

# 1 Grundstrukturen der Mathematik

**1.1 Mathematische Logik** Eine Formel oder ein Satz der Alltagssprache heißt *Aussage*, wenn sie wahr oder falsch sein kann. Einige Beispiele:

$$2 < 5, \quad 3 = 5, \quad \text{Sokrates hatte eine Glatze.}$$

Einer Aussage kann man daher prinzipiell einen *Wahrheitswert* „wahr“ oder „falsch“ zuordnen, wobei es im Beispiel rechts vermutlich nie mehr möglich sein wird, über den Wahrheitswert dieser Aussage zu entscheiden. Denn Platon hat viele Gespräche und Reden des Sokrates aufgezeichnet, aber niemals kam es zu einem Satz wie „Sokrates fasste sich an die Glatze“.

*Logische Konjunktionen* verbinden Aussagen, die einfachsten sind

$$\begin{aligned} (A \text{ oder } B) \text{ wahr} &\Leftrightarrow A \text{ wahr oder } B \text{ wahr oder beide wahr}, \\ (A \text{ und } B) \text{ wahr} &\Leftrightarrow A \text{ wahr und } B \text{ wahr}. \end{aligned}$$

Für *oder* und *und* sowie für die Verneinung schreiben wir

$$\begin{aligned} \vee &\quad \text{oder (nicht ausschließend) (von lat. vel)} \\ \wedge &\quad \text{und} \\ \neg &\quad \text{Verneinungszeichen} \end{aligned}$$

In Tafelform können wir die obigen Regeln so schreiben:

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

Bei der Verneinung kehrt sich der Wahrheitswert einfach um. Ferner gelten die Verneinungsregeln für Aussagen  $A, B$

$$\begin{aligned} \neg(A \vee B) &\Leftrightarrow \neg A \wedge \neg B, \\ \neg(A \wedge B) &\Leftrightarrow \neg A \vee \neg B, \end{aligned}$$

Schwieriger ist die Implikation, weil es hier Unterschiede zwischen den natürlichen Sprachen (auch untereinander) und der Sprache der Mathematik gibt. Klar ist noch das Beispiel

„Wenn es regnet, dann ist die Straße nass.“

Ob es nun regnet oder nicht, der Satz ist immer wahr und er behauptet nichts, wenn es nicht regnet. Genau so wird die Implikation in der Mathematik verwendet: Ist die Voraussetzung (hier: Es regnet) nicht erfüllt, so ist es gleichgültig, was in der Behauptung (hier: Die Straße ist nass) steht, die Implikation selber ist in diesem Fall wahr.

Zur Kollision zwischen Mathematik und natürlichen Sprachen kommt es in Sätzen wie

„Wenn Albert Einstein den Nobelpreis nicht bekommen hätte,  
dann wäre er an Hänschen Klein verliehen worden.“

Was würden Sie sagen, ist der Satz wahr oder falsch? Oder wir betrachten den Fall, dass es für einen Preis zwei heiße Kandidaten E. und K. gibt. Nach der Preisverleihung lesen wir in der Zeitung

„Wenn E den Preis nicht bekommen hätte,  
dann wäre er an K verliehen worden.“

Ist der Satz wahr oder falsch? Offenbar sind die beiden letzten Sätze logisch vollkommen identisch,

sie besitzen aber einen unterschiedlichen Kontext. In der Mathematik muss der Wahrheitswert einer Konjunktion aus den Wahrheitswerten der beteiligten Aussagen bestimmbar sein, der Kontext darf keine Rolle spielen. Daher orientiert sich die Wahrheitstafel für die Implikation an der Aussage „Wenn es regnet, dann ist die Straße nass“:

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

Eine Implikation ist daher immer wahr, wenn die Voraussetzung falsch ist. Damit sind alle Aussagen über mögliche Preisträger wahr.

Aufgrund der Wahrheitstafeln ist

$$(A \Rightarrow B) \Leftrightarrow (\neg A \vee B)$$

Daraus folgt für die Verneinung der Implikation

$$\neg(A \Rightarrow B) \Leftrightarrow (A \wedge \neg B)$$

Zwei Aussagen  $A$  und  $B$  heißen *äquivalent*, Schreibweise  $A \Leftrightarrow B$ , wenn  $A \Rightarrow B$  und  $A \Leftarrow B$ .

Ein mathematischer Beweis besteht aus einer Folge von Aussagen, die entweder von vorneherein als richtig angesehen werden oder aus der folgenden Schlussregel, dem *modus ponens*, abgeleitet werden können:

$$\begin{array}{ll} \text{„Wenn es regnet, dann ist die Straße nass“} & A \Rightarrow B \\ \text{„Es regnet“} & A \\ \hline \text{„Die Straße ist nass“} & B \end{array}$$

Wir nehmen an, es gibt nur Lügner, die immer lügen, und Wahrheitssprecher, die immer die Wahrheit sagen.

Ein Kreter sagt: „Alle Kreter lügen“.

Dies ist die *Antinomie des Epimenides*. Allgemeiner wird eine Aussage (?) Antinomie genannt, wenn die Zuweisung eines der Wahrheitswerte wahr oder falsch in jedem Fall zu einem Widerspruch führt. Die Antinomie des Epimenides stammt aus der Bibel, der Apostel Paulus behauptet in einem Brief: Die Kreter sind alle Lügner und faule Bäuche, das sagt sogar ihr eigener Prophet. Erst später erkannten die Kirchenlehrer, dass es sich hier um eine problematische logische Konstruktion handelt und sie versuchten, diesen kretischen Propheten zu ermitteln. Anscheinend existieren nur wenige bekannte Kreter. Sie mussten etwa 500 Jahre v.Ch. zurückgehen, um auf Epimenides zu treffen, über den fast nichts bekannt ist.

Warum handelt es sich hier um eine Antinomie? Ist die Aussage „Alle Kreter lügen“ wahr, so behauptet der Kreter, dass er ein Lügner ist. Ist die Aussage falsch, so ist er demnach gar kein Lügner. Was ist an dieser Argumentation falsch?

Die Antinomie beruht darauf, dass in der Alltagslogik der Satz

„Alle Kreter lügen“.

verneint wir durch

„Alle Kreter lügen nicht“.

Die Alltagslogik ist daher *Aussagenlogik*. Wir stellen uns zu einer Aussage den Kosmos der Möglichkeiten vor, in diesem Fall 3000 Kreter, von denen ein jeder ein Wahrheitssprecher oder ein Lügner sein kann. Jede Aussage über kretische Wahrheitssprecher und Lügner greift eine Teilmenge dieses

Kosmos heraus. Die korrekte Verneinung einer solchen Aussage muss das Komplement der durch die Aussage festgelegten Teilmenge beschreiben. Die Aussage „Alle Kreter lügen“ greift eine Teilmenge heraus, die nur aus einem Element besteht, dass nämlich die Kreter 1, 2, ..., 3000 alle Lügner sind. Das Komplement dieser Teilmenge wird dadurch charakterisiert, dass es dort mindestens einen Wahrheitsprecher gibt. Damit liegt gar keine Antinomie vor: Es gibt sowohl Lügner als auch Wahrheitssprecher. Der Satz „Alle Kreter lügen“ ist daher falsch und der Kreter, der ihn ausspricht, ist ein Lügner.

Interessant ist natürlich, dass die Antinomie des Epimenides, die ja vom Lügen handelt, weder eine Antinomie ist, noch von Epimenides stammt.

Die Grundidee der Antinomie kann aber gerettet werden. Wenn jemand sagt „Ich lüge“ oder präziser „Der Satz, den ich gerade ausspreche, ist falsch“, so lässt sich diesen Sätzen in der Tat kein Wahrheitswert zuordnen.

Die Logik der Mathematik und Informatik ist eine *Prädikatenlogik*. Ein (einstelliges) Prädikat ist von der Form  $A(x)$  mit einer Variablen  $x$  aus einem Definitionsbereich  $D$ . Wenn wir in  $A(x)$  ein konkretes  $x \in D$  einsetzen, so muss eine Aussage entstehen. Ein Beispiel ist

$$x \text{ ist ein Mensch}, \quad D = \text{Lebewesen}.$$

Für  $x$  können wir hier Sokrates oder den Hund Lupo einsetzen, in jedem Fall entsteht eine Aussage.

Wir verwenden

$$\begin{aligned} \forall & \text{ „für alle“} \\ \exists & \text{ „es existiert“}. \end{aligned}$$

Ist  $A(x)$  ein einstelliges Prädikat, so gelten die Verneinungsregeln

$$\begin{aligned} \neg \forall x A(x) & \Leftrightarrow \exists x \neg A(x), \\ \neg \exists x A(x) & \Leftrightarrow \forall x \neg A(x). \end{aligned}$$

Man mache sich diese Regeln an Hand von lügenden oder wahrheitssprechenden Kretern klar.

**1.2 Satz und Beweis** Ein mathematischer Satz ist meist von der Form  $A \Rightarrow B$  und besteht aus einer Voraussetzung  $A$  und einer Behauptung  $B$ . Der *Beweis* des Satzes besteht aus einer Folge von wahren Aussagen, deren letzte die Behauptung ist.

**Beispiel 1.1** Man zeige für  $a, b \geq 0$

$$\frac{a+b}{2} \geq \sqrt{ab}.$$

Das geschieht häufig so:

$$\begin{aligned} \Rightarrow a + b & \geq 2\sqrt{ab} \Rightarrow (a+b)^2 \geq 4ab \\ \Rightarrow a^2 - 2ab + b^2 & \geq 0 \Rightarrow (a-b)^2 \geq 0. \end{aligned}$$

Hier hat man den Fehler gemacht, dass man von der zu beweisenden Aussage ausgeht und so lange folgert, bis eine wahre Aussage entsteht. Korrekt ist natürlich die umgekehrte Reihenfolge.  $\square$

Aus den Wahrheitstafeln folgt

$$A \Rightarrow B \Leftrightarrow \neg A \vee B$$

und entsprechend

$$\neg(A \Rightarrow B) \Leftrightarrow A \wedge \neg B$$

Im *indirekten Beweis* zeigen wir, dass  $A \wedge \neg B$  falsch ist. Dazu nehmen wir die Verneinung der Behauptung als wahr an und zeigen, dass nicht gleichzeitig die Voraussetzung wahr ist.

**Beispiel 1.2** Man zeige für  $a, b \geq 0$

$$\frac{a+b}{2} \geq \sqrt{ab}.$$

Als Voraussetzung können wir hier  $(a-b)^2 \geq 0$  nehmen. Um mit dem indirekten Beweis zu beginnen, nehmen wir die Verneinung der Behauptung an,

$$\frac{a+b}{2} < \sqrt{ab}$$

und erhalten

$$\begin{aligned} \Rightarrow a+b &< 2\sqrt{ab} \Rightarrow (a+b)^2 < 4ab \\ \Rightarrow a^2 - 2ab + b^2 &< 0 \Rightarrow (a-b)^2 < 0. \end{aligned}$$

mit Widerspruch zur Voraussetzung. Wie so oft ist hier der indirekte Beweis komplizierter als der direkte. Daher: Immer zuerst direkt versuchen!  $\square$

Aus der Wahrheitstafel für die Implikation folgt

$$(A \Rightarrow B) \Leftrightarrow (\neg B \Rightarrow \neg A).$$

Man nennt  $\neg B \Rightarrow \neg A$  die *Kontraposition* zu  $A \Rightarrow B$ .

**Beispiel 1.3** Man zeige:

Wenn  $n^2$  gerade ist, so ist auch  $n$  gerade.

Die Kontraposition ist:

Wenn  $n$  ungerade ist, so ist auch  $n^2$  ungerade.

Beweis der Kontraposition:

$$n = 2k + 1 \Rightarrow n^2 = 4k^2 + 4k + 1.$$

$\square$

**1.3 Mengen, Relationen, Abbildungen** Unter einer Menge verstehen wir eine Zusammenfassung von Gegenständen. Wir schreiben  $a \in A$ , wenn  $a$  in der Menge  $A$  liegt, ansonsten schreiben wir  $a \notin A$ . Für die Menge  $A_2 = \{1, 2\}$  gilt beispielsweise  $1 \in A_2$  und  $3 \notin A_2$ .

Die *natürlichen Zahlen*

$$\mathbb{N} = \{1, 2, 3, \dots\}$$

werden später noch genauer untersucht. Für kompliziertere Mengen gibt es eine Vielzahl von Beschreibungen. So lässt sich die Menge  $G$  der geraden natürlichen Zahlen schreiben als

$$2\mathbb{N}, \quad G = \{2n : n \in \mathbb{N}\}, \quad G = \{n \in \mathbb{N} : n \text{ ist gerade}\}.$$

In der *leeren Menge*  $\emptyset = \{\}$  begegnet uns der horror vacui: Die Aussage  $a \in \emptyset$  ist für jedes  $a$  falsch.

Wir setzen

$$\begin{aligned} A = B &\Leftrightarrow (x \in A \Leftrightarrow x \in B) \\ A \subset B &\Leftrightarrow (x \in A \Rightarrow x \in B) \quad (\text{Teilmenge}), \\ A \cap B &\Leftrightarrow \{x : x \in A \wedge x \in B\} \quad (\text{Schnittmenge}), \\ A \cup B &\Leftrightarrow \{x : x \in A \vee x \in B\} \quad (\text{Vereinigungsmenge}), \\ A \setminus B &\Leftrightarrow \{x : x \in A, x \notin B\} \quad (\text{Komplement}). \end{aligned}$$

Die Definition der Teilmenge ist wörtlich zu nehmen. Es gilt  $B \subset B$  und  $\emptyset \subset B$  für alle Mengen  $B$ . Denn die Voraussetzung  $x \in \emptyset$  ist falsch und daher ist  $x \in \emptyset \Rightarrow x \in B$  wahr. Sind die zu untersuchenden Mengen alle Teilmengen einer Menge  $M$ , so schreibt man auch  $A^c$  an Stelle von  $M \setminus A$ .

Man beachte den Unterschied zwischen  $A = \{a\}$  und  $A' = \{\{a\}\}$ . Es gilt  $\{a\} \in A'$ , aber  $a \notin A'$ . Die *Potenzmenge* einer Menge  $A$  ist

$$\mathcal{P}(A) = 2^A = \{B : B \subset A\}.$$

Für die Menge  $A_2 = \{1, 2\}$  gilt beispielsweise

$$\mathcal{P}(A_2) = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}.$$

$\mathcal{P}(A_2)$  besteht daher aus 4 Elementen, die alle selber Mengen sind.

Für eine nichtleere Menge  $I$ , die hier *Indexmenge* genannt wird, gebe es für jedes  $i \in I$  eine Menge  $A_i$ . Dann kann man Durchschnitt und Vereinigung eines solchen Mengensystems ähnlich wie zuvor definieren:

$$\cap_{i \in I} A_i = \{x : \forall i \in I : x \in A_i\}, \quad \cup_{i \in I} A_i = \{x : \exists i \in I : x \in A_i\}.$$

Für zwei Elemente  $a, b$  heißt  $(a, b)$  *geordnetes Paar*. Im Unterschied zur Menge  $A = \{a, b\} = \{b, a\}$  kommt es hier auf die Reihenfolge an. Es gilt  $(a, b) = (a', b')$  genau dann, wenn  $a = a'$  und  $b = b'$ . Analog ist das (geordnete)  $n$ -tupel  $(a_1, a_2, \dots, a_n)$  definiert. Für Mengen  $A_1, \dots, A_n$  ist das *kartesische Produkt*

$$A_1 \times A_2 \times \dots \times A_n = \{(a_1, a_2, \dots, a_n) : a_i \in A_i \text{ für } i = 1, 2, \dots, n\}$$

definiert.

Für eine Menge  $A$  heißt  $R \subset A \times A$  *Relation*. Ist  $(a, b) \in R$ , so schreibt man meist  $aRb$  und sagt, dass  $a$  und  $b$  in der Relation  $R$  stehen. Als einfaches Beispiel können wir die natürlichen Zahlen nehmen, die nach Größe geordnet werden. Die zugehörige Relation ist dann

$$R = \{(a, b) \in \mathbb{N} \times \mathbb{N} : a \leq b\}.$$

Statt  $aRb$  schreibt man dann gleich  $a \leq b$ .

Eine Relation  $R$  heißt

- (a) *reflexiv*, wenn  $aRa$  für alle  $a \in A$ ,
- (b) *symmetrisch*, wenn mit  $aRb$  auch  $bRa$  gilt,
- (c) *antisymmetrisch*, wenn aus  $aRb$  und  $bRa$  folgt, dass  $a = b$ ,
- (d) *transitiv*, wenn aus  $aRb$  und  $bRc$  folgt, dass  $aRc$ .

Eine Relation  $R$  heißt *Ordnungsrelation* oder *Halbordnung*, wenn sie reflexiv, antisymmetrisch und transitiv ist. Eine Ordnungsrelation heißt *total*, wenn zusätzlich gilt: Für alle  $a, b \in A$  gilt  $aRb$  oder  $bRa$ .

Aus einer Ordnungsrelation  $\leq$  erhält man eine *strenge Ordnungsrelation*, indem man setzt

$$a < b \Leftrightarrow a \leq b \text{ und } a \neq b.$$

Ordnungsrelationen auf Mengen von Zahlen wie die oben beschriebene sind meist totale Ordnungsrelationen.

**Beispiele 1.4** (i) Sei  $B$  eine nichtleere Menge. Auf der Potenzmenge  $\mathcal{P}(B)$  ist die Teilmengenbeziehung  $\subset$  eine Ordnungsrelation, die aber, sofern  $B$  mehr als ein Element enthält, keine totale Ordnung ist.

(ii) Auf einer Menge  $A$  existiere eine totale Ordnung  $\leq$ . Wie kann man die  $n$ -tupel  $(a_1, a_2, \dots, a_n) \in A^n = A \times \dots \times A$  sinnvoll ordnen? Eine Möglichkeit ist die *komponentenweise Ordnung*  $\leq_k$

$$(a_1, \dots, a_n) \leq_k (b_1, \dots, b_n) \Leftrightarrow a_i \leq b_i \text{ für } i = 1, \dots, n.$$

Diese ist zwar eine Ordnungsrelation, aber bereits für die Menge  $A_2 = \{1, 2\}$  und  $n = 2$  nicht total, wie die Unvergleichbarkeit von  $(1, 2)$  und  $(2, 1)$  zeigt.

Eine totale Ordnung ist die *lexikographische Ordnung*

$$(a_1, \dots, a_n) <_l (b_1, \dots, b_n) \Leftrightarrow \exists i_0 \ a_i = b_i \text{ für } i = 1, \dots, i_0 - 1 \text{ und } a_{i_0} < b_{i_0},$$

Wörterbücher verwenden eine leicht modifizierte lexikographische Ordnung, leicht modifiziert deshalb, weil die Wörter unterschiedlich lang sind.  $\square$

Sei  $A$  mit einer Halbordnung  $\leq$  versehen. Wir nennen  $m \in A$  *minimal*, wenn es kein  $a \in A \setminus \{m\}$  gibt mit  $a \leq m$ .  $m \in A$  heißt *Minimum* von  $A$ , wenn  $m \leq a$  für alle  $a \in A$  gilt. In diesem Fall schreiben wir  $m = \min A$ . Das Minimum ist, falls es existiert, eindeutig bestimmt und minimal. Jede endliche, total geordnete Menge  $A$  besitzt ein Minimum.

$T \subset A$  heißt *nach unten beschränkt in  $A$* , wenn es ein  $u \in A$  gibt mit  $u \leq t$  für alle  $t \in T$ . In diesem Fall heißt  $u$  *untere Schranke* von  $T$ . Die ganzen Zahlen  $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$  mit der natürlichen Ordnung sind nach unten unbeschränkt.

Die Begriffe maximal, Maximum, obere Schranke sind analog definiert. Wenn  $T$  nach unten und nach oben beschränkt in  $A$  ist, so heißt  $T$  *beschränkt in  $A$* .

Eine reflexive, symmetrische und transitive Relation heißt *Äquivalenzrelation*. In diesem Fall schreiben wir meist  $\sim$  statt  $R$ . Jedem  $a \in A$  ordnen wir die *Äquivalenzklasse*

$$\bar{a} = \{x \in A : a \sim x\}$$

zu. Wegen der Reflexivität ist  $a \in \bar{a}$ . Weiter folgt aus der Transitivität, dass  $\bar{a} = \bar{b}$  genau dann, wenn  $a \sim b$ . Hier deutet sich schon an, dass  $A$  in Äquivalenzklassen zerfällt. Das wollen wir im Folgenden präzisieren.

Sei  $A$  nicht leer. Eine Menge von Teilmengen  $A_i \subset A$ ,  $A_i \neq \emptyset$ ,  $i \in I$ , heißt *Partition* oder *disjunkte Zerlegung* von  $A$ , wenn  $A_i \cap A_j = \emptyset$  für  $i \neq j$  (disjunkt!) und  $A = \bigcup_{i \in I} A_i$  (Zerlegung!). Es gilt

**Satz 1.5** Die Begriffe „Äquivalenzrelation“ und „disjunkte Zerlegung“ sind im folgenden Sinne äquivalent:

(a) Zu einer disjunkten Zerlegung  $\{A_i : i \in I\}$  von  $A$  definiere

$$a \sim b \Leftrightarrow \exists i \in I \text{ mit } a, b \in A_i.$$

Dann ist  $\sim$  Äquivalenzrelation mit Äquivalenzklassen  $A_i$ .

(b) Die Äquivalenzklassen bilden eine disjunkte Zerlegung von  $A$ .

*Beweis:* (a)  $a \sim a$  gilt wegen  $A = \bigcup_i A_i$ ,  $a$  muss daher zu einem  $A_i$  gehören.  $a \sim b \Leftrightarrow b \sim a$  ist offensichtlich.  $a \sim b$  bedeutet  $a, b \in A_i$  und  $b \sim c$  bedeutet  $b, c \in A_j$ . Da die Mengen  $A_i$  und  $A_j$  disjunkt sind und  $b$  zu beiden Mengen gehört, muss  $i = j$  gelten.

(b) Wegen  $a \in \bar{a}$  sind die Äquivalenzklassen nicht leer. Mit  $\bigcup_{a \in A} \{a\} = A$  ist auch  $\bigcup_{a \in A} \bar{a} = A$ . Ist  $\bar{a} \cap \bar{b} \neq \emptyset$ , so gibt es ein  $c$  mit  $c \in \bar{a}$ ,  $c \in \bar{b}$ . Damit folgt  $a \sim c$  und  $b \sim c$ , mit der Transitivität auch  $a \sim b$ .  $\square$

Das Konzept der Relation lässt sich auf vielfältige Weise erweitern. Zunächst kann man auch Relationen zwischen verschiedenen Mengen definieren, also  $R \subset A \times B$ .  $A$  könnte hier eine Menge von Personen sein und  $B$  eine Menge von Firmen.  $aRb$  könnte dann bedeuten, dass  $a$  ein Kunde der Firma  $b$  ist. Auch der Übergang zu mehr als zweistelligen Relationen ist manchmal sinnvoll.

Seien  $A$  und  $B$  nichtleere Mengen. Eine *Abbildung*  $f$  zwischen diesen Mengen ordnet jedem Element  $a \in A$  genau ein Element  $b \in B$  zu. Wir schreiben für diese Zuordnung  $f(a) = b$  und  $a \mapsto b$  sowie  $f : A \rightarrow B$ .  $A$  heißt *Definitions-* und  $B$  *Werte- oder Zielbereich* der Abbildung  $f$ . Man kann eine Abbildung auch als Relation auffassen vermöge der Beziehung

$$f(a) = b \Leftrightarrow (a, b) \in G_f \subset A \times B.$$

$G_f$  heißt *Graph* von  $f$ .

Abbildungen zwischen Zahlbereichen werden oft als *Funktionen* bezeichnet, was hauptsächlich historische Gründe hat.

Für  $A' \subset A$  setzen wir

$$f(A') = \{f(a) : a \in A'\}.$$

$f(A')$  heißt *Bild von  $A'$*  und ist auch im Fall  $A' = A$  eine sinnvolle Bezeichnung. Es gilt für  $A_1, A_2 \subset A$

$$f(A_1 \cup A_2) = f(A_1) \cup f(A_2), \quad f(A_1 \cap A_2) \subset f(A_1) \cap f(A_2).$$

Solche Beziehungen beweist man durch Rückgriff auf die einzelnen Elemente. Als Beispiel beweisen wir die Aussage rechts. Ist  $b \in f(A_1 \cap A_2)$ , so gibt es ein  $a \in A_1 \cap A_2$  mit  $f(a) = b$ . Dann gilt aber auch  $b \in f(A_1)$  und  $b \in f(A_2)$ . Als Gegenbeispiel, dass keine Gleichheit herrschen muss, nehmen wir  $A_1 = \{a_1\}$ ,  $A_2 = \{a_2\}$  und  $f(a_1) = f(a_2) = b$ .

Für  $B' \subset B$  heißt

$$(1.1) \quad f^{-1}(B') = \{a \in A : f(a) \in B'\}$$

das *Urbild von  $B'$* . Für  $B_1, B_2 \subset B$  gilt

$$f^{-1}(B_1 \cup B_2) = f^{-1}(B_1) \cup f^{-1}(B_2), \quad f^{-1}(B_1 \cap B_2) = f^{-1}(B_1) \cap f^{-1}(B_2).$$

Sei  $f : A \rightarrow B$  eine Abbildung.  $f$  heißt *surjektiv*, wenn  $f(A) = B$ .  $f$  heißt *injektiv*, wenn aus  $f(a_1) = f(a_2)$  folgt, dass  $a_1 = a_2$ . Anders ausgedrückt werden verschiedene Elemente des Urbildbereichs auf verschiedene Elemente des Zielbereichs abgebildet. Eine Abbildung heißt *bijektiv*, wenn sie injektiv und surjektiv ist. Anschaulich hat in diesem Fall jedes Element  $a \in A$  genau einen Partner  $b = f(a) \in B$  und umgekehrt hat jedes  $b \in B$  genau einen Partner  $a \in A$ . Damit existiert die *Umkehrabbildung* oder *Invertierung* von  $f$

$$f^{-1} : B \rightarrow A \text{ mit } f^{-1}(f(a)) = a \text{ und } f(f^{-1}(b)) = b.$$

Diese Umkehrabbildung unterscheidet sich in der Notation nicht von der Definition des Urbilds in (1.1), im Gegensatz zu letzterer ist sie aber eine echte Abbildung.

Bei endlichen Mengen  $A$  haben die Selbstabbildungen  $f : A \rightarrow A$  eine Besonderheit. In diesem Fall gilt nämlich

$$(1.2) \quad f \text{ ist surjektiv} \Leftrightarrow f \text{ ist injektiv} \Leftrightarrow f \text{ ist bijektiv.}$$

Ist  $g : A \rightarrow B$  und  $f : B \rightarrow C$ , so ist die *Hintereinanderausführung* oder *Verkettung* definiert durch

$$f \circ g : A \rightarrow C, \quad a \mapsto f(g(a)).$$

## 1.4 Mathematische Strukturen

lassen sich in der Form

$$\mathbb{S} = \{S, e_1, \dots, e_l, f_1, \dots, f_m, R_1, \dots, R_n\}$$

schreiben mit

- $S$  Grundmenge
- $e_i$  ausgezeichnete (meist neutrale) Elemente ,
- $f_j$  Abbildungen (meist zweistellige Operationen wie +),
- $R_k$  (meist zweistellige) Relationen.

Dies ist redundant, weil man alle Abbildungen auch als Relationen schreiben kann.

**1.5 Gruppen** Eine *Gruppe*  $\mathbb{G} = (G, e, \circ)$  besteht aus einer Menge  $G$ , einer zweistelligen Operation  $\circ$  mit  $z = x \circ y \in G$ , und einem ausgezeichneten Element  $e \in G$ , so dass:

(G1) (Assoziativgesetz) Für alle  $x, y, z \in G$  gilt

$$(x \circ y) \circ z = x \circ (y \circ z).$$

(G2) (Neutrales Element) Für alle  $x \in G$  gilt

$$e \circ x = x \circ e = x.$$

(G3) (Inverses Element) Zu jedem  $x \in G$  gibt es ein  $x^{-1} \in G$  mit

$$x^{-1} \circ x = x \circ x^{-1} = e.$$

Die Axiome sind in dieser Form redundant. Z.B. genügt es, an Stelle von (G2) nur  $x \circ e = x$  zu fordern, was in der Literatur manchmal geschieht. In diesem Fall muss man  $e \circ x = x$  mit Hilfe der anderen Axiome explizit beweisen, was wir uns ersparen möchten.

