kann. Ein Beispiel ist $\int xe^{-x^2} dx$. Das erweist sich der Produktintegration gegenüber als resistent. Mit der sogenannten *Substitutionsmethode* kann es gelöst werden.

12.2.2 Die Substitutionsmethode ist die Integrationstechnik, die zur Kettenregel korrespondiert. Sie beruht auf folgender einfachen Beobachtung:

Ist F die Stammfunktion von f , also gilt $F' = f$, dann ist $F \circ u$ die Stammfunktion von $(f \circ u) \cdot u'$. Denn nach der Kettenregel haben wir

$$\frac{d}{dx} F \circ u(x) = \frac{d}{dx} F(u(x)) = F'(u(x)) \cdot u'(x) = f(u(x)) \cdot u'(x).$$

Also folgt

$$\int f(u(x)) \cdot u'(x) dx = F(u(x)). \quad (12.13)$$

Für das bestimmte Integral ergibt sich daraus

$$\int_a^b f(u(x)) \cdot u'(x) dx = \int_a^b \frac{d}{dx} F(u(x)) dx = [F(u(x))]_a^b$$

$$= F(u(b)) - F(u(a)) = \left[F(t) \right]_{u(a)}^{u(b)} = \int_{u(a)}^{u(b)} f(t) dt,$$

also

$$\int_a^b f(u(x)) \cdot u'(x) dx = \int_{u(a)}^{u(b)} f(t) dt. \quad (12.14)$$

Wir machen den Gebrauch dieser Methode wieder an einem Beispiel deutlich. Wir suchen die Stammfunktion

$$\int \frac{3x}{2x^2 + 4} dx.$$

Um Gleichung (12.13) anwenden zu können, müssen wir f und u bestimmen. Wir wählen $f(t) = \frac{1}{t}$ (also $F(t) = \ln(|t|)$) und $u(x) = 2x^2 + 4$. Es ist $f(u(x)) = \frac{1}{2x^2+4}$, also bis auf den Faktor $3x$ der Integrand unseres Integrals. $u'(x) = 4x$, so daß $f(u(x))u'(x) = \frac{4x}{2x^2+4}$. Das ist nicht genau unser Integrand, denn der Faktor 4 im Zähler stimmt nicht. Aber das lässt sich reparieren:

$$\int \frac{3x}{2x^2 + 4} dx = \frac{3}{4} \int \frac{4x}{2x^2 + 4} dx = \frac{3}{4} \int f(u(x))u'(x) dx = \frac{3}{4} F(u(x)) = \frac{3}{4} \ln(2x^2 + 4).$$

Das Betragszeichen in $\ln(\dots)$ ist hier wegen $2x^2 + 4 > 0$ überflüssig.

Üblicherweise geschieht die Anwendung der Substitutionsmethode in formalerer Weise. Der Leitfaden dafür ist die folgende formale Rechnung:

$$\int f(u(x))u'(x) dx = \int f(u) \frac{du}{dx} dx = \int f(u) du.$$

D.h., wir versuchen in unserem Ausgangsintegral einen bestimmtem Ausdruck durch die Variable u zu ersetzen, also aus dem Integral in der Variablen x ein Integral in der Variablen u zu machen. Dabei muß insbesondere dx in du umgerechnet werden. Vergleichen wir die linke und die rechte Seite in obiger Gleichung, dann muß $u'(x) dx$ durch du ersetzt werden. Das formalisiert man, indem man für u' den Ausdruck $\frac{du}{dx}$ verwendet und diesen nach du auflöst. Dieser Vorgang ist der eigentlich formale, denn dx und dy sind symbolische Objekte, keine Zahlen, mit denen wir so ohne weiteres rechnen können. Trotzdem liefert der beschriebene Vorgang die richtige Merkregel. Für das oben angeführte Beispiel sieht das folgendermaßen aus:

In $\int \frac{3x}{2x^2 + 4} dx$ substituieren wir $u = 2x^2 + 4$. Dann ist $\frac{du}{dx} = 4x$, also $4x dx = du$. Vergleichen wir mit unserem Integral, so sehen wir, daß $3x dx$ schon vorgebildet ist. Das können wir durch $\frac{3}{4} du$ ersetzen. Damit haben wir insgesamt

$$\int \frac{3x}{2x^2 + 4} dx = \frac{3}{4} \int \frac{1}{u} du = \frac{3}{4} \ln(|u|) = \frac{3}{4} \ln(|2x^2 + 4|) = \frac{3}{4} \ln(2x^2 + 4).$$

Ist das Ausgangsintegral ein bestimmtes Integral, also etwa $\int_0^2 \frac{3x}{2x^2 + 4} dx$, so stehen uns zwei Wege zur Lösung offen. Der erste besteht darin, wie oben beschrieben die Stammfunktion auszurechnen und danach die Grenzen einzusetzen. Das ist üblicherweise zu empfehlen, denn man hat dabei Rechenfehlern gegenüber die bessere Kontrolle. Man kann aber auch gleich das bestimmte Integral berechnen. Dabei müssen wir dann jedoch auch die x -Grenzen des Ausgangsintegrals in die u -Grenzen der substituierten Version umrechnen. Das ist nicht schwer, denn

wenn $x = 0$ ist, ist u einfach durch Einsetzen zu erhalten: $u(0) = 4$. Genauso gehört zu $x = 2$ die Grenze $u(2) = 12$. Also

$$\int_0^4 \frac{3x}{2x^2 + 4} dx = \frac{3}{4} \int_4^{12} \frac{1}{u} du = \frac{3}{4} \left[\ln(|u|) \right]_4^{12} = \frac{3}{4} (\ln(12) - \ln(4)) = \frac{3}{4} \ln(3).$$

Das ist die Rechenmethode, die zu (12.14) gehört.

i) $\int xe^{-x^2} dx$: Wir substituieren $u = -x^2$. Dann gilt $u' = \frac{du}{dx} = -2x$, also $x dx = -\frac{1}{2} du$. Daher ist $\int xe^{-x^2} dx = -\frac{1}{2} \int e^u du = -\frac{1}{2} e^u = -\frac{1}{2} e^{-x^2}$.

ii) $\int \frac{1}{t} \ln(t) dt$: Wir setzen $u = \ln(t)$ und erhalten mit $\frac{du}{dt} = \frac{1}{t}$, also $du = \frac{1}{t} dt$, das Integral $\int \frac{1}{t} \ln(t) dt = \int u du = \frac{1}{2} u^2 = \frac{1}{2} \ln^2(t)$.

iii) $\int \frac{\sqrt{x^2 - 1}}{x} dx$: Nach einigem Probieren findet man eine günstige Substitution, nämlich $u = \sqrt{x^2 - 1}$. Mit Hilfe der Kettenregel erhalten wir $\frac{du}{dx} = \frac{x}{\sqrt{x^2 - 1}}$, also $dx = \frac{u}{x} du$. Für unser Integral benötigen wir $\frac{1}{x} dx = \frac{u}{x^2} du$. Quadrieren wir $u = \sqrt{x^2 - 1}$ und lösen nach x^2 auf: $x^2 = u^2 + 1$. Damit ist $\frac{1}{x} dx = \frac{u}{u^2 + 1} du$, und das Integral wird zu

$$\begin{aligned} \int \frac{\sqrt{x^2 - 1}}{x} dx &= \int \frac{u^2}{u^2 + 1} du = \int \frac{u^2 + 1 - 1}{u^2 + 1} du = \int du - \int \frac{1}{u^2 + 1} du = u - \arctan(u) \\ &= \sqrt{x^2 - 1} - \arctan(\sqrt{x^2 - 1}). \end{aligned}$$

Es ist eine gute Übung, durch Ableiten nachzurechnen, daß es sich hierbei tatsächlich um eine Stammfunktion von $x \mapsto \frac{\sqrt{x^2 - 1}}{x}$ handelt.

12.2.3 Die Logarithmus-Regel

$$\int \frac{f'(x)}{f(x)} dx = \ln(|f(x)|).$$

Wir bestätigen sie, indem wir die rechte Seite mit Hilfe der Kettenregel ableiten. Für alle $x \in \mathbb{R}$ mit $f(x) \neq 0$ gilt: $\ln(|f(x)|) = \frac{1}{2} \ln(f^2(x))$. Die Ableitung ist daher

$$\frac{d}{dx} \ln(|f(x)|) = \frac{1}{2} \frac{d}{dx} \ln(f^2(x)) = \frac{1}{2} \frac{2f(x)f'(x)}{f^2(x)} = \frac{f'(x)}{f(x)}.$$

Die Möglichkeiten dieser Formel sieht man wieder am besten an den Beispielen.

i) $\int \frac{3x}{4x^2 - 2} dx = \frac{3}{8} \int \frac{8x}{4x^2 - 2} dx = \frac{3}{8} \ln(|4x^2 - 2|)$.

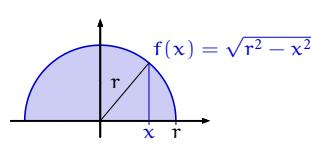
Der Zähler $3x$ des Bruchs war nicht genau die Ableitung $8x$ des Nenners $4x^2 - 2$. Allerdings haben wir das entscheidende x vorgefunden, so daß wir den Vorfaktor korrigieren konnten, indem wir die Zahl 8 ergänzten.

ii) $\int \tanh(x) dx = \int \frac{e^x - e^{-x}}{e^x + e^{-x}} dx = \ln(e^x + e^{-x})$.

$$\begin{aligned} \text{iii)} \quad & \int \tan(x) dx = - \int \frac{-\sin(x)}{\cos(x)} dx = -\ln(|\cos(x)|) = -\frac{1}{2} \ln(\cos^2(x)). \\ \text{iv)} \quad & \int \frac{5}{2-4e^{-4x}} dx = \int \frac{5}{2-4e^{-4x}} \cdot \frac{e^{4x}}{e^{4x}} dx = \int \frac{5e^{4x}}{2e^{4x}-4} dx = \frac{5}{8} \int \frac{8e^{4x}}{2e^{4x}-4} dx = \frac{5}{8} \ln(|2e^{4x}-4|). \end{aligned}$$

Dieses Integral haben wir in zwei Schritten angepaßt. Im ersten Schritt haben wir durch Erweitern mit e^{4x} dafür gesorgt, daß im Zähler bis auf einen Vorfaktor die Ableitung des Nenners steht. Im zweiten Schritt haben wir auf die übliche Weise, also wie in (i), den Vorfaktor korrigiert.

In manchen Situationen ist es vorteilhaft, wenn man in einem Integral $\int f(x) dx$ nicht einen Ausdruck durch eine Variable u , sondern x durch einen Ausdruck ersetzt, um so z. B. eine bestimmte algebraische Relation auszunutzen. Wir verdeutlichen das an einem Beispiel.