Endliche Gruppen gibt man mit einer *Gruppentafel* an, in der die Ergebnisse von  $x \circ y$  eingetragen werden. Wir bezeichnen die Gruppenelemente mit  $0, 1, 2, \dots$ , wobei  $0$  das neutrale Element ist. Die Gruppe mit 3 Elementen ist eindeutig bestimmt:

$\circ$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

Vierelementige Gruppen gibt es schon mehrere:

$\circ$	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

$\circ$	0	1	2	3
0	0	1	2	3
1	1	0	3	2
2	2	3	0	1
3	3	2	1	0

Gruppen mit unendlicher Grundmenge sind die ganzen und die rationalen Zahlen

$$\mathbb{G} = (\mathbb{Z}, 0, +), \quad \mathbb{G} = (\mathbb{Q}, 0, +), \quad \mathbb{G} = (\mathbb{Q} \setminus \{0\}, 1, \cdot),$$

die in den nächsten Kapiteln besprochen werden. Dagegen bilden die natürlichen Zahlen mit der Addition keine Gruppe, weil wir die positiven Zahlen nicht invertieren können.

Also: Konkrete Gruppen können alles mögliche sein. Daher ist es hier wie meist in der Algebra wichtig, dass die Beweise streng aus den Axiomen folgen. Als Beispiel zeigen wir den folgenden

**Satz 1.6** *In jeder Gruppe sind die Gleichungen  $x \circ a = b$  und  $a \circ x = b$  eindeutig nach  $x$  auflösbar.*

*Beweis:* Man muss hier vorsichtig sein, weil das Kommutativgesetz  $x \circ y = y \circ x$  nicht unbedingt gelten muss. Als Lösung von  $x \circ a = b$  vermuten wir  $x = b \circ a^{-1}$ ,

$$x \circ a = (b \circ a^{-1}) \circ a \stackrel{(G1)}{=} b \circ (a^{-1} \circ a) \stackrel{(G3)}{=} b \circ e \stackrel{(G2)}{=} b.$$

Für den Beweis der Eindeutigkeit nehmen wir an, dass die Gleichung  $x \circ a = b$  von zwei Gruppen-elementen  $x, x'$  gelöst wird. Aus  $x \circ a = x' \circ a$  folgt durch Multiplikation von rechts mit  $a^{-1}$ , dass  $x = x'$ .

Die eindeutige Lösbarkeit von  $a \circ x = b$  zeigt man ganz analog.  $\square$

Aufgrund dieses Satzes sind das neutrale und das inverse Element eindeutig bestimmt. Ferner sind die Gruppentafeln *Lateinische Quadrate*, bei denen in jeder Zeile und in jeder Spalte jedes Element genau einmal vorkommt.

Eine Gruppe heißt *abelsch* oder *kommutativ*, wenn zusätzlich das Kommutativgesetz gilt:

(G4) Für alle  $x, y \in G$  gilt

$$x \circ y = y \circ x.$$

Bei einer kommutativen Gruppe schreibt man meist  $+$  statt  $\circ$  mit dem neutralen Element 0. Dies erinnert an die ganzen Zahlen  $\mathbb{Z} = (Z, 0, +)$ , die ja eine kommutative Gruppe bilden.

**Satz 1.7** *Sei  $A$  eine nichtleere Menge. Dann bilden die bijektiven Selbstabbildungen  $f : A \rightarrow A$  zusammen mit der Verkettung  $f \circ g$  eine Gruppe mit neutralem Element  $id_A : A \rightarrow A$ ,  $id_A(a) = a$ . Insbesondere ist mit  $f, g$  bijektiv auch  $f \circ g$  bijektiv und*

$$(f \circ g)^{-1} = g^{-1} \circ f^{-1}.$$

*Beweis:* Die angegebene Formel für die Inverse von  $f \circ g$  rechnet man nach.  $f \circ id_A = id_A \circ f = f$  ist klar. Das inverse Element zu  $f$  ist die von uns definierte Umkehrabbildung. Ist  $a \xrightarrow{h} b \xrightarrow{g} c \xrightarrow{f} d$ , so gilt gleichgültig wie man klammert, immer  $f(g(h(a))) = d$ .  $\square$

Hat die Menge  $A$  drei Elemente oder mehr, so ist die Gruppe der bijektiven Selbstabbildungen nicht kommutativ. Als Beispiel nehmen wir für  $A_3 = \{1, 2, 3\}$

$$f(1) = 2, \quad f(2) = 3, \quad g(1) = 2, \quad g(2) = 1 \Rightarrow f \circ g(1) = 3, \quad g \circ f(1) = 1.$$

**1.6 Ringe und Körper** Ein *Ring* besteht aus einer Grundmenge  $R$ , zwei ausgezeichneten Elementen 0, 1 und zwei zweistelligen Operationen „+“ und „·“, kurz  $\mathbb{R} = (R, 0, 1, +, \cdot)$ . Die Ringaxiome sind so gefasst, dass man im Ring ähnlich rechnen kann wie mit Zahlen sonst auch. Genauer muss gelten:

(R1)  $(R, 0, +)$  ist eine kommutative Gruppe.

(R2)  $(R, 1, \cdot)$  ist eine Halbgruppe mit neutralem Element 1, d.h. es gilt das Assoziativgesetz sowie  $x \cdot 1 = 1 \cdot x = x$ .

(R3) Es gelten die Distributivgesetze

$$(x + y) \cdot z = x \cdot z + y \cdot z, \quad x \cdot (y + z) = x \cdot y + x \cdot z.$$

Ist die Operation „·“ zusätzlich kommutativ, so heißt der Ring *kommutativ*.

Der bekannteste kommutative Ring sind die ganzen Zahlen  $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$  mit der üblichen Addition und Multiplikation.

Ein *Körper*  $\mathbb{K} = (K, 0, 1, +, \cdot)$  ist ein kommutativer Ring, in dem zusätzlich  $(K \setminus \{0\}, 1, \cdot)$  eine kommutative Gruppe ist. Zu jedem  $x \neq 0$  gibt es daher ein  $x^{-1}$  mit  $xx^{-1} = 1$ .

Der einfachste Körper ist  $\mathbb{Z}_2$ , der nur aus den beiden neutralen Elementen 0 und 1 besteht. 0 ist ja neutral bezüglich der Addition und 1 ist neutral bezüglich der Multiplikation. Darüberhinaus setzen wir  $1 + 1 = 0$  und  $0 \cdot 0 = 0$ . Damit ist 1 zu sich selbst invers bezüglich Addition und Multiplikation.

## 2 Die natürlichen Zahlen und vollständige Induktion

### 2.1 Einführung

Mit  $\mathbb{N}$  bezeichnen wir die Menge der natürlichen Zahlen

$$\mathbb{N} = \{1, 2, 3, \dots\}.$$

Manche Autoren lassen die natürlichen Zahlen auch mit der Null beginnen, wir schreiben dafür  $\mathbb{N}_0 = \{0, 1, 2, \dots\} = \mathbb{N} \cup \{0\}$ .

Wir wollen die folgende Formel für die Summe der ersten  $n$  ungeraden Zahlen beweisen

$$(A_n) \quad 1 + 3 + 5 + \dots + (2n - 1) = n^2, \quad n = 1, 2, 3, \dots$$

Für  $n = 1$  erhalten wir auf der linken Seite 1 und auf der rechten  $1^2 = 1$ . Überprüfen wir ferner den Fall  $n = 2$ : Links steht  $1 + 3 = 4$  und rechts  $2^2 = 4$ . Da wir auch den Fall  $n = 3$  leicht im Kopf berechnen können, ist die Formel also für  $n = 1, 2, 3$  richtig. Ein Physiker wäre mit diesem Argument vielleicht schon zufrieden und würde hieraus kühn auf die Richtigkeit von  $(A_n)$  für alle  $n$  folgern. Wir nennen dies einen Induktionsschluss, weil eine allgemeine Behauptung durch Nachweis von endlich vielen Fällen aufgestellt wird. Dem Physiker bleibt freilich nichts anderes übrig: Er kann ein von ihm postulierte Gesetz nur in endlich vielen Fällen experimentell nachweisen, obwohl es in unendlich vielen Fällen gültig sein soll. In der Mathematik muss die Behauptung  $(A_n)$  dagegen für jedes  $n$  bewiesen werden.

Bei der Überprüfung von  $(A_n)$  kann man auf bereits Berechnetes zurückgreifen:

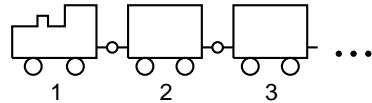
$$\begin{aligned} (2.1) \quad 1 + 3 + 5 &= (1 + 3) + 5 = 4 + 5 = 9 \\ 1 + 3 + 5 + 7 &= (1 + 3 + 5) + 7 = 9 + 7 = 16 \\ 1 + 3 + 5 + 7 + 9 &= (1 + 3 + 5 + 7) + 9 = 16 + 9 = 25 \end{aligned}$$

Wie wir gleich sehen werden, kann man hieraus einen vollständigen Beweis machen, es fehlt nur noch ein Schema, das diese Rechnung allgemeingültig macht.

Wir können  $(A_1), (A_2), \dots$  mit Hilfe des *Prinzips der vollständigen Induktion* beweisen. Dazu beweist man zwei Dinge:

- (i)  $(A_1)$  (=Induktionsanfang oder Induktionsverankerung),
- (ii)  $(A_n) \Rightarrow (A_{n+1})$  für alle  $n \in \mathbb{N}$  (=Induktionsschritt).

Der „Beweis“ von  $(A_1)$  ist nichts anderes als dass man nachrechnet, dass  $(A_1)$  eine wahre Aussage ist, was wir bereits getan haben. Der zweite Schritt lässt sich so interpretieren: Unter der Voraussetzung, dass wir schon wissen, dass die *Induktionsvoraussetzung*  $(A_n)$  richtig ist, können wir auch die Richtigkeit von  $(A_{n+1})$  nachweisen. Warum ist mit diesen beiden Schritten tatsächlich der Nachweis von  $(A_n)$  für jedes  $n \in \mathbb{N}$  erfolgt?



Wir betrachten den unendlich langen Zug in obiger Abbildung. Die Aussage „ $(A_n)$  ist richtig“ soll in diesem Bild bedeuten „Der Waggon  $n$  fährt“. Wir nehmen zunächst an, dass die Waggons nicht miteinander gekoppelt sind. Wenn also  $(A_1)$  bewiesen ist, so fährt die Lokomotive los – allerdings allein, weil nichts aneinandergekoppelt ist. Haben wir „ $(A_1) \Rightarrow (A_2)$ “ bewiesen, so haben wir die Wahrheit von  $(A_2)$  an die Wahrheit von  $(A_1)$  gekoppelt: Mit  $(A_1)$  wahr, ist auch  $(A_2)$  wahr. Fährt die Lokomotive los, so auch Waggon 2. Im Induktionsschritt sind sogar alle

Waggons miteinander gekoppelt. Fährt nun die Lokomotive aufgrund der Induktionsverankerung los, so auch der unendlich lange Zug.

Nun können wir  $(A_n)$  beweisen.  $(A_1)$  ist ja richtig. Zum Nachweis von  $(A_{n+1})$  dürfen wir nun  $(A_n)$  verwenden. Wir schauen  $(A_{n+1})$  tief in die Augen und kommen dann mit dem gleichen Verfahren wie bei (2.1) auf

$$\begin{aligned} & 1 + 3 + \dots + (2n - 1) + (2(n + 1) - 1) \\ &= \left(1 + 3 + \dots + (2n - 1)\right) + (2(n + 1) - 1) \\ &= n^2 + (2(n + 1) - 1) \\ &= n^2 + 2n + 1 = (n + 1)^2. \end{aligned}$$

Damit ist  $(A_{n+1})$  bewiesen.

Anders ausgedrückt wird der modus ponens aus Abschnitt 1.1 unendlich oft angewendet.  $(A_1)$  ist die Induktionsvoraussetzung, dann wird der Induktionsschritt für  $n = 1$  angewendet, also ist  $(A_1) \Rightarrow (A_2)$  ebenfalls bewiesen, nach dem modus ponens daher auch  $(A_2)$ . Durch fortgesetzte Anwendung des Induktionsschritts begleitet vom modus ponens erhält man den Beweis von  $(A_n)$  für alle  $n$ .

**Beispiel 2.1**  $n$  Autos stehen auf einer Kreislinie. Die Autos besitzen zusammen so viel Benzin, um damit einmal um den Kreis herumzufahren. Man zeige, dass es ein Auto gibt, das den Kreis einmal umrunden kann, wenn es das Benzin der Autos, bei denen es vorbeikommt, mitnehmen darf.

Der Einfachheit halber nehmen wir an, dass das Umrunden des Kreises eine Entfernungseinheit beträgt und dass man dazu eine Einheit Benzin benötigt. Das Auto  $i$  erhält  $t_i$  Benzin mit  $\sum t_i = 1$ .

Der Induktionsanfang  $n = 1$  ist klar, weil in diesem Fall das Auto 1  $t_1 = 1$  Benzin bekommt. Sei also die Behauptung für  $n$  Autos bewiesen. Bei  $n + 1$  Autos überlegen wir als erstes, dass es ein Auto geben muss, das zumindest das nächste Auto erreichen kann. Da die Summe der Entfernungen zwischen den Autos 1 ist und die Summe des Benzins ebenfalls 1, muss dies für ein Auto möglich sein. Sei die Nummer dieses Autos  $n$  und das mit dem Benzin von  $n$  erreichbare Auto habe die Nummer  $n + 1$ . Wir geben das Benzin von  $n + 1$  dem Auto  $n$  und lassen das Auto  $n + 1$  weg. Nach Induktionsvoraussetzung gibt es ein Auto, das die Umrundung schafft. Dieses Auto schafft die Umrundung aber auch in der unmodifizierten Situation: Wenn es bei  $n$  vorbeikommt, reicht das Benzin, um bis zum Auto  $n + 1$  zu kommen und das Benzin von  $n + 1$  mitzunehmen.  $\square$

## 2.2 Die Fibonacci-Zahlen

Die *Fibonacci-Zahlen*  $F_n$  sind definiert durch die Anfangsvorgaben

$$F_0 = 0, \quad F_1 = 1,$$

sowie durch die *Rekursion*

$$F_{n+1} = F_n + F_{n-1} \quad \text{für alle } n \in \mathbb{N}.$$

Wir bekommen die Folge  $F_0, F_1, \dots$  der Fibonacci-Zahlen, indem wir die letzte Formel sukzessive für  $n = 1, 2, \dots$  anwenden. Für  $n = 1$  ergibt sich also  $F_2 = F_1 + F_0 = 1 + 0 = 1$ . Allgemeiner ist jede Fibonacci-Zahl die Summe ihrer beiden Vorgänger. In der Definition haben wir also ein verallgemeinertes Induktionsprinzip kennengelernt: Da jede Fibonacci-Zahl  $F_{n+1}$  von ihren beiden Vorgängern  $F_n, F_{n-1}$  abhängt, benötigen wir *zwei „Induktionsanfänge“*  $F_0$  und  $F_1$ . Damit sind die Fibonacci-Zahlen für alle natürlichen Zahlen definiert und lassen sich, da nur die beiden vorherigen Fibonacci-Zahlen addiert werden müssen, leicht hinschreiben:

$$F_0 = 0, \quad F_1 = 1, \quad F_2 = 1, \quad F_3 = 2, \quad F_4 = 3, \quad F_5 = 5, \quad F_6 = 8, \quad F_7 = 13, \quad F_8 = 21, \quad F_9 = 34.$$

Erfunden hat die Fibonacci-Zahlen der Mathematiker Leonardo von Pisa (ca 1170-1240), der später Fibonacci genannt wurde. Mit den Fibonacci-Zahlen soll die Kaninchenaufgabe gelöst werden, also wie viele Kaninchen im Laufe einer Zeitspanne aus einem Paar entstehen. Es wird angenommen, dass jedes Paar allmonatlich ein neues Paar in die Welt setzt, das wiederum nach *zwei* Monaten ein weiteres Paar produziert. Man nimmt also an, dass die neugeborenen Kaninchen nicht sofort geschlechtsreif sind. Todesfälle werden nicht berücksichtigt. Hat man im ersten Monat ein neugeborenes Paar (N), so im zweiten Monat ein geschlechtsreifes Paar (G) und im dritten Monat 2 Paare, nämlich 1N+1G. Im 4. Monat hat man 3 Paare, nämlich 1N+2G. Bezeichnet man mit  $F_n$  die Anzahl der Kaninchenaare im Monat  $n$ , so kommen im Monat  $n + 1$  gerade  $F_{n-1}$  hinzu:

$$\begin{array}{ccccccccc} F_{n+1} & = & F_n & + & F_{n-1} \\ \text{Paare in } n+1 & & \text{Paare in } n & & \text{geschlechtsreife Paare in } n \end{array}$$

Wäre jedes neugeborene Paar sofort geschlechtsreif, so hätte man stattdessen die Rekursion  $F_{n+1} = 2F_n$ , was eine Verdoppelung der Paare in jedem Monat bedeuten würde. Die Berücksichtigung der Geschlechtsreife führt dagegen zu einem langsameren Wachstum der Population, nämlich

$$\begin{aligned} \frac{F_6}{F_5} &= \frac{8}{5} = 1,6, & \frac{F_7}{F_6} &= \frac{13}{8} = 1,625, & \frac{F_8}{F_7} &= \frac{21}{13} = 1,615\dots, \\ \frac{F_9}{F_8} &= \frac{34}{21} = 1,619\dots, & \frac{F_{10}}{F_9} &= \frac{55}{34} = 1,617\dots. \end{aligned}$$

Das sieht recht geheimnisvoll aus: Die Quotienten scheinen um einen nicht offensichtlichen Wert zu oszillieren, der in der Nähe von 1,618 liegt.

Nun wollen wir die verwandte Frage diskutieren, für welche positiven Zahlen  $a$  die Abschätzung

$$F_n \leq a^n$$

für alle  $n \in \mathbb{N}_0$  richtig ist. Um erst einmal die Struktur des Beweises zu verstehen, machen wir es uns einfach und beweisen die Aussagen

$$(D_n) \quad F_n \leq 2^n \quad \text{für alle } n \in \mathbb{N}_0.$$

Wir wollen das Induktionsprinzip verwenden, haben aber Schwierigkeiten, weil in  $F_{n+1} = F_n + F_{n-1}$  sowohl  $F_n$  als auch  $F_{n-1}$  vorkommen. Wir zeigen daher

- (i)  $(D_0)$  und  $(D_1)$  (= Induktionsanfang),
- (ii)  $(D_{n-1}), (D_n) \Rightarrow (D_{n+1})$  für alle  $n \in \mathbb{N}$  (= Induktionsschritt).

Wir können leicht durchprobieren, dass damit die Behauptung für alle  $n \in \mathbb{N}_0$  bewiesen ist.  $(D_0)$  und  $(D_1)$  sind nach dem ersten Schritt richtig. Zum Beweis von  $(D_2)$  setzen wir im zweiten Schritt  $n = 1$ , und erhalten, da  $(D_0)$  und  $(D_1)$  richtig sind, die Behauptung  $(D_2)$ . Für die größeren  $n$  geht das ganz genauso.

Der Beweis von  $(D_0)$  und  $(D_1)$  ist

$$F_0 = 0 \leq 1 = 2^0, \quad F_1 = 1 \leq 2 = 2^1.$$

Zum Nachweis von  $(D_{n+1})$  dürfen wir die Induktionsvoraussetzung

$$F_n \leq 2^n, \quad F_{n-1} \leq 2^{n-1}$$

verwenden. Demnach gilt

$$(2.2) \quad F_{n+1} = F_n + F_{n-1} \leq 2^n + 2^{n-1} \leq 2 \cdot 2^n = 2^{n+1}.$$

Damit ist  $(D_{n+1})$  bewiesen.

Kommen wir nun zur Ausgangsfrage zurück, für welche  $a > 0$  die Abschätzung

$$F_n \leq a^n$$

für alle  $n$  in  $\mathbb{N}_0$  richtig ist. Gleichzeitig soll hier gezeigt werden, dass mit dem Prinzip der vollständigen Induktion nicht nur vermutete Aussagen bewiesen, sondern auch völlig neue Erkenntnisse hergeleitet werden können, wenn man mit dem Prinzip kreativ umgeht. Der Beweis der neuen Aussage läuft genauso wie vorher. Der Induktionsanfang  $F_0 \leq a^0$  und  $F_1 \leq a$  ist für jedes  $a \geq 1$  richtig. Die Hauptschwierigkeit ist der Schritt (2.2), den wir ganz analog durchführen wollen:

$$F_{n+1} = F_n + F_{n-1} \leq a^n + a^{n-1} \stackrel{!}{\leq} a^{n+1}.$$

Das Ausrufezeichen bedeutet hier, dass wir diejenigen  $a$  herausfinden müssen, für die

$$a^n + a^{n-1} \leq a^{n+1}$$

richtig ist. Da  $a \geq 1$  wegen des Induktionsanfangs, können wir hier kürzen und erhalten

$$(2.3) \quad a + 1 \leq a^2$$

und somit

$$a \geq \Phi = \frac{1}{2} + \frac{\sqrt{5}}{2} = 1.618033\dots,$$

was im Einklang mit den obigen Untersuchungen von  $F_{n+1}/F_n$  steht. Die Zahl  $\Phi$  heißt *goldener Schnitt* und löst folgendes Problem: Gesucht ist das Verhältnis der Seitenlängen  $a, b$  eines Rechtecks mit

$$\frac{a}{b} = \frac{a+b}{a} \Leftrightarrow \frac{\text{lange Seite}}{\text{kurze Seite}} = \frac{\text{Summe der Seiten}}{\text{lange Seite}}.$$

Mit  $\Phi = a/b$  folgt hieraus  $\Phi = 1 + \Phi^{-1}$  und  $\Phi^2 = \Phi + 1$ , was gerade die mit (2.3) verbundene quadratische Gleichung ist.

**2.3 Mächtigkeit der Potenzmenge** Wir hatten die Potenzmenge  $\mathcal{P}(A)$  einer Menge  $A$  definiert als die Menge aller Teilmengen von  $A$ , wobei auch  $\emptyset$  und  $A$  Elemente von  $\mathcal{P}(A)$  sind.

Wir wollen die Anzahl der Teilmengen der Menge  $A_n = \{1, 2, \dots, n\}$  bestimmen. Dazu bietet sich vollständige Induktion über  $n$  an, allerdings müssen wir erst einmal wissen, *was* wir beweisen sollen – die Induktion sagt uns das ja nicht. Durch Probieren stellen wir zunächst eine Hypothese auf:

$A_1 : \emptyset, \{1\}$	2
$A_2 : \emptyset, \{1\}, \{2\}, \{1, 2\}$	4
$A_3 : \emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$	8

Die Vermutung ist also:  $A_n$  besitzt  $2^n$  Teilmengen.

Für  $n = 1$  ist die Behauptung richtig (=Induktionsanfang). Sei  $2^n$  die Anzahl der Teilmengen von  $A_n$  (=Induktionsvoraussetzung). Die Beweisidee bei solchen kombinatorischen Problemen ist die Strukturierung der zu zählenden Objekte nach dem Motto „Teile und Herrsche“. Wir zerlegen die Teilmengen von  $A_{n+1}$  in zwei Gruppen:

I : Teilmengen, die  $n+1$  nicht enthalten,

II : Teilmengen, die  $n+1$  enthalten.

Gruppe I enthält genau die Teilmengen von  $A_n$ , das sind nach Induktionsvoraussetzung  $2^n$ . In den Teilmengen von Gruppe II können wir das Element  $n + 1$  weglassen und wir erhalten eine Teilmenge von  $A_n$ . Umgekehrt können wir jede Teilmenge von  $A_n$  durch Anfügen von  $n + 1$  zu einer Teilmenge von Gruppe II machen. Damit enthält auch Gruppe II genau  $2^n$  Teilmengen, zusammen also  $2^n + 2^n = 2^{n+1}$ , wie zu beweisen war. Wir haben damit gezeigt:

**Satz 2.2** *Die Anzahl der Teilmengen einer  $n$ -elementigen Menge ist  $2^n$ .*

**2.4 Permutationen und Fakultät** Eine *Permutation* von  $(1, 2, \dots, n)$  ist eine Umstellung der Zahlen  $1, \dots, n$ . Beispielsweise besitzt  $(1, 2, 3)$  die Permutationen

$$(1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1).$$

Alternativ kann man die Permutationen als bijektive Abbildungen der Menge  $A_n = \{1, 2, \dots, n\}$  in sich definieren. Beispielsweise gehört zur Permutation  $(2, 3, 1)$  die Abbildung mit  $f(1) = 2$ ,  $f(2) = 3$  und  $f(3) = 1$ . Wir stellen uns dabei vor, dass  $(\cdot, \cdot, \cdot)$  aus nummerierten Kästchen besteht, in denen wir die Werte von  $f$  hineinschreiben.

Für eine Zahl  $n \in \mathbb{N}$  ist  $n!$  (gesprochen:  $n$  Fakultät) definiert durch

$$n! = 1 \cdot 2 \cdots n.$$

Die Fakultäten wachsen sehr schnell in  $n$ ,

$$3! = 6, \quad 4! = 24, \quad 5! = 120, \quad 6! = 720, \quad 20! = 2.43 \dots \times 10^{18}.$$

Rein aus praktischen Gründen setzt man  $0! = 1$ .

**Satz 2.3** *Die Anzahl der Permutationen von  $(1, 2, \dots, n)$  ist  $n!$ .*

*Beweis:* Man kann das durch vollständige Induktion über  $n$  beweisen. Einfacher ist die Überlegung, auf wie viele Arten man die Zahlen  $1, 2, \dots, n$  auf  $n$  nummerierte Kästchen verteilen kann. Für die Zahl 1 hat man  $n$  Möglichkeiten, für die Zahl 2 sind es  $n - 1$ , für die letzte Zahl  $n$  verbleibt nur noch eine Möglichkeit.  $\square$

**2.5 Binomialkoeffizienten und binomische Formel** Für  $n \in \mathbb{N}_0$  sind die *Binomialkoeffizienten* folgendermaßen definiert

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad \text{für } 0 \leq k \leq n.$$

Dass all diese Werte natürliche Zahlen sind, werden wir später sehen. Wichtig sind im Folgenden die Fälle

$$(2.4) \quad \binom{n}{0} = \frac{n!}{0!n!} = 1, \quad \binom{n}{n} = \frac{n!}{n!0!} = 1.$$

Wir beweisen die technische Formel

**Lemma 2.4**

$$\binom{n}{k-1} + \binom{n}{k} = \binom{n+1}{k}.$$

*Beweis:* Wir bringen die linke Seite auf den Hauptnenner,

$$\begin{aligned} \binom{n}{k-1} + \binom{n}{k} &= \frac{n!}{(k-1)!(n-k+1)!} + \frac{n!}{k!(n-k)!} \\ &= \frac{n!k}{k!(n-k+1)!} + \frac{n!(n-k+1)}{k!(n-k+1)!} = \frac{(n+1)!}{k!(n+1-k)!} = \binom{n+1}{k}. \end{aligned}$$

□

Man interpretiert das Lemma durch das *Pascalsche Dreieck*:

n=0	1					
n=1	1      1					
n=2	1      2      1					
n=3	1      3      3      1					
n=4	1      4      6      4      1					
n=5	1      5      10     10     5      1					

Jede neue Zeile wird rechts und links um 1 ergänzt, was den Werten  $\binom{n}{0}$  und  $\binom{n}{n}$  in (2.4) entspricht, die übrigen Einträge erhält man aus dem Lemma, jeder Eintrag ist die Summe der links und rechts über ihm stehenden Zahlen.

**Satz 2.5** Die Zahl der  $k$ -elementigen Teilmengen einer  $n$ -elementigen Menge ist  $\binom{n}{k}$ .

*Beweis:* Wir zeigen dies durch vollständige Induktion über  $n$ . Für  $n = 0$  ist die Behauptung richtig, denn die leere Menge enthält nur sich selbst als Teilmenge. Die Behauptung ist auch richtig für  $k = 0$  und  $k = n$ , in beiden Fällen haben wir nur eine Teilmenge, die leere Menge bzw. die Menge selbst, was mit den Werten in (2.4) übereinstimmt. Nach dem Prinzip „Teile und Herrsche“ strukturieren wir die  $k$ -elementigen Teilmengen der Menge  $A_{n+1} = \{1, 2, \dots, n+1\}$  in zwei Gruppen:

I :  $k$ -elementige Teilmengen, die  $n+1$  nicht enthalten,

II :  $k$ -elementige Teilmengen, die  $n+1$  enthalten.

Gruppe I besteht genau aus den  $k$ -elementigen Teilmengen der Menge  $A_n = \{1, 2, \dots, n\}$ , nach Induktionsvoraussetzung sind das  $\binom{n}{k}$ .

In den Teilmengen der Gruppe II können wir das Element  $n+1$  weglassen und erhalten eine  $k-1$ -elementige Teilmenge von  $A_n$ . Umgekehrt können wir jede  $k-1$ -elementige Teilmenge von  $A_n$  um das Element  $n+1$  ergänzen und erhalten eine Teilmenge von Gruppe II. Nach Induktionsvoraussetzung ist die Zahl der Teilmengen in Gruppe II gerade  $\binom{n}{k-1}$ . Für die Gesamtzahl der Teilmengen gilt daher mit obigem Lemma

$$\text{Gruppe I} + \text{Gruppe II} = \binom{n}{k} + \binom{n}{k-1} = \binom{n+1}{k}.$$

Damit ist der Induktionsbeweis erfolgreich abgeschlossen. □

**Beispiel 2.6** Beim Lotto „6 aus 49“ ist die Wahrscheinlichkeit, alle sechs Zahlen richtig getippt zu haben, gleich der Wahrscheinlichkeit, aus der Gesamtheit der 6-elementigen Teilmengen von  $\{1, 2, \dots, 49\}$  die „richtige“ herausgefunden zu haben. Die Zahl der Möglichkeiten ist

$$\binom{49}{6} = \frac{49!}{6! 43!} = \frac{49 \cdot (2 \cdot 4 \cdot 6) \cdot 47 \cdot 46 \cdot (3 \cdot 5 \cdot 3) \cdot 44}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6} = 49 \cdot 47 \cdot 46 \cdot 3 \cdot 44 = 13\,983\,816.$$

Die Wahrscheinlichkeit für sechs Richtige ist daher ungefähr 1 : 14 Millionen.  $\square$

Wir betrachten nun einen kommutativen Ring  $(R, 0, 1, +, \cdot)$ . Wir definieren Potenzen

$$a^n = \underbrace{a \cdot a \cdot \dots \cdot a}_{n\text{-mal}}, \quad a^0 = 1.$$

Für  $m, n \in \mathbb{N}_0$  gelten die Potenzgesetze

$$(2.5) \quad a^{m+n} = a^m \cdot a^n, \quad a^n b^n = (ab)^n, \quad (a^m)^n = a^{mn},$$

die sich leicht durch vollständige Induktion beweisen lassen.

**Satz 2.7** Für  $a, b \in R$  und  $n \in \mathbb{N}_0$  gilt die binomische Formel

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

*Beweis:* Wir verwenden Induktion über  $n$ . Für  $n = 0$  ist die Formel richtig wegen  $a^0 = b^0 = 1$ . Unter der Annahme, dass sie für  $n$  richtig ist, folgt

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

Mit Ummumerierung erhalten wir für den ersten Summanden

$$\sum_{i=0}^n \binom{n}{i} a^{n-i+1} b^i = a^{n+1} + \sum_{i=0}^{n-1} \binom{n}{i+1} a^{n-i} b^{i+1},$$

daher

$$(a+b)^{n+1} = a^{n+1} + \sum_{i=0}^{n-1} \left( \binom{n}{i+1} + \binom{n}{i} \right) a^{n-i} b^{i+1} + b^{n+1}.$$

Die Behauptung folgt aus der Additionseigenschaft der Binomialkoeffizienten in Lemma 2.4.  $\square$

**2.6 Modelle** Eine konkrete Menge (mit zugehörigen ausgezeichneten Elementen, Operationen und Relationen), in der die Axiome einer mathematischen Struktur gelten, heißt *Modell* dieser Struktur.

Alles, was wir als Beispiele von Gruppen bezeichnet haben, sind Modelle der Gruppe. Modelle sind daher konkret, haben philosophisch gesprochen ein eigenes Sein in der mathematischen Welt. Dagegen ist eine mathematische Struktur i.A. abstrakt. Das Axiomensystem der Gruppe definiert gleichzeitig, was eine Gruppe ist.

In der Mathematik gibt es zwei Arten von Strukturen:

1. Strukturen mit unterschiedlichen Modellen wie Gruppen, Ringe, Körper. Diese werden durch die jeweiligen Axiome definiert, die man kennen und für die Beweise nutzen muss.

2. „Eindeutige“ Strukturen, die im Wesentlichen nur ein Modell besitzen wie etwa die natürlichen Zahlen. Diese beginnen mit einer Wurzel, meist 0 oder 1 genannt, und bestehen aus den Nachfolgern der Wurzel. Abgesehen davon, dass man den Zahlen unterschiedliche Namen geben kann, ist diese Struktur immer dieselbe. Es gibt keine wirklich verschiedenen Modelle wie etwa bei den Gruppen. Weitere Strukturen dieses Typs sind die ganzen, die rationalen und die reellen Zahlen, die später eingeführt werden. Hier brauchen wir eigentlich keine Axiome, es ist legitim, sich einfach auf den Standpunkt zu stellen, dass man diese Strukturen kennt.

**2.7 Die Peanoschen Axiome für die natürlichen Zahlen** Die Axiome für die natürlichen Zahlen müssen so formuliert werden, dass es nur ein einziges Modell gibt – eine sportliche Herausforderung, die der Mathematiker Giuseppe Peano Ende des 19. Jahrhunderts erfolgreich angenommen hat.

In moderner Schreibweise sind die natürlichen Zahlen eine Struktur  $\mathbb{N} = (N, 1, f)$  mit dem ausgezeichneten Element 1 und einer einstelligen Abbildung  $f : N \rightarrow N$ , die als Nachfolger interpretiert wird. Die Axiome sind dann

- (P1) Für alle  $m, n$ : Wenn  $f(m) = f(n)$ , so gilt  $m = n$ ,
- (P2) Es gibt kein  $n \in N$  mit  $f(n) = 1$ ,
- (P3) Für alle Teilmengen  $M \subset N$  gilt:

Ist  $1 \in M$  und folgt aus  $n \in M$ , dass auch  $f(n) \in M$ , so  $M = N$ .

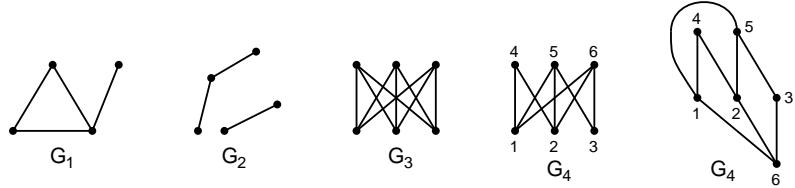
(P1) bedeutet, dass  $f$  injektiv ist, dass es also höchstens ein Urbild zu jedem  $n \in N$  gibt. (P2) besagt, dass 1 nicht im Bild  $f(N)$  von  $f$  liegt, insbesondere ist  $f$  nicht surjektiv. Was lässt sich daraus für die Modelle von (P1) und (P2) (ohne (P3)) schließen? Wäre  $N$  eine endliche Menge, so müsste wegen der Injektivität der Bildbereich  $f(N)$  genauso viel Elemente enthalten wie der Urbildbereich  $N$ , also  $N = f(N)$ . Andererseits darf  $f$  nicht surjektiv sein, womit wir einen Widerspruch erhalten (siehe (1.2)). Die Axiome (P1) und (P2) zwingen die Modelle dazu, unendlich viele Elemente zu besitzen.

Um die Rolle von (P3) zu erläutern, betrachten wir folgendes Modell von (P1) und (P2). Neben den normalen natürlichen Zahlen  $N_1 = \{1, 2, 3, \dots\}$  mit 1 als ausgezeichnetem Element soll die Grundmenge noch die Elemente der Menge  $N_2 = \{a, b\}$  enthalten. Der Nachfolger auf  $N_1$  ist wie üblich als  $f(n) = n + 1$  definiert. Auf  $N_2$  setzen wir  $f(a) = b$  und  $f(b) = a$ . Damit ist  $N = N_1 \cup N_2$  zusammen mit der so definierten Nachfolgerabbildung ein Modell von (P1) und (P2). Denn nach wie vor ist  $f$  injektiv und 1 liegt nicht im Bild von  $f$ . (P3) ist aber nicht erfüllt: Wir wählen  $M = N_1$  und es ist jetzt in der Tat  $1 \in N_1$  und aus  $n \in N_1$  folgt auch  $f(n) = n + 1 \in N_1$ , aber  $N_1$  stimmt nicht mit  $N$  überein.

Das Axiom (P3) der vollständigen Induktion sorgt also dafür, dass unter allen Modellen von (P1) und (P2) das minimale genommen wird:  $N$  soll nur aus  $1, f(1), f(f(1)), \dots$  bestehen.

**2.8 Induktion über den rekursiven Aufbau - Eulersche Polyederformel** Ein (*ungerichteter*) Graph besteht aus einer Knotenmenge  $V$  (engl. vertex) und einer Kantenmenge  $E$  (engl. edge). Anschaulich verbindet eine Kante zwei verschiedene Knoten, wobei es auf die geometrische Form der Kanten meist nicht ankommt.

Ein Graph heißt *zusammenhängend*, wenn je zwei Knoten durch einen Kantenzug miteinander verbunden werden können. Ein Graph heißt *planar*, wenn er auf der Ebene so gezeichnet werden kann, dass seine Kanten sich nicht überkreuzen. Ein Graph heißt *nichtleer*, wenn er mindestens einen Knoten besitzt.



$G_2$  ist nicht zusammenhängend, die anderen Graphen aber schon.  $G_4$  ist planar, weil er kreuzungsfrei gezeichnet werden kann.

Ein zusammenhängender planarer Graph unterteilt die Ebene in  $f$  Flächen, wobei die Außenfläche mitgezählt wird.

**Satz 2.8 (Eulersche Polyederformel)** Sei  $G$  ein nichtleerer, planarer, zusammenhängender Graph mit  $f$  Flächen,  $e$  Kanten und  $v$  Knoten. Dann gilt

$$v - e + f = 2.$$

Für den Graphen  $G_4$  erhalten wir

$$v = 6, \quad e = 8, \quad f = 4 \quad \Rightarrow \quad v - e + f = 2.$$

*Beweis:* Der Induktionsanfang ist der Graph, der nur aus einem Knoten besteht. In diesem Fall ist  $v = 1$ ,  $e = 0$  und  $f = 1$ , also  $v - e + f = 2$ .

Zum Zeichnen des Graphen benötigen wir die Operationen:

1. Setzen eines neuen Knotens und Verbinden dieses Knotens mit einem alten Knoten. In diesem Fall ändert sich  $v$  um +1 und  $e$  um +1.
2. Verbinden zweier Knoten. Hier ändert sich  $e$  um +1 und  $f$  um +1.

Wir können jeden nichtleeren, planaren, zusammenhängenden Graphen beginnend mit dem Graphen, der nur aus einem Knoten besteht, mit den beiden genannten Operationen aufbauen. Die Beziehung  $v - e + f = 2$  ist für den Anfangsgraphen richtig und bleibt nach jedem Schritt bestehen.

□



Für einen Polyeder mit  $v$  Knoten,  $e$  Kanten und  $f$  Seitenflächen ist die Eulersche Polyederformel ebenso richtig. Wir können nämlich einen Polyeder an einer Fläche aufschneiden und auf die Ebene klappen. Daher wird in der Polyederformel die Außenfläche mitgezählt, weil sie einer Fläche des Polyeders entspricht.

Wie aus dem Beweis der Polyederformel ersichtlich ist, hat die Formel  $v - e + f = 2$  einen „statischen“ Charakter: Ändern wir einen nichtleeren, planaren, zusammenhängenden Graphen so ab, beispielsweise indem wir eine Kante weglassen, dass er immer noch nichtleer, planar und zusammenhängend ist, so bleibt die Formel erhalten. Eine solche Formel, die in einer sich dynamisch verändernden Struktur erhalten bleibt, nennt man *Invariante* der Struktur. Dazu das instruktive

**Beispiel 2.9** Es gibt 11 rote, 4 blaue und 6 gelbe Chamäleons. Treffen zwei Chamäleons verschiedener Farbe aufeinander, so nehmen sie die dritte Farbe an. Z.B. entstehen aus einem roten und einem blauen Chamäleon zwei gelbe. Ist es möglich, dass am Ende alle Chamäleons die gleiche Farbe besitzen?

Wir schreiben rote, blaue und gelbe Chamäleons als 3-tupel  $(r, b, g)$ . Den beschriebenen Farbwechseln entsprechen dann die drei Operationen

$$(r, b, g) \rightarrow (r - 1, b - 1, g + 2), \quad (r, b, g) \rightarrow (r - 1, b + 2, g - 1), \quad (r, b, g) \rightarrow (r + 2, b - 1, g - 1).$$

Eine Invariante dieser Farbwechsel ist zunächst die Anzahl der Chamäleons, die sich nicht ändert. Leider nutzt uns diese Invariante nichts bei der Beantwortung der Frage. Wir brauchen eine Invariante, die die Verteilung der Farben beschreibt. Versuchen wir einmal, eine Invariante zu finden, die man aus zwei Farben bestimmen kann:

$$(r, b) \rightarrow (r - 1, b - 1), \quad (r, b) \rightarrow (r - 1, b + 2), \quad (r, b) \rightarrow (r + 2, b - 1).$$

Die Differenz  $r - b$  bleibt gleich oder verändert sich um  $\pm 3$ . Bei  $r = 11, b = 4$  erhalten wir  $11 - 4 = 7$ . Ist die geforderte Endposition  $r = 21$ , so ist  $21 - 0 = 21$ . Ist sie  $g = 21$ , so  $0 - 0 = 0$ . Man kann also die Endposition mit 21 gleichfarbigen Chamäleons nicht erreichen.  $\square$

### 3 Algebra und Zahlentheorie

**3.1 Grundlegende Sätze der elementaren Zahlentheorie** Mit  $a, b$  usw. bezeichnen wir, solange nichts anderes gesagt wird, immer ganze Zahlen. Wir schreiben  $a|b$ , wenn  $a$  ein Teiler von  $b$  ist, wenn also  $b = aq$  für eine ganze Zahl  $q$ .

$p \in \mathbb{N}$  heißt *Primzahl* genau dann, wenn  $p$  genau zwei Teiler hat, nämlich 1 und  $p$ . Damit ist 1 keine Primzahl, weil sie nur einen Teiler besitzt.

Aus der Schule ist der folgende Satz über die eindeutige Zerlegung einer natürlichen Zahl in ihre Primfaktoren bekannt.

**Satz 3.1 (Fundamentalsatz der Arithmetik)** Ist  $p_1 = 2, p_2 = 3, p_3 = 5, \dots$  die Folge der Primzahlen, so gibt es zu jeder natürlichen Zahl  $a > 1$  eindeutige Exponenten  $r_1, \dots, r_k \in \mathbb{N}_0$  mit

$$a = p_1^{r_1} p_2^{r_2} \cdots p_k^{r_k}, \quad r_k > 0.$$

Der Satz ist hoffentlich intuitiv klar. Wenn wir vor der rechten Seite ein Minuszeichen setzen dürfen, gilt er auch für alle  $a \in \mathbb{Z}$  mit  $|a| > 1$ .

Unter den vielen Anwendungen dieses Satzes erwähnen wir: Ist eine Primzahl  $p$  Teiler von  $ab$ , so ist  $p|a$  oder  $p|b$ . Schauen wir uns nämlich die Primfaktorzerlegungen von  $a$  und  $b$  an, so muss  $p$  in einer der beiden vorkommen.

Wir sagen,  $a$  ist *kongruent zu  $b$  modulo  $m$* , wenn die natürliche Zahl  $m$  ein Teiler von  $b - a$  ist, und schreiben dafür  $a \equiv b \pmod{m}$ . Die Zahl  $m$  heißt *Modul* der Kongruenz. Die Differenz zweier gerader Zahlen ist gerade, sie sind daher kongruent modulo 2. Ebenso sind zwei ungerade Zahlen kongruent modulo 2, weil ihre Differenz ebenfalls geradzahlig ist.

Sind zwei Zahlen kongruent modulo  $m$ , so muss die Differenz der beiden Zahlen ein ganzzahliges Vielfaches von  $m$  sein, daher

$$(3.6) \quad a \equiv b \pmod{m} \Leftrightarrow m|a - b \Leftrightarrow a = b + qm \text{ für ein } q \in \mathbb{Z}.$$

Ist  $a$  eine natürliche Zahl, so hinterlässt sie beim Teilen durch  $m$  einen Rest in der Menge  $\{0, 1, \dots, m - 1\}$ . Zwei natürliche Zahlen  $a, b$  sind genau dann kongruent modulo  $m$ , wenn sie beim Teilen durch  $m$  den gleichen Rest besitzen, denn dieser Rest fällt ja in  $b - a$  heraus. Dieses Prinzip lässt sich auch auf negative Zahlen ausdehnen, wenn wir  $m$  auf den Rest addieren.

Aus (3.6) entnimmt man direkt die Rechenregeln

$$a \equiv b \pmod{m}, \quad c \equiv d \pmod{m} \Rightarrow a \pm c \equiv b \pm d \pmod{m} \quad \text{und} \quad ac \equiv bd \pmod{m},$$

insbesondere auch

$$a \equiv b \pmod{m} \Rightarrow a^k \equiv b^k \pmod{m}.$$

Diese Regeln lassen sich folgendermaßen zusammenfassen: Ist  $p(x)$  ein Polynom mit ganzzahligen Koeffizienten, so gilt

$$a \equiv b \pmod{m} \Rightarrow p(a) \equiv p(b) \pmod{m}.$$

Zwei natürliche Zahlen heißen *teilerfremd*, wenn sie nur 1 als gemeinsamen Teiler besitzen.

Vorsicht ist bei der Division in der Kongruenzrelation geboten. Es gilt  $m \equiv 2m \pmod{m}$ , aber  $1 \not\equiv 2 \pmod{m}$ . Daher

$$ac \equiv bc \pmod{m}, \quad c \text{ und } m \text{ teilerfremd} \Rightarrow a \equiv b \pmod{m}.$$

Man beweist diese Regel, indem man für  $ac \equiv bc \pmod{m}$  die äquivalente Form  $m|(b - a)c$  betrachtet. Sind  $m$  und  $c$  teilerfremd, so kommen in den Primfaktorzerlegungen von  $m$  und  $c$  nur verschiedene Primzahlen vor. Damit muss  $m$  ein Teiler von  $b - a$  sein.

**Satz 3.2 (Kleiner Satz von Fermat)** Sei  $a$  positiv und  $p$  eine Primzahl. Dann gilt

$$a^p \equiv a \pmod{p}.$$

Ist  $p$  kein Teiler von  $a$ , folgt hieraus

$$a^{p-1} \equiv 1 \pmod{p}.$$

*Beweis:* Wir verwenden vollständige Induktion über  $a$ . Für  $a = 1$  ist  $p \mid 1^p - 1$  richtig. Als Induktionsvoraussetzung nehmen wir an, dass die Behauptung für  $a$  richtig ist, dass also  $p \mid a^p - a$ . Wir müssen zeigen, dass

$$p \mid (a+1)^p - (a+1).$$

Mit der binomischen Formel Satz 2.7 erhalten wir

$$\begin{aligned} (a+1)^p - (a+1) &= \sum_{i=0}^p \binom{p}{i} a^i - (a+1) \\ &= a^p + 1 + \sum_{i=1}^{p-1} \binom{p}{i} a^i - (a+1) \\ (3.7) \quad &= a^p - a + \sum_{i=1}^{p-1} \binom{p}{i} a^i. \end{aligned}$$

Auf der rechten Seite ist  $a^p - a$  aufgrund der Induktionsvoraussetzung durch  $p$  teilbar. Die Binomialkoeffizienten

$$\binom{p}{i} = \frac{p!}{i!(p-i)!}$$

sind ganzzahlig. Ist  $p$  eine Primzahl, so kann der Faktor  $p$  im Zähler für  $i \neq 0$  und  $i \neq p$  nicht herausgekürzt werden. Da die Binomialkoeffizienten in (3.7) durch  $p$  teilbar sind, ist auch die linke Seite von (3.7) durch  $p$  teilbar.  $\square$

**Beispiel 3.3** Zeigen Sie, dass für jede positive Zahl  $n$  gilt  $30 \mid n^5 - n$ .

*Lösung:* Die Teilbarkeit durch 5 folgt aus dem Fermatschen Satz. Wegen

$$n^5 - n = (n-1)n(n+1)(n^2+1)$$

ist  $n^5 - n$  außerdem durch 2 und durch 3 teilbar, denn beide Zahlen müssen Teiler einer Zahl in der Folge  $n-1, n, n+1$  sein.  $\square$

Aus der Elementarmathematik gut bekannt ist die „Division mit Rest“: Sind  $a \in \mathbb{N}_0$  und  $b \in \mathbb{N}$ , so gibt es eindeutig bestimmte Zahlen  $m, r \in \mathbb{N}_0$  mit

$$(3.8) \quad a = mb + r, \quad 0 \leq r < b.$$

Dazu überlegt man sich, dass jede nichtnegative ganze Zahl in genau einem Intervall  $[0, b), [b, 2b), \dots$  liegen muss. Daher sind sowohl  $m$  als auch  $r$  eindeutig bestimmt.

Wir können (3.8) auch für ganzzahliges  $a$  übernehmen. In diesem Fall existieren eindeutige  $b \in \mathbb{Z}$  und  $0 \leq r < m$  mit

$$(3.9) \quad a = mb + r.$$

Wir bekommen damit die *Ganzzahldivision*  $a \text{ div } m = b$ , die auch in den meisten Programmiersprachen implementiert ist. Man beachte  $15 \text{ div } 7 = 2$ , aber, weil  $r \geq 0$  gefordert wird,  $-15 \text{ div } 7 = -3$ .

Den in (3.9) auftretenden Rest  $r \in \{0, \dots, m - 1\}$  bezeichnen wir als *Rest von  $a$  modulo  $m$*  und schreiben dafür  $a \bmod m$ . Mit diesen Bezeichnungen können wir (3.9) in der Form

$$a = m \cdot (a \text{ div } m) + (a \bmod m)$$

schreiben.

$d$  heißt *größter gemeinsamer Teiler* von  $a \in \mathbb{N}$  und  $b \in \mathbb{N}$ , wenn  $d | a, b$  und wenn aus  $t | a$  und  $t | b$  folgt, dass  $t | d$ . Wir schreiben dafür  $d = \text{ggT}(a, b)$ . Für teilerfremde Zahlen gilt  $\text{ggT}(a, b) = 1$ . Den größten gemeinsamen Teiler kann man aus den Primfaktorzerlegungen der Zahlen  $a$  und  $b$  bestimmen, indem man das Produkt der gemeinsamen Primfaktoren bildet. Dieses Verfahren ist allerdings sehr langsam, so dass man besser auf den im Beweis des nächsten Satzes dargestellten *erweiterten Euklidischen Algorithmus* zurückgreift.

**Satz 3.4 (Satz vom größten gemeinsamen Teiler, Lemma von Bézout)** *Für  $a, b \in \mathbb{N}$  existiert genau ein größter gemeinsamer Teiler  $d \in \mathbb{N}$ . Ferner gibt es Zahlen  $\alpha, \beta \in \mathbb{Z}$  mit*

$$d = \alpha a + \beta b.$$

*Beweis:* Wir dürfen  $a > b$  annehmen. Wir wenden fortgesetzte Division mit Rest nach folgendem Schema solange an, bis der Rest 0 entsteht:

$$\begin{aligned} a &= b \cdot q_1 + r_1, & 0 < r_1 < b, \\ b &= r_1 \cdot q_2 + r_2, & 0 < r_2 < r_1, \\ r_1 &= r_2 \cdot q_3 + r_3, & 0 < r_3 < r_2, \\ &\vdots \\ r_{k-4} &= r_{k-3} \cdot q_{k-2} + r_{k-2}, & 0 < r_{k-2} < r_{k-3}, \\ r_{k-3} &= r_{k-2} \cdot q_{k-1} + r_{k-1}, & 0 < r_{k-1} < r_{k-2}, \\ r_{k-2} &= r_{k-1} \cdot q_k + r_k, & 0 < r_k < r_{k-1}, \\ r_{k-1} &= r_k \cdot q_k. \end{aligned}$$

Da die Folge der Reste nichtnegativ und streng monoton fallend ist, kommen wir nach endlich vielen Schritten zum Rest 0. Wir zeigen nun, dass die Zahl  $r_k$  der größte gemeinsame Teiler von  $a$  und  $b$  ist. Liest man nämlich die Gleichungen von unten nach oben, so kommt man auf die Beziehungen

$$r_k | r_{k-1}, r_k | r_{k-2}, \dots, r_k | b, r_k | a,$$

womit  $r_k$  ein gemeinsamer Teiler von  $b$  und  $a$  ist. Für einen beliebigen gemeinsamen Teiler  $t$  von  $a$  und  $b$  kommt man, wenn man die Gleichungen von oben nach unten liest, auf

$$t | r_1, t | r_2, \dots, t | r_k.$$

Damit ist in der Tat  $r_k = \text{ggT}(a, b)$ .

Zum Nachweis von  $r_k = \alpha a + \beta b$  gehen wir die obigen Gleichungen nochmals von unten nach oben durch. Aus der vorletzten Gleichung ergibt sich

$$r_k = r_{k-2} - r_{k-1}q_k$$

und mit der darüberstehenden Gleichung folgt

$$r_k = (1 + q_{k-1}q_k)r_{k-2} - q_k r_{k-3}.$$

Auf die gleiche Weise kann man hier  $r_{k-2}$  durch eine Kombination von  $r_{k-4}$  und  $r_{k-3}$  darstellen und verbleibt am Ende mit

$$r_k = \alpha a + \beta b.$$

□

**Beispiel 3.5** Das im letzten Beweis dargestellte Verfahren ist deshalb so effektiv, weil sich die  $r_i$  in jedem Schritt mindestens halbieren. Für  $a = 38$  und  $b = 10$  erhält man

$$38 = 10 \cdot 3 + 8$$

$$10 = 8 \cdot 1 + 2$$

$$8 = 2 \cdot 4,$$

also  $\text{ggT}(38, 10) = 2$ .  $\alpha$  und  $\beta$  bestimmt man aus

$$\begin{aligned} 2 &= 10 - 1 \cdot 8 \\ &= 10 - 1 \cdot (38 - 10 \cdot 3) = 4 \cdot 10 - 1 \cdot 38, \end{aligned}$$

also  $\alpha = -1$  und  $\beta = 4$ .  $\square$

**3.2 Stellenwertsysteme** Sei  $g \in \mathbb{N} \setminus \{1\}$ . Die *g-adische Darstellung* einer natürlichen Zahl  $n$  ist von der Form

$$(3.10) \quad n = a_0 \cdot g^0 + a_1 \cdot g^1 + \dots + a_s g^s = \sum_{k=0}^s a_k g^k$$

mit „Ziffern“  $a_k \in \{0, 1, \dots, g-1\}$ . Für die *Basis*  $g$  hat sich im täglichen Gebrauch  $g = 10$  durchgesetzt, wir schreiben ja  $a_s \dots a_0$  für eine solche Dezimalzahl. Relikte anderer Basen sind bei uns noch erkennbar: Stunden, Minuten und Sekunden sind im 60er System strukturiert, das Dutzend und das Gros erinnern an die Basis 12.

Für die Darstellung in (3.10) schreiben wir

$$n = a_s a_{s-1} \dots a_0 g.$$

Diese Darstellung ist offenbar eindeutig und es gilt

$$a_k = (n \text{ div } g^k) \mod g \quad \text{für } k = 0, 1, \dots$$

Für die praktische Rechnung dividiert man fortgesetzt ganzzahlig durch  $g$  und nimmt anschließend die Ergebnisse modulo  $g$ . Z.B. für  $n = 50$  und  $g = 2$  gilt

$$50 \text{ div } 1 = 50, \quad 50 \text{ div } 2 = 25, \quad 25 \text{ div } 2 = 12, \quad 12 \text{ div } 2 = 6, \quad 6 \text{ div } 2 = 3, \quad 3 \text{ div } 2 = 1,$$

daher  $50_{10} = 110010_2$ .

Kommen wir nun zur Darstellung ganzer Zahlen im Rechner. Stehen uns  $s+1$  Bits im Binärsystem  $g = 2$  zur Verfügung, so geht ein Bit für das Vorzeichen verloren. Es ist aber ungünstig, explizit das Vorzeichen zu codieren, weil das bei der Addition zu Fallunterscheidungen führt, denn die Vorzeichen der beiden zu addierenden Zahlen entscheiden darüber, ob addiert oder subtrahiert wird. Besser ist es daher, *Zweierkomplemente* zu verwenden, nämlich

$$[a_s a_{s-1} \dots a_0]_2 = a_{s-1} \dots a_0 - 2^s a_s.$$

Bei  $a_s = 0$  werden die nichtnegativen ganzen Zahlen von  $0 = [0 \dots 0]_2$  bis  $2^s - 1 = [011 \dots 1]_2$  codiert. Die negativen Zahlen laufen von  $-1 = [11 \dots 1]_2$  bis  $-2^s = [10 \dots 0]_2$ . Bei der Addition solcher Zahlen führt man eine normale binäre Addition durch, ohne die besondere Bedeutung der Stelle  $s$  zu berücksichtigen. Allerdings fällt ein Übertrag von der Stelle  $s$  unter den Tisch. Solange sich die Zahlen im angegebenen Bereich bewegen, ist diese Addition korrekt:

**Beispiel 3.6** Für  $s = 3$  können die Zahlen von  $-2^3 = -8$  bis  $2^3 - 1 = 7$  dargestellt werden. Es ist klar, dass zwei nichtnegative Zahlen korrekt addiert werden, solange die 7 nicht überschritten wird. Andernfalls erhalten wir z.B.  $4 + 4 = [0100]_2 + [0100]_2 \stackrel{?}{=} [1000]_2 = -8$ . Bei der Summe zweier negativer Zahlen darf die Summe nicht kleiner als  $-8$  werden, z.B.  $-1 - 1 = [1111]_2 + [1111]_2 = [1110]_2 = -2$ , aber  $-4 - 5 = [1100]_2 + [1011]_2 \stackrel{?}{=} [0111]_2 = 7$ .  $\square$

**3.3 Untergruppen und der Satz von Lagrange** Wir hatten  $\mathbb{G} = (G, e, \circ)$  eine Gruppe (siehe Abschnitt 1.5) genannt, wenn die zweistellige Operation  $\circ$  assoziativ ist, das neutrale Element  $e$  besitzt und es zu jedem  $x$  ein  $x^{-1}$  gibt mit  $x \circ x^{-1} = x^{-1} \circ x = e$ .