Die Berechnung der Kreisfläche führt auf das Integral

$$2 \int_{-r}^r \sqrt{r^2 - x^2} dx.$$

Die algebraische Relation, die wir ausnutzen wollen, ist die zentrale Beziehung $\sin^2(u) + \cos^2(u) = 1$ zwischen Sinus und Kosinus. Dafür setzen wir $x = r \sin(u)$. Dann ist $\sqrt{r^2 - x^2} = \sqrt{r^2 - r^2 \sin^2(u)} = r\sqrt{1 - \sin^2(u)} = r\sqrt{\cos^2(u)}$. Betrachten wir die Grenzen: Für $x = -r$ muß $-r = r \sin(u)$, also $-1 = \sin(u)$ gelten, d.h., $u = \arcsin(-1) = -\frac{\pi}{2}$ (vergl. Abbildung 12.5). Für $x = r$ folgt die u -Grenze ebenso: $u = \frac{\pi}{2}$. In dem Bereich von $u = -\frac{\pi}{2}$ bis $u = \frac{\pi}{2}$ ist $\cos(u)$ positiv, so daß wir aus $\sqrt{\cos^2(u)}$ die positive Wurzel $\cos(u)$ ziehen können. Nun müssen wir noch dx in du umrechnen. Dazu leiten wir x nach u ab: $\frac{dx}{du} = r \cos(u)$, also $dx = r \cos(u) du$. Jetzt haben wir alles beisammen, um die Substitution durchzuführen. Für die Kreisfläche erhalten wir unter Verwendung von (12.11)

$$\begin{aligned} 2 \int_{-r}^r \sqrt{r^2 - x^2} dx &= 2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} r \cos(u) \cdot r \cos(u) du = 2r^2 \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^2(u) du \\ &= r^2 \left[u + \sin(u) \cos(u) \right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} = r^2 \left[\frac{\pi}{2} - \left(-\frac{\pi}{2} \right) \right] = \pi r^2. \end{aligned}$$

12.2.4 A Zeigen Sie die Abschätzung

$$e \cdot \left(\frac{n}{e} \right)^n \leq n! \leq \frac{e^2}{4} \cdot \left(\frac{n+1}{e} \right)^{n+1} \quad (12.15)$$

indem Sie $\ln(n!)$ mittels geeigneter Unter- bzw. Obersummen gegen bestimmte Integrale von \ln abschätzen. Zeigen Sie damit

$$\lim_{n \rightarrow \infty} \frac{\sqrt[n]{n!}}{n} = \frac{1}{e}, \quad \lim_{n \rightarrow \infty} \frac{\sqrt[n]{(an)!}}{n^a} = \frac{a^a}{e^a}, \quad a \in \mathbb{N}. \quad (12.16)$$

12.3 Anwendungen

12.3.1 TAYLOR-Entwicklung Die TAYLOR-Entwicklung stellt eine Methode dar, mit deren Hilfe es möglich ist, die Funktionswerte vieler Funktionen, wie sin, cos, exp in beliebiger Genauigkeit zu berechnen. Wir verschaffen uns die TAYLOR-Formel mittels Produktintegration. Unser Ziel ist es dabei, die Funktionswerte $f(x)$ einer Funktion f in einer Umgebung eines geeigneten Punktes x_0 beliebig genau durch Summen von Potenzen $(x - x_0)^n$ auszudrücken, so daß er durch Anwendung von Grundrechenarten zugänglich wird. Anders gesagt: Wir versuchen $f(x)$ durch eine Potenzreihe $\sum_{k=0}^{\infty} a_k (x - x_0)^k$ darzustellen. Wir starten mit der Formel

$$f(x) - f(x_0) = \int_{x_0}^x f'(t) dt,$$

also

$$f(x) = f(x_0) + \int_{x_0}^x f'(t) dt. \quad (12.17)$$

Nun formen wir das Integral $\int_{x_0}^x f'(t) dt = \int_{x_0}^x 1 \cdot f'(t) dt$ mittels Produktintegration um. Dabei wählen wir für $u'(t)$ die Zahl 1 und für die von (12.9) geforderten Stammfunktion u die Funktion $t - x$. $v(t)$ ist dann natürlich $f'(t)$. Dabei machen wir uns noch einmal klar, daß wir bzgl. der Variablen t integrieren, so daß x wie eine gewöhnliche Konstante behandelt wird. Hier haben wir sie als die Zahl genommen, die wir zu jeder Stammfunktion hinzufügen dürfen:

$$\int_{x_0}^x f'(t) dt = \left[(t - x)f'(t) \right]_{x_0}^x - \int_{x_0}^x (t - x)f''(t) dt = -(x_0 - x)f'(x_0) - \int_{x_0}^x (t - x)f''(t) dt.$$

In der nächsten Runde wählen wir $u'(t) = (t - x)$ und $u(t) = \frac{1}{2}(t - x)^2$, sowie $v(t) = f''(t)$. Wir setzen alles in (12.17) ein und erhalten

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0)f'(x_0) - \left[\frac{1}{2}(t - x)^2 f''(t) \right]_{x_0}^x + \frac{1}{2} \int_{x_0}^x (t - x)^2 f'''(t) dt \\ &= f(x_0) + (x - x_0)f'(x_0) + \frac{1}{2}(x_0 - x)^2 f''(x_0) + \frac{1}{2} \int_{x_0}^x (t - x)^2 f'''(t) dt. \end{aligned}$$

Wir fahren auf diese Weise fort: Jetzt ist $u'(t) = \frac{1}{2}(t - x)^2$ und $u(t) = \frac{1}{2 \cdot 3}(t - x)^3$. Wir erhalten damit bereits

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0)f'(x_0) + \frac{1}{2}(x - x_0)^2 f''(x_0) \\ &\quad + \frac{1}{2 \cdot 3} \left[(t - x)^3 f'''(t) \right]_{x_0}^x - \frac{1}{2 \cdot 3} \int_{x_0}^x (t - x)^3 f^{(4)}(t) dt \\ &= f(x_0) + (x - x_0)f'(x_0) + \frac{1}{2}(x - x_0)^2 f''(x_0) - \frac{1}{3!}(x_0 - x)^3 f'''(x_0) \\ &\quad - \frac{1}{3!} \int_{x_0}^x (t - x)^3 f^{(4)}(t) dt. \end{aligned}$$

Die nächste Runde wird uns das Gesetz verraten, nach dem die weiteren Summanden zu bilden sind. Wir wählen $u'(t) = \frac{1}{3!}(t - x)^3$ und $u(t) = \frac{1}{4!}(t - x)^4$. $v(t)$ ist inzwischen die 4.te Ableitung $f^{(4)}(t)$. Wir erhalten:

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{1}{2}(x - x_0)^2 f''(x_0) + \frac{1}{3!}(x - x_0)^3 f'''(x_0)$$

$$\begin{aligned}
& -\frac{1}{4!} \left[(t-x)^4 f^{(4)}(t) \right]_{x_0}^x + \frac{1}{4!} \int_{x_0}^x (t-x)^4 f^{(5)}(t) dt \\
& = f(x_0) + (x-x_0)f'(x_0) + \frac{1}{2}(x-x_0)^2 f''(x_0) + \frac{1}{3!}(x-x_0)^3 f'''(x_0) \\
& \quad + \frac{1}{4!}(x-x_0)^4 f^{(4)}(x_0) + \frac{1}{4!} \int_{x_0}^x (t-x)^4 f^{(5)}(t) dt.
\end{aligned}$$

Der n -te Summand wird also durch $\frac{1}{n!}(x-x_0)^n f^{(n)}(x_0) = \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n$ gegeben sein und das letzte Integral durch

$$R_{f,n}(x, x_0) := \frac{1}{n!} \int_{x_0}^x (x-t)^n f^{(n+1)}(t) dt.$$

Das Ergebnis unserer Überlegungen besteht nun in der folgenden Formel:

$$\begin{aligned}
f(x) &= f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \frac{f'''(x_0)}{3!}(x-x_0)^3 + \cdots + \\
&\quad \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + R_n(x_0, x).
\end{aligned} \tag{12.18}$$

Das ist die *TAYLOR-Entwicklung der Funktion f um den Punkt x_0 herum bis zur n-ten Ordnung*. Das Polynom

$$T_{f,n}(x) := f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n \tag{12.19}$$

ist das *TAYLOR-Polynom n-ter Ordnung*. Der Ausdruck $R_{f,n}(x, x_0)$ wird *Restglied* der Entwicklung genannt. Für die uns interessierenden Funktionen hat es die Eigenschaft im Grenzwert $n \rightarrow \infty$ gleichmäßig bzgl. x aus einem Intervall $[x_0 - R, x_0 + R]$ gegen Null zu konvergieren. Das bedeutet, daß $f(x)$ auf diesem Intervall durch die TAYLOR-Reihe gegeben ist:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} (x-x_0)^k.$$

Daher erhält man bei großem n eine gute Näherung für den Funktionswert $f(x)$, wenn man das Restglied einfach wegläßt:

$$f(x) \approx f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n.$$

Die meisten Funktionen werden um die Stelle $x_0 = 0$ herum entwickelt. Dann lautet die TAYLOR-Formel etwas einfacher:

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + R_n(0, x), \tag{12.20}$$

bzw.

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k. \tag{12.21}$$

Wie wendet man diese Formel an? Für eine konkrete Funktion f verlangt sie von uns, daß wir die Ableitungen bis zu der Ordnung bestimmen, bis zu der wir die Funktion entwickeln wollen. Dann haben wir nur noch die Stelle x_0 in die Ableitungen einzusetzen. Wenn wir in der Lage sind, die allgemeine Form des n -ten Summanden zu bestimmen, dann können wir die TAYLOR-Entwicklung bequem auf jede Ordnung ausdehnen.

- i) $f(x) = e^x$. Wir wissen bereits $f'(x) = f''(x) = \dots = f^{(n)}(x) = e^x$, so daß $f(0) = f'(0) = \dots = f^{(n)}(0) = e^0 = 1$ gilt. Nun müssen wir nur noch in (12.21) einsetzen und erhalten die uns schon bekannte Exponentialreihe (10.55):

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{3!}x^3 + \dots + \frac{1}{n!}x^n + \dots = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

- ii) $f(x) = \sin(x)$. Dann ist $f'(x) = \cos(x)$, $f''(x) = -\sin(x)$, $f'''(x) = -\cos(x)$ und $f^{(4)}(x) = \sin(x)$. Ab hier wiederholen sich die Ableitungen in genau derselben Reihenfolge, wie die ersten vier. Wir erhalten $f(0) = 0$, $f'(0) = 1$, $f''(0) = 0$, $f'''(0) = -1$, $f^{(4)}(0) = 0$, $f^{(5)}(0) = 1$, $f^{(6)}(0) = 0$, ... Versuchen wir das Bildungsgesetz aufzustellen: Offensichtlich verschwinden alle geraden Ableitungen an der Stelle 0. Die ungeraden Ableitungen sind abwechselnd 1 und -1 . Setzen wir in die TAYLOR-Formel ein, so erhalten wir

$$\sin(x) = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \dots + \frac{(-1)^k}{(2k+1)!}x^{2k+1} + \dots = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!}x^{2k+1}.$$

Die Approximation von \sin durch die TAYLOR-Polynome $T_n := T_{\sin,n}$ bis zur Ordnung $n = 21$ ist auf Seite 395 wiedergegeben.