$U \subset G$  heißt *Untergruppe* von  $G$ , wenn  $(U, e, \circ)$  ebenfalls eine Gruppe ist mit der gleichen Operation „ $\circ$ “ eingeschränkt auf  $U \times U$ . In diesem Fall schreiben wir  $U \leq G$  und, falls  $U \neq G$ ,  $U < G$ .

**Satz 3.7 (Untergruppenkriterium)**  $(G, e, \circ)$  sei eine Gruppe.  $U \subset G$  ist genau dann eine Untergruppe von  $G$ , wenn

- (a)  $U \neq \emptyset$ ,
- (b) Mit  $x, y \in U$  ist auch  $x \circ y \in U$ .
- (c) Zu jedem  $x \in U$  existiert  $x^{-1} \in U$ .

*Beweis:* Eine Untergruppe erfüllt (a),(b),(c). Wegen (a) gibt es ein  $x \in U$ , das nach (c) ein inverses Element  $x^{-1} \in U$  besitzt. Nach (b) ist dann auch  $x \circ x^{-1} = e \in U$ . Das Assoziativgesetz gilt in  $U$ , weil es in  $G$  gilt.  $\square$

$U = \{e\}$  und  $U = G$  sind immer Untergruppen einer Gruppe  $G$ , man nennt sie die *trivialen Untergruppen*. Weitere Beispiele:

- Die ganzen Zahlen sind mit der üblichen Addition eine Untergruppe der rationalen Zahlen. Die geraden Zahlen sind wiederum eine Untergruppe der ganzen Zahlen.
- Die Menge der Permutationen von  $A_n = \{1, 2, \dots, n\}$  mit  $p(1) = 1$  ist eine Untergruppe, die die gleiche Struktur wie die Permutationen der Menge  $A_{n-1}$  besitzt.

**Satz 3.8 (Satz von Lagrange)** Sei  $G$  eine endliche Gruppe. Ist  $U$  eine Untergruppe von  $G$ , so ist ihre Kardinalität  $|U|$  ein Teiler von  $|G|$ .

*Beweis:* Sei  $U$  Untergruppe der endlichen Gruppe  $G$ . Für jedes  $x \in G$  betrachten wir die Nebenklasse

$$xU = \{xy : y \in U\}.$$

Ist  $xy_1 = xy_2$  für  $y_1, y_2 \in U$ , so folgt  $y_1 = y_2$ . Damit sind alle Nebenklassen gleich groß und haben  $|U|$  viele Elemente. Haben zwei Nebenklassen  $x_1U, x_2U$  ein Element  $x_1y_1 = x_2y_2$  gemeinsam, so sind die Nebenklassen gleich wegen

$$x_1U = x_1(y_1U) = (x_1y_1)U = x_2y_2U = x_2U.$$

Wegen  $x = xe \in xU$  kommt jedes  $x \in G$  in einer Nebenklasse vor. Daher unterteilen die Nebenklassen die Menge  $G$  in endlich viele disjunkte Teilmengen mit  $|U|$  Elementen, womit  $|G|$  ein ganzzahliges Vielfaches von  $|U|$  sein muss.  $\square$

**3.4 Restklassenkörper und der Satz von Wilson** Wir hatten in Abschnitt 1.6  $\mathbb{K} = (K, 0, 1, +, \cdot)$  Körper genannt, wenn  $(K, 0, +)$  und  $(K \setminus 0, 1, \cdot)$  abelsche Gruppen sind und das Distributivgesetz  $a \cdot (b + c) = a \cdot b + a \cdot c$  gilt. Das inverse Element von  $a$  bezüglich der Addition schreiben wir als  $-a$ , das der Multiplikation als  $a^{-1}$ . Üblicherweise verwendet man  $a - b$  statt  $a + (-b)$  und  $ab$  statt  $a \cdot b$ . Weiter gilt die bekannte Regel „Punktrechnung geht vor Strichrechnung“.

Hieraus lassen sich alle Rechenregeln ableiten, die wir von den reellen Zahlen kennen:

**Satz 3.9** Sei  $(K, 0, 1, +, \cdot)$  ein Körper. Dann gilt:

- (a) Die neutralen Elemente der Addition und der Multiplikation sind eindeutig bestimmt.
- (b) Das inverse Element  $-a$  der Addition und das inverse Element  $a^{-1}$ ,  $a \neq 0$ , der Multiplikation sind eindeutig bestimmt.
- (c) Es gilt  $a \cdot 0 = 0$ ,  $(-1)a = -a$ ,  $(-a)b = -ab$ .
- (d) Ist  $a \neq 0$ , so folgt aus  $ab = ac$ , dass  $b = c$ .
- (e) Ein Körper ist nullteilerfrei, d.h. aus  $ab = 0$  folgt  $a = 0$  oder  $b = 0$ .

*Beweis:* (a) und (b) folgen aus Satz 1.6.

(c) Aus  $a0 = a(0+0) = a0 + a0$  folgt  $a0 = 0$ . Aus  $0 = 0a = (1 + (-1))a = a + (-1)a$  folgt  $(-1)a = -a$ . Mit  $(-1)a = -a$  folgt  $(-a)b = (-1)ab = (-1)(ab) = -ab$ .

(d) Dies ist wieder Satz 1.6.

(e) Ist  $ab = 0$  und  $b \neq 0$ , so  $a = abb^{-1} = 0b^{-1} = 0$  wegen (c).  $\square$

Sei  $n > 1$  eine natürliche Zahl. Dann ist die auf  $\mathbb{Z} \times \mathbb{Z}$  erklärte Relation  $a \equiv b \pmod{n}$  eine Äquivalenzrelation. Denn sie ist reflexiv und symmetrisch sowie transitiv wegen

$$a \equiv b \pmod{n}, b \equiv c \pmod{n} \Rightarrow a = b + qm, b = c + q'm \Rightarrow a = c + (q + q')m.$$

Zwei ganze Zahlen sind daher äquivalent, wenn sie bei der Division durch  $n$  den gleichen Rest modulo  $n$  besitzen. Die zugehörigen Äquivalenzklassen besitzen daher die natürlichen Vertreter  $0, 1, \dots, n-1$ . Die Menge

$$\mathbb{Z}_n = \{\bar{0}, \bar{1}, \dots, \bar{n-1}\}$$

bildet eine Partition von  $\mathbb{Z}$ .

Auf  $\mathbb{Z}_n$  können wir die Operationen

$$\bar{a} + \bar{b} = \overline{a+b}, \quad \bar{a} \cdot \bar{b} = \overline{a \cdot b}$$

definieren. Wir beweisen die Korrektheit dieser Definitionen, also die Unabhängigkeit von den Vertretern der jeweiligen Äquivalenzklasse. Ist  $a' \in \bar{a}$ ,  $b' \in \bar{b}$ , so  $a' = a + pn$ ,  $b' = b + qn$ . Dann

$$a' + b' = a + b + (p+q)n \in \overline{a+b}, \quad a' \cdot b' = ab + aqn + bpn + pqn^2 \in \overline{a \cdot b}.$$

Alternativ wird auch

$$\mathbb{Z}_n = \{0, 1, \dots, n-1\}$$

geschrieben. Man addiert und multipliziert diese Zahlen „normal“ in  $\mathbb{N}_0$  und ordnet das Ergebnis der zugehörigen Äquivalenzklasse beziehungsweise ihrem Vertreter in  $\mathbb{Z}_n$  zu. Um nicht in Konfusion mit den üblichen Operationen zu kommen, schreiben wir dann  $+_n$  und  $\cdot_n$  für die so definierten Operationen. Beispielsweise gilt in  $\mathbb{Z}_4$   $2 \cdot 3 = 6 \equiv 2 \pmod{4}$ , daher  $2 \cdot_4 3 = 2$ .

Wir erhalten für  $n = 2$  die Tafeln

$+_2$	0	1		$\cdot_2$	0	1	
0	0	1		0	0	0	
1	1	0		1	0	1	

Für  $n = 4$ :

$+_4$	0	1	2	3	$\cdot_4$	0	1	2	3
0	0	1	2	3	0	0	0	0	0
1	1	2	3	0	1	0	1	2	3
2	2	3	0	1	2	0	2	0	2
3	3	0	1	2	3	0	3	2	1

Welche algebraischen Eigenschaften haben die so definierten Operationen? Zunächst ist klar, dass beide Operationen assoziativ und kommutativ sind. Ferner ist 0 neutral bezüglich der Addition. Zu  $a \in \mathbb{Z}_n$  ist  $n-a$  das inverse Element bezüglich der Addition, denn es gilt  $a+(n-a) = n \equiv 0 \pmod{n}$ . Damit ist  $(\mathbb{Z}_n, 0, +)$  eine kommutative Gruppe. Das Distributivgesetz wird von der Rechnung mit ganzen Zahlen geerbt und ist daher ebenfalls gültig. 1 ist neutrales Element der Multiplikation, was  $\mathbb{Z}_n$  zu einem kommutativen Ring macht (siehe Abschnitt 1.6).

Wie die Tafel oben rechts zeigt, gibt es für die 2 bei  $n = 4$  kein inverses Element. Allgemein ist für zusammengesetztes  $n = kl$  die Struktur kein Körper wegen  $k \cdot_n l = n \equiv 0 \pmod{n}$ , sie ist damit nicht nullteilerfrei.

Bei Primzahlen  $p$  haben wir dagegen:

**Satz 3.10** *Ist  $p$  eine Primzahl, so ist  $\mathbb{Z}_p$  zusammen mit den Operationen  $+_p$  und  $\cdot_p$  ein Körper, der Restklassenkörper modulo  $p$  genannt wird. Für  $a \neq 0$  gilt  $-a = p - a$  sowie  $a^{-1} \equiv a^{p-2} \pmod{p}$ . Genau die Elemente 1 und  $p - 1$  sind zu sich selbst invers bezüglich der Multiplikation  $\cdot_p$ , alle anderen Elemente  $\neq 0$  lassen sich zu Paaren  $a, a'$ ,  $a \neq a'$ , zusammenfassen mit  $a \cdot_p a' = 1$ .*

*Beweis:* Nach dem kleinen Satz von Fermat 3.2 gilt  $a^{p-1} \equiv 1 \pmod{p}$  für alle  $a \in \{1, \dots, p-1\}$ . Somit  $a \cdot a^{p-2} \equiv 1 \pmod{p}$  und die Restklasse modulo  $p$  von  $a^{p-2}$  ist das inverse Element von  $a$  bezüglich  $\cdot_p$ .

Aus  $a^2 \equiv 1 \pmod{p}$  folgt  $(a-1)(a+1) \equiv 0 \pmod{p}$ , was genau für  $a = 1$  oder  $a = p-1$  erfüllt ist.  $\square$

**Satz 3.11 (Wilson)** *Für jede Primzahl  $p$  gilt*

$$(p-2)! \equiv 1 \pmod{p}, \quad (p-1)! \equiv -1 \pmod{p}.$$

*Beweis:* Es gilt  $(p-2)! = 2 \cdot \dots \cdot (p-2)$ . Nach dem letzten Satz wird dieses Produkt von Paaren mit  $aa' \equiv 1 \pmod{p}$  gebildet, daher  $(p-2)! \equiv 1 \pmod{p}$ . Wir multiplizieren dies mit  $p-1$  und erhalten den zweiten Teil der Behauptung.  $\square$

Es gilt auch die Umkehrung dieses Satzes: Ist  $(p-1)! \equiv -1 \pmod{p}$ , so ist  $p$  eine Primzahl.

**3.5 Geheimcodes** dienen dazu, Nachrichten so zu verschlüsseln, dass sie nur vom Empfänger lesbar gemacht werden können. Wir untersuchen zunächst klassische Verschlüsselungen und behandeln dann die moderne RSA-Technik.

**Die Substitution** besteht darin, jeden Buchstaben eines Textes durch einen anderen zu ersetzen. Im Folgenden verwenden wir kleine Buchstaben für den zu verschlüsselnden Text (=Klartext) und große Buchstaben für die verschlüsselte Nachricht (=Geheimtext). Verwenden wir die Zuordnung

Klartextalphabet: a b c d e f g h i j k l m n o p q r s t u v w x y z

Geheimtextalphabet: J L P A W I Q B C T R Z Y D S K E G F X H U O N V M

so erhalten wir beispielsweise

Klartext: gehen wir aus?

Geheimtext: QWBWD OCG JHF?

Die Anzahl der auf diese Weise erzeugten Geheimcodes ist gleich der Anzahl der Permutationen der 26 Buchstaben, das sind  $26! \sim 4 \cdot 10^{26}$ . Obwohl diese Zahl zu groß ist, um alle Möglichkeiten auch mit Hilfe eines Rechners durchzuprobieren, sind solche Codes leicht zu entschlüsseln, wenn nur der Text genügend lang ist. Man macht sich dabei die Tatsache zu Nutze, dass in jeder Sprache die Buchstaben unterschiedlich häufig vorkommen. Im Deutschen gilt für die Häufigkeiten:

Buchstabe	Häufigkeit in %	Buchstabe	Häufigkeit in %
a	6,51	n	9,78
b	1,89	o	2,51
c	3,06	p	0,79
d	5,08	q	0,02
e	17,40	r	7,00
f	1,66	s	7,27
g	3,01	t	6,15
h	4,76	u	4,35
i	7,55	v	0,67
j	0,27	w	1,89
k	1,21	x	0,03
l	3,44	y	0,04
m	2,53	z	1,13

Neben dem im Deutschen leicht zu identifizierendem e kann man sich an den Wörtern mit drei Buchstaben orientieren: Sie bezeichnen meist einen der Artikel der, die, das, ein.

**Vignère-Verschlüsselungen** Bei der Vignère-Verschlüsselung nimmt man für jeden Buchstaben in Abhängigkeit seiner Position im Klartext einen anderen Schlüssel. Im einfachsten Fall vereinbart man ein Schlüsselwort, beispielsweise LICHT, das wiederholt über den Klartext geschrieben wird.

Schlüsselwort	LICHTLICHTLICHTLICHTL
Klartext	truppenabzugnachosten
Geheimtext	EZWWIPVCISFOEHVSWUAXY

Der Buchstabe des Schlüsselworts gibt an, wieweit der Buchstabe des Klartextes im Alphabet verschoben werden muss. Im obigen Beispiel ist L der 12. Buchstabe des Alphabets und man verschiebt das t des Klartexts um  $12 - 1 = 11$  Positionen nach rechts modulo 26, das ist gerade E. Der nächste Buchstabe r wird wegen des an 9. Position stehenden I um 8 Positionen nach rechts verschoben, das ist Z.

Damit wird jeder Buchstabe auf 5 verschiedene Arten verschlüsselt, eine Häufigkeitsanalyse der Buchstaben ist zur Entschlüsselung nicht mehr möglich. Allerdings kann bei kurzen Schlüsselwörtern eine Häufigkeitsanalyse nach Sequenzen vorgenommen werden wie etwa nach dem häufigsten dreibuchstabigen Wort „die“. Auch nach Verschlüsselung werden die zugehörigen verschlüsselten Sequenzen immer noch häufig sein und führen somit auf das Schlüsselwort.

Man kann die Vignère-Verschlüsselung dahingehend verbessern, dass an Stelle eines Schlüsselwortes ein ganzer Text vereinbart wird, beispielsweise ein Abschnitt eines Romans. In diesem Fall muss der Entschlüssler den Text kennen. Eine moderne Version dieser Technik verwendet einen Zufallsgenerator an Stelle eines Textes. Vor der Verschlüsselung müssen daher nur die Daten des Generators festgelegt werden.

Eine Variante der Vignère-Verschlüsselung wurde von Deutschland mit dem Enigma-Gerät im 2. Weltkrieg verwendet. In der einfachsten Version besteht die Enigma aus einer Tastatur und mindestens drei Rotoren sowie einigen Steckverbindungen, die eine involutorische Permutation der Buchstaben bewirken. Dabei heißt eine Permutation  $p$  *involutoriisch*, wenn  $p^2 = p \circ p = id$ . Wird eine Buchstabentaste gedrückt, so fließt ein Strom durch die Steckverbindungen und Rotoren, der den zugehörigen Buchstaben des Geheimtextes erscheinen lässt. Die Rotoren haben jeweils 26 Positionen, die mit den Buchstaben des Alphabets beschriftet sind und sich nach jeder Eingabe eines

Buchstabens ändern. Zudem fließt der Strom nach Durchlaufen der Rotoren auf eine „Umkehrwalze“, die eine involutorische Permutation  $u$  darstellt. Anschließend fließt der Strom durch die Rotoren und die Steckverbindungen zurück. Insgesamt entsteht bei einem Zustand  $z$  der Rotoren eine Permutation der 26 Buchstaben der Form

$$b_i = R_z(a_i) = r^{-1} \circ p_z^{-1} \circ u \circ p_z \circ r(a_i), \quad i = 1, \dots, 26,$$

wobei  $p_z$  die Permutation bezeichnet, die von den Rotoren herührt und vom aktuellen Zustand  $z$  der Rotoren abhängt.  $r$  ist die Permutation, die aus den Steckverbindungen hervorgeht und sich während der Nachrichtenübermittlung nicht ändert. Wegen  $u^2 = id_{A_{26}}$  gilt

$$(r^{-1} \circ p_z^{-1} \circ u \circ p_z \circ r) \circ (r^{-1} \circ p_z^{-1} \circ u \circ p_z \circ r) = r^{-1} \circ p_z^{-1} \circ u \circ u \circ p_z \circ r = id_{A_{26}}.$$

$R_z$  ist dadurch ebenfalls involutorisch, was den Vorteil hat, dass das Dechiffrieren mit dem selben Gerät erfolgen kann, wenn die Anfangsstellung der Rotoren und der Steckverbindungen bekannt ist. Insgesamt kommt man auf eine Verschlüsselung mit einer Bitlänge von etwa 70 – ein auch für die heutige Zeit kaum knackbarer Code. Diese anscheinend hohe Zahl relativiert sich aber, denn die Rotoren liefen nach einem teilweise bekannten Algorithmus und bestimmte Wörter wie oberkommandoderwehrmacht oder wetterbericht kamen in fast jeder Nachricht vor. Jedenfalls konnten die Alliierten die meiste Zeit alle Funksprüche der Deutschen dechiffrieren, was den 2. Weltkrieg sicherlich abkürzte.

Alle diese Verschlüsselungsmethoden eignen sich nicht für eine moderne Kommunikation zwischen wechselnden Partnern über Handy oder Internet, da zuvor der Schlüssel ausgetauscht werden muss. Dies geschieht unverschlüsselt und kann daher abgehört werden.

**Die RSA-Verschlüsselung** beruht auf zwei Sätzen, die mit den uns zur Verfügung stehenden Methoden leicht bewiesen werden können.

**Satz 3.12 (Existenz der modularen Inversen)** *Sind  $a$  und  $n$  teilerfremde natürliche Zahlen, so gibt es eine ganze Zahl  $b$  mit der Eigenschaft*

$$ab \equiv 1 \pmod{n}.$$

*Beweis:* Nach dem Satz über den größten gemeinsamen Teiler 3.4 gibt es Zahlen  $\alpha, \beta \in \mathbb{Z}$  mit

$$1 = ggT(a, n) = \alpha a + \beta n,$$

also  $\alpha a \equiv 1 \pmod{n}$ . Die Zahl  $b = \alpha$  erfüllt daher die Behauptung.  $\square$

**Satz 3.13** *Seien  $p$  und  $q$  zwei verschiedenen Primzahlen und sei  $a$  teilerfremd zu  $pq$ . Dann gilt*

$$a^{(p-1)(q-1)} \equiv 1 \pmod{pq}.$$

*Beweis:* Mit  $a$  teilerfremd zu  $q$  ist auch  $a^{p-1}$  teilerfremd zu  $q$ . Mit dem kleinen Satz von Fermat 3.2 folgt

$$a^{(p-1)(q-1)} \equiv 1 \pmod{q} \Leftrightarrow a^{(p-1)(q-1)} = kq + 1$$

Auf die gleiche Weise folgt  $a^{(p-1)(q-1)} = lp + 1$ , daher  $kq = lp$ . Also ist  $kq = lp$  sowohl durch  $q$  als auch durch  $p$  teilbar. Somit  $kq = lp = mpq$  und  $a^{(p-1)(q-1)} = mpq + 1$  oder  $a^{(p-1)(q-1)} \equiv 1 \pmod{pq}$ .  $\square$

Die RSA-Verschlüsselung ist asymmetrisch. Wer mir eine verschlüsselte Nachricht senden will, verschlüsselt sie mit einem öffentlichen Schlüssel, den ich beispielsweise im Internet zur Verfügung stelle. Das Entschlüsseln geschieht mit einer geheimen Zahl, die nicht versendet werden muss und auch dem Sender der Nachricht unbekannt ist. Genauer geht man folgendermaßen vor:

- Es werden zwei verschiedene Primzahlen  $p$  und  $q$  gewählt und  $n = pq$  berechnet.
- Mit einer weiteren frei gewählten Zahl  $e$ , die teilerfremd zu  $(p - 1)(q - 1)$  ist, wird  $d$  so berechnet, dass

$$ed \equiv 1 \pmod{(p - 1)(q - 1)} \text{ oder } ed = 1 + k(p - 1)(q - 1).$$

Dies ist die modulare Inverse aus Satz 3.12.  $d$  kann mit Hilfe des erweiterten euklidischen Algorithmus aus Satz 3.4 effektiv berechnet werden.

- Öffentlicher Schlüssel:  $e$  und  $n$ .
- Privater Schlüssel:  $d$  (kann größer als Null gewählt werden).
- $p, q$  und  $(p - 1)(q - 1)$  werden nicht mehr benötigt und sollten sicherheitshalber vernichtet werden.

Nun gibt man die Zahlen  $n$  und  $e$  öffentlich bekannt. Die „geheime“ Zahl  $d$  wird nicht bekannt gegeben. Will jemand eine Nachricht  $m < n$  an uns senden, so übermittelt er

$$c \equiv m^e \pmod{n}.$$

Die Zahl  $c$  wird entschlüsselt durch

$$m' \equiv c^d \pmod{n}.$$

**Satz 3.14 (Korrektheit der RSA-Verschlüsselung)** *Mit obigem Verschlüsselungsverfahren gilt  $m' = m$ .*

*Beweis:* Aus  $a \equiv r \pmod{n}$  folgt  $a^d \equiv r^d \pmod{n}$ . Für  $a = m^e$  ergibt das

$$m^e \equiv c \pmod{n} \Leftrightarrow m^{ed} \equiv c^d \equiv m' \pmod{n}.$$

Wir müssen daher zeigen, dass  $m^{ed} \equiv m \pmod{n}$  gilt. Nach Definition von  $e$  und  $d$  ist  $ed = 1 + k(p - 1)(q - 1)$ . Daraus folgt

$$m^{ed} = m^{1+k(p-1)(q-1)} \equiv m \cdot m^{k(p-1)(q-1)} \pmod{n}.$$

Wegen  $m < n$  gilt  $m \equiv m \pmod{n}$  und wegen Satz 3.13  $m^{k(p-1)(q-1)} \equiv 1 \pmod{n}$ . Bilden wir das Produkt dieser Kongruenzen, so

$$m \cdot m^{k(p-1)(q-1)} \equiv m \cdot 1 \equiv m \pmod{n}.$$

Somit ergibt sich nach dem Dechiffrieren mit dem privaten Schlüssel tatsächlich wieder  $m$ .  $\square$

Im Gegensatz zu den in den vorigen Abschnitten beschriebenen Verfahren werden keine Schlüssel ausgetauscht. Jeder kann mir eine verschlüsselte Nachricht senden, wenn er sich die von mir bekannt gegebenen Zahlen  $n$  und  $e$  verschafft. Das Verfahren ist daher abhörsicher.

Die RSA-Verschlüsselung beruht auf dem Glauben, dass aus den öffentlichen Zahlen  $n$  und  $e$  der Schlüssel  $d$  nicht in vernünftiger Zeit rekonstruiert werden kann, wenn  $n$  genügend groß gewählt wurde. In der Tat kann  $d$  nur über die Faktoren in  $n = pq$  bestimmt werden. Man ist sich ziemlich sicher, dass diese Faktorisierung nicht „schnell“ gelingt.

## 4 Zahlen

**4.1 Körper – Potenzen und geometrische Summenformel** Wir hatten in Abschnitt 2.5 die Potenzen  $a^n$  in einem kommutativen Ring definiert. In einem Körper  $\mathbb{K}$  können wir für  $a \neq 0$  auch negative Potenzen  $a^{-n} = (a^n)^{-1}$  erklären. Die in Abschnitt 2.5 angegebenen Potenzgesetze

$$a^{m+n} = a^m \cdot a^n, \quad a^n b^n = (ab)^n, \quad (a^m)^n = a^{mn},$$

gelten nun auch für  $m, n \in \mathbb{Z}$ , sofern  $a, b \neq 0$ .

**Satz 4.1** In einem Körper  $\mathbb{K}$  gilt die geometrische Summenformel

$$\sum_{i=0}^n q^i = \frac{1 - q^{n+1}}{1 - q}, \quad \text{für alle } q \neq 1.$$

*Beweis:* Man verwendet den „Teleskopeffekt“

$$\sum_{i=0}^n q^i (1 - q) = \sum_{i=0}^n q^i - \sum_{i=0}^n q^{i+1} = 1 - q^{n+1}.$$

□

**4.2 Angeordnete Körper und die rationalen Zahlen** Eine Struktur  $\mathbb{K} = (K, 0, 1, +, \cdot, \leq)$  heißt *angeordneter Körper*, wenn  $(K, 0, 1, +, \cdot)$  ein Körper und  $\leq$  eine totale Ordnung ist, die zudem mit den beiden algebraischen Operationen verträglich ist.

Wir hatten eine Relation  $\leq$  eine totale Ordnung genannt, wenn

- (O1)  $a \leq a$
- (O2)  $a \leq b$  und  $b \leq c \Rightarrow a \leq c$ .
- (O3) Für  $a, b$  gilt genau eine der folgenden Relationen

$$a < b, \quad a > b, \quad a = b.$$

Hier bedeutet  $a < b$ , dass  $a \leq b$  und  $a \neq b$  erfüllt ist.

Unter der Verträglichkeit von  $\leq$  verstehen wir, dass die beiden *Anordnungsaxiome*

- (A1)  $a \leq b \Rightarrow a + c \leq b + c$ ,
- (A2)  $a \leq b$  und  $c \geq 0 \Rightarrow ac \leq bc$ ,

erfüllt sind.

**Satz 4.2** Im angeordneten Körper gelten die folgenden Rechenregeln

- (a)  $a \leq b \Leftrightarrow -b \leq -a$ .
- (b)  $ab \geq 0 \Leftrightarrow a, b \geq 0$  oder  $a, b \leq 0$ .

Insbesondere ist  $a^2 > 0$  für  $a \neq 0$  sowie  $1 = 1^2 > 0$ .

- (c) Ist  $a \leq b$ , so gilt  $ac \geq bc$  für  $c \leq 0$ .
- (d) (a)-(c) bleiben richtig, wenn man  $\leq$  durch  $<$  und  $\geq$  durch  $>$  ersetzt.

*Beweis:* (a) Auf  $a \leq b$  addieren wir auf beiden Seiten  $-a - b$ .

(b) Für  $a, b \geq 0$  folgt aus (A2)  $ab \geq 0$ . Ferner folgt für  $a \leq 0$  aus (a), dass  $-a \geq 0$ . Für  $a, b \leq 0$  daher  $0 \leq (-a)(-b) = ab$ . Ist  $a \leq 0, b \geq 0$  (oder umgekehrt), so folgt mit analoger Argumentation  $ab \leq 0$ .

(c) Auch hier ist wieder nach (A2)  $a(-c) \leq b(-c)$ , also  $-ac \leq -bc$  und nach (a)  $bc \leq ac$ .  $\square$

Kombinieren wir  $1 > 0$  mit (A1), so erhalten wir  $0 < 1 < 1 + 1 < \dots$ . Wir können diese Folge mit den natürlichen Zahlen identifizieren, also  $\mathbb{N} \subset \mathbb{K}$  für jeden angeordneten Körper  $\mathbb{K}$ . Mit  $n \in K$  ist auch  $-n \in K$ , womit auch die ganzen Zahlen  $\mathbb{Z}$  in  $K$  sind. Ferner ist  $m/n \in K$  für  $m \in \mathbb{Z}$  und  $n \in \mathbb{Z} \setminus \{0\}$ . Damit finden wir auch den aus ganzzahligen Brüchen bestehenden Körper  $\mathbb{Q}$  der rationalen Zahlen in jedem angeordneten Körper wieder. Gleichzeitig ist  $\mathbb{Q}$  der minimale angeordnete Körper. Insbesondere können alle endlichen Körper nicht angeordnet werden.