- iii) $f(x) = \cos(x)$. Dasselbe Verfahren ergibt hier

$$\cos(x) = 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \dots + \frac{(-1)^k}{(2k)!}x^{2k} + \dots = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!}x^{2k}.$$

- iv) $f(x) = \ln(1+x)$. Die TAYLOR-Entwicklung einer Funktion ist eine Potenzreihe. Da diese eindeutig sind, ist es egal, auf welche Weise man sie gewinnt. Wir könnten, gemäß (12.21) alle Ableitungen bilden und dann versuchen, ein Bildungsgesetz zu erkennen (was in diesem Fall tatsächlich nicht schwierig ist). Wir wollen an diesem Beispiel aber einmal zeigen, daß die Entwicklung mitunter auch auf anderem Wege zu finden sein kann. Der Schlüssel dazu ist die geometrische Reihe und

$$\frac{d}{dx} \ln(1+x) = \frac{1}{1+x} = \sum_{k=0}^{\infty} (-1)^k x^k, \quad x \in (-1, 1).$$

Erinnern wir uns daran, daß eine Potenzreihe gliedweise integriert werden kann, ohne daß sich der Konvergenzradius ändert (man überlege sich, daß eine mögliche additive Konstante hier Null sein muß):

$$\ln(1+x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k+1} x^{k+1} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k}.$$

An diesem Beispiel sehen wir auch, daß diese Entwicklung im Allgemeinen nicht für alle $x \in \mathbb{R}$ sinnvoll ist. Offensichtlich ist $x = -1$ auf der linken Seite nicht erlaubt, denn 0 liegt nicht im Definitionsbereich von \ln . Für diesen Wert kann also auch die rechte Seite keine Annäherung an den Funktionswert ergeben, da dieser ja gar nicht vorhanden ist.

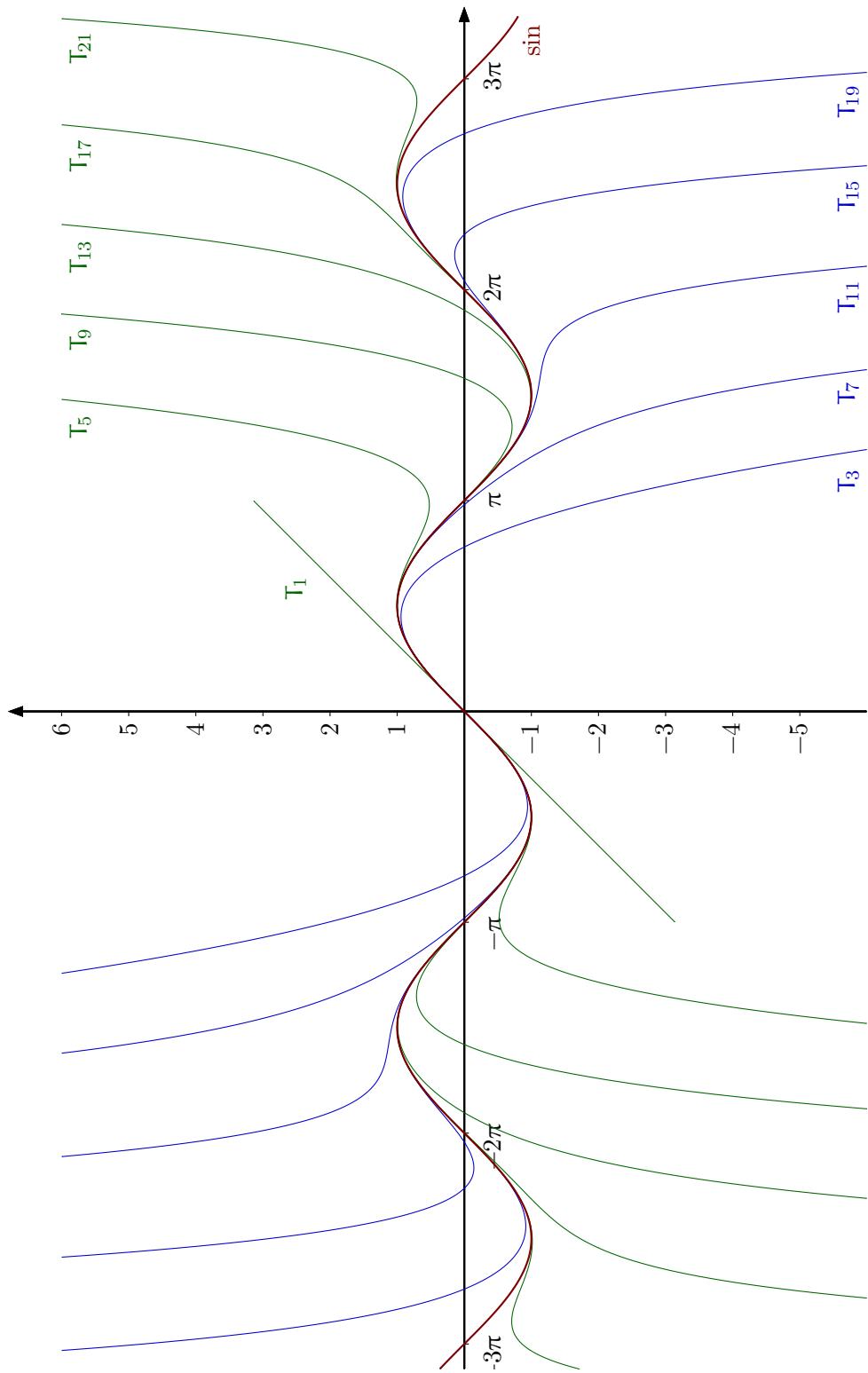
Tatsächlich ergibt die rechte Seite für $x = -1$ (bis auf ein gemeinsames Vorzeichen) die harmonische Reihe, die gegen $+\infty$ divergiert. Für $x = 1$ dagegen entsteht die LEIBNIZ-Reihe (10.38), die zwar konvergiert, aber nicht absolut konvergent ist. Als Nebenprodukt unserer Überlegungen erhalten wir den Wert dieser Reihe durch Einsetzen von 1 in $\ln(1 + x)$:

$$\ln(2) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{1}{k}.$$

Allerdings konvergiert sie nur sehr langsam. So ist für $n = 1000$ der genäherte Wert der Reihenentwicklung 0.6926474305598223, gegenüber $\ln(2) = 0.6931471805599453 \dots$

- v) $f(x) = e^{-x^2}$. Die TAYLOR-Entwicklung für diese Funktion erhalten wir einfach dadurch, daß wir $-x^2$ statt x in die TAYLOR-Entwicklung von e^x einsetzen. Das ergibt:

$$e^{-x^2} = \sum_{n=0}^{\infty} (-1)^n \frac{1}{n!} x^{2n}.$$

Abbildung 12.2 Approximation von \sin durch die TAYLOR-Polynome T_1, T_3, \dots, T_{21}

12.3.2 Das Volumen eines Rotationskörpers

Bisher haben wir das Integral zur Berechnung von Flächeninhalten verwendet. Tatsächlich hat es aber viel weiterreichende Anwendungsmöglichkeiten. Um für eine gegebenes Problem den richtigen Integralausdruck zu finden, gehen wir den Weg über die RIEMANN-Summen. Wir führen das hier am Beispiel des Volumens eines Rotationskörpers vor. Dazu denken wir uns den Graphen einer Funktion f um die x -Achse rotierend. Dabei beschreibt er die Oberfläche eines Rotationskörpers (siehe nebenstehende Skizze). Das Volumen bestimmen wir näherungsweise, indem wir den Körper in Scheiben mit der Breite Δx_i und dem Radius $f(x_i)$ zerlegen. Dabei ist $a = x_0 < x_1 < x_2 < \dots < x_i < x_{i+1} < \dots < x_n < x_{n+1} = b$ eine Zerlegung \mathcal{Z} des Basisintervalls $[a, b]$ mit der Feinheit $|\mathcal{Z}|$ (vergl. Seite 378). Das Volumen $\pi f(x_i)^2 \Delta x_i$ einer solchen Scheibe ist das eines Zylinders mit Radius $f(x_i)$ und Höhe Δx_i . Zählen wir die Volumina dieser Scheiben zusammen, so erhalten wir den Näherungswert

$$V \approx \pi(f(x_0)^2 \Delta x_0 + f(x_1)^2 \Delta x_1 + f(x_2)^2 \Delta x_2 + \dots + f(x_n)^2 \Delta x_n) = \pi \sum_{i=0}^n f^2(x_i) \Delta x_i.$$

Abb. 12.3 Volumen eines Rotationskörpers

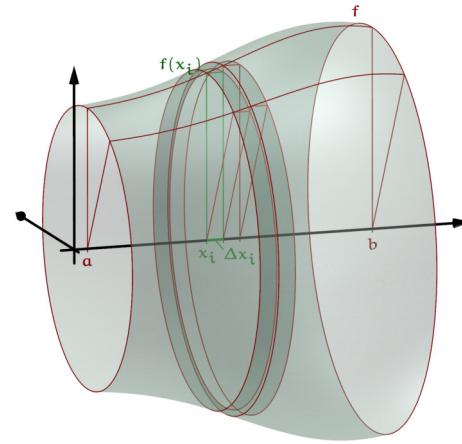
Um den genauen Wert des Volumens zu erhalten, führen wir den Grenzwert $|\mathcal{Z}| \rightarrow 0$ beliebig feiner Zerlegungen durch:

$$V = \lim_{|\mathcal{Z}| \rightarrow 0} \pi \sum_{i=0}^n f^2(x_i) \Delta x_i.$$

Wenn wir das mit (12.1) vergleichen, so sehen wir, daß das Volumen V der Grenzwert von RIEMANN-Summen für die Funktion πf^2 ist. Das Volumen des Rotationskörpers ist demnach durch das Integral

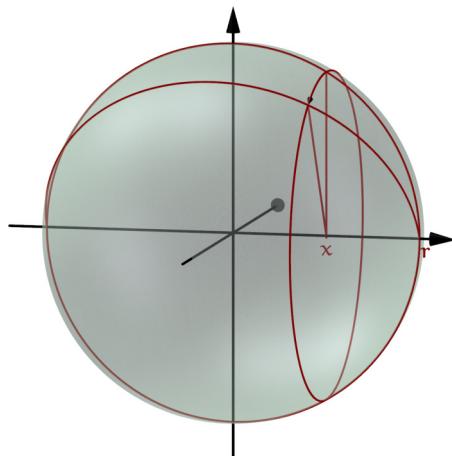
$$V = \pi \int_a^b f^2(x) dx \quad (12.22)$$

gegeben.



Als Beispiel berechnen wir das Volumen einer Kugel. Dafür lassen wir einen Halbkreis um die x -Achse rotieren. Seine Gleichung ist $f(x) = \sqrt{r^2 - x^2}$. Damit ergibt sich für das Kugelvolumen:

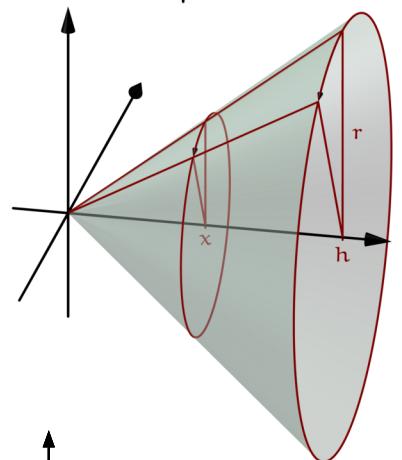
$$\begin{aligned} V &= \pi \int_{-r}^r f^2(x) dx = \pi \int_{-r}^r (r^2 - x^2) dx \\ &= \pi \left[r^2 x - \frac{1}{3} x^3 \right]_{-r}^r \\ &= \pi \left(r^3 - \frac{1}{3} r^3 \right) - \pi \left(-r^3 + \frac{1}{3} r^3 \right) \\ &= \frac{4}{3} \pi r^3. \end{aligned}$$



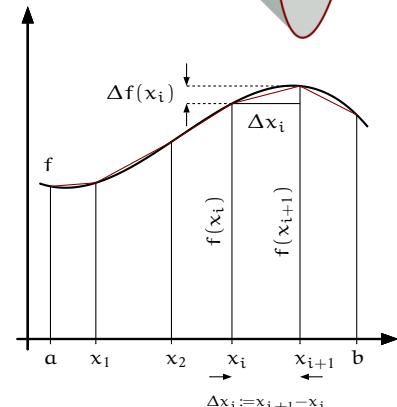
Das Volumen eines Kreiskegels erhalten wir durch Rotation der Gerade $f(x) = mx$ um die x -Achse:

$$\begin{aligned} V &= \pi \int_0^h f^2(x) dx = \pi \int_0^h m^2 x^2 dx \\ &= \pi \left[\frac{1}{3} m^2 x^3 \right]_0^h = \frac{1}{3} \pi (mh)^2 h = \frac{1}{3} \pi r^2 h, \end{aligned}$$

denn der Radius r des Grundkreises ist $r = mh$.