Wir können also die Grundrechenarten innerhalb von  $\mathbb{Q}$  uneingeschränkt ausführen, trotzdem lässt  $\mathbb{Q}$  noch einige Wünsche offen:

**Satz 4.3** *Es gibt keine rationale Zahl  $r$  mit  $r^2 = 2$ .*

*Beweis:* Angenommen, für teilerfremde  $m, n \in \mathbb{N}$  wäre  $(\frac{m}{n})^2 = 2$ . Dann folgt  $m^2 = 2n^2$ . Da die rechte Seite durch 2 teilbar ist, muss auch die linke durch 2 teilbar sein, also  $m = 2k$  und  $4k^2 = 2n^2$ . In dieser Identität kann durch 2 geteilt werden,  $2k^2 = n^2$ . Mit der gleichen Argumentation wie vorher folgt, dass auch  $n$  durch 2 teilbar ist. Da  $m$  und  $n$  durch 2 teilbar sind, erhalten wir einen Widerspruch zur Annahme, dass  $m$  und  $n$  teilerfremd sind.  $\square$

Man kann sich der „Lückenhaftigkeit“ der rationalen Zahlen auch durch Dezimalzahlen der Form

$$(4.1) \quad n + \sum_{k=1}^{\infty} a_k 10^{-k}, \quad n \in \mathbb{N}_0 \text{ und } a_k \in \{0, 1, \dots, 9\},$$

nähern. Der Einfachheit halber betrachten wir hier und im Folgenden nur nichtnegative Zahlen.

**Satz 4.4** *Eine Zahl ist genau dann rational, wenn ihre Dezimalentwicklung periodisch ist, wenn der Dezimalbruch also von der Form*

$$0, a_1 \dots a_k \overline{b_1 \dots b_l}$$

ist.

*Beweis:* Sei

$$x = 0, a_1 \dots a_k \overline{b_1 \dots b_l}.$$

Dann ist

$$10^k x = a_1 \dots a_k, \overline{b_1 \dots b_l}, \quad 10^{k+l} x = a_1 \dots a_k b_1 \dots b_l, \overline{b_1 \dots b_l}$$

und

$$(10^{k+l} - 10^k)x = a_1 \dots a_k b_1 \dots b_l - a_1 \dots a_k$$

Damit ist  $x$  rational. Insbesondere ist  $0, \overline{9} = 1$ .

Für die umgekehrte Richtung erinnern wir an den Divisionsalgorithmus für die Division zweier natürlicher Zahlen. In jedem Teilschritt von  $m : n$  führen wir eine Division mit Rest aus:  $m' \text{ div } n = a$  mit Rest  $r \in \{0, 1, \dots, n-1\}$ . Entweder wir gelangen irgendwann zum Rest  $r = 0$ , dann ist die Zahl nichtperiodisch, oder wir kommen zu einem Rest, den wir bereits hatten. Von da an ist die Dezimalzahl periodisch.  $\square$

Der Beweis lässt sich in jedem Stellenwertsystem mit Grundzahl  $g > 1$  auf die gleiche Weise durchführen. Allerdings hängt es von  $g$  ab, ob  $m : n$  eine endliche  $g$ -adische Darstellung besitzt oder nicht.

Für eine rationale Zahl  $x$  bezeichnen wir mit  $\lfloor x \rfloor$  die größte ganze Zahl  $\leq x$ , z.B.  $\lfloor 1,3 \rfloor = 1$ ,  $\lfloor -1,3 \rfloor = -2$ . Für  $0 \leq x < 1$  berechnen wir die Ziffern in  $x = 0, a_1 a_2 \dots_g$  durch folgenden Algorithmus. Setze  $x_0 = x$  und bestimme für  $k = 0, 1, \dots$

$$y_{k+1} = g \cdot x_k, \quad a_{k+1} = \lfloor y_{k+1} \rfloor, \quad x_{k+1} = y_{k+1} - a_{k+1}.$$

**Beispiel 4.5** Wir stellen  $x = 0,1$  im Dreiersystem dar und schreiben die Ziffern in Klammern

$$0,1 \rightarrow 0,3(0) \rightarrow 0,9(0) \rightarrow 2,7(2) \rightarrow 0,7 \rightarrow 2,1(2) \rightarrow 0,1,$$

daher  $0,1 = 0,\overline{022}_3$ .  $\square$

**4.3 Reelle Zahlen** Die reellen Zahlen bestehen aus allen Dezimalbrüchen der Form (4.1). Im Gegensatz zu den rationalen Zahlen ist die Definition der reellen Zahlen mit einem Grenzübergang verbunden, der hier weiter erläutert wird. Dem reellen Dezimalbruch  $x = 0, a_1 a_2 \dots$  können wir rationale Zahlen

$$x_n = \sum_{k=1}^n a_k 10^{-k} = 0, a_1 a_2 \dots a_n, \quad y_n = x_n + 10^{-n}$$

zuordnen. Anschaulich liegt die reelle Zahl „zwischen“  $x_n$  und  $y_n$ . Die Paare  $(x_n, y_n)_{n \in \mathbb{N}}$  bilden eine *Intervallschachtelung*: Die Folge  $(x_n)_{n \in \mathbb{N}}$  ist steigend, die Folge  $(y_n)_{n \in \mathbb{N}}$  ist fallend und für die Differenz gilt  $y_n - x_n \leq 10^{-n}$ .

Die reellen Zahlen sind unabhängig von der Grundzahl  $g$ . Wie in Beispiel 4.5 können wir jeden Dezimalbruch in einen  $g$ -adischen Bruch umwandeln.

Um noch einen anderen Zugang zu den reellen Zahlen anzugeben, wiederholen wir einige Begriffe aus Abschnitt 1.3. Eine Teilmenge  $A$  in einer angeordneten Menge  $K$  heißt *nach unten (oben) beschränkt*, wenn es ein  $\xi \in K$  gibt mit  $\xi \leq a$  ( $\xi \geq a$ ) für alle  $a \in A$ .  $\xi$  heißt in diesem Fall *untere (obere) Schranke* von  $A$ . Ist die Menge  $A$  sowohl nach unten als auch nach oben beschränkt, so heißt  $A$  *beschränkt*.

$\xi$  heißt *größte untere Schranke* von  $A$  oder *Infimum* von  $A$ , wenn für jede andere untere Schranke  $\xi'$  gilt  $\xi' \leq \xi$ . Entsprechend heißt  $\xi$  *kleinste obere Schranke* oder *Supremum* von  $A$ , wenn für jede andere obere Schranke  $\xi'$  gilt  $\xi' \geq \xi$ . In diesen Fällen schreiben wir  $\xi = \inf A$  beziehungsweise  $\xi = \sup A$ . Gehört ein Infimum (Supremum) selber zu  $A$ , so heißt es *Minimum (Maximum)*.

Infimum und Supremum einer Menge  $A$  sind eindeutig bestimmt, denn wären beispielsweise  $\xi$  und  $\eta$  Infima, so würde aufgrund der Definition sowohl  $\xi \leq \eta$  als auch  $\eta \leq \xi$  gelten, was wegen des Trichotomiegesetzes  $\xi = \eta$  impliziert.

**Beispiel 4.6** Wir bestimmen, sofern vorhanden, Supremum, Infimum, Maximum und Minimum der Menge

$$M = \{2^{-m} + n^{-1} : m, n \in \mathbb{N}\}.$$

Offenbar ist 0 eine untere Schranke. Da sowohl  $2^{-m}$  als auch  $n^{-1}$  für große  $m, n$  beliebig klein werden, ist 0 auch die größte untere Schranke. Diese wird aber in der Menge nicht angenommen, ein Minimum gibt es daher nicht. Das maximale Element erhält man für  $m = n = 1$ , womit  $3/2$  das Maximum von  $M$  ist.  $\square$

Sei  $\mathbb{K}$  ein angeordneter Körper. Ist zusätzlich noch

(V) Vollständigkeitsaxiom: Jede nichtleere, nach oben beschränkte Menge besitzt ein Supremum. erfüllt, so heißt  $\mathbb{K}$  der *Körper der reellen Zahlen* und wird mit  $\mathbb{R}$  bezeichnet.

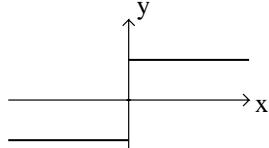
Der Leser hat vielleicht bemerkt, dass im Gegensatz zu den bisherigen Definitionen die letzte auch eine Aussage enthält, dass nämlich der Körper  $\mathbb{R}$  *eindeutig* durch die Axiome festgelegt wird.

Im Körper  $\mathbb{Q}$  gilt das Vollständigkeitsaxiom nicht, denn wir hatten bereits in Satz 4.3 gesehen, dass  $x^2 = 2$  in  $\mathbb{Q}$  keine Lösung besitzt. Damit hat die nach oben beschränkte Menge  $M = \{x : x^2 < 2\}$  kein Supremum in  $\mathbb{Q}$ .  $\mathbb{R}$  enthält neue Zahlen wie eben  $\sqrt{2}$ , die wir *irrationale Zahlen* nennen.

**Satz 4.7** Ist  $a \geq 0$  und  $n \in \mathbb{N}$ , so besitzt die Gleichung  $x^n = a$  genau eine reelle Lösung  $x$  mit  $x \geq 0$ . Diese Lösung wird mit  $\sqrt[n]{a}$  bezeichnet und die  $n$ -te Wurzel aus  $a$  genannt.

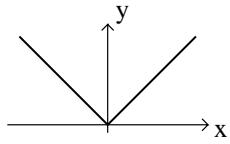
Der Beweis wird später nachgetragen. Man beachte, dass wir die Wurzel nur aus nichtnegativen Zahlen ziehen und dass diese Wurzel eine nichtnegative Zahl ist, obwohl  $(-2)^2 = 4$  und  $(-3)^3 = -27$ .

Zu einer reellen Zahl heißt



$$\operatorname{sgn} a = \begin{cases} 1 & \text{für } a > 0 \\ 0 & \text{für } a = 0 \\ -1 & \text{für } a < 0 \end{cases}$$

das Vorzeichen (=Signum) von  $a$ . Ferner heißt  $|a| = a \operatorname{sgn} a$  oder



$$|a| = \begin{cases} a & \text{für } a \geq 0 \\ -a & \text{für } a < 0 \end{cases}$$

der Betrag von  $a$ .

Der Betrag hat die drei grundlegenden Eigenschaften

- (a) Für  $a \neq 0$  ist  $|a| > 0$ .
- (b) Es gilt  $|ab| = |a||b|$ .
- (c)  $|a + b| \leq |a| + |b|$  (Dreiecksungleichung).

Die Beweise von (a) und (b) folgen direkt aus der Definition. (c) ergibt sich aus  $\pm a \leq |a|$ ,  $\pm b \leq |b|$  und daher

$$a + b \leq |a| + |b|, \quad -(a + b) \leq |a| + |b|.$$

Seien  $a, b \in \mathbb{R}$  mit  $a < b$ . Man nennt

$[a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$	abgeschlossenes Intervall
$(a, b) = \{x \in \mathbb{R} : a < x < b\}$	offenes Intervall
$[a, b) = \{x \in \mathbb{R} : a \leq x < b\}$	(nach rechts) halboffenes Intervall
$(a, b] = \{x \in \mathbb{R} : a < x \leq b\}$	(nach links) halboffenes Intervall

Unbeschränkte Intervalle werden mit Hilfe der Symbole  $\infty$  und  $-\infty$  definiert. Für  $a \in \mathbb{R}$  heißen die Mengen

$$(-\infty, a) = \{x \in \mathbb{R} : x < a\}, \quad (a, \infty) = \{x \in \mathbb{R} : x > a\}$$

offene und die Mengen

$$(-\infty, a] = \{x \in \mathbb{R} : x \leq a\}, \quad [a, \infty) = \{x \in \mathbb{R} : x \geq a\}$$

abgeschlossene Intervalle. Die Menge  $\mathbb{R}$  wird auch als Intervall  $(-\infty, \infty)$  angesehen und sowohl als offen als auch als abgeschlossen definiert.

Die positiven reellen Zahlen werden mit  $\mathbb{R}_+$ , die negativen mit  $\mathbb{R}_-$  bezeichnet. Die Mengen  $\mathbb{R}_+ = (0, \infty)$  und  $\mathbb{R}_- = (-\infty, 0)$  sind demnach offene Intervalle.

Das Symbol  $\infty$  für unendlich wird in der Analysis häufig benutzt, allerdings immer in einem genau präzisierten Sinn. Das Intervall  $(a, \infty)$  ist nur die Kurzbezeichnung für die angegebene Punktmenge, weitreichende philosophische Gedanken sollte man sich nicht machen.

#### 4.4 Das Rechnen mit reellen Zahlen

**Beispiele 4.8** (i) Wir bestimmen Infimum und Supremum der Menge

$$M = \left\{ x + \frac{1}{x} : \frac{1}{2} < x \leq 2 \right\}$$

und untersuchen, ob  $M$  Minimum und Maximum besitzt. Aus  $(x - 1)^2 \geq 0$  folgt für  $x > 0$

$$x^2 - 2x + 1 \geq 0 \Leftrightarrow x + \frac{1}{x} \geq 2.$$

Da der Wert 2 für  $x = 1$  angenommen wird, ist das Infimum 2 auch Minimum. Für  $x \in [\frac{1}{2}, 2]$  gilt

$$\left| x - \frac{5}{4} \right| \leq \frac{3}{4}$$

und nach Quadrieren und Division durch  $x$  folgt

$$x + \frac{1}{x} \leq \frac{5}{2}.$$

Damit ist das Supremum  $\frac{5}{2}$ , das für  $x = 2$  angenommen wird.  $\frac{5}{2} \in M$  ist daher auch das Maximum von  $M$ .

(ii) Wir bestimmen die Menge

$$M = \left\{ x \in \mathbb{R} : \frac{x+4}{x-2} < x \right\}.$$

Um den Bruch umzuformen, unterscheiden wir die Fälle  $x > 2$  und  $x < 2$ ,

$$M = M_1 \cup M_2, \quad M_1 = \{x > 2 : 0 < x^2 - 3x - 4\}, \quad M_2 = \{x < 2 : 0 > x^2 - 3x - 4\}.$$

Mit  $x^2 - 3x - 4 = (x+1)(x-4)$  folgt dann

$$M_1 = \{x > 4\}, \quad M_2 = \{-1 < x < 2\}, \quad M = (4, \infty) \cup (-1, 2).$$

(iii) Für  $a, b > 0$  wollen wir die Ungleichung

$$\frac{a}{\sqrt{b}} + \frac{b}{\sqrt{a}} \geq \sqrt{a} + \sqrt{b}$$

beweisen. Bevor man solche Ungleichungen in seitensweisen Rechnungen umformt, sollte man sie zu vereinfachen suchen. In diesem Fall ist es naheliegend, beide Seiten durch  $\sqrt{b}$  zu teilen, was zu einer äquivalenten Ungleichung in  $x = \sqrt{a}/\sqrt{b}$  führt,

$$x^2 + \frac{1}{x} \geq x + 1, \quad x > 0.$$

Auf diese Weise sind wir sowohl eine Variable als auch die Wurzel losgeworden. In dieser Gleichung dürfen wir sogar  $x \geq 1$  annehmen, denn andernfalls teilen wir die Ausgangsungleichung durch  $\sqrt{a}$ . Wir setzen nun  $x = 1 + y$  mit  $y \geq 0$  und erhalten mit Hilfe der binomischen Formel

$$x^3 - x^2 - x + 1 = (y^3 + 3y^2 + 3y + 1) - (y^2 + 2y + 1) - (y + 1) + 1 = y^3 + 2y^2 + y \geq 0.$$

Damit ist die behauptete Ungleichung bewiesen.

(iv) Für  $a, b \in \mathbb{R}$  und  $\varepsilon > 0$  gilt

$$\left( \sqrt{\varepsilon}a \pm \frac{1}{\sqrt{\varepsilon}}b \right)^2 \geq 0$$

und es folgt die *Youngsche Ungleichung*

$$|ab| \leq \frac{\varepsilon}{2}a^2 + \frac{1}{2\varepsilon}b^2.$$

□

## 4.5 Komplexe Zahlen

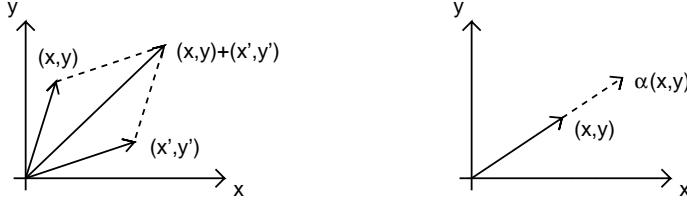
Sei

$$\mathbb{R}^2 = \{(x, y) : x, y \in \mathbb{R}\}.$$

$\mathbb{R}^2$  können wir als Punkte in der Ebene oder als Vektoren mit Komponenten  $x$  und  $y$  auffassen. Für  $(x, y), (x', y') \in \mathbb{R}^2$  definieren wir die Summe durch

$$(x, y) + (x', y') = (x + x', y + y').$$

Dies ist die übliche Addition zweier ebener Vektoren: Wir verschieben  $(x', y')$  so, dass sein Fußpunkt auf dem Endpunkt von  $(x, y)$  steht, der Endpunkt des so verschobenen Vektors zeigt dann auf den Endpunkt der Summe (siehe Abbildung links).



Für  $\alpha \in \mathbb{R}$  und  $(x, y) \in \mathbb{R}^2$  ist die *Skalarmultiplikation* definiert durch

$$\alpha(x, y) = (\alpha x, \alpha y).$$

Für  $\alpha \geq 0$  ist der Ergebnisvektor die Verlängerung oder Verkürzung um das  $\alpha$ -fache (siehe Abbildung rechts). Bei  $\alpha < 0$  kehrt sich zusätzlich die Orientierung um.

Nach dem Satz des Pythagoras ist die Länge eines Vektors  $(x, y)$

$$|(x, y)| = \sqrt{x^2 + y^2}.$$

Bis hierin haben wir nur die üblichen Operationen für Vektoren definiert, was in anderen Raumdimensionen genauso geht. Die Vektoren bilden mit der Addition und dem Vektor  $(0, 0)$  eine abelsche Gruppe, die Inverse von  $(x, y)$  ist  $(-x, -y)$ .

Mit Hilfe der Multiplikation

$$(x, y) \cdot (x', y') = (xx' - yy', xy' + yx')$$

kann man, wie wir gleich sehen werden, auf den ebenen Vektoren einen Körper definieren (siehe Abschnitt 3.4). Diese etwas geheimnisvolle Definition ist diesem Ziel geschuldet: Im Wesentlichen gibt es nur diese eine Möglichkeit, aus den Vektoren einen Körper zu machen und sie funktioniert nur im ebenen Fall. Das Element  $(1, 0)$  ist neutral bezüglich dieser Multiplikation und die Inverse von  $(x, y) \neq (0, 0)$  ist

$$(x, y)^{-1} = \left( \frac{x}{x^2 + y^2}, \frac{-y}{x^2 + y^2} \right)$$

wegen

$$\begin{aligned} (x, y) \cdot (x, y)^{-1} &= (x, y) \left( \frac{x}{x^2 + y^2}, \frac{-y}{x^2 + y^2} \right) \\ &= \left( \frac{x^2}{x^2 + y^2} - \frac{-y^2}{x^2 + y^2}, \frac{-xy}{x^2 + y^2} + \frac{xy}{x^2 + y^2} \right) = (1, 0). \end{aligned}$$

Da die übrigen Körperaxiome sich leicht nachrechnen lassen, ist der  $\mathbb{R}^2$  zusammen mit den so definierten Operationen ein Körper, den wir den *Körper der komplexen Zahlen* nennen und mit  $\mathbb{C}$  bezeichnen.

Wir können die Elemente von  $\mathbb{C}$  der Form  $(x, 0)$  mit der reellen Zahl  $x$  identifizieren, denn es gilt

$$(x, 0) + (y, 0) = (x + y, 0)$$

$$(x, 0) \cdot (y, 0) = (xy - 0 \cdot 0, x \cdot 0 + y \cdot 0) = (xy, 0).$$

Die komplexe Zahl  $i = (0, 1)$  heißt *imaginäre Einheit*. Es gilt

$$i^2 = (0, 1) \cdot (0, 1) = (0 \cdot 0 - 1 \cdot 1, 0 \cdot 1 + 0 \cdot 1) = (-1, 0) = -1.$$

Statt  $z = (x, y)$  schreiben wir  $z = x + iy$  und können unter Beachtung von  $i^2 = -1$  „normal“ rechnen ( $z' = x' + iy'$ )

$$z + z' = (x + iy) + (x' + iy') = (x + x') + i(y + y'),$$

$$z \cdot z' = (x + iy) \cdot (x' + iy') = xx' - yy' + i(xy' + yx').$$

Der Leser sollte sich davor hüten, die imaginäre Einheit zu verrätseln, weil sich das Wort imaginär so rätselhaft anhört. Nach wie vor sind die komplexen Zahlen die ebenen Vektoren, auf denen eine Multiplikation definiert ist, die sie zu einem Körper machen. Und die ebenen Vektoren sind genauso wenig imaginär wie alles andere in der Mathematik auch.

Für  $z = x + iy$  setzen wir ferner

$$\bar{z} = x - iy \quad \text{komplexe Konjugation von } z,$$

$$|z| = \sqrt{x^2 + y^2} \quad \text{Absolutbetrag von } z,$$

wobei  $|z|$  mit der zuvor definierten Länge  $|(x, y)|$  des Vektors  $(x, y)$  übereinstimmt. Die komplexe Konjugation bedeutet geometrisch die Spiegelung des Vektors an der  $x$ -Achse.

Ferner definieren wir *Real-* und *Imaginärteil* einer komplexen Zahl  $z = x + iy$  durch

$$\operatorname{Re} z = x, \quad \operatorname{Im} z = y.$$

Kommen wir nun zu den Rechenregeln für komplexe Zahlen:

**Satz 4.9** Für komplexe Zahlen  $z, z'$  gilt:

- (a)  $z^{-1} = \frac{\bar{z}}{|z|^2}$  für  $z \neq 0$ .
- (b)  $|z|^2 = z\bar{z}$ .
- (c)  $(\bar{z} \pm \bar{z}') = (\bar{z} \pm \bar{z}')$ ,  $\bar{z}\bar{z}' = \bar{z}\bar{z}'$ ,  $\overline{\left(\frac{z}{z'}\right)} = \frac{\bar{z}}{\bar{z}'} \text{ für } z' \neq 0$ .
- (d)  $|\bar{z}| = |z|$ ,  $|zz'| = |z||z'|$ ,  $\left|\frac{z}{z'}\right| = \frac{|z|}{|z'|}$ .
- (e)  $\operatorname{Re} z = \frac{1}{2}(z + \bar{z})$ ,  $\operatorname{Im} z = \frac{1}{2i}(z - \bar{z})$ .
- (f)  $|\operatorname{Re} z| \leq |z|$ ,  $|\operatorname{Im} z| \leq |z|$ .
- (g)  $|z + z'| \leq |z| + |z'|$ ,  $||z| - |z'|| \leq |z - z'|$ .

*Beweis:* Die Beweise folgen aus den Definitionen, es muss allerdings nachgerechnet werden. (b) folgt aus

$$z\bar{z} = (x + iy)(x - iy) = x^2 + y^2 = |z|^2$$

und daraus bekommen wir (a) durch Erweiterung des Bruchs

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

Der erste Teil von (c) und (d) folgt direkt aus der Definition der komplexen Konjugation. Die Produktregel in (c) erhalten wir aus

$$\overline{zz'} = \overline{(x+iy)(x'+iy')} = \overline{xx' - yy' + i(yx' + xy')} = (x-iy)(x'-iy') = \bar{z}\bar{z'}.$$

Mit (b) folgt die Produktregel in (d)

$$|zz'|^2 = zz' \overline{zz'} = z\bar{z}z'\bar{z'}.$$

Genauer brauchen wir uns nur noch die Dreiecksungleichung (g) anzuschauen, die wir mit (b)-(f) beweisen

$$\begin{aligned} |z+z'|^2 &= (z+z')(\bar{z}+\bar{z}') = z\bar{z} + z'\bar{z}' + z\bar{z}' + z'\bar{z} \\ &= |z|^2 + |z'|^2 + 2\operatorname{Re} z\bar{z}' \leq |z|^2 + |z'|^2 + 2|z\bar{z}'| \\ &= |z|^2 + |z'|^2 + 2|z||z'| = (|z| + |z'|)^2. \end{aligned}$$

Für die zweite Ungleichung in (g), *inverse Dreiecksungleichung* genannt, verwenden wir die erste

$$|z| = |z-z'+z'| \leq |z-z'| + |z'|.$$

Das umgekehrte Vorzeichen bekommt man, wenn man hier die Rollen von  $z$  und  $z'$  vertauscht.  $\square$

Zu jedem reellen Vektor  $(x, y)$  mit  $x^2 + y^2 = 1$  gibt es genau ein  $\phi \in [0, 2\pi)$  mit

$$x = \cos \phi, \quad y = \sin \phi.$$

$\phi$  ist dabei der im Gegenuhrzeigersinn gemessene Winkel zwischen der positiven reellen Achse und dem Strahl vom Nullpunkt zum Punkt  $(x, y)$ . Aus diesem Grund können wir eine komplexe Zahl  $z = x + iy$  mit  $z \neq 0$  eindeutig in der Form

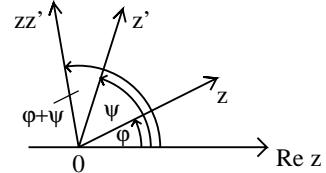
$$z = r(\cos \phi + i \sin \phi) \quad \text{mit } 0 \leq \phi < 2\pi, \quad r = |z| > 0,$$

schreiben.  $r$  ist der von uns bereits definierte Absolutbetrag und  $\phi = \arg z$  heißt *Argument* von  $z$ .

Für das Produkt der beiden Zahlen  $z = r(\cos \phi + i \sin \phi)$  und  $z' = s(\cos \psi + i \sin \psi)$  ergibt sich wegen der Additionstheoreme für Sinus und Kosinus

$$\begin{aligned} (4.2) \quad z \cdot z' &= rs(\cos \phi \cos \psi - \sin \phi \sin \psi + i(\sin \phi \cos \psi + \cos \phi \sin \psi)) \\ &= rs(\cos(\phi + \psi) + i \sin(\phi + \psi)). \end{aligned}$$

Der Ortsvektor  $zz'$  besitzt demnach die Länge  $|zz'|$  und zeigt in Richtung  $\phi + \psi$ . Beim Produkt zweier komplexer Zahlen werden die Beträge multipliziert und die Argumente addiert.



**Beispiel 4.10** Für  $z = 1 + i$  gilt  $|z| = \sqrt{2}$  und damit

$$1+i = \sqrt{2} \left( \cos \frac{\pi}{4} + i \sin \frac{\pi}{4} \right), \quad (1+i)^2 = 2 \left( \cos \frac{\pi}{2} + i \sin \frac{\pi}{2} \right) = 2(0+i \cdot 1) = 2i.$$

$\square$

## 5 Lineare Vektorräume

**5.1 Der Raum  $\mathbb{K}^n$**  Sei  $\mathbb{K}$  ein Körper. Der Raum  $\mathbb{K}^n$  besteht aus  $n$ -tupeln in  $\mathbb{K}$ , die spaltenweise angeordnet werden

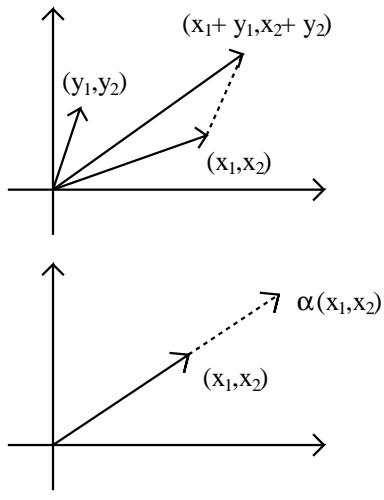
$$u = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad x_i \in \mathbb{K}.$$

Um weniger Platz zu verbrauchen, schreiben wir dafür auch

$$u = (x_1, x_2, \dots, x_n)^T.$$

Die Elemente von  $\mathbb{K}^n$  bezeichnen wir als *Vektoren*, die von  $\mathbb{K}$  als *Skalare*.

Addition und Skalarmultiplikation definiert man komponentenweise



$$u + v = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}, \quad \alpha u = \begin{pmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{pmatrix}$$

Im  $\mathbb{R}^2$  können wir einen Vektor zunächst als Punkt in ein Koordinatensystem einzeichnen und diesen dann mit dem Nullpunkt des Koordinatensystems verbinden. Der am Ende in den Punkt eingezeichnete Pfeil gibt die Orientierung des Vektors an. Die Addition zweier Vektoren verläuft anschaulich wie im nebenstehenden Bild. Wir verschieben  $(y_1, y_2)$  so, dass sein Fußpunkt auf dem Endpunkt von  $(x_1, x_2)$  steht, der Endpunkt des so verschobenen Vektors zeigt dann auf den Endpunkt der Summe.

Für  $\alpha \geq 0$  ist der Ergebnisvektor die Verlängerung oder Verkürzung um das  $\alpha$ -fache. Bei  $\alpha < 0$  kehrt sich zusätzlich die Orientierung um.

Mit

$$(5.1) \quad e_1 = (1, 0, 0, \dots, 0)^T, \quad e_2 = (0, 1, 0, \dots, 0)^T, \quad \dots$$

können wir schreiben

$$(5.2) \quad u = (x_1, x_2, \dots, x_n)^T = \sum_{i=1}^n x_i e_i.$$

Man bezeichnet  $\{e_i\}_{i=1,\dots,n}$  auch als *kanonische Basis*.

Viele mathematische Objekte lassen sich mit einer offensichtlichen Identifikation auf den  $\mathbb{K}^n$  zurückführen:

**Beispiele 5.1** (i) Polynome in einer Variablen  $x \in \mathbb{R}$  sind von der Form

$$p(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n = \sum_{i=0}^n a_i x^i, \quad a_i \in \mathbb{R} \text{ für } i = 0, 1, \dots, n.$$

Da wir hier beliebige  $x \in \mathbb{R}$  einsetzen können, definiert jedes Polynom eine Abbildung  $p : \mathbb{R} \rightarrow \mathbb{R}$ . Polynome addiert man komponentenweise und multipliziert sie komponentenweise mit Elementen  $\alpha \in \mathbb{R}$ : Für  $p(x) = \sum_i a_i x^i$ ,  $p'(x) = \sum_i a'_i x^i$  sowie  $\alpha \in \mathbb{R}$  ist

$$p(x) + p'(x) = \sum_i (a_i + a'_i) x^i, \quad \alpha p(x) = \sum_i \alpha a_i x^i.$$

Wir sagen, das Polynom  $p$  besitzt den Grad  $k$ , wenn  $a_k \neq 0$  und wenn  $a_i = 0$  für alle  $i > k$ . Der Raum der Polynome vom Grad  $\leq n$  wird mit  $\mathbb{P}_n$  bezeichnet.  $\mathbb{P}_n$  lässt sich vermöge der Identifikation

$$p(x) = \sum_{i=1}^n a_i x^i \leftrightarrow (a_0, a_1, \dots, a_n)^T \in \mathbb{R}^{n+1}$$

mit dem  $\mathbb{R}^{n+1}$  identifizieren. Denn zum einen ist dies eine bijektive Abbildung zwischen den angegebenen Räumen. Weiter erhält diese Abbildung die beiden hier definierten algebraischen Operationen Addition und Skalarmultiplikation: Bezeichnen wir die angegebene Abbildung mit  $I : \mathbb{P}_n \rightarrow \mathbb{R}^{n+1}$ , also  $Ip = (a_0, a_1, \dots, a_n)$ , so gilt

$$I(p + p') = Ip + Ip', \quad I(\alpha p) = \alpha I(p) \quad \forall p, p' \in \mathbb{P}_n \quad \forall \alpha \in \mathbb{R}.$$

(ii) Hier betrachten wir  $(m \times n)$ -Schemata der Form

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}, \quad a_{ij} \in \mathbb{K} \text{ für } 1 \leq i \leq m, 1 \leq j \leq n.$$

Diese Schemata kommen in Form von Tabellen überall vor. Spezieller bezeichnen wir ein solches Schema als  $(m \times n)$ -Matrix, wenn zusätzlich noch Addition und Skalarmultiplikation definiert sind: Für  $(m \times n)$ -Matrizen  $A$  und  $B$  mit erzeugenden Koeffizienten  $a_{ij}$  bzw.  $b_{ij}$  sowie  $\alpha \in \mathbb{K}$  setzen wir

$$A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{pmatrix}, \quad \alpha A = \begin{pmatrix} \alpha a_{11} & \alpha a_{12} & \dots & \alpha a_{1n} \\ \alpha a_{21} & \alpha a_{22} & \dots & \alpha a_{2n} \\ \vdots & \vdots & & \vdots \\ \alpha a_{m1} & \alpha a_{m2} & \dots & \alpha a_{mn} \end{pmatrix}.$$

Man beachte, dass diese Addition nur definiert ist, wenn die beiden Dimensionsgrößen  $m$  und  $n$  für die beiden Matrizen die selben sind. Für eine  $(m \times n)$ -Matrix schreiben wir kürzer  $A = (a_{ij})_{1 \leq i \leq m, 1 \leq j \leq n}$  oder manchmal, wenn die Dimensionierung aus dem Zusammenhang klar ist, noch kürzer  $A = (a_{ij})$ . Damit gilt für  $A = (a_{ij})$ ,  $B = (b_{ij})$  einfach  $A + B = (a_{ij} + b_{ij})$ ,  $\alpha A = (\alpha a_{ij})$ .

Ähnlich wie im vorigen Beispiel verfahren wir für den Raum  $\mathbb{K}^{m \times n}$  der  $(m \times n)$ -Matrizen und setzen

$$I : \mathbb{K}^{m \times n} \rightarrow \mathbb{K}^{mn}, \quad IA = (a_{11}, \dots, a_{1n}, a_{21}, \dots, a_{2n}, \dots, a_{mn})^T,$$

d.h. wir stellen die Zeilen der Matrix  $A$  von oben nach unten als Vektor des  $\mathbb{K}^{mn}$  zusammen. Diese Abbildung ist bijektiv zwischen den angegebenen Räumen und erhält Addition und Skalarmultiplikation,  $I(A + B) = IA + IB$ ,  $I(\alpha A) = \alpha IA$ . Wir können den Raum der  $(m \times n)$ -Matrizen daher komplett mit dem  $\mathbb{K}^{mn}$  identifizieren. Der kanonischen Basis für den  $\mathbb{K}^{mn}$  entsprechen die kanonischen Basismatrizen für  $1 \leq i \leq m$ ,  $1 \leq j \leq n$

$$A_{ij} = (\delta_{ij}) \quad \text{mit } \delta_{ij} = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{falls } i \neq j \end{cases}.$$

□

**5.2 Allgemeine lineare Vektorräume** Wir betrachten eine Menge  $V$  mit einem ausgezeichneten Element  $0 \in V$  und einem Körper  $\mathbb{K}$ . Auf  $(V, \mathbb{K})$  sollen zwei Operationen  $+ : V \times V \rightarrow V$  und  $\cdot : \mathbb{K} \times V \rightarrow V$  definiert sein, wobei meist kürzer  $\alpha \cdot u = \alpha u$  geschrieben wird. Die Menge  $V$  heißt *linearer Vektorraum über dem Körper  $\mathbb{K}$* , und die Elemente von  $V$  dann Vektoren, wenn man mit  $V$  und  $\mathbb{K}$  so rechnen kann, wie wir es vom  $\mathbb{K}^n$  gewohnt sind. Also:

(A1)  $(V, 0, +)$  bildet mit der Operation  $+$  eine kommutative Gruppe mit dem neutralen Element  $0 \in V$ .

(A2) Es gelten die beiden distributiven Gesetze für  $\alpha, \beta \in \mathbb{K}$ ,  $u, v \in V$ ,

$$\alpha(u + v) = \alpha u + \alpha v, \quad (\alpha + \beta)u = \alpha u + \beta u.$$

(A3) Es gilt ein Assoziativgesetz für die Skalarmultiplikation

$$(\alpha\beta)u = \alpha(\beta u), \quad \alpha, \beta \in \mathbb{K}, \quad u \in V.$$

(A4) Für die 1 des Körpers  $\mathbb{K}$  gilt  $1 \cdot u = u$  für alle  $u \in V$ .

Das wichtigste Beispiel für einen linearen Vektorraum ist der im vorigen Abschnitt eingeführte  $\mathbb{K}^n$  mit dem Nullvektor  $0 = (0, 0, \dots, 0)$  und den dort definierten Operationen. Aus den Axiomen lassen sich leicht die Rechenregeln

$$0 \cdot u = 0, \quad \alpha \cdot 0 = 0$$

ableiten. Die Ungenauigkeit unserer Notation, nämlich nicht zwischen  $0 \in \mathbb{K}$  und  $0 \in V$  zu unterscheiden, wird dadurch ein wenig abgemildert. Die erste Gleichheit folgt aus

$$0 \cdot u = (0 + 0) \cdot u = 0 \cdot u + 0 \cdot u,$$

und, da  $(V, +)$  eine Gruppe ist,  $0 \cdot u = 0$ . Die zweite Gleichung folgt genauso mit Hilfe des anderen Distributivgesetzes:  $\alpha \cdot 0 = \alpha \cdot (0 + 0) = \alpha \cdot 0 + \alpha \cdot 0$ . Zwei weitere aus dem  $\mathbb{K}^n$  bekannte Gesetze gelten ebenfalls in allgemeinen Vektorräumen

$$(-1) \cdot u = -u, \quad \alpha u = 0 \Rightarrow \alpha = 0 \text{ oder } u = 0.$$

Das linke Gesetz folgt leicht mit

$$(-1) \cdot u + u = (-1 + 1) \cdot u = 0 \cdot u = u,$$

also ist  $u$  invers zu  $(-1) \cdot u$ . Das zweite Gesetz beweist man so: Ist  $\alpha = 0$ , so ist in der Tat  $\alpha \cdot u = 0$ . Ist  $\alpha \neq 0$ , so können wir beide Seiten mit  $\alpha^{-1}$  multiplizieren und aus  $\alpha^{-1}(\alpha u) = 0$  folgt mit dem Assoziativgesetz  $u = 0$ .

Wir nennen eine Teilmenge  $U$  eines linearen Vektorraums *Unterraum* von  $V$ , wenn  $U$  selber ein linearer Vektorraum über  $\mathbb{K}$  ist.

**Satz 5.2** Sei  $V$  ein Vektorraum über  $\mathbb{K}$  und  $U \subset V$  eine Teilmenge von  $V$ . Dann ist  $U$  genau dann ein Unterraum von  $V$ , wenn er abgeschlossen bezüglich den beiden Operationen Addition und Skalarmultiplikation ist, wenn also

$$u + u' \in U \text{ und } \alpha u \in U \text{ für alle } u, u' \in U \text{ und für alle } \alpha \in \mathbb{K}.$$

Wir sagen daher auch, dass  $U$  die Axiome (A1)-(A4) von  $V$  „erbt“.

*Beweis:* Viel ist hier nicht zu zeigen. Wegen  $0u = 0$  ist auch  $0 \in U$  und wegen  $-u = (-1)u \in U$  ist auch das inverse Element bezüglich der Addition in  $U$ . Die weiteren Axiome gelten in  $U$ , weil sie bereits in  $V$  gelten.  $\square$

**Beispiele 5.3** (i) Die Menge der Abbildungen von  $\mathbb{K}$  nach  $\mathbb{K}$  bilden einen linearen Vektorraum über  $\mathbb{K}$  mit der punktweisen Addition und Skalarmultiplikation,

$$(f + g)(x) = f(x) + g(x), \quad (\alpha f)(x) = \alpha f(x).$$

Der Nullvektor ist die Nullabbildung  $x \mapsto 0$  und das inverse Element zu  $f$  ist  $-f$ . Die Axiome folgen aus den Rechenregeln für  $\mathbb{K}$ . Jeder Polynomraum  $\mathbb{P}_n$  ist demnach ein Unterraum dieses Raumes.

(ii) Neben dem trivialen Unterraum  $U = \{0\}$  und dem ganzen Raum, der immer Unterraum von sich selbst ist, gibt es im  $\mathbb{R}^2$  als Unterräume nur noch die Geraden, die durch den Nullpunkt laufen. Sie werden von einem Vektor  $u \in \mathbb{R}^2 \setminus \{0\}$  erzeugt durch  $U_u = \{\alpha u : \alpha \in \mathbb{R}\}$ . Dieses Beispiel zeigt auch, dass im Allgemeinen  $U_1 \cup U_2$  keine Unterraumstruktur besitzt. Liegen  $u_1 \neq 0$  und  $u_2 \neq 0$  nicht auf einer Geraden, so ist  $u_1 + u_2 \notin U_1 \cup U_2$ .  $\square$

**5.3 Linearkombinationen und erzeugende Systeme** Sei ab nun  $V$  ein linearer Vektorraum über dem Körper  $\mathbb{K}$ . Für Vektoren  $u_1, \dots, u_k$  und Skalare  $\alpha_1, \dots, \alpha_k$  heißt

$$u = \alpha_1 u_1 + \alpha_2 u_2 + \dots + \alpha_k u_k = \sum_{i=1}^k \alpha_i u_i$$

eine *Linearkombination* der Vektoren  $u_1, \dots, u_k$ . Wir können den Vektoren  $u_1, \dots, u_k$  die Menge der mit ihnen erzeugten Linearkombinationen zuordnen

$$U = \left\{ u = \sum_{i=1}^k \alpha_i u_i : \alpha_i \in \mathbb{K} \text{ für } 1 \leq i \leq k \right\}$$

$U$  ist Unterraum, denn wenn  $u = \sum_i \alpha_i u_i$  und  $u' = \sum_i \alpha'_i u_i$ , so ist auch  $u+u' = \sum_i (\alpha_i + \alpha'_i) u_i \in U$  und auch  $\alpha u = \sum_i \alpha \alpha_i u_i \in U$ .  $U$  heißt der von  $u_1, \dots, u_k$  aufgespannte Unterraum und wird auch mit

$$U = \text{span} \{u_1, \dots, u_k\}$$

bezeichnet.

Ist  $U = V$  so heißt  $\{u_i\}_{i=1, \dots, k}$  erzeugendes System von  $V$ .

**5.4 Basis und Dimension** Für eine Folge von Vektoren  $u_1, u_2, \dots$  können wir die Folge von aufgespannten Unterräumen betrachten

$$U_k = \text{span} \{u_1, \dots, u_k\}.$$

Klar ist  $U_k$  eine aufsteigende Folge von Unterräumen, aber wann wird  $U_{k+1}$  echt größer als  $U_k$ ? Wenn beispielweise  $u_{k+1}$  bereits in  $U_k$  enthalten ist,

$$u_{k+1} = \sum_{i=1}^k \beta_i u_i,$$

so kommt in den Linearkombinationem mit  $u_{k+1}$  gegenüber  $U_k$  nichts Neues hinzu wegen

$$\sum_{i=1}^{k+1} \alpha_i u_i = \sum_{i=1}^k (\alpha_i + \alpha_{k+1} \beta_i) u_i,$$

also gilt in diesem Fall  $U_{k+1} = U_k$ .

Wir sagen, die Menge von Vektoren  $u_1, \dots, u_k$  ist *linear unabhängig* (kurz: l.u.), wenn

$$\sum_{i=1}^k \alpha_i u_i = 0 \quad \Rightarrow \quad \alpha_1 = \alpha_2 = \dots = \alpha_k = 0.$$

In diesem Fall kann keiner der Vektoren  $u_i$  als Linearkombination der anderen Vektoren dargestellt werden, denn dann hätten wir ja  $u_j = \sum_{i \neq j} \alpha_i u_i$ , also  $\sum_{i \neq j} \alpha_i u_i - u_j = 0$  und die Vektoren wären nicht linear unabhängig. Man kann das dahinterstehene Prinzip noch etwas markanter formulieren:

**Lemma 5.4** Die Vektoren  $u_1, \dots, u_k$  sind genau dann linear unabhängig, wenn jedes  $u \in U = \text{span} \{u_1, \dots, u_k\}$  sich eindeutig als Linearkombination der  $u_i$  darstellen lässt.

*Beweis:* Angenommen, es gäbe zwei Darstellungen von  $u$ ,

$$u = \sum_{i=1}^k \alpha_i u_i = \sum_{i=1}^k \alpha'_i u_i.$$

Dann folgt  $0 = \sum_i (\alpha_i - \alpha'_i) u_i$ . Die Eigenschaft  $\alpha_i = \alpha'_i$  für alle  $i$  ist daher äquivalent zur linearen Unabhängigkeit der Vektoren  $u_i$ .  $\square$

Sind die Vektoren  $u_1, \dots, u_k$  nicht l.u., so heißen sie *linear abhängig* (kurz: l.a.). In diesem Fall gibt es eine nichttriviale Linearkombination zur 0, also  $\sum_i \alpha_i u_i = 0$  mit mindestens einem  $\alpha_{i_0} \neq 0$ . In diesem Fall können wir nach  $u_{i_0} = -\sum_{i \neq i_0} \alpha_i / \alpha_{i_0} u_i$  auflösen. Kurz: Die Vektoren  $u_1, \dots, u_k$  sind genau dann l.a., wenn man zum Aufspannen des Unterraums  $U_k = \text{span} \{u_1, \dots, u_k\}$  nicht alle Vektoren  $u_1, \dots, u_k$  benötigt.

Eine nichtleere Menge  $M \subset V$  heißt linear unabhängig, wenn alle endlichen Teilmengen von  $M$  linear unabhängig sind. Andernfalls heißt  $M$  linear abhängig.

**Beispiele 5.5** (i) Sei  $V = \mathbb{K}^n$  und seien  $e_i, i = 1, \dots, n$ , die kanonischen Basisvektoren aus (5.1). Wegen

$$u = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix} + \dots + x_n \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

sind die Vektoren  $\{e_i\}_{i=1, \dots, n}$  l.u.: Jeder Vektor  $e_i$  ist sozusagen für die  $i$ -te Komponente zuständig.

(ii) Enthält eine Menge den Nullvektor, ist sie linear abhängig, denn der Nullvektor ist bereits selber l.a.

(iii) Sind  $u_1, \dots, u_k$  l.u. und sind  $\lambda_1, \dots, \lambda_{k-1}$  beliebige Skalare, so sind auch die Vektoren  $u_1 - \lambda_1 u_k, \dots, u_{k-1} - \lambda_{k-1} u_k, u_k$  l.u.. Es gilt nämlich

$$\sum_{i=1}^{k-1} \alpha_i (u_i - \lambda_i u_k) + \alpha_k u_k = 0 \Rightarrow \sum_{i=1}^{k-1} \alpha_i u_i + \left( \alpha_k - \sum_{i=1}^{k-1} \lambda_i \alpha_i \right) u_k = 0,$$

und wegen der linearen Unabhängigkeit folgt zunächst  $\alpha_1, \dots, \alpha_{k-1} = 0$  und schließlich  $\alpha_k = 0$ .

(iv) (Aufgabe) Eine Menge  $v_1, \dots, v_n$  sei l.u. in einem Vektorraum  $V$ . Für  $\alpha_1, \dots, \alpha_n \in \mathbb{K}$  sei  $w = \sum_{i=1}^n \alpha_i v_i$ . Man formuliere eine notwendige und hinreichende Bedingung an die  $\alpha_i$ , so dass auch die Vektoren

$$x_i = v_i - w, \quad i = 1, \dots, n$$

l.u. sind.

Aus  $\sum_{i=1}^n \lambda_i x_i = 0$  folgt

$$\begin{aligned} 0 &= \sum_{i=1}^n \lambda_i \left( v_i - \sum_{k=1}^n \alpha_k v_k \right) = \sum_{i=1}^n \lambda_i v_i - \sum_{i=1}^n \sum_{k=1}^n \lambda_i \alpha_k v_k \\ &= \sum_{i=1}^n \lambda_i v_i - \sum_{i=1}^n \sum_{k=1}^n \lambda_k \alpha_i v_i = \sum_{i=1}^n \left( \lambda_i - \alpha_i \sum_{k=1}^n \lambda_k \right) v_i. \end{aligned}$$

Aus der linearen Unabhängigkeit der  $v_i$  folgt

$$(5.3) \quad \lambda_i - \alpha_i \sum_{k=1}^n \lambda_k = 0, \quad i = 1, \dots, n.$$

Wir summieren bezüglich  $i$ ,

$$\sum_{i=1}^n \lambda_i - \sum_{i=1}^n \alpha_i \sum_{k=1}^n \lambda_k = 0$$

und erhalten  $\sum_i \lambda_i = 0$ , sofern  $\sum_i \alpha_i \neq 1$ . Aus (5.3) folgt dann  $\lambda_i = 0$  für alle  $i$ . Damit sind die Vektoren  $x_i$  l.u. Falls  $\sum_i \alpha_i = 1$ , so ist  $\lambda_i = \alpha_i$  eine nichttriviale Lösung des Gleichungssystems (5.3). Damit sind die Vektoren in diesem Fall l.a.  $\square$

Wir sagen, der lineare Vektorraum  $V$  wird *endlich erzeugt*, wenn eine endliche Menge von Vektoren den Raum  $V$  aufspannen.

**Lemma 5.6**  *$V$  werde von den  $n$  Vektoren  $v_1, \dots, v_n$  erzeugt. Dann ist jede  $n+1$ -elementige Menge  $M \subset V$  l.a..*

*Beweis:* Angenommen, die Menge  $M = \{w_1, \dots, w_{n+1}\}$  ist l.u.. Wir tauschen sukzessive ein Element von  $M$  gegen ein Element von  $N = \{v_1, \dots, v_n\}$  aus. Da die Elemente von  $N$  den Raum  $V$  erzeugen, gibt es  $\alpha_i \in \mathbb{K}$  mit  $w_1 = \sum_{i=1}^n \alpha_i v_i$ . Da  $w_1$  nicht der Nullvektor ist, gibt es ein  $\alpha_{i_0} \neq 0$ , sagen wir  $i_0 = 1$ . Damit

$$(5.4) \quad v_1 = \frac{1}{\alpha_1} \left( w_1 - \sum_{i=2}^n \alpha_i v_i \right).$$

Wir tauschen in  $N$   $v_1$  gegen  $w_1$  aus und erhalten die modifizierte Menge  $N' = \{w_1, v_2, \dots, v_n\}$ .  $N'$  ist nach wie vor erzeugend, denn in jeder Linearkombination von  $v_1, \dots, v_n$  können wir  $v_1$  nach (5.4) durch  $w_1$  und die übrigen  $v_i$  ausdrücken. Auf diese Weise fahren wir fort und tauschen nach und nach die anderen Elemente von  $N$  aus. Dieser Austauschprozess kommt zum Erliegen, wenn in  $w_{k+1} = \sum_{i=1}^{k-1} \alpha_i w_i + \sum_{i=k}^n \alpha_i v_i$  die  $\alpha_i$  mit  $i \geq k$  alle verschwinden. In diesem Fall ist  $w_{k+1}$  eine Linearkombination der  $w_i$  und die Ausgangsmenge  $M$  l.a.. Geht der Austauschprozess bis zum Ende durch, so ist  $N$  vollständig ersetzt durch  $\{w_1, \dots, w_n\}$  und  $w_{n+1}$  lässt sich als Linearkombination der  $w_i$  für  $i \leq n$  darstellen. Auch in diesem Fall sind die Vektoren in  $M$  l.a..  $\square$

Nun kommt der wichtigste Begriff der linearen Algebra: Ein linear unabhängiges erzeugendes System von  $V$  heißt *Basis* von  $V$ . Aus dem letzten Lemma folgt: Besitzt ein Vektorraum eine Basis mit endlich vielen Elementen, so haben alle Basen dieses Raumes dieselbe endliche Zahl von Elementen.

**Satz 5.7** *Jeder endlich erzeugte Vektorraum besitzt eine Basis.*

*Beweis:* Sei  $\{u_1, \dots, u_k\}$  ein erzeugendes System des Vektorraums  $V$ . Sind die  $u_i$  l.u., so sind wir fertig. Andernfalls gilt für einen Index  $i_0$

$$u_{i_0} = \sum_{i \neq i_0} \alpha_i u_i.$$

Wir können das  $u_{i_0}$  aus der Menge  $\{u_1, \dots, u_k\}$  entfernen, die Menge bleibt erzeugend. Denn in jeder Linearkombination der  $u_i$  kann mit der letzten Gleichung das  $u_{i_0}$  eliminiert werden. Nach endlich vielen Schritten dieser Konstruktion erhalten wir eine Basis von  $V$ .  $\square$

Ein endlich erzeugter Vektorraum  $V$  heißt *endlich dimensional*. Die Mächtigkeit  $n \in \mathbb{N}$  der Basis heißt *Dimension* von  $V$ . Wir schreiben dann  $\dim V = n$  und setzen für den etwas pathologischen Fall  $\dim\{0\} = 0$ . Ist  $V$  nicht endlich erzeugt, so schreiben wir  $\dim V = \infty$ .

**Beispiele 5.8** (i) Es gilt  $\dim \mathbb{K}^n = n$  und die einfachste Basis ist die kanonische Basis  $\{e_i\}_{i=1,\dots,n}$  wie in (5.1),(5.2).

(ii)  $\mathbb{C} = \mathbb{C}^1$  ist Vektorraum über  $\mathbb{C}$  und es gilt natürlich  $\dim \mathbb{C} = 1$ . Wir können  $\mathbb{C}$  aber auch als  $\mathbb{R}$ -Vektorraum auffassen und in diesem Fall gilt  $\dim \mathbb{C} = 2$  mit den kanonischen Basisvektoren  $1$  und  $i = \sqrt{-1}$ . Dies entspricht unserer „reellen“ Vorstellungswelt, in der die komplexen Zahlen als ebene Vektoren dargestellt werden.

(iii) Auch unendlich dimensionale Vektorräume besitzen eine Basis, allerdings gibt es i.A. kein Verfahren, um eine solche Basis zu konstruieren. Eine Ausnahme bildet der einfachste unendlich dimensionale Vektorraum  $c_{00}$ , der aus den endlichen Folgen besteht. Gedanklich verlängern wir eine endliche Folge durch Nullen zu einer unendlichen Folge. Auf diesen verlängerten Folgen können wir

wie gewohnt komponentenweise addieren und mit Skalaren multiplizieren. In beiden Fällen verbleibt der Ergebnisvektor im Raum der endlichen Folgen. Die kanonischen Einheitsvektoren  $\{e_i\}$  bilden wieder eine Basis dieses Raumes, diesmal allerdings für  $i = 1, 2, \dots$ . Damit ist  $\dim c_{00} = \infty$  und der Raum ist abzählbar dimensional.  $\square$

**Satz 5.9** Sei  $\dim V = n$  und für  $s < n$  seien  $b_1, \dots, b_s$  l.u.. Dann gibt es  $b_{s+1}, \dots, b_n \in V$ , so dass  $b_1, \dots, b_n$  eine Basis von  $V$  ist.

**Bemerkung 5.10** Der Satz kann auch verwendet werden, um eine Basis zu konstruieren. In diesem Fall startet man mit einem beliebigen  $b_1 \in V \setminus \{0\}$ .  $\square$

*Beweis:* Da alle Basen die gleiche Kardinalität besitzen, ist  $V_s = \text{span}\{b_1, \dots, b_s\}$  echt in  $V$  enthalten. Es gibt daher ein  $b_{s+1} \in V \setminus V_s$ . Wäre in

$$\sum_{i=1}^{s+1} \alpha_i b_i = 0$$

$\alpha_{s+1} \neq 0$ , so  $b_{s+1} = -\sum_{i=1}^s \alpha_i / \alpha_{s+1} b_i \in V$ . Daher ist  $\alpha_{s+1} = 0$  und aus der linearen Unabhängigkeit von  $b_1, \dots, b_s$  folgt  $\alpha_i = 0$  für die übrigen  $i$ . Damit sind auch die Vektoren  $b_1, \dots, b_{s+1}$  l.u. und mit dieser Konstruktion erreicht man schließlich das gewünschte Ziel.  $\square$

**Beispiel 5.11** (Aufgabe) Seien  $U, V$  Unterräume eines endlich dimensionalen Vektorraums  $W$ . Man konstruiere eine Basis von

$$U + V = \{z = u + v : u \in U, v \in V\}.$$

Ist  $U \subset V$  oder  $V \subset U$ , so ist  $U + V = V$  oder  $U + V = U$  und es ist nichts zu zeigen.