12.3.3 Die Länge einer Kurve Die Länge des Bogens eines Funktionsgraphen bestimmen wir zunächst näherungsweise, indem wir ihn durch einanderstoßende Geradenstücke, einen sog. *Polygonzug*, ersetzen. Dazu führen wir wieder die Zerlegung \mathcal{Z} , $a = x_0 < x_1 < x_2 < \dots < x_i < x_{i+1} < \dots < x_n < x_{n+1} = b$, des Basisintervalls $[a, b]$ mit der Feinheit $|\mathcal{Z}|$ ein. Dann ersetzen wir den tatsächlichen Kurvenverlauf auf jedem der kleinen Teilintervalle $[x_i, x_{i+1}]$ durch ein Geradenstück, das die Randpunkte $[x_i, f(x_i)]$ und $[x_{i+1}, f(x_{i+1})]$ des Kurvenbogens über $[x_i, x_{i+1}]$ miteinander verbindet. Die Länge eines solchen Geradenstücks ist nach dem Satz von PYTHAGORAS:



$$\ell_i = \sqrt{(\Delta x_i)^2 + (\Delta f(x_i))^2} = \sqrt{1 + \left(\frac{\Delta f(x_i)}{\Delta x_i} \right)^2} \cdot \Delta x_i.$$

Die Länge L des Kurvenbogens zwischen a und b erhalten wir dann ungefähr, wenn wir die Längen ℓ_i aufsummieren:

$$L \approx \sum_{i=0}^n \ell_i = \sum_{i=0}^n \sqrt{1 + \left(\frac{\Delta f(x_i)}{\Delta x_i} \right)^2} \cdot \Delta x_i.$$

Den genauen Wert für L hoffen wir wieder durch den Grenzwert $|\mathcal{Z}| \rightarrow 0$ verschwindender Feinheit der Zerlegung \mathcal{Z} zu gewinnen. Dafür müssen wir die richtige RIEMANN-Summe finden, die uns verrät, welches Integral wir zur Berechnung von L benutzen können. In obiger Summe steht unter der Wurzel der Differenzenquotient $\frac{\Delta f(x_i)}{\Delta x_i}$, der für $|\mathcal{Z}| \rightarrow 0$, also für kleines Δx_i , durch die Ableitung $f'(x_i)$ ersetzt werden kann. Dadurch erhalten wir nun tatsächlich eine RIEMANN-Summe für die Approximation von L :

$$L \approx \sum_{i=0}^n \sqrt{1 + f'(x_i)^2} \cdot \Delta x_i.$$

Wir vergleichen das wieder mit (12.1) und sehen, daß die Länge L durch das Integral

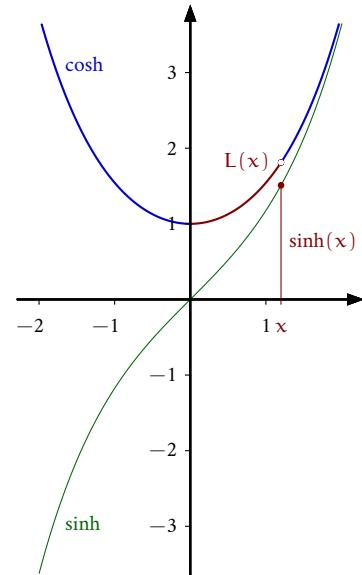
$$L = \int_a^b \sqrt{1 + f'^2(x)} dx \quad (12.23)$$

gegeben sein muß.

Als Anwendung bestimmen wir die Länge der sog. *Kettenlinie*, die durch die Funktion \cosh gegeben ist. Sie hat die Ableitung und die Stammfunktion \sinh . Der Name *Kettenlinie* für \cosh stammt daher, daß diese Funktion den Verlauf einer Kette beschreibt, die zwischen zwei Punkten lose aufgehängt ist.

Jetzt können wir die Länge $L(x)$ der Kettenlinie im Bereich von 0 bis x berechnen (dabei verwenden wir $1 + \sinh^2 = \cosh^2$):

$$\begin{aligned} L(x) &= \int_0^x \sqrt{1 + \cosh'^2(t)} dt = \int_0^x \sqrt{1 + \sinh^2(t)} dt \\ &= \int_0^x \sqrt{\cosh^2(t)} dt = \int_0^x \cosh(t) dt = [\sinh(t)]_0^x \\ &= \sinh(x). \end{aligned}$$



Die Funktion $x \mapsto \sinh(x)$ gibt also nicht nur die Ableitung und die Stammfunktion von \cosh wieder, sondern auch die Länge des Kurvenbogens dieser Funktion über dem Intervall $[0, x]$.

12.3.4 Transzendenz von e^* Eine reelle Zahl x heißt *algebraisch*, falls sie die Lösung einer *algebraischen Gleichung* ist, also einer Gleichung

$$a_0 + a_1x + a_2x^2 + \cdots + a_nx^n = 0$$

mit ganzzahligen Koeffizienten a_k . Andernfalls heißt x *transzendent*.

Zu entscheiden, ob eine Zahl transzendent ist oder nicht, ist meist eine wirklich schwere Aufgabe. Für jede Zahl ist dabei jeweils eine ganz eigenständige zündende Idee nötig. Im Laufe der Zeit sind die Beweise für so wichtige Zahlen wie π und e , gegenüber den Originalbeweisen einfacher geworden. Ein besonders eleganter Beweis für die Transzendenz von e stammt von HILBERT. Wie bei solchen Beweisen üblich, geht er davon aus, daß e doch die Lösung einer algebraischen Gleichung

$$a_0 + a_1e + a_2e^2 + \cdots + a_ne^n = 0 \quad (*)$$

ist und führt das zu einem Widerspruch. Wir können $a_0 > 0$ annehmen. Wir gehen von den Integralen $J_k := \int_k^\infty x^q [(x-1)(x-2)\cdots(x-n)]^{q+1} e^{-x} dx$ für $k = 0, 1, \dots, n$ aus und zeigen, daß $\frac{1}{q!} J_0$ und $\frac{e^k}{(q+1)!} J_k$ für $k = 1, \dots, n$ jeweils ganze Zahlen sind. Wir multiplizieren Gleichung (*) mit $\frac{1}{q!} J_0$ durch und spalten die auftretenden Terme $a_k \frac{e^k}{q!} J_0$ folgendermaßen auf

$$a_k \frac{e^k}{q!} J_0 = a_k \frac{e^k}{q!} \int_0^\infty \cdots dx = a_k \frac{e^k}{q!} \int_k^\infty \cdots dx + a_k \frac{e^k}{q!} \int_0^k \cdots dx \quad (**)$$

Wir erhalten

$$\sum_{k=0}^n a_k \frac{e^k}{q!} J_k + \sum_{k=0}^n a_k \frac{e^k}{q!} \int_0^k \cdots dx = 0.$$

Die erste Summe ist für jedes $q \in \mathbb{N}$ eine ganze Zahl, die für ausreichend große Primzahlen $q+1$ nicht 0 ist. Die zweite Summe kann für genügend große q beliebig klein gemacht werden, so daß sie die erste Summe nicht mehr kompensieren kann. Das ist der gesuchte Widerspruch.

Bevor wir uns mit der Durchführung dieses Plans befassen, stellen wir ein Hilfsmittel bereit: Für jedes Polynom p vom Grad m gilt, wenn wir, wie üblich, mit $p^{(k)}$ die k -te Ableitung von p bezeichnen:

$$\int p(x)e^{-x} dx = - \sum_{k=0}^m p^{(k)}(x)e^{-x}. \quad (12.24)$$

Die Ableitung der rechten Seite ergibt nämlich

$$\sum_{k=0}^m p^{(k)}(x)e^{-x} - \sum_{k=0}^{m-1} p^{(k+1)}(x)e^{-x} = \sum_{k=0}^m p^{(k)}(x)e^{-x} - \sum_{\ell=1}^m p^{(\ell)}(x)e^{-x} = p(x)e^{-x}.$$

Insbesondere ist

$$\int x^m e^{-x} dx = -m! \sum_{k=0}^m \frac{x^k}{k!} e^{-x}$$

und daher

$$\int_0^\infty x^m e^{-x} dx = -m! \lim_{t \rightarrow \infty} \left[\sum_{k=0}^m \frac{x^k}{k!} e^{-x} \right]_0^t = m!. .$$

Für jedes Polynom $p(x) = \sum_{k=0}^m b_k x^k$ mit ganzzahligen Koeffizienten b_k folgt

$$\int_0^\infty p(x)e^{-x} dx = \sum_{k=0}^m b_k \cdot k! \in \mathbb{Z}$$

und

$$\frac{1}{q!} \int_0^\infty x^q \cdot p(x)e^{-x} dx = \sum_{k=0}^m b_k \cdot \frac{(k+q)!}{q!} = \sum_{k=0}^m b_k \cdot k! \binom{k+q}{k} \in \mathbb{Z} \quad (***)$$

In dieser Summe enthalten alle Summanden, außer eventuell dem ersten, den Faktor $q + 1$.

J_0 ist ein Integral dieses Typs. Das Polynom $p(x) := [(x - 1) \cdots (x - n)]^{q+1}$ hat den ersten Koeffizienten $b_0 = \pm(n!)^{q+1}$. Wählen wir $q + 1$ prim und größer als n , dann sind nach Korollar 3.1.10 $n!$ und $q + 1$ teilerfremd. Daher ist

$$\frac{1}{q!} J_0 = \pm(n!)^{q+1} \mod q + 1 \neq 0.$$

Die Integrale J_k für $k > 0$ werden durch die Substitution $y = x - k$ auf den Typ (***)) zurückgeführt:

$$\begin{aligned} \frac{e^k}{(q+1)!} J_k &= \frac{1}{(q+1)!} \int_k^\infty x^q [(x-1) \cdots (x-k) \cdots (x-n)]^{q+1} e^{-(x-k)} dx \\ &= \frac{1}{(q+1)!} \int_0^\infty (y+k)^q (y+k-1)^{q+1} \cdots y^{q+1} \cdots (y+k-n)^{q+1} e^{-y} dy \in \mathbb{Z}. \end{aligned}$$

Daher ist

$$\sum_{k=0}^n a_k \frac{e^k}{q!} J_k = \frac{1}{q!} a_0 J_0 + (q+1) \sum_{k=0}^n a_k \frac{e^k}{(q+1)!} J_k = \pm(n!)^{q+1} a_0 \mod q + 1 \neq 0.$$

wenn wir auch noch dafür sorgen, daß $q + 1 > a_0$ gilt.