Das Problem ist, dass eine Basis von  $U$  zusammen mit einer Basis von  $V$  zwar den Raum  $U + V$  erzeugen, aber im Falle  $U \cap V \neq \{0\}$  nicht l.u. sind. Deshalb nehmen wir zunächst eine Basis  $w_1, \dots, w_r$  von  $U \cap V$  und ergänzen sie mit dem letzten Satz zu Basen von  $U$  beziehungsweise  $V$ . Seien also  $w_1, \dots, w_r, u_1, \dots, u_s, w_1, \dots, w_r, v_1, \dots, v_t$  Basen von  $U$  beziehungsweise  $V$ . Dann ist  $w_1, \dots, w_r, u_1, \dots, u_s, v_1, \dots, v_t$  ein erzeugendes System von  $U + V$ , denn in  $w = u + v$  können wir  $u$  und  $v$  mit diesen Vektoren darstellen. Mit  $U' = \text{span}\{u_1, \dots, u_s\}$  und  $V' = \text{span}\{v_1, \dots, v_t\}$  lässt sich jedes  $w \in U + V$  auch eindeutig in der Form

$$w = s + u' + v', \quad s \in U \cap V, \quad u' \in U', \quad v' \in V',$$

schreiben. Damit lässt sich  $w$  eindeutig mit den angegebenen Vektoren darstellen.  $\square$

## 6 Lineare Abbildungen und Matrizen

**6.1 Lineare Abbildungen** Seien  $V, W$  Vektorräume über  $\mathbb{K}$ . Eine Abbildung  $f : V \rightarrow W$  heißt *linear*, wenn sie die beiden linearen Operationen Addition und Skalarmultiplikation erhält, wenn also

$$f(u + v) = f(u) + f(v), \quad f(\alpha u) = \alpha f(u) \quad \text{für alle } u, v \in V, \alpha \in \mathbb{K}.$$

Das Grundprinzip der modernen Mathematik besteht darin, zunächst Strukturen und dann strukturhaltende Abbildungen zu definieren. In diesem Fall ist die Struktur der lineare Vektorraum zusammen mit Addition und Skalarmultiplikation.

Man kann die Linearität einer Abbildung äquivalent mit der Bedingung  $f(\alpha u + \beta v) = \alpha f(u) + \beta f(v)$  definieren. Mehrfache Anwendung der Linearitätsbedingung liefert

$$f\left(\sum_{i=1}^k \alpha_i v_i\right) = \sum_{i=1}^k \alpha_i f(v_i).$$

Aus  $f(0) = f(0 \cdot 0) = 0f(0)$  folgt  $f(0) = 0$ .

Der *Nullraum* oder *Kern* einer linearen Abbildung ist

$$\text{Kern } f = \{v \in V : f(v) = 0\} \subset V.$$

Er ist Unterraum von  $V$  weil für alle  $v, v' \in \text{Kern } f$  und alle  $\alpha \in \mathbb{K}$  gilt  $f(v + v') = f(v) + f(v') = 0$ ,  $f(\alpha v) = \alpha f(v) = 0$ , womit das Unterraumkriterium aus Satz 5.2 erfüllt ist.

Eine einfache, aber oft verwendete Eigenschaft des Kerns: Eine lineare Abbildung  $f$  ist genau dann injektiv, wenn  $\text{Kern } f = \{0\}$ . Enthält der Kern noch eine weiteres Element, so werden dieses und die Null auf die Null abgebildet. Enthält der Kern nur die Null, so kann es nicht sein, dass ein  $w$  zwei verschiedene Urbilder  $v, v'$  besitzt wegen  $0 = f(v - v')$  und daher  $v = v'$ .

Der *Bildraum* oder das *Bild* einer linearen Abbildung ist

$$\text{Bild } f = \{w \in W : \exists v \in V \text{ mit } f(v) = w\} \subset W.$$

Das Bild ist Unterraum von  $W$ , denn wenn  $w, w' \in \text{Bild } f$ , so gibt es  $v, v' \in V$  mit  $f(v) = w$  und  $f(v') = w'$ . Damit ist  $w + w' = f(v) + f(v') = f(v + v')$  und  $w + w'$  ist aus dem Bild von  $f$ . Für die Skalarmultiplikation zeigt man das genauso.

**Beispiele 6.1** (i) Die Abbildung in die Null  $v \mapsto 0$  ist immer eine lineare Abbildung zwischen den Räumen  $V$  und  $W$ . In diesem Fall ist  $\text{Kern } f = V$  und  $\text{Bild } f = \{0\}$ .

(ii) Die Identität  $Id : V \rightarrow V$  ist linear mit  $\text{Kern } f = \{0\}$  und  $\text{Bild } f = V$ .

(iii) Die *orthogonalen Transformationen* des  $\mathbb{R}^2$ , das sind Drehungen und Spiegelungen an einer Geraden, die durch den Nullpunkt läuft, sind linear. Auf die Konstruktion solcher Abbildungen werden wir später eingehen.

(iv) Dagegen ist die *Translation* um einen Vektor  $v_0 \neq 0$ , das ist  $f(v) = v_0 + v$ , keine lineare Selbstabbildung des  $\mathbb{R}^2$ . Man erkennt das schon daran, dass  $f(0) = v_0 \neq 0$ .  $\square$

**Satz 6.2** (a) *Die Komposition linearer Abbildungen ist linear.*

(b) *Ist  $f : V \rightarrow W$  linear und bijektiv, so ist auch die Inverse  $f^{(-1)} : W \rightarrow V$  linear.*

(c) *Sind  $f, g : V \rightarrow W$  linear, so ist die punktweise Summe  $(f + g)(x) = f(x) + g(x)$  und die Multiplikation mit Skalaren  $(\alpha f)(x) = \alpha f(x)$  linear.*

*Beweis:* (a) Ist  $g : V \rightarrow W$  linear sowie  $f : W \rightarrow X$  linear, so gilt

$$f(g(v + v')) = f(g(v) + g(v')) = f(g(v)) + f(g(v')).$$

Für die Skalarmultiplikation läuft das genauso.

(b) Das folgt aus

$$f(v) = w, \quad f(v') = w', \quad f(v + v') = f(v) + f(v'),$$

wenn man in der letzten Gleichung auf beiden Seiten  $f^{(-1)}$  anwendet,

$$v + v' = f^{(-1)}(f(v) + f(v')) \Rightarrow f^{(-1)}(w) + f^{(-1)}(w') = f^{(-1)}(w + w').$$

Für die Skalarmultiplikation folgt das ebenso einfach.

(c) Dazu brauchen wir eine sehr einfache Rechnung

$$\begin{aligned} (f + g)(\beta v + \beta' v') &= f(\beta v + \beta' v') + g(\beta v + \beta' v') = \beta f(v) + \beta' f(v') + \beta g(v) + \beta' g(v') \\ &= \beta(f + g)(v) + \beta'(f + g)(v'). \end{aligned}$$

Für  $\alpha f$  beweist man das ganz analog.  $\square$

Eine bijektive lineare Abbildung  $f : V \rightarrow W$  heißt *Isomorphismus*. In diesem Fall heißen die beiden Räume  $V$  und  $W$  *isomorph* und man schreibt  $V \cong W$ . Nach Satz 6.2(b) ist die Inverse eines Isomorphismusses selber linear, insbesondere ist  $f^{(-1)} : W \rightarrow V$  ein Isomorphismus.

Die Komposition von Isomorphismen ist ein Isomorphismus, denn nach Satz 6.2(a) ist sie linear und bekanntlich ist die Komposition bijektiver Abbildungen bijektiv.

Damit ist die Isomorphie eine Äquivalenzrelation. Es gilt  $Id : V \rightarrow V$  linear und damit  $V \cong V$ . Da die Umkehrung eines Isomorphismus ebenfalls ein Isomorphismus ist, gilt  $V \cong W$  genau dann, wenn  $W \cong V$ . Aus „Komposition von Isomorphismen ist wieder Isomorphismus“ folgt die Transitivität von  $\cong$ .

Damit können isomorphe Vektorräume als Vektorräume nicht voneinander unterschieden werden. Sofern eine Aussage nur aus den beiden linearen Operationen aufgebaut ist, gilt sie in  $V$  genau dann, wenn sie auch in  $W$  gilt.

**Satz 6.3** Alle endlich dimensionalen Vektorräume über  $\mathbb{K}$  der Dimension  $n$  sind zueinander isomorph. Insbesondere ist jeder  $n$ -dimensionale Vektorraum isomorph zu  $\mathbb{K}^n$ .

*Beweis:* Wir nehmen eine beliebige Basis von  $V$ , sagen wir  $v_1, \dots, v_n$ , und definieren die *Koordinatenabbildung*

$$v = \sum_{i=1}^n \alpha_i v_i \in V \mapsto (\alpha_1, \dots, \alpha_n)^T \in \mathbb{K}^n.$$

Die so definierte Abbildung  $f : V \rightarrow \mathbb{K}^n$ ,  $f(v) = (\alpha_1, \dots, \alpha_n)^T$  ist linear. Denn wenn  $v = \sum_i \alpha_i v_i$ ,  $v' = \sum_i \alpha'_i v_i$ , so folgt

$$f(v + v') = (\alpha_1 + \alpha'_1, \dots, \alpha_n + \alpha'_n)^T = (\alpha_1, \dots, \alpha_n)^T + (\alpha'_1, \dots, \alpha'_n)^T = f(v) + f(v').$$

Es gilt  $f(v_i) = e_i$  mit den kanonischen Einheitsvektoren  $e_i$  des  $\mathbb{K}^n$ . Da die  $e_i$  eine Basis des  $\mathbb{K}^n$  bilden, ist  $f(v) = f(\sum_i \alpha_i v_i) = \sum_i \alpha_i e_i = 0$ , genau dann, wenn alle  $\alpha_i = 0$ . Damit ist  $f$  injektiv.  $f$  surjektiv ist noch offensichtlicher, weil jeder Punkt  $\sum_i \alpha_i e_i$  im Bild von  $f$  liegt.

Da Isomorphie eine Äquivalenzrelation ist, sind alle  $\mathbb{K}$ -Vektorräume der Dimension  $n$  zueinander isomorph.  $\square$

**Beispiel 6.4** Wir können jetzt die Konstruktionen in den Beispielen 5.1 mathematisch präziser fassen.

Wir hatten in 5.1(i) einem Polynom  $p(x) = \sum_i \alpha_i x^i \in \mathbb{P}_n$  den Koeffizientenvektor  $(\alpha_0, \dots, \alpha_n) \in \mathbb{R}^{n+1}$  zugeordnet. Dies definiert eine lineare Abbildung  $I : \mathbb{P}_n \rightarrow \mathbb{R}^{n+1}$ . Die Linearität hatten wir

nachgewiesen, bijektiv ist  $I$  offenbar auch. Damit ist  $I$  ein Isomorphismus zwischen den angegebenen Räumen.

Genauso hatten wir im Beispiel 5.1(ii) den Matrizenraum  $\mathbb{K}^{m \times n}$  linear und bijektiv auf den Raum  $\mathbb{K}^{mn}$  abgebildet. Auch diese Räume sind daher isomorph.  $\square$

Mit  $\mathcal{L}(V, W)$  bezeichnen wir die Menge der linearen Abbildungen zwischen den Vektorräumen  $V$  und  $W$ . Wie in Satz 6.2(c) gezeigt wurde, sind auf  $\mathcal{L}(V, W)$  die punktweise Addition und Skalarmultiplikation, also  $(f + g)(v) = f(v) + g(v)$  sowie  $(\alpha f)(v) = \alpha f(v)$  erklärt, die  $\mathcal{L}(V, W)$  ebenfalls zu einem  $\mathbb{K}$ -Vektorraum machen.

Sei  $\dim V = n$  und  $v_1, \dots, v_n$  eine Basis von  $V$ . Dann ist wegen

$$(6.1) \quad f\left(\sum_{i=1}^n \alpha_i v_i\right) = \sum_{i=1}^n \alpha_i f(v_i)$$

jede lineare Abbildung durch die Werte auf einer Basis eindeutig bestimmt. Ferner wird das Bild durch eine Linearkombination dieser  $n$  Bildvektoren  $f(v_i)$  aufgespannt. Daher ist das Bild eines  $n$ -dimensionalen Vektorraums endlich dimensional mit  $\dim \text{Bild } f \leq n$ .

Die Dimension des Bildes einer linearen Abbildung  $f$  heißt *Rang* von  $f$ , geschrieben  $\text{rang } f$ . Der folgende Satz wird auch *Rangformel* genannt:

**Satz 6.5** *Sei  $V$  endlich dimensional und  $f \in \mathcal{L}(V, W)$ . Dann gilt*

$$\dim V = \dim \text{Kern } f + \text{rang } f.$$

*Beweis:* Da  $\text{Kern } f$  endlich dimensional ist, gibt es Basen  $u_1, \dots, u_r$  von  $\text{Kern } f$  und  $w_1, \dots, w_s$  von Bild  $f$  mit Urbildern  $v_1, \dots, v_s$ . Wegen  $\sum_i \alpha_i f(v_i) = \sum_i \alpha_i w_i$  spannt das Bild von  $V_B = \text{span}\{v_1, \dots, v_s\}$  das gesamte Bild auf. Die Vektoren  $\{v_i\}$  müssen daher l.u. sein und damit  $\dim V_B = s = \text{rang } f$ . Für beliebiges  $v \in V$  gilt  $f(v) = \sum_i \alpha_i w_i$ . Setze daher

$$v = v_B + v_K, \quad v_B = \sum_{i=1}^s \alpha_i v_i \Rightarrow f(v - v_B) = 0 \Rightarrow f(v_K) = 0 \Rightarrow v_K = \sum_{i=1}^r \beta_i u_i.$$

Nach Konstruktion sind die Koeffizienten  $\alpha_1, \dots, \alpha_s, \beta_1, \dots, \beta_r$  eindeutig bestimmt, daher die Vektoren  $u_1, \dots, u_r, v_1, \dots, v_s$  linear unabhängig. Da sich jedes  $v$  nach diesen Vektoren entwickeln lässt, handelt es sich um eine Basis von  $V$  und es gilt  $n = r + s$ .  $\square$

**Korollar 6.6** *Sind  $V, W$  endlich dimensional mit  $\dim V = \dim W$ , so gilt für jede lineare Abbildung  $f \in \mathcal{L}(V, W)$*

$$f \text{ ist injektiv} \Leftrightarrow f \text{ ist surjektiv} \Leftrightarrow f \text{ ist bijektiv.}$$

*Beweis:* Nach der Rangformel impliziert jede dieser Bedingungen, dass  $\dim \text{Bild } f = \dim V$ .  $\square$

Halten wir noch einmal die wichtigsten Aussagen dieses Abschnitts für ein  $n$ -dimensionales  $V$  und  $f \in \mathcal{L}(V, W)$  fest:

*Die Abbildung  $f$  ist bereits durch die Werte auf einer beliebigen Basis von  $V$  eindeutig bestimmt.*

Dies folgt aus (6.1), die gleichzeitig die wichtigste Formel der linearen Algebra ist. (6.1) zeigt auch, dass das Bild von  $f$  von den  $f(v_i)$  aufgebaut wird. Demnach ist  $\dim \text{Bild } f \leq \dim V$ . Eine lineare Abbildung kann also höchstens einen  $n$ -dimensionalen Bildraum aufspannen.

*Für eine Basis  $v_1, \dots, v_n$  von  $V$  kann man  $f(v_i) \in W$  beliebig vorgeben. Durch diese Vorgaben ist die lineare Abbildung  $f$  eindeutig bestimmt.*

$f$  ist demnach für  $v = \sum_i \alpha_i v_i$  definiert durch  $f(v) = \sum_i \alpha_i f(v_i)$ . Dass ein so definiertes  $f$  linear ist, weist man ohne Mühe nach.

**6.2 Darstellungsmatrizen linearer Abbildungen** Eine  $(m \times n)$ -Matrix  $A$  hatten wir als rechteckiges Schema  $A = (a_{ij})_{i=1,\dots,m, j=1,\dots,n}$  definiert zusammen mit komponentenweiser Addition und Skalarmultiplikation.  $\mathbb{K}^{m \times n}$  ist mit diesen Operationen ein linearer Vektorraum über  $\mathbb{K}$  mit  $\dim \mathbb{K}^{m \times n} = mn$ .

Mit Hilfe von Matrizen lassen sich lineare Abbildungen  $f : \mathbb{K}^n \rightarrow \mathbb{K}^m$  anschaulich beschreiben. Wir hatten im letzten Abschnitt eingesehen, dass  $f$  durch die Werte auf einer Basis von  $\mathbb{K}^n$  eindeutig bestimmt ist. Naheliegend ist es, hier die kanonische Basis  $\{e_j\}_{j=1,\dots,n}$  zu wählen. Es gilt also  $f(e_j) = a_j \in \mathbb{K}^m$ . Wir schreiben die Spaltenvektoren  $(a_j)_{j=1,\dots,n}$  hintereinander und erhalten so eine  $(m \times n)$ -Matrix

$$A_f = (a_1 | a_2 | \dots | a_n) = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \quad \text{mit } a_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix},$$

die als *Darstellungsmatrix* der linearen Abbildung  $f$  bezeichnet wird. Für  $u = \sum_{j=1}^n x_j e_j$  gilt dann

$$f(u) = \sum_{j=1}^n x_j f(e_j) = \sum_{j=1}^n x_j a_j = \sum_{j=1}^n \sum_{i=1}^m x_j a_{ij} e_i = \begin{pmatrix} \sum_{j=1}^n a_{1j} x_j \\ \sum_{j=1}^n a_{2j} x_j \\ \vdots \\ \sum_{j=1}^n a_{mj} x_j \end{pmatrix}.$$

Man bestimmt also  $f(u)$  nach der Regel „Zeile  $\times$  Spalte“. Um die  $i$ -te Komponente von  $f(u)$  zu bekommen, nimmt man die  $i$ -te Zeile der Matrix, das ist  $a_{i1}, \dots, a_{in}$  und multipliziert sie mit den entsprechenden Einträgen des Spaltenvektors  $u$ , das ist  $x_1, \dots, x_n$  und addiert schließlich das Ganze, also

$$(f(u))_i = a_{i1}x_1 + \dots + a_{in}x_n.$$

**Beispiel 6.7** Sei  $f : \mathbb{R}^3 \rightarrow \mathbb{R}^2$  mit  $f(e_1) = (1, 1)^T$ ,  $f(e_2) = (2, 1)^T$ ,  $f(e_3) = (-1, 3)^T$ . Zu  $f$  gehört demnach die Darstellungsmatrix

$$A_f = \begin{pmatrix} 1 & 2 & -1 \\ 1 & 1 & 3 \end{pmatrix}.$$

Für  $u = (1, 2, -4)^T$  bestimmen wir  $f(u)$  nach der Regel Zeile mal Spalte

$$f(u) = \begin{pmatrix} 1 \cdot 1 + 2 \cdot 2 + (-1) \cdot (-4) \\ 1 \cdot 1 + 1 \cdot 2 + 3 \cdot (-4) \end{pmatrix} = \begin{pmatrix} 9 \\ -9 \end{pmatrix}.$$

Die Regel Zeile mal Spalte soll aber nicht vergessen lassen, dass wir eine Linearkombination der Spaltenvektoren bilden, in diesem Fall

$$f(u) = 1 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} + 2 \cdot \begin{pmatrix} 2 \\ 1 \end{pmatrix} - 4 \cdot \begin{pmatrix} -1 \\ 3 \end{pmatrix} = \begin{pmatrix} 9 \\ -9 \end{pmatrix}.$$

□

Nun untersuchen wir, wie die Komposition linearer Abbildungen mit den zugehörigen Darstellungsmatrizen realisiert werden kann. Seien  $f : \mathbb{K}^m \rightarrow \mathbb{K}^l$  und  $g : \mathbb{K}^n \rightarrow \mathbb{K}^m$  mit Matrixdarstellungen  $A \in \mathbb{K}^{l \times m}$  von  $f$  und  $B \in \mathbb{K}^{m \times n}$  von  $g$ . Mit  $e_1^m, \dots, e_m^m$  bezeichnen wir die kanonische Basis von  $\mathbb{K}^m$  und mit  $e_1^l, \dots, e_l^l$  die kanonische Basis von  $\mathbb{K}^l$ . Dann gilt

$$f(e_k^m) = \begin{pmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{lk} \end{pmatrix} = \sum_i a_{ik} e_i^l, \quad g(x) = \begin{pmatrix} \sum_j b_{1j} x_j \\ \sum_j b_{2j} x_j \\ \vdots \\ \sum_j b_{mj} x_j \end{pmatrix} = \sum_{kj} b_{kj} x_j e_k^m,$$

und daher

$$\begin{aligned}
f(g(x)) &= f\left(\sum_{kj} b_{kj} x_j e_k^m\right) = \sum_{kj} b_{kj} x_j f(e_k^m) = \sum_{kji} b_{kj} x_j a_{ik} e_i^l \\
&= \sum_{ji} \underbrace{\sum_{k=1}^m a_{ik} b_{kj}}_{\text{Zeile } \times \text{ Spalte} = c_{ij}} x_j e_i^l = \sum_{ij} c_{ij} x_j e_i^l = Cx, \quad C = (c_{ij}).
\end{aligned}$$

Um die Darstellungsmatrix für die Komposition zweier linearer Abbildungen zu bekommen gilt also wieder die Regel „Zeile  $\times$  Spalte“. Ist  $A = (a_{ik})$ ,  $B = (b_{kj})$  und  $C = (c_{ij})$  die Darstellungsmatrix für die Komposition, so gilt

$$c_{ij} = a_{i1} b_{1j} + a_{i2} b_{2j} + \dots + a_{im} b_{mj}.$$

**Beispiel 6.8** Wir nehmen als  $f$  die gleiche lineare Abbildung wie in Beispiel 6.7 und für  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^3$  die Abbildung mit  $g(e_1) = (2, 0, 1)^T$  und  $g(e_2) = (2, 1, 0)^T$ ,

$$A_f = \begin{pmatrix} 1 & 2 & -1 \\ 1 & 1 & 3 \end{pmatrix}, \quad B_g = \begin{pmatrix} 2 & 2 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Wir bestimmen das Produkt nach der Regel Zeile mal Spalte

$$C = \begin{pmatrix} 1 \cdot 2 + 2 \cdot 0 + (-1) \cdot 1 & 1 \cdot 2 + 2 \cdot 1 + (-1) \cdot 0 \\ 1 \cdot 2 + 1 \cdot 0 + 3 \cdot 1 & 1 \cdot 2 + 1 \cdot 1 + 3 \cdot 0 \end{pmatrix} = \begin{pmatrix} 1 & 4 \\ 5 & 3 \end{pmatrix}$$

□

**6.3 Der Matrizenkalkül** Der Raum  $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^m)$  mit den Operationen der punktweisen Addition  $(f+g)(x) = f(x) + g(x)$  und der punktweisen Skalarmultiplikation  $(\alpha f)(x) = \alpha f(x)$  sind zum Matrizenraum  $\mathbb{K}^{m \times n}$  isomorph mit dem Isomorphismus  $I : f \mapsto A_f$ . Der punktweisen Addition im Raum  $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^m)$  entspricht genau die Matrizenaddition, bei der Skalarmultiplikation ist es genauso. Ferner können wir den Raum  $\mathbb{K}^n$  mit dem Raum der Spaltenmatrizen  $\mathbb{K}^{n \times 1}$  identifizieren. Im  $\mathbb{K}^n$  haben wir damit zwei völlig äquivalente Begriffe, nämlich Vektoren und lineare Abbildungen auf der einen Seite und Matrizen auf der anderen Seite. Wir können daher die Herkunft aus der linearen Algebra vergessen und eine reine Matrizenrechnung betreiben, zumal wir jetzt nicht mehr zwischen Vektoren und Matrizen unterscheiden müssen. Letztere sind jetzt ebenfalls Matrizen und der Auswertung  $f(u)$  entspricht das Matrizenprodukt  $Ax$  zwischen der Darstellungsmatrix  $A$  und der Matrix  $x \in \mathbb{K}^{n \times 1}$ . Dennoch werden wir zur besseren Unterscheidbarkeit von anderen Matrizen die  $\mathbb{K}^{n \times 1}$ -Matrizen als Vektoren bezeichnen.

Auf dem Matrizenraum  $\mathbb{K}^{m \times n}$  sind daher Addition und Skalarmultiplikation sowie das Matrizenprodukt zwischen Elementen aus  $\mathbb{K}^{l \times m}$  und  $\mathbb{K}^{m \times n}$  definiert mit Ergebnis im Raum  $\mathbb{K}^{l \times n}$ .

**Satz 6.9** Sofern die Operationen definiert sind, gelten die folgenden Rechenregeln:

(a) Das Matrizenprodukt ist assoziativ

$$(AB)C = A(BC).$$

(b) Es gelten die Distributivgesetze

$$A(B+C) = AB + AC, \quad (A+B)C = AC + BC.$$

(c) Matrix- und skalare Multiplikation sind homogen

$$\alpha \cdot AB = (\alpha A)B = A(\alpha B).$$

*Beweis:* (a) Hier greift man besser auf die Herkunft des Matrizenprodukts als Komposition linearer Abbildungen zurück. Letztere ist bekanntlich assoziativ.

(b) Das gilt ebenfalls für lineare Abbildungen, folgt aber auch direkt aus der Definition „Zeile  $\times$  Spalte“.

(c) Das ist trivial, ist doch  $\alpha A = (\alpha a_{ij})$   $\square$

Im Fall  $m = n$  sprechen wir von *quadratischen Matrizen*  $A \in \mathbb{K}^{n \times n}$ . Die Matrizenprodukte  $AB$  und  $BA$  sind hier zwar definiert, stimmen in der Regel aber nicht überein wie das folgende Beispiel zeigt:

**Beispiel 6.10** Es ist

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

$\square$

Die *Einheitsmatrix* von  $\mathbb{K}^{n \times n}$  ist

$$E_n = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & & 1 \end{pmatrix}$$

und entspricht der Identität im Raum  $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ . Es gilt  $AE_n = E_nA = A$ .

Eine Matrix  $A \in \mathbb{K}^{n \times n}$  heißt *regulär* oder *invertierbar*, wenn es eine Matrix  $A^{-1} \in \mathbb{K}^{n \times n}$  gibt mit  $AA^{-1} = E_n$ . Eine reguläre Matrix ist die Darstellungsmatrix eines Isomorphismus des  $\mathbb{K}^n$ ,  $A^{-1}$  ist demnach die Darstellungsmatrix des inversen Isomorphismus. Daher gilt im Falle einer regulären Matrix auch  $A^{-1}A = E_n$ . Die regulären Matrizen bilden mit der Matrix-Multiplikation eine Gruppe mit neutralem Element  $E_n$ , die mit  $GL(n, \mathbb{K})$  notiert und allgemeine lineare Gruppe (engl: general linear group) genannt wird.

Wir können viele Begriffsbildungen für lineare Abbildungen auf Matrizen  $A \in \mathbb{K}^{m \times n}$  übertragen: Das *Bild* von  $A$  ist die Menge der durch  $Ax$  erzeugten Elemente

$$\text{Bild } A = \{y = Ax : x \in \mathbb{K}^n\} = \{y = \sum_{i=1}^n x_i a_i, x_i \in \mathbb{K}\},$$

wobei  $a_1, \dots, a_n \in \mathbb{K}^m$  die Spalten von  $A$  bezeichnet. Der *Rang* von  $A$  ist die Dimension des Bildraums, also die Anzahl der linear unabhängigen Spalten von  $A$ . Da wir auch von der Anzahl der linear unabhängigen Zeilen sprechen können, wird der Rang im Zusammenhang mit Matrizen auch als *Spaltenrang* bezeichnet.

Der *Kern* ist die Menge der Vektoren, die auf die Null abgebildet werden,

$$\text{Kern } A = \{x : Ax = 0\}.$$

Die Rangformel für  $A \in \mathbb{K}^{m \times n}$  ist dann

$$(6.2) \quad n = \dim \text{Kern } A + \text{rang } A.$$

**Lemma 6.11** Es gilt  $\text{rang } A \leq \min\{m, n\}$ .

*Beweis:* Für  $m \geq n$  ist das richtig, wir haben ja nur  $n$  Spalten zur Verfügung. Andererseits ist der Bildraum ein Teilraum des  $\mathbb{K}^m$ . Dort kann es höchstens  $m$  linear unabhängige Vektoren geben.  $\square$

Eine quadratische Matrix  $A \in \mathbb{K}^{n \times n}$  ist genau dann regulär, wenn  $\text{rang } A = n$ . Nur in diesem Fall ist die zugehörige Selbstabbildung surjektiv und damit bijektiv.

**Beispiel 6.12** Wir untersuchen die folgenden Matrizen in  $\mathbb{K} = \mathbb{C}$ ,

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & i & 1 \\ 0 & 1 & -i \\ 1 & 0 & 0 \end{pmatrix}.$$

In der Matrix  $A$  sind die Spaltenvektoren  $a_1$  und  $a_3$  l.u.. Ferner ist  $a_4 = a_1 - a_3$ . Damit gilt  $\text{Bild } A = \text{span}\{a_1, a_3\}$ . Nach der Rangformel ist  $\dim \text{Kern } A = 2$ . Als Basis des Kerns können wir die Vektoren  $e_2, (1, 0, -1, -1)^T \in \mathbb{C}^4$  nehmen.

In der Matrix  $B$  gilt  $b_2 = ib_3$ , damit ist  $\text{Bild } A = \text{span}\{b_1, b_2\}$ . Der Kern ist daher eindimensional, offenbar ist  $(0, 1, -i)^T \in \text{Kern } B$ .  $\square$

**Satz 6.13** (a) Für Matrizen  $A : \mathbb{K}^{l \times m}$ ,  $B : \mathbb{K}^{m \times n}$  gilt

$$\text{rang } AB \leq \min\{\text{rang } A, \text{rang } B\}.$$

(b) Sei  $A : \mathbb{K}^{m \times n}$ ,  $B : \mathbb{K}^{n \times m}$ ,  $C : \mathbb{K}^{n \times n}$  mit  $\text{rang } B = m$ ,  $\text{rang } C = n$ . Dann gilt

$$\text{rang } BA = \text{rang } A, \quad \text{rang } AC = \text{rang } A.$$

*Beweis:* (a) Ist  $\text{rang } A \leq \text{rang } B$ , so gilt  $\text{Bild } AB \subset \text{Bild } A$ . Ist umgekehrt  $\text{rang } B \leq \text{rang } A$ , so bildet  $A$  das Bild von  $B$  auf einen Unterraum der Dimension  $\leq \text{rang } B$  ab.