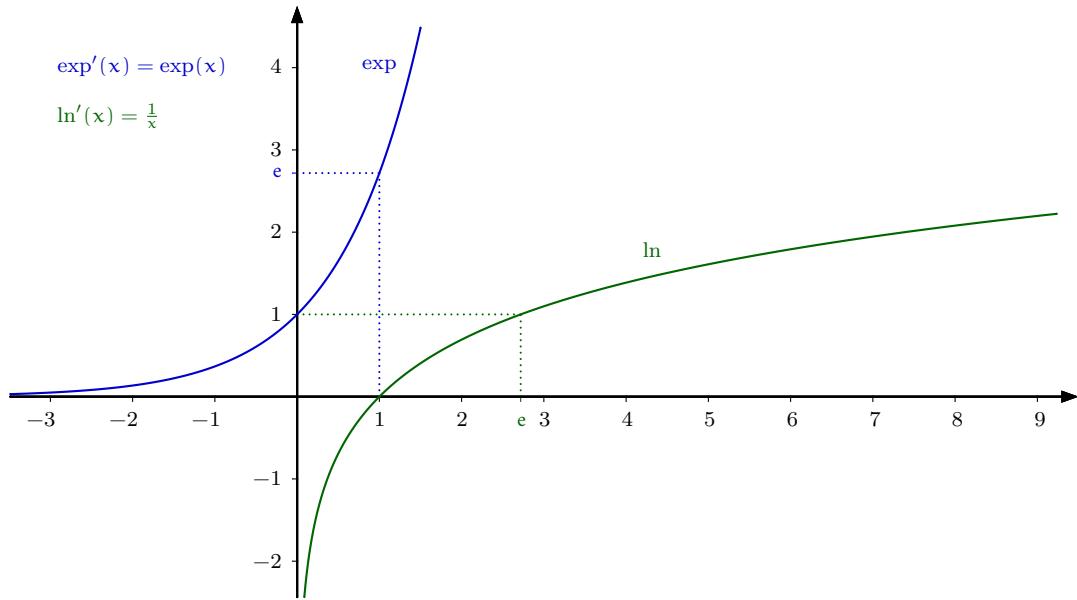
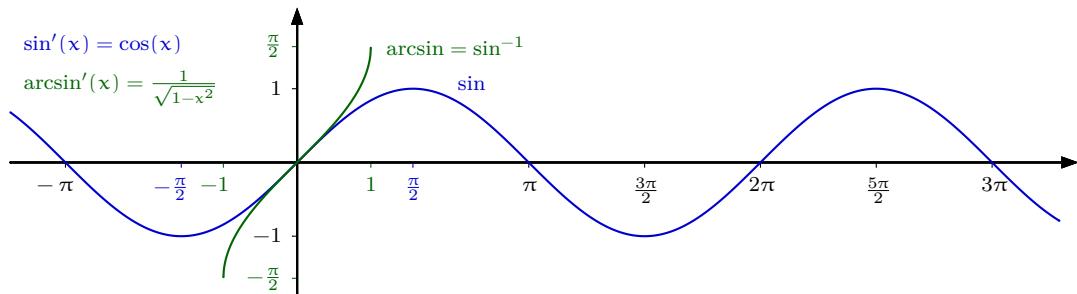
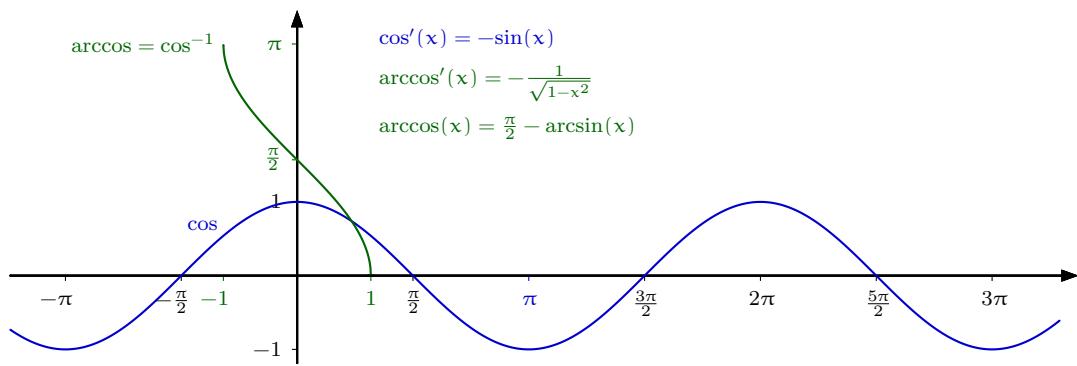
Nun zur zweiten Summe in (**). Aus dem Mittelwertsatz der Integralrechnung erhalten wir ein $\xi \in (0, k)$, mit dem wir folgendermaßen abschätzen können:

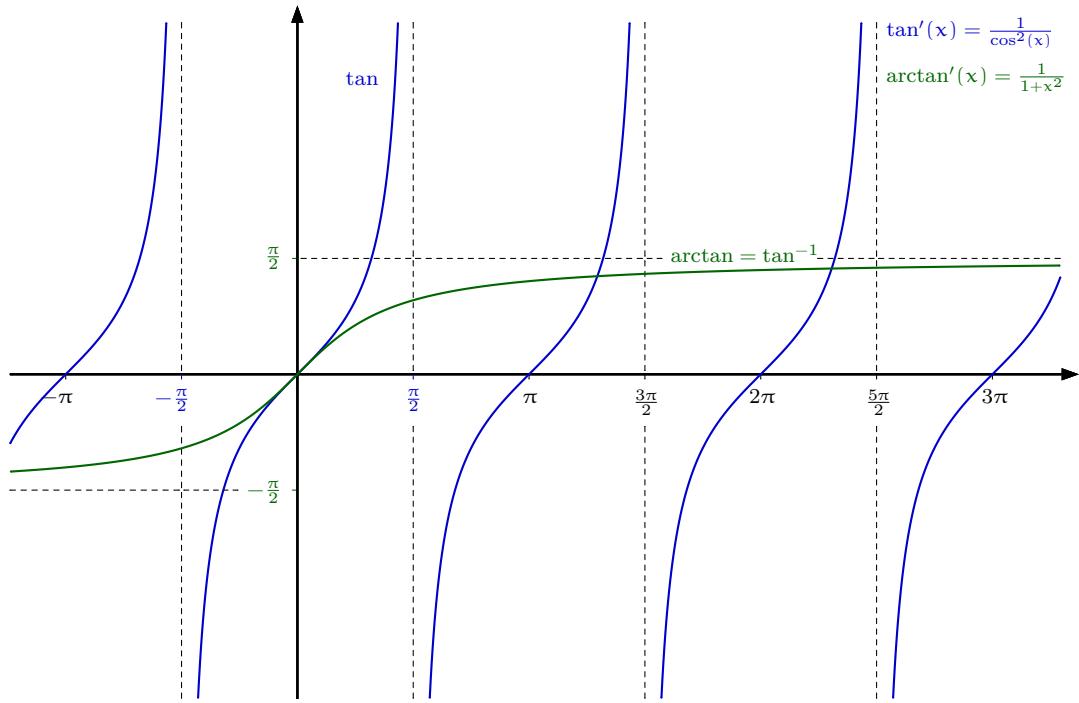
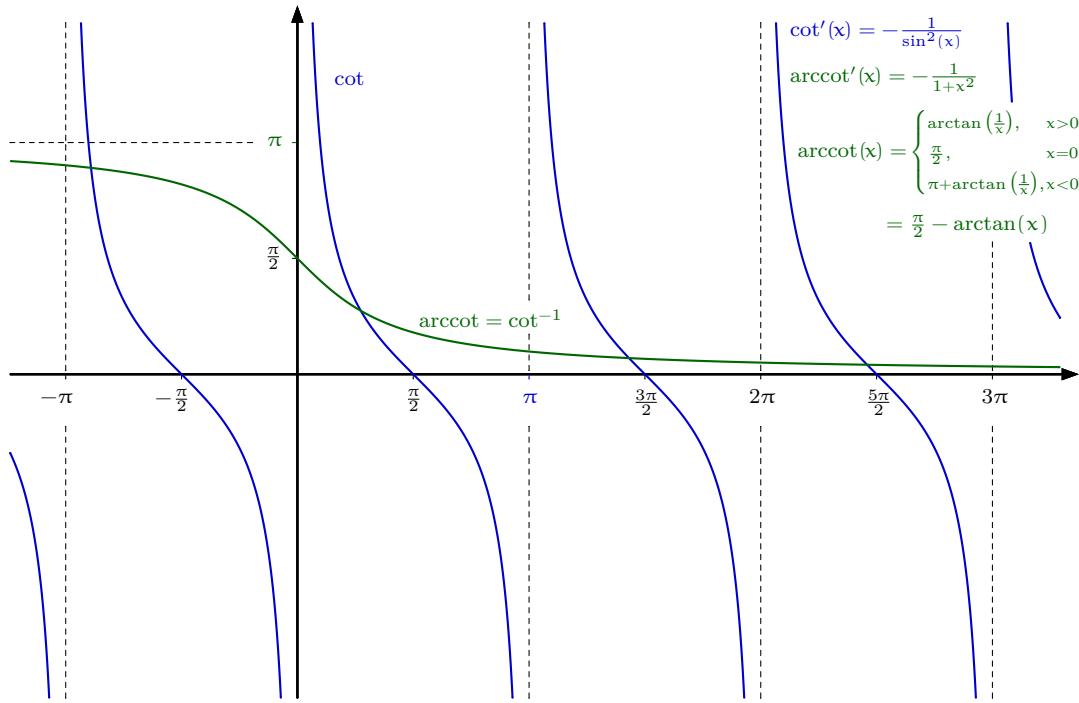
$$\begin{aligned} \left| \frac{e^k}{q!} \int_0^k x^q [(x-1) \cdots (x-n)]^{q+1} e^{-x} dx \right| &= k \frac{e^k}{q!} \left| \xi^q [(\xi-1) \cdots (\xi-n)]^{q+1} e^{-\xi} \right| \\ &\leq \frac{e^n}{q!} n^{(n+1)(q+1)} = e^n n^{n+1} \frac{(n^{n+1})^q}{q!} \xrightarrow[q \rightarrow \infty]{10.13} 0. \end{aligned}$$

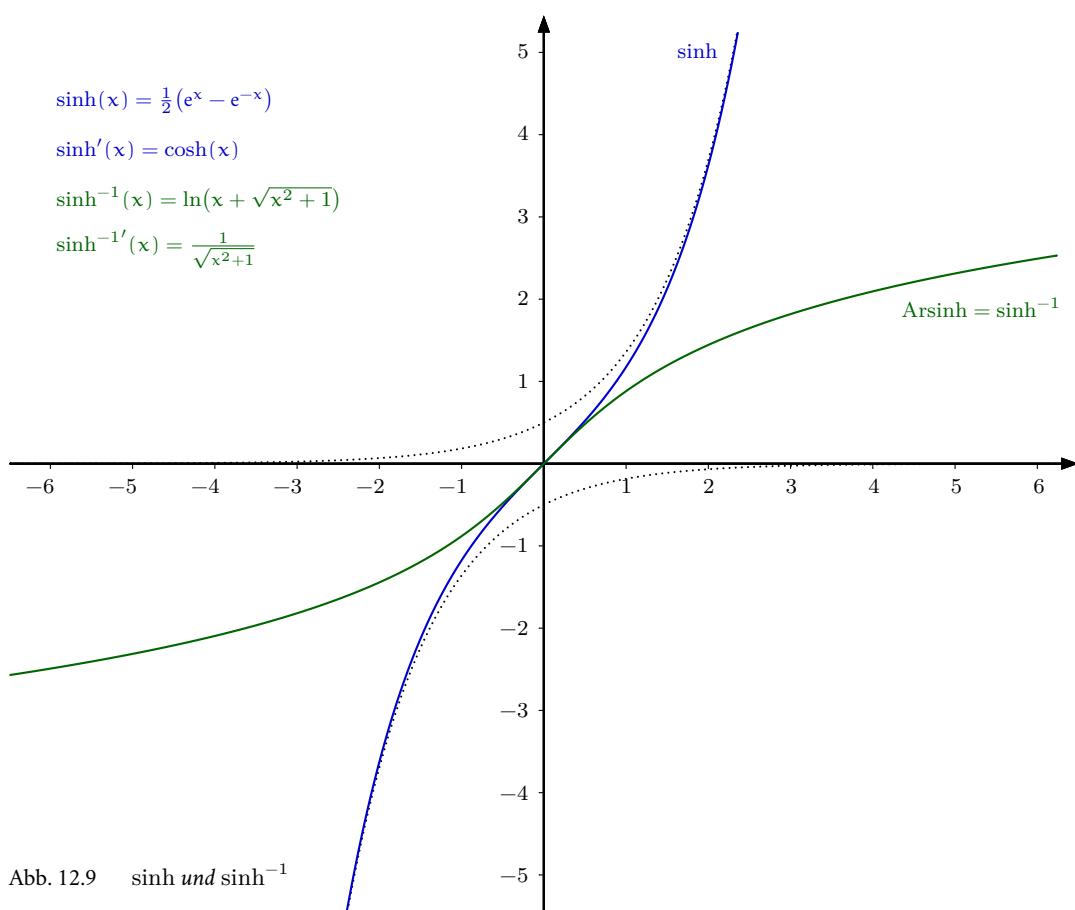
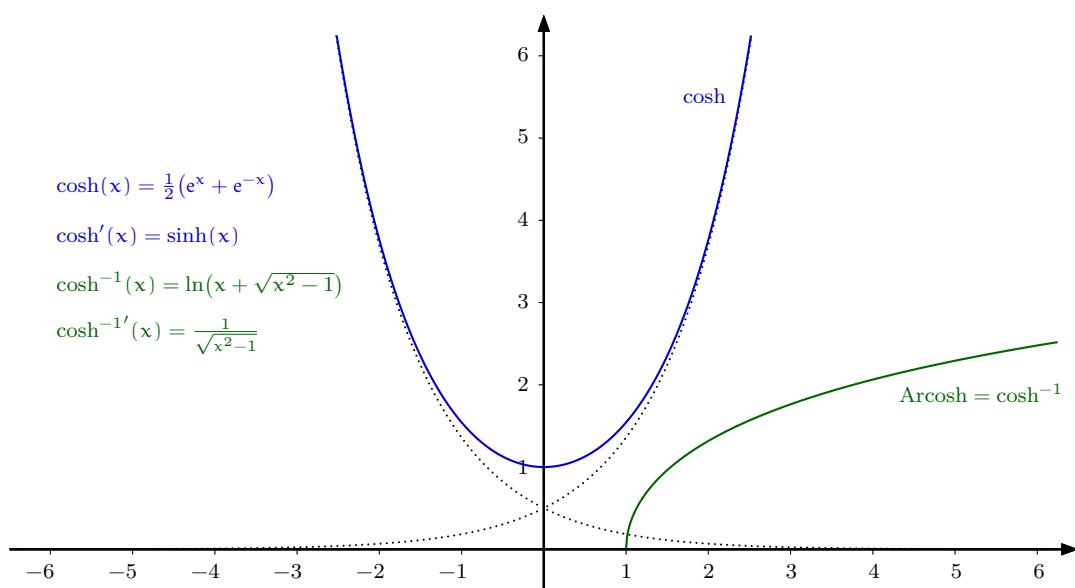
Die zweite Summe in (**) besteht aus n Summanden dieser Art, so daß sie für ausreichend großes q beliebig klein gemacht werden kann.

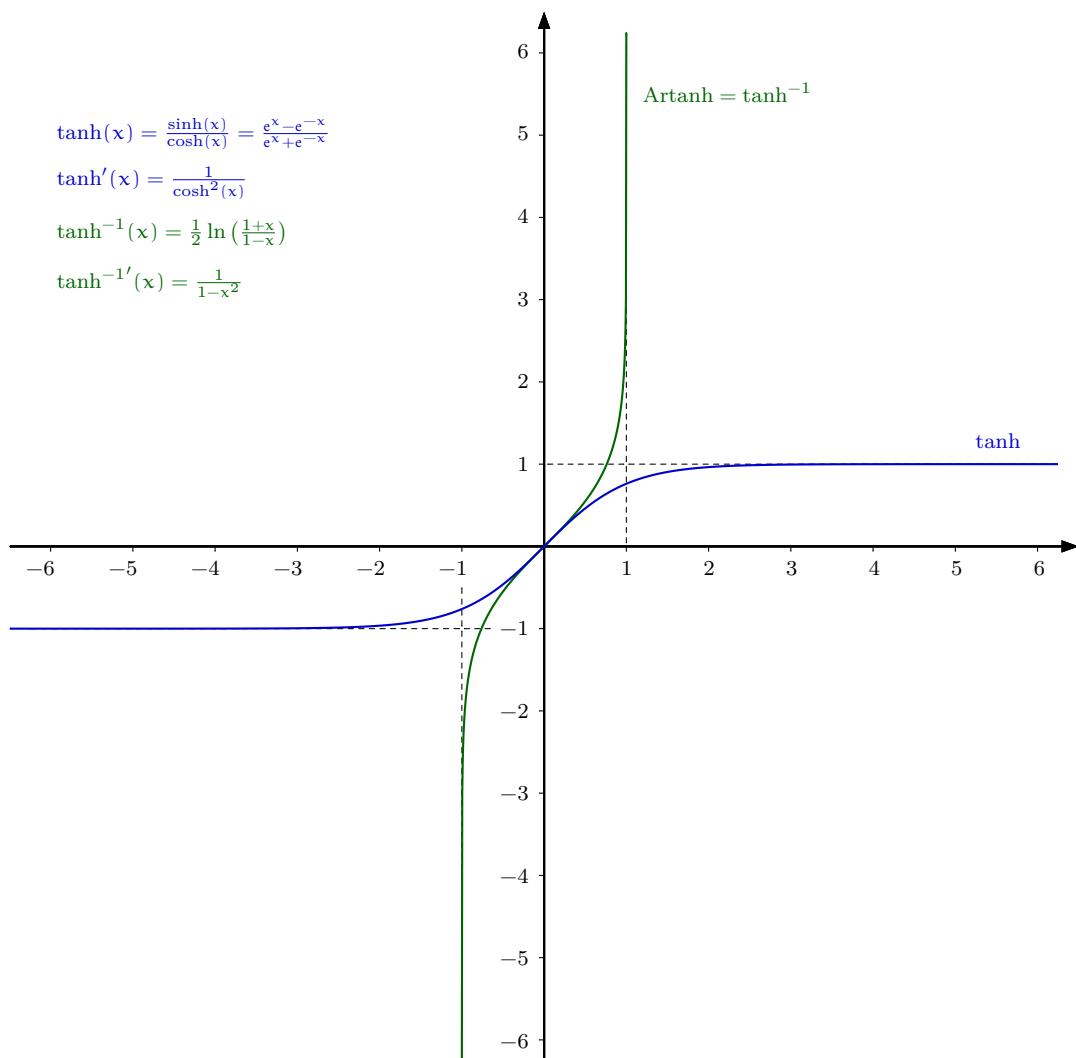
Damit haben wir alle Schritte unseres Plans durchgeführt und den behaupteten Widerspruch nachgewiesen. Als Ergebnis ist e als transzendent erkannt.

12.4 Schaubilder der elementaren Funktionen

Abb. 12.4 \exp und \ln Abb. 12.5 \sin und \arcsin Abb. 12.6 \cos und \arccos

Abb. 12.7 \tan und \arctan Abb. 12.8 \cot und arccot

Abb. 12.9 \sinh und \sinh^{-1} Abb. 12.10 \cosh und \cosh^{-1}

Abb. 12.11 \tanh und \tanh^{-1}

Nomenclature

f' , $\frac{d}{dx}$	Ableitung	284
A^*	adjungierte Matrix	144
\Leftrightarrow	äquivalent	3
\sim	Äquivalenzrelation	22
B_k	BERNOULLI-Zahlen	329
$ z $	Betrag	78
$\text{im } f$	Bild, Wertebereich	25
$\binom{\alpha}{k}$	Binomialkoeffizient, verallgemeinerter	327
$\binom{n}{k}$	Binomialkoeffizient	15
D_f	Definitionsbereich	25
\det	Determinante	73
$A \setminus B$	Differenzmenge	10
\dim	Dimension	122
\oplus	direkte Summe	135
C_1	Komplexer Einheitskreis	80
1	Einheitsmatrix	135
U_ε	ε -Umgebung	227
\exists	es existiert	5
$(x_n)_{n \in \mathbb{N}}, (x_n)$	Folge	209
\Rightarrow	folgt	2
$x \mapsto f(x)$	Funktion	25
\forall	für alle	5
$[y]$	GAUSS-Klammer	245
ggT	größter gemeinsamer Teiler	36
φ, Φ	goldener Schnitt	205
grad	Grad eines Polynoms	337
$\lim_{x \rightarrow a^\pm}$	Grenzwert, links-/rechtsseitiger	264
$\lim_{n \rightarrow \infty}, \xrightarrow{n \rightarrow \infty}$	Grenzwert	212
id	identische Funktion	26
i	imaginäre Einheit	76

(a, b) , $[a, b]$	Intervalle	8
R^{-1}	inverse Relation	24
\mathcal{E}	kanonische Basis $\{ e_1, \dots, e_n \}$	123
\ker	Kern	138
kgV	kleinstes gemeinsames Vielfaches	48
\leqslant	kleiner oder gleich	21
A^c	Komplement einer Menge	10
\mathbb{C}	komplexe Zahlen	77
\bar{z}	konjugiert komplexe Zahl	78
$x \otimes y$	Kreuzprodukt	68
δ_{ik}	KRONECKER-Symbol	129
\mathbb{K}	Körper	90
\emptyset	leere Menge	10
$\liminf_{n \rightarrow \infty}, \underline{\lim}_{n \rightarrow \infty}$	Limes Inferior	248
$\limsup_{n \rightarrow \infty}, \overline{\lim}_{n \rightarrow \infty}$	Limes Superior	248
lh	lineare Hülle	120
$\ \cdot\ _p$	L^p -Norm	97
L^p	L^p -Raum	98
$M_n(\mathbb{K})$, M_n	$n \times n$ -Matrizen	140
\max, \min	Maximum, Minimum	219
$x = y \pmod p, =_p$	Gleichheit modulo p	40
$\binom{n}{k_1, k_2, \dots, k_r}$	Multinomialkoeffizient	53
\mathbb{N}_n	natürliche Zahlen $\leq n$	30
\mathbb{N}, \mathbb{N}_0	natürliche Zahlen	12
$\neg A, \overline{A}$	Negation	2
$\ x\ $	Norm, Länge	65
o. B. d. A.	ohne Beschränkung der Allgemeinheit	8
V^\perp, x^\perp	orthogonaler Teilraum	136
S_n	Permutationen, symmetrische Gruppe	30
P_n	Polynome vom grad $\leq n$	347
\mathbb{P}	Primzahlen	9
$A \times B$	Produktmenge	10
\prod	Produktzeichen	15
$[P]$	Projektionsraum	177
ran	Rang einer Matrix	172
\mathbb{Q}	rationale Zahlen	9

$\operatorname{Re}(z)$, $\operatorname{Im}(z)$	Realteil, Imaginärteil	77
aRb , $R(a_1, \dots, a_n)$	Relation	19
$R_{f,n}$	Restglied	319
$\mathbb{Z}/p\mathbb{Z}$	Restklassen modulo p	24
\mathbb{R}^n	\mathbb{R}^n	10
\cap, \cup	Schnitt, Vereinigung von Mengen	10
sgn	Vorzeichen einer Permutation	32
sgn	Vorzeichen einer reellen Zahl	88
$\langle x y \rangle$	Skalarprodukt	65
$\sigma(A)$	Spektrum	173
\sum	Summenzeichen	15
\sup, \inf	Supremum, Infimum	219
$T_{f,n}$, T_f	TAYLOR-Polynom, TAYLOR-Reihe	319
$a b$	a teilt b	35
\subseteq, \supseteq	Teilmenge, Obermenge	7
A^t	Transponierte	105
t_{ij}	Transposition	30
f^{-1}	Umkehrfunktion	28
\wedge, \vee	und, oder	2
$f^{-1}(B)$	Urbild	26
$x = [x_1, \dots, x_n]^t$	Vektor	61
$g \circ f$	Verkettung von Funktionen	25
$S \circ R$	Verkettung von Relationen	24
$\sqrt[n]{\quad}$	n-te Wurzel	221
\mathbb{Z}_p^*	endlicher Körper zu $p \in \mathbb{N}$	41
$[a]_~, [a]$	Äquivalenzklasse	22

Index

- Abbildung
 - lineare, 132
 - längentreue, 146
 - normale, 181
 - positive, 179
 - schiefsymmetrische, 183
 - selbstadjungierte, 175
 - winkeltreue, 146
- ABELSche Summation, 314
- ABELSches Kriterium, 315
- Ableitung, 284
 - Kettenregel der, 287
 - Produkregel der, 286
 - Quotientenregel der, 287
 - Summenregel der, 286
- Ableitungsregeln, 286
- absolute Konvergenz, 241
 - einer Doppelreihe, 256
- Absolutglied, 337
- Abstand
 - Punkt zu Ebene, 70
 - Punkt zu Gerade, 75
- Additionssätze
 - der Trigonometrie, 79, 135, 272
- Adjungierte
 - Matrix, 144, 145
- Adjunkte, 169
- algebraische Gleichung, 399
- algebraische Zahl, 399
- Algorithmus
 - euklidischer, 36
 - erweiterter, 39, 46
- analytische Funktion, 324
- Ankathete, 272
- antilinear, 89, 93
- Antisymmetrie, 93
- arccos, 309, 401
- arccot, 402
- Archimedisches Axiom, 220
- Arcosh, 403
- arcsin, 309, 401
- arctan, 309, 402
- Arcuscosinus, 309, 401
- Arcuscotangens, 309, 402
- Arcussinus, 309, 401
- Arcustangens, 309, 402
- Areacosinus, 310
- Areasinus, 310
- arithmetisches Mittel, 375
- Arsinh, 403
- Artanh, 404
- Asympote, 347
- Aufspann, x
- Aussageform, 5, 8
- Aussagen, 1
- Aussagenlogik, 1
- Banachraum, 95
- Basis, 120
 - duale, 147, 149
 - Entwicklung nach einer, 124, 147
 - kanonische, 123
- Basiswechsel, 147
- BERNOULLI-Polynome, 329
- BERNOULLI-Ungleichung, 216, 373
 - umgekehrte, 373
- BERNOULLI-Zahlen, 329
- bestimmt divergent, 301
- Betrag
 - einer komplexen Zahl, 78
- Betragsfunktion, 380

- Beweis
 direkter, 3
 durch Widerspruch, 4
- bijektiv, 27
- Binom
 drittes, 29, 189
- Binomialkoeffizient, 15, 44, 53
 verallgemeinerter, 327
- binomische Reihe, 327
- binomischer Lehrsatz, 15
- Blockmatrix, 161
- Brennpunkt, 185, 187, 189
- BROWNSche Bewegung, 255
- Cardanische Formeln, 76, 84
- CATALAN Zahlen, 255
- CAUCHY-Bedingung, 232
- CAUCHY-Folge, 232
 gleichmäßige, 239, 280
- CAUCHY-Kriterium, 232
- CAUCHY-Produkt
 endliches, 340
- CAUCHY-Regel, 350
- CAUCHY-SCHWARZ-Ungleichung, 67, 93, 94, 97
- charakteristisches Polynom, 174
- Chinesischer Restsatz, 47
- \cos, \cos^{-1} , 271, 309, 401
- \cosh, \cosh^{-1} , 310, 403
- \cot, \cot^{-1} , 271, 309, 402
- DE L'HOSPITAL
 Regel von, 300
- Definitheit, 93
- δ_{ik} , 129, 135
- DESCARTES
 Regel von, 359
- Determinante, 73, 157
 Summendarstellung, 159
- Determinanten-Produktsatz, 161
- Dezimalentwicklung, 244
- Diagramm
 kommutierendes, 154
- Differentialrechnung, 283
- Differenzenquotient, 284
- differenzierbar, 284
- Dimension
 eines Vektorraums, 122
- direkte Summe, 135
- DIRICHLET-Kriterium, 315
- disjunkt, 10
- Disjunktion, 2
- Diskriminante, 76, 84, 207
- divergent, 209
- Division
 von Hand, 246
- Doppelfakultät, 328
- Doppelfolge, 235
- Doppellimes, 235
- Doppelreihensatz, 256
- Drehung, 181–183
 ebene, 134
- Dreibein, 72
- Dreieck, 377
- Dreiecks-Blockmatrix
 obere/untere, 162
- Dreiecksform
 obere, 104, 116
- Dreiecksmatrix
 obere/untere, 160
- Dreiecksungleichung, 78, 93
 umgekehrte, 78, 96
- Dreieckszahl, 9
- 3-er Probe, 41
- duale Basis, 147, 149
- Eigenvektor, 173
- Eigenwert, 173
- Eigenwertgleichung, 173
- Einheitskreis, 80
- Einheitsmatrix, 135
- Einheitswurzel
 n -te, 81
- $\mathbb{1}$, 135
- 11-er Probe, 42
- Ellipse, 185
 Tangente, 185
- Ellipsoid, 194

- Entfernungsmodul, 308
- Entwicklungskoeffizienten, 124
- Entwickelpunkt, 252, 318
- ε - δ -Kriterium, 268
- ε -Schlauch, 280
- ε -Umgebung, 96, 227
- erzeugend, 120
- EULER-Formel, 79
- exp, 259, 401
- Exponentialfunktion, 259, 401
 - Funktionalgleichung der, 259
 - Stetigkeit der, 282
- Exponentialreihe, 210, 242, 250, 259
- Extremstelle, 363
- Extremum, 363
- Fakultät, 15
- fast alle, 227
- Faßkreisbogen
 - Satz vom, 149
- Feinheit, 378
- FERMAT
 - kleiner Satz von, 44
- FIBONACCI-Folge, 204, 206, 256
- Flachpunkt, 365
- Flächenproblem, 377
- Folge, 209
 - mit Parameter, 238
 - zulässige, 263
- Fundamentalsatz der Algebra, 338, 353
- Funktion, 21, 25
 - analytische, 324
 - Einschränkung der, 26
 - glatte, 325
 - identische, 26
 - konkave / konvexe, 371
 - stetige, 263
 - trigonometrische, 271
 - von TAKAGI, 294
- Funktionenfolge, 279
 - gleichmäßige Konvergenz, 280
 - punktweise Konvergenz, 279
- Funktionswert, 25
- GAUSS-Klammer, 245
- GAUSS-Verfahren, 100, 107
 - erweitertes, 141
 - komplexes, 125
- GAUSSsche Zahlenebene, 77, 78
- Gegenkathete, 272
- geometrische Reihe, 240
- geometrisches Mittel, 375
- glatte Funktion, 325
- Gleichheit
 - modulo, 40
- gleichmäßige Konvergenz, 238, 280
- Gleichung
 - kubische, 76, 84
- Gleichungssystem
 - in/homogenes, 100, 106
 - lineares, 99
- GOLDBACHSche Vermutung, 1
- goldener Schnitt, 205, 256
- Grad, 337
- GRAM-SCHMIDT-Verfahren, 129
- Graph
 - von \cos, \cos^{-1} , 401
 - von \cosh, \cosh^{-1} , 403
 - von \cot, \cot^{-1} , 402
 - von \exp, \ln , 401
 - von \sin, \sin^{-1} , 401
 - von \sinh, \sinh^{-1} , 403
 - von \tan, \tan^{-1} , 402
 - von \tanh, \tanh^{-1} , 404
 - von f^{-1} , 277
- Graph einer Funktion, 25
- GRASSMANN-Identität, 183
- GRASSMANN-PLÜCKER-Identität, 184
- Grenzwert, 212
 - iterierter, 235
 - links/rechtsseitiger, 264, 267
- Gruppe, 43
 - symmetrische, 30
- Gruppenhomomorphismus, 32
- Hauptsatz
 - der Differential- und Integralrechnung, 381

- Helligkeit
 absolute, scheinbare, 308
- HERON-Verfahren, 224
- HERONS Formel, 80
- HESSE-Form, 71
- HESSE-Form, 70
- Hilbertraum, 95, 129
- homogene Lösung, 106
- HORNER-Schema, 348
- DE L'HOSPITAL
 Regel von, 300
- Hyperbel, 187
 Tangente, 187
- Hyperboloid, 194
- Hypotenuse, 272
- Häufungspunkt, 228
 einer Menge, 264
- Höhe
 im Dreieck, 150
- HÖLDER-Ungleichung, 97
- identische Funktion, 275
- identische Relation, 24
- Identitätssatz
 für Polynome, 345
- Implikation, 2
- Indexfolge, 223
- Induktion
 vollständige, 13
- Induktionsprinzip, 13
- inhomogene Lösung, 106
- inhomogenes LGS
 Lösungsstruktur, 106
- injektiv, 27, 275
- Inkreis, 80
- Inkreismittelpunkt, 119
- Inkreisradius, 80
- Integral
 bestimmtes, 379
 unbestimmtes, 383
- Integral-Restglied, 320, 392
- Integration
 partielle, 385
- Interpolationspolynom, 347
- Intervall, 8
 abgeschlossenes, offenes, 8
 halboffenes, 8
- Intervallschachtelung, 229
- Inverse, 139
 Matrix, 140, 141, 168, 170
 modulo p, 42
 inverse Relation, 24
- Inversion, 30
- JENSENSche-Ungleichung, 374
- kanonische Basis, 123
- Kegelschnitt, 185
 Polardarstellung, 191
- Kegelvolumen, 397
- Kern
 einer linearen Abbildung, 138
- Kettenlinie, 398
- Kettenregel, 287
- Koeffizient, 99
- Koeffizientenvergleich, 340
- Kombinatorik, 51
- kommutierendes Diagramm, 154
- konjugiert komplex, 78
- Konjunktion, 2
- konkav, 371
- konvergent, 209
- Konvergenz
 gleichmäßige, 237, 238, 280
 punktweise, 238, 279
- konvex, 371
- Konvexitätskombination, 374
- Koordinatenform
 der Kugelgleichung, 74
 einer Ebene, 71
- Koordinatentransformation, 154
- Koordinatenvektor
 bzgl. einer Basis, 153
- Kosinus, 271, 401
 Additionssatz, 272
 Stetigkeit des, 273
- Kosinushyperbolicus, 310, 398, 403
- Kosinusreihe, 322

- Kosinussatz, 66
- Kotangens, 402
 - Additionssatz, 272
- Kreisfläche, 377, 390
- Kreuzprodukt, 68, 172, 183
- KRONECKER-Symbol, 129, 135
- kubische Gleichung, 76, 84
- Kugel
 - offene, abgeschlossene, 96
- Kugelgleichung, 74
- Kugelvolumen, 397
- Körper, 43
- l. a., 114
- l. u., 114
- LAGRANGE-Identität, 183
- leere Menge, 10
- LEIBNIZ-Reihe, 247, 248, 323, 394
- Leitgerade, 189, 191
- Leitkoeffizient, 337
- Lemma
 - von EUKLID, 37, 38, 46
- Leuchtkraft, 307
- lh, 120
- Limes, 212
 - Limes Inferior, 248
 - Limes Superior, 248
- linear, 106
 - linear abhängig, 114
 - linear unabhängig, 74, 114
- lineare Abbildung
 - orthogonale, 146
 - unitäre, 146
- lineare Abbildung, 132
 - unitäre, 181
- lineare Hülle, 120
- lineares Gleichungssystem, 99
- Linearfaktor, 337
- Linearform, 97
- Linearkombination, 63, 91
- Linkskurve, 371
- ln, 306, 401
- Logarithmus, 307
 - natürlicher, 306, 401
- Rechengesetze, 306
- Logarithmus-Regel, 389
- \log_b , 307
- Lotto, 59
- L^p -Norm, 96, 97
- L^p -Raum, 98
- Länge
 - einer Kurve, 397
 - eines Vektors, 65
- magisches Quadrat, 109
- Magnitude, 307
- Majorante, 242
- Matrix, 105
 - adjungierte, 144, 145
 - inverse, 140, 141, 168, 170, 172
 - längentreue, 146
 - orthogonale, 146
 - quadratische, 140
 - unitäre, 146
 - winkeltreue, 146
- Matrixprodukt, 134
- MINKOWSKI-Ungleichung, 98
- Minoren, 167
- Mittel
 - arithmetisches, 375
 - geometrisches, 375
- Mittelpunkt
 - einer Kugel, 74
- Mittelwertsatz, 290
 - verallgemeinerter, 290
- Mitternachtsformel, 76, 207, 338
- modulo, 40
- Monom, 116, 337
- DE MORGANSche Regeln
 - für Mengen, 11
- DE MORGANSche Regeln, 4
- multilinear, 156
- Multilinearform
 - alternierende, 156
- Multinomialkoeffizient, 53
- Multiplikation
 - Matrix-Matrix, 134
 - Matrix-Vektor, 106

- IN, 12
- n -te Wurzel, 221
- \mathbb{N}_0 , 13
- natürliche Zahlen, 12
- Nebenwinkel, 191
- 9-er Probe, 41
- NEWTON-Verfahrens, 331
- Norm, 65, 93
 - auf \mathbb{C}^n , 89
 - auf \mathbb{R}^n , 75
 - euklidische, 65, 75, 89
- Normalenform
 - einer Ebene, 69
- Normalenvektor, 69
- normiert, 67
- normierter Raum, 95
- Nullabbildung, 135
- Nullfolge, 214
- Nullkombination, 114
 - triviale, 114
- Nullmatrix, 135
- Nullpolynom, 117, 337
- Nullstelle
 - einfache, 337
 - mehrfache, 337
- nullteilerfrei, 78
- Nullvektor, 61
- Näherungskurve, 347
- o. B. d. A., 8
- ONB, 129
- Ordnungsrelation, 21
- Orientierung
 - von Vektoren, 73
- orthogonal, 67
 - Matrix, 146
- Orthonormalbasis, 129
- Parabel, 189
- Parameterdarstellung
 - einer Ebene, 65
- Parsek, 308
- Partialsumme, 240
- PASCALSches Dreieck, 16
- PAULI-Matrizen, 176
- Permutation, 30, 52, 158
 - (un)gerade, 32
- Pfadbild
 - der Determinante, 159
- Pfadregel, 160, 161
- Polardarstellung, 79
- Polarisationsgleichung, 146, 180
- Polygonzug, 397
- Polynom, 337
 - normiertes, 337
 - rationale Nullstelle, 342
 - un/gerades, 337
- Polynomdivision, 29, 344
- positiv orientiert, 73
- Potenzfunktion, 28
- Potenzreihe, 252
 - Ableitung der, 299
- pq-Formel, 207
- Primzahl, 9
- Produktintegration, 385
- Produktmenge, 9
- Produktregel, 286
- Produktzeichen, 15
- Projektion, 176
 - eindimensionale, 178
 - nicht triviale, 177
 - Spektrum der, 177
- Projektionsraum, 176
- Projektor, 176
- PYTHAGORAS
 - Satz des, 65, 271
- Quader, 10
- quadratische Form, 192
- Quadratwurzel, 88
- Quadrik, 192, 195
- Quantor, 5
- Quersumme, 41
 - alternierende, 42
- Quotientenkriterium, 249
- Quotientenregel, 287
- Radius, 74

- Rang
 - einer Matrix, 172
 - einer Matrix, 170
 - einer Menge, 170
- rationale Zahlen, 9
- Raum
 - mit Skalarprodukt, 95
 - normierter, 95
- Rechte-Hand-Regel, 73
- Rechteck, 377
- Regel
 - von CAUCHY, 350
 - von DESCARTES, 359
 - von DE L'HOSPITAL, 300
- Reihe, 209, 240
 - absolut konvergente, 241
 - alternierende, 247
 - binomische, 327
 - geometrische, 202, 240
 - harmonische, 210
- Rekurrenzgleichung, 201, 204
 - 1. Ordnung, 202
 - 2. Ordnung, 204, 205
- Rekursionsgleichung, 201, 204
- Relation, 19
 - bijektive, 21
 - funktionale, 21
 - homogene, 21
 - identische, 24
 - injektive, 21
 - inverse, 24
 - linkstotale, 21
 - rechtstotale, 21
 - reflexive, 21
 - surjektive, 21
 - symmetrische, 21
 - transitive, 21
- Repräsentant
 - einer Äquivalenzrelation, 22
- Restglied, 392
 - nach CAUCHY, 320
 - nach LAGRANGE, 320
 - nach SCHLÖMILCH, 319
- Richtungsvektor, 64
- RIEMANN-Summe, 378, 379
- Rotationskörper, 396
- RSA-Verschlüsselung, 45
- Rückwärtsentwicklung, 201
- Sandwich-Prinzip, 214
- Sattelpunkt, 364
- Satz
 - vom Faßkreisbogen, 149
 - vom Maximum, 269
 - von BOLZANO-WEIERSTRASS, 230
 - von RIESZ, 97
 - von TAYLOR, 320
 - von PYTHAGORAS, 65, 271
- Schlüssel
 - öffentlicher/privater, 45
- Schnitt, 23
- Sehne, 371
- selbstadjungiert, 175
- sesquilinear, 93
- Sesquilinearform, 93
- Signum, 88
- sin, sin⁻¹, 271, 309, 401
- sinh, sinh⁻¹, 310, 403
- Sinus, 271, 395, 401
 - Additionssatz, 272
 - Stetigkeit des, 273
- Sinushyperbolicus, 310, 398, 403
- Sinusreihe, 321
- Skalarmultiplikation, 61
- Skalarprodukt, 65, 93
 - auf \mathbb{C}^n , 89
 - auf \mathbb{R}^n , 75
 - euklidisches, 75, 89
- Skalarprodukt-Raum, 95
- Spaltenvektor, 105
- Spatprodukt, 71
- Spektrum, 173
- Spiegelung, 181
- Stammfunktion, 313, 380
- Stellenverdopplung, 333
- stetige Funktion, 263
- Stetigkeit
 - links/rechtsseitige, 264

- Strahlungsfluß, 307
 STURMSche Kette, 354
 Stützvektor, 64
 Substitutionsmethode, 387
 Summation
 ABELSche, 314
 Summationsindex, 15
 Summenzeichen, 15
 surjektiv, 27
 symmetrische Gruppe, 30
- TAKAGI
 Funktion von, 294
 tan, \tan^{-1} , 271, 309, 402
 Tangens, 271, 402
 Additionssatz, 272
 Tangenshyperbolicus, 311, 404
 Tangentengleichung, 285
 Tangentialebene
 an die Kugel, 75
 tanh, \tanh^{-1} , 311, 404
 Tautologie, 3
 TAYLOR-Entwicklung, 391, 392
 TAYLOR-Formel, 320, 392
 TAYLOR-Polynom, 318, 392
 TAYLOR-Reihe, 318
 Teilen mit Rest, 35, 246, 343
 Teiler, 35, 345
 größter gemeinsamer, 36, 345
 teilerfremd, 36
 Teilfolge, 223
 Teilmenge, 7
 einer n -elementigen Menge, 53
 Teilraum, 91
 orthogonaler, 136
 Teleskop-Summe, 242, 261
 Torus, 11
 transponieren, 105
 Transposition, 30
 transzendent, 399
 Trapez, 377
 Tupel, 17
 Turm von Hanoi, 201
- Umkehrfunktion, 28, 275
 Ableitung der, 305
 Stetigkeit der, 276
 Umkreis, 149
 Radius, 149
 Umordnung
 einer Doppelreihe, 256
 einer Reihe, 243
 unbestimmte Ausdrücke, 299
 unendlichdimensional, 122
 Ungleichung
 BERNOULLI-, 216, 373
 CAUCHY-SCHWARZsche, 93, 94, 97
 JENSENSche, 374
 MINKOWSKI-, 98
 HÖLDERSche, 97
 YOUNGsche, 98, 376
 umgekehrte BERNOULLI-, 373
- unitär
 Matrix, 146
- Unterdeterminante, 167
 Urbild, 25, 26
- VANDERMONDSche Identität, 58, 59
 Vektor, viii, 61, 90
 Vektorraum, viii, 90
 Verkettung
 von Funktionen, 25
 von Relationen, 24
 Verschlüsselung
 RSA-, 45
- Vielfaches
 kleinstes gemeinsames, 48
- Vielfachheit, 337
 VIETAScher Wurzelsatz, 360
 vollständige Induktion, 13
 Vorzeichenflip, 228
 Vorzeichenfunktion, 88
 Vorzeichenmatrix, 168
 Vorzeichenmethode, 364
- Wahrheitswert, 1
 Wendepunkt, 364, 369
 Wendestelle, 364

- Widerspruchsbeweis, 4
- Winkelhalbierende, 191
 - äußere, 191
- wohldefiniert, 23, 282
- Wurzel
 - n-te, 29, 221
 - einer linearen Abbildung, 180
- Wurzelkriterium, 249
- YOUNG-Ungleichung, 98, 376
- $\mathbb{Z}/p\mathbb{Z}$, 24
- Zahl
 - komplexe, 77
 - konjugiert komplexe, 78
- Zahlen
 - algebraische, 399
- natürliche, 12
- rationale, 9
- transzendenten, 399
- Zeilenumformung, 107
 - äquivalente, 109
- Zerlegung, 378
 - einer Zahl, 54, 55
- \mathbb{Z}_p , 41
- Zwei-Quadrat-Satz, 78
- Zwischenwertsatz, 269
- Zylinder, 10
- Äquivalenzklasse, 40
- Äquivalenzrelation, 21, 22, 40
- Überdeckung
 - disjunkte, 22