(b) Wir können die Matrizen als Darstellungsmatrizen der zugehörigen linearen Abbildungen interpretieren. Dann sind  $B$  und  $C$  Isomorphismen zwischen den angegebenen Räumen.  $B$  bildet das Bild von  $A$  auf einen Unterraum gleicher Dimension ab. Das Bild von  $C$  spannt den ganzen  $\mathbb{K}^n$  auf. In diesem Fall gilt sogar  $\text{Bild } A = \text{Bild } AC$ .  $\square$

Beispiel 6.10 zeigt, dass man in (a) nicht mehr zeigen kann. In diesem Beispiel haben die Matrizen  $A$  und  $B$  jeweils den Rang 1 und für die Produkte gilt  $\text{rang } AB = 1$ , aber  $\text{rang } BA = 0$ .

Für eine Matrix  $A = (a_{ij}) \in \mathbb{K}^{m \times n}$  ist die *transponierte Matrix*  $A^T = (a_{ij}^T) \in \mathbb{K}^{n \times m}$  definiert durch  $a_{ij}^T = a_{ji}$  für alle  $1 \leq i \leq m$ ,  $1 \leq j \leq n$ . Anschaulich klappt man die Matrix  $A$  von links unten nach rechts oben:

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}, \quad A^T = \begin{pmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \\ 4 & 8 \end{pmatrix}.$$

Sind  $A, B$  Matrizen über  $\mathbb{K}$ , für die das Produkt  $AB$  erklärt ist, so gilt

$$(AB)^T = B^T A^T.$$

Mit  $C = AB$  haben wir nämlich

$$c_{ij} = \sum_k a_{ik} b_{kj} \Rightarrow c_{ij}^T = c_{ji} = \sum_k a_{jk} b_{ki} = \sum_k b_{ik}^T a_{kj}^T = (B^T A^T)_{ij}.$$

Eine ähnliche Formel gilt für die Inverse des Produkts zweier regulärer Matrizen  $A, B \in \mathbb{K}^{n \times n}$ :

$$(AB)^{-1} = B^{-1} A^{-1},$$

wegen

$$(AB)(B^{-1} A^{-1}) = A(B^{-1} B)A^{-1} = E_n.$$

Für eine reguläre Matrix  $A \in \mathbb{K}^{n \times n}$  ist es gleichgültig, ob man zuerst transponiert und dann invertiert oder umgekehrt:

$$(A^{-1})^T = (A^T)^{-1}$$

wegen

$$(A^{-1})^T A^T = (A A^{-1})^T = E_n,$$

also ist  $A^T$  die Inverse von  $(A^{-1})^T$ . Wir schreiben daher auch  $A^{-T}$  anstatt  $(A^{-1})^T$  oder  $(A^T)^{-1}$ .

## 7 Lineare Gleichungssysteme und Determinanten

**7.1 Dreiecks- und Diagonalmatrizen** Linke untere bzw. rechte obere *Dreiecksmatrizen* sind quadratische Matrizen der Gestalt

$$L = \begin{pmatrix} * & & & \\ & \ddots & 0 & \\ & * & \ddots & \\ & & & * \end{pmatrix}, \quad R = \begin{pmatrix} * & & & \\ & \ddots & * & \\ 0 & & \ddots & \\ & & & * \end{pmatrix}.$$

Genauer heißt eine quadratische Matrix  $L = (l_{ij})$  linke untere Dreiecksmatrix, wenn  $l_{ij} = 0$  für  $i < j$ .  $R = (r_{ij})$  heißt rechte obere Dreiecksmatrix, wenn  $r_{ij} = 0$  für  $i > j$ .

Der Raum der linken unteren bzw. rechten oberen Dreiecksmatrizen ist abgeschlossen gegenüber allen bisher definierten Operationen. Sind  $L_1, L_2$  linke untere Dreiecksmatrizen gleicher Dimension und  $\alpha \in \mathbb{K}$ , so sind auch

$$L_1 + L_2, \quad \alpha L, \quad L_1 L_2, \quad L_1^{-1} \text{ falls } L_1 \text{ regulär,}$$

linke untere Dreiecksmatrizen.

**Lemma 7.1** Eine rechte obere (linke untere) Dreiecksmatrix ist genau dann regulär, wenn  $r_{ii} \neq 0$  ( $l_{ii} \neq 0$ ) für alle  $1 \leq i \leq n$ .

*Beweis:* Wir zeigen das nur für eine rechte obere Dreiecksmatrix. Ist  $r_{11} = 0$ , so ist die erste Spalte Null und die Matrix besitzt  $e_1$  als nichttrivialen Vektor im Kern. Sind  $r_{11}, \dots, r_{kk} \neq 0$ , aber  $r_{k+1, k+1} = 0$ , so sind die Spaltenvektoren  $r_1, \dots, r_{k+1}$  l.a., denn diese Vektoren liegen de facto in einem  $\mathbb{K}^k$ , wenn man nur die oberen  $k$  Komponenten betrachtet.

Die umgekehrte Richtung beweisen wir in Beispiel 7.4.  $\square$

Quadratische Matrizen  $D = (d_{ij})$  mit  $d_{ij} = 0$  für  $i \neq j$  heißen *Diagonalmatrizen*. Wir schreiben auch kurz  $\text{diag } D = (d_1, \dots, d_n)$ , geben also nur die Elemente auf der Hauptdiagonalen an, z.B.:

$$D = \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix} \Leftrightarrow \text{diag } D = (d_1, d_2, d_3).$$

Auch der Raum der Diagonalmatrizen ist abgeschlossen gegenüber Addition, Skalarmultiplikation und Matrizenmultiplikation. Das letzte Lemma gilt auch für Diagonalmatrizen, sind diese doch gleichzeitig rechte obere und linke untere Dreiecksmatrizen.

**7.2 Problemstellung** Ein lineares Gleichungssystem (LGS) über einem Körper  $\mathbb{K}$  besteht aus einer Matrix  $A \in \mathbb{K}^{m \times n}$  und einem Vektor  $b \in \mathbb{K}^m$ , der rechte Seite genannt wird. Gesucht wird ein  $x \in \mathbb{K}^n$  mit

$$(7.1) \quad Ax = b$$

oder ausgeschrieben

$$a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1$$

$$a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2$$

$$\vdots \qquad \vdots$$

$$a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m.$$

Die Gleichungen müssen alle erfüllt sein, nur dann sprechen wir davon, dass das LGS lösbar ist.

**Satz 7.2** (a) Das Gleichungssystem ist genau dann lösbar, wenn  $b \in \text{Bild } A$ .

(b) Dies ist genau dann der Fall, wenn

$$\text{rang } A = \text{rang } (A|b),$$

wobei die Matrix  $(A|b) \in \mathbb{K}^{m \times (n+1)}$  von der Form  $(a_1 | \dots | a_n | b)$  ist. Die Spalten von  $A$  werden durch  $b$  als Spalte  $n+1$  ergänzt.

(c) Ist das Gleichungssystem lösbar, so ist die Lösungsmenge von der Form  $x + \text{Kern } A$ , wobei  $x$  eine beliebige Lösung von  $Ax = b$  ist.

*Beweis:* (a) Das Gleichungssystem wird durch eine Linearkombination der Form  $\sum_{i=1}^n x_i a_i = b$  gelöst, wobei  $a_i \in \mathbb{R}^m$  die Spaltenvektoren von  $A$  bezeichnen. Dies ist aber gleichbedeutend damit, dass  $b \in \text{Bild } A$  ist.

(b) Werden die Spaltenvektoren um den Vektor  $b$  ergänzt, dann vergrößert er  $\text{span } \{a_1, \dots, a_n\}$  oder er tut das nicht. Im ersten Fall ist  $\dim \text{Bild } (A|b) > \dim \text{Bild } A$ , also  $\text{rang } (A|b) > \text{rang } A$ , und  $b$  liegt nicht in  $\text{Bild } A$ . Im zweiten Fall ist  $\text{rang } (A|b) = \text{rang } A$  und  $b$  liegt in  $\text{Bild } A$ .

(c) Wir können auf eine Lösung  $x$  einen beliebigen Vektor  $y \in \text{Kern } A$  addieren und es gilt  $A(x+y) = Ax + Ay = b$ . Haben wir zwei Lösungen  $x, x'$ , also  $Ax = Ax' = b$ , so folgt  $A(x-x') = 0$ . Damit unterscheiden sich zwei Lösungen nur um einen Vektor im Kern. Die Lösungsmenge ist von der Form  $x + \text{Kern } A$  wie angegeben.  $\square$

**Korollar 7.3** Das LGS (7.1) ist genau dann eindeutig lösbar, wenn

$$\text{rang } A = \text{rang } (A|b) = n = \text{Anzahl der Spalten von } A.$$

*Beweis:* Das erste Gleichheitszeichen ist das Lösbarkeitskriterium aus dem letzten Satz. Wenn  $\text{rang } A = n$ , so folgt aus der Rangformel (6.2), dass  $\dim \text{Kern } A = 0$ , also  $\text{Kern } A = \{0\}$ .  $\square$

**7.3 Der Gauß-Algorithmus** Im Folgenden ist immer  $A \in \mathbb{K}^{m \times n}$ ,  $b \in \mathbb{K}^m$  und zu lösen ist das LGS  $Ax = b$  für  $x \in \mathbb{K}^n$ . Die Idee des Algorithmus besteht darin, durch Umformungen mittels regulärer Matrizen  $B_i \in \mathbb{K}^{m \times m}$  das LGS auf eine einfachere Gestalt zu bringen. Es gilt ja

$$Ax = b \Leftrightarrow B_i Ax = B_i b,$$

die Lösungsmenge ändert sich nicht, sofern die Matrizen  $B_i$  regulär sind.

Ziel der Umformungen ist es, eine *Zeilenstufenform* (ZSF) für die umgeformte Matrix zu erreichen. Dabei liegt eine Matrix in ZSF vor, wenn für alle  $i = 1, \dots, m-1$  gilt:

Zeile  $i+1$  besitzt mehr führende Nullen als Zeile  $i$ , es sei denn, dass Zeile  $i$  eine Nullzeile ist. In diesem Fall muss auch Zeile  $i+1$  eine Nullzeile sein.

**Beispiele 7.4** In den folgenden beiden Beispielen von ZSF-Matrizen bezeichnen wir mit  $*$  ein Element  $\neq 0$  und mit  $a$  ein beliebiges Element.

$$A = \begin{pmatrix} * & a & a & a \\ 0 & a & * & a \\ 0 & 0 & 0 & a \end{pmatrix}, \quad R = \begin{pmatrix} * & a & a \\ 0 & * & a \\ 0 & 0 & * \end{pmatrix}.$$

In der Matrix  $A$  ist  $a_{11} \neq 0$ . Da alle anderen Zeilen mehr führende Nullen besitzen müssen, verschwindet die ganze erste Spalte unterhalb von  $a_{11}$ .

Die Lösung von  $Rx = b$  bestimmt man von unten nach oben, beginnt also mit der  $n$ -ten Gleichung  $r_{nn}x_n = b_n$ . Daraus bestimmt man  $x_n = b_n/r_{nn}$  und setzt diesen Wert in die oberen Gleichungen ein. Dann geht man zur Gleichung  $n-1$  und bestimmt daraus  $x_{n-1}$ . Wir können daher immer eindeutig lösen, was den Beweis von Lemma 7.1 komplettiert.  $\square$

Beim Gauß-Algorithmus für quadratische Matrizen erreicht man immer eine solche rechte obere Dreiecksmatrix, wenn die Ausgangsmatrix regulär ist. Aus dem hier beschriebenen Verfahren zur Lösung von  $Rx = b$  wird klar, dass man bei einer allgemeinen Matrix in ZSF ähnlich verfahren kann, was diese Form erstrebenswert macht.

Wir besprechen nun die zwei Typen von Äquivalenzumformungen, die im Gauß-Algorithmus benötigt werden:

Typ I: Vertauschung zweier Zeilen  $i$  und  $j$ , wird geleistet mit der Multiplikation von links mit der *Permutationsmatrix*

$$P_{ij} = \begin{pmatrix} E_{i-1} & & & \\ & 0 & & 1 \\ & & E_{j-i-1} & \\ 1 & & & 0 \\ & & & E_{m-i} \end{pmatrix}.$$

Anders ausgedrückt: In der Einheitsmatrix wird  $d_{ii}, d_{jj} = 0$  gesetzt sowie  $d_{ij} = d_{ji} = 1$ .  $P_{ij}$  ist regulär, weil die Spalten von  $P_{ij}$  aus den kanonischen Einheitsvektoren bestehen.

Typ II: Addition des  $\alpha$ -fachen der Zeile  $j$  auf die Zeile  $i > j$ , wird geleistet durch Multiplikation von links mit der Matrix

$$G_{ij,\alpha} = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & \alpha & \ddots \\ & & & 1 \end{pmatrix} \quad \text{für } \alpha \text{ an Position } i, j.$$

Als linke untere Dreiecksmatrix mit nichtverschwindenden Diagonalelementen ist  $G_{ij,\alpha}$  nach Lemma 7.1 regulär.

Wir erläutern den Gauß-Algorithmus an Hand des Beispiels in  $\mathbb{K} = \mathbb{R}$

$$A = \begin{pmatrix} 0 & 4 & 8 \\ 1 & 4 & 4 \\ 2 & 4 & 0 \end{pmatrix}, \quad b_1 = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}, \quad b_2 = \begin{pmatrix} 12 \\ 8 \\ 6 \end{pmatrix}.$$

Wir wollen simultan die beiden Gleichungssysteme  $Ax_1 = b_1$ ,  $Ax_2 = b_2$  lösen. Zur Durchführung des Gauß-Algorithmus schreiben wir die um die rechten Seiten erweiterte Matrix  $(A|b_1|b_2) \in \mathbb{R}^{3 \times 5}$  auf:

$$(A|b_1|b_2) = \left( \begin{array}{ccc|cc} 0 & 4 & 8 & 4 & 12 \\ 1 & 4 & 4 & 5 & 8 \\ 2 & 4 & 0 & 6 & 6 \end{array} \right) \quad \rightarrow \quad \left( \begin{array}{ccc|cc} 2 & 4 & 0 & 6 & 6 \\ 1 & 4 & 4 & 5 & 8 \\ 0 & 4 & 8 & 4 & 12 \end{array} \right) \quad \rightarrow \quad \left( \begin{array}{ccc|cc} 2 & 4 & 0 & 6 & 6 \\ 0 & 2 & 4 & 2 & 5 \\ 0 & 4 & 8 & 4 & 12 \end{array} \right).$$

Wir nehmen uns Spalte für Spalte vor, um im unteren Bereich der Spalte möglichst viele Nullen zu produzieren. Ist eine Spalte komplett Null, gehen wir zur nächsten über. Für die erste Spalte bedeutet diese Vorgehensweise, dass nach erfolgter Umformung höchstens  $a_{11} \neq 0$  ist. In unserem konkreten Fall müssen wir Zeile 2 oder Zeile 3 mit Zeile 1 vertauschen. In einem endlichen Körper wäre es gleichgültig, welche Zeile wir vertauschen. Bei Rechnung in  $\mathbb{R}$  oder  $\mathbb{C}$  durch ein Computerprogramm fallen Rundungsfehler an. In diesem Fall soll man das betragsmäßig größte Element der Spalte nach oben bringen. Wir tun das auch hier (=Typ I) und erhalten die Matrix oben in der Mitte.

Im nächsten Schritt ziehen wir das 1/2-fache der ersten Zeile von der zweiten ab (=Typ II). Weil man das betragsgrößte Element zur Elimination genommen hat, hält man die Elemente klein, die auf die zweite Zeile addiert werden (Matrix oben rechts).

$$(7.2) \quad \left( \begin{array}{ccc|cc} 2 & 4 & 0 & 6 & 6 \\ 0 & 4 & 8 & 4 & 12 \\ 0 & 2 & 4 & 2 & 5 \end{array} \right) \quad \rightarrow \quad \left( \begin{array}{ccc|cc} 2 & 4 & 0 & 6 & 6 \\ 0 & 4 & 8 & 4 & 12 \\ 0 & 0 & 0 & 0 & -1 \end{array} \right)$$

Da man die Nullen in der ersten Spalte erhalten möchte, wird nun mit dem zweiten Element in der zweiten Spalte eliminiert. Auch hier vertauschen wir die Zeile 2 mit 3, um das betragsgrößte Element nach oben zu bringen. Schließlich wird die zweite Spalte eliminiert, indem das  $-1/2$ -fache der zweiten Zeile auf die dritte Zeile addiert wird. Da in der neuen Matrix  $a_{33} = 0$  gilt, besitzt die Matrix nur Rang 2 und die beiden ersten Spalten bilden eine Basis des Bildes.

Für das LGS  $Ax_1 = b_1$  liegt die rechte Seite im Bild der ersten beiden Spaltenvektoren. Wir können daher  $x_{1,3} = 0$  setzen und erhalten mit  $x_{1,2} = x_{1,2} = 1$  eine Lösung. Wegen  $\text{rang } A = 2$  ist  $\dim \text{Kern } A = 1$ . Da die dritte Spalte von den ersten beiden abhängig ist, gilt  $a_3 = \alpha_1 a_1 + \alpha_2 a_2$ . Wir finden daher eine Kernfunktion, indem wir mit  $y_{h,3} = 1$  eine Lösung des homogenen Problems  $Ay_h = 0$  bestimmen. Aus der zweiten Gleichung (mit rechter Seite 0) folgt  $y_{h,2} = -2$  und aus der ersten schließlich  $y_{h,1} = 4$ . Der Lösungsraum ist daher

$$\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \alpha \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix}, \quad \alpha \in \mathbb{R}.$$

Das Gleichungssystem  $Ax_2 = b_2$  ist nach (7.2) unlösbar, weil einer Nullzeile das Element  $-1$  auf der rechten Seite gegenübersteht.

In dem hier betrachteten Beispiel war die Sache natürlich sehr übersichtlich. Für kompliziertere Situationen verwendet man den folgenden Satz.

**Satz 7.5** Für eine ZSF-Matrix  $Z \in \mathbb{K}^{m \times n}$  gilt

$$\text{rang } Z = \text{Anzahl der Zeilen mit nichtverschwindenden Elementen.}$$

*Beweis:* Wir streichen aus  $Z$  alle Zeilen, die aus lauter Nullen bestehen. Es verbleibt eine Matrix  $Z' \in \mathbb{K}^{r \times n}$ , die den gleichen Rang wie  $Z$  besitzt. Aus den Spalten von  $Z'$  wählen wir diejenigen aus, die ein nichtverschwindendes Element in einer Zeile haben, das führend in dieser Zeile ist, die zugehörige Zeile also von der Form  $(0, \dots, 0, *, \dots)$  ist.  $*$  ist dann das führende Element und die zugehörige Spalte wird ausgewählt. Es entsteht eine  $(r \times r)$ -Matrix  $Z''$ , die eine rechte obere Dreiecksmatrix ist mit nichtverschwindenden Elementen in der Haupdiagonalen. Diese Matrix ist regulär, also ist  $\text{rang } Z' \geq r$ . Die  $(r \times n)$ -Matrix  $Z'$  kann aber höchstens Rang  $r$  haben, also ist  $\text{rang } Z' = r$ .  $\square$

**Beispiel 7.6** Wir betrachten in  $\mathbb{Z}_2$  das lineare Gleichungssystem

$$Zx = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} x = b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \Rightarrow Z'' = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 1 & 3 & 4 \end{pmatrix}$$

Hier liegt die Systemmatrix  $Z$  bereits in ZSF vor, rechts ist die Matrix  $Z''$  aus dem Beweis des letzten Satzes angegeben, zusammen mit der Spaltennummer, die jeder Spaltenvektor in  $Z$  besitzt. Die führenden Einser liegen ja in der ersten Zeile in Spalte 1, in der zweiten Zeile in Spalte 3 und schließlich in der letzten Zeile in Spalte 4. Die Spalten der Matrix  $Z''$  bilden eine Basis des Bildes. Im allgemeinen Fall ist das LGS bereits durch eine Linearkombination der Spalten von  $Z''$  lösbar oder überhaupt unlösbar. Im vorliegenden Fall ist  $\text{rang } Z = \text{rang } Z'' = 3$ . Wir haben also Vollrang und das LGS ist für jede rechte Seite lösbar. Das LGS  $Z''y = b$  besitzt die eindeutige Lösung  $y = (0, 1, 1)^T$ . Wir müssen nun noch beachten, zu welchen Spalten von  $Z$  die Komponenten dieses Vektors gehören, und wir erhalten mit  $x = (0, 0, 1, 1, 0, 0)^T$  eine Lösung von  $Zx = b$ .

Nach der Rangformel benötigen wir drei linear unabhängige Kernfunktionen, die wir den Spaltenindizes von  $Z$  zuordnen, die nicht in  $Z'$  auftreten. In unserem Fall sind das 2, 5, 6. Wir garantieren

die lineare Unabhängigkeit der Kernfunktionen, indem wir eine dieser Komponenten = 1 setzen und die anderen = 0. Wir bringen die zugehörige Spalte von  $Z$  auf die andere Seite und lösen dann  $Z'y_i = -z_i$  für  $i = 2, 5, 6$ . Damit haben wir die rechten Seiten und Lösungen (man beachte  $1 + 1 = 0$ )

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \Rightarrow y_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \Rightarrow y_5 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \Rightarrow y_6 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

Insgesamt erhalten wir die Lösungsmenge

$$\left\{ x = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \alpha_1 \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + \alpha_2 \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad \alpha_1, \alpha_2, \alpha_3 \in \mathbb{F}_2 \right\}$$

□

Viele weitere Probleme lassen sich mit Hilfe des Gauß-Algorithmus lösen:

**Testen auf linear Unabhängigkeit, Rangbestimmung** Haben wir Vektoren  $v_1, \dots, v_k \in \mathbb{K}^m$  und wollen sie auf lineare Unabhängigkeit testen, so stellen wir sie zu einer  $(m \times k)$ -Matrix  $V = (v_1 | v_2 | \dots | v_k)$  zusammen. Wir bringen sie auf ZSF und bestimmen den Rang der zugehörigen ZSF-Matrix mit dem letzten Satz.

**Inverse einer Matrix** Wollen wir zu einer Matrix  $\in \mathbb{K}^{n \times n}$  die Inverse  $A^{-1}$  bestimmen, so führen wir den simultanen Gauß-Algorithmus mit Matrix  $A$  und rechten Seiten  $e_1, \dots, e_n$  durch. Die Lösungen  $v_i$  mit  $Av_i = e_i$  stellen wir (wie immer als Spaltenvektoren) zur Matrix  $A^{-1} = (v_1 | \dots | v_n)$  zusammen. Wie man aus der Regel Zeile mal Spalte erkennt, gilt nun in der Tat  $AA^{-1} = E_n$ .

**Zeilenrang=Spaltenrang** Wir bringen eine beliebige Matrix  $A \in \mathbb{K}^{m \times n}$  auf ZSF,  $Z = BA$  mit einer regulären  $(m \times m)$ -Matrix  $B$ .

$$Z = \begin{pmatrix} * & a & a & a & a \\ 0 & 0 & * & a & a \\ 0 & 0 & 0 & * & a \end{pmatrix} \quad \rightarrow \quad Z' = ZC = \begin{pmatrix} * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & 0 & 0 \\ 0 & 0 & 0 & * & 0 \end{pmatrix}$$

In der Beispielmatrix links bezeichnen die Sterne Elemente ungleich 0, die mit  $a$  bezeichneten Elemente sind beliebig. Wir arbeiten nun zeilenweise von oben nach unten und beginnen mit der ersten Zeile. Durch Addition eines Vielfachen der ersten Spalte auf die zweite Spalte erzielen wir auf der Position (1, 2) eine Null. Dies ist aber gleichbedeutend mit der Multiplikation einer Matrix vom Typ II von rechts. Auf diese Weise erzeugen wir lauter Nullen rechts von  $*$  in der ersten Zeile. Für die übrigen Zeilen verfahren wir genauso und erhalten schließlich die Matrix  $Z'$  rechts. Es gilt  $Z' = ZC = BAC$  mit regulären Matrizen  $B, C$ , Nach Satz 6.13(b) haben  $Z'$  und  $A$  den gleichen Rang. Die Matrix  $Z'^T$  hat ebenfalls ZSF. Auch hier können wir Satz 7.5 anwenden. Wir haben bewiesen:

**Satz 7.7** Für eine beliebige Matrix  $A \in \mathbb{K}^{m \times n}$  gilt

$$\text{rang } A = \text{Anzahl der l.u Zeilen von } A = \text{Zeilenrang von } A,$$

oder anders ausgedrückt  $\text{rang } A = \text{rang } A^T$ .

**7.4 Permutationen** Eine *Permutation*  $\pi$  ist nach Abschnitt 2.4 eine bijektive Abbildung der Menge  $\{1, \dots, n\}$  auf sich selbst. Statt  $\pi(i) = a_i \in \{1, \dots, n\}$  schreiben wir kürzer  $(a_1, a_2, \dots, a_n)$  und stellen uns dabei vor, dass wir die Zahlen  $a_1, \dots, a_n$  auf die „Fächer“  $1, 2, \dots, n$  verteilen.

Die Vertauschung zweier benachbarter Fächer nennen wir eine elementare Permutation. Jede Permutation lässt sich durch eine Folge elementarer Permutationen aus der Identität  $(1, 2, \dots, n)$  erzeugen. Wir bringen als erstes das Element  $a_1$  an die erste Position und verfahren mit den folgenden Elementen genauso. Beispielsweise erzeugen wir  $(2, 4, 1, 3)$  durch

$$(1, 2, 3, 4) \rightarrow (2, 1, 3, 4) \rightarrow (2, 1, 4, 3) \rightarrow (2, 4, 1, 3).$$

Die Anzahl der *Fehlstellen* einer Permutation  $\pi$  ist

$$F(\pi) = |\{(i, j) : i < j \text{ und } \pi(i) > \pi(j) \text{ für } 1 \leq i < j \leq n\}|$$

Wie immer bezeichnen wir mit  $|M|$  die Kardinalität der Menge  $M$ . Die Anzahl der Fehlstellen der Identität  $(1, 2, \dots, n)$  ist Null, die Anzahl der Fehlstellen der Permutation  $(n, n-1, \dots, 1)$  ist  $\frac{1}{2}n(n-1)$ .

Das *Signum* oder *Vorzeichen* einer Permutation  $\pi$  ist

$$\text{sign}(\pi) = \begin{cases} 1 & \text{falls } F(\pi) \text{ gerade} \\ -1 & \text{falls } F(\pi) \text{ ungerade} \end{cases}.$$

Wir nennen eine Permutation gerade, wenn  $\text{sign}(\pi) = 1$ , andernfalls ungerade. Jede Permutation lässt sich auf vielfältige Weise durch Hintereinanderschaltung von einfachen Permutationen, bei denen nur  $\pi(i)$  und  $\pi(j)$  vertauscht werden, erzeugen. Es ist interessant und zunächst gar nicht offensichtlich, dass man bei einer geraden Permutation immer eine gerade Zahl von einfachen Permutationen benötigt, um sie zu erzeugen. Jede einfache Permutation lässt sich nur durch eine ungerade Zahl von elementaren Permutationen erzeugen. Bei einer elementaren Permutation ändert sich die Anzahl der Fehlstellen um  $\pm 1$ , gleiches gilt demnach auch für eine einfache Permutation.

**7.5 Determinanten** Die Determinante ist eine Abbildung  $\det : \mathbb{K}^{n \times n} \rightarrow \mathbb{K}$ . Historisch gesehen hat es verschiedene äquivalente Definitionen gegeben. Wir wählen hier die Leibnizsche Definition aus, mit der man die Determinante unmittelbar berechnen kann.

Wir definieren die Determinante durch

$$(7.3) \quad \det A = \sum_{\pi} \text{sign}(\pi) \prod_{i=1}^n a_{i\pi(i)}.$$

In jedem einzelnen Summanden kommt aus jeder Zeile und jeder Spalte der Matrix nur ein Element vor.

Bei  $n = 2$  gibt es nur zwei Permutationen, nämlich die Identität mit positivem Signum und die Vertauschung von 1 und 2 mit negativem Signum. Daher gilt für  $A = (a_{ij})_{1 \leq i, j \leq 2}$

$$\det A = a_{11}a_{22} - a_{12}a_{21}.$$

Für  $n = 3$  haben wir die drei einfachen Permutationen mit negativem Signum  $(1, 3, 2), (3, 2, 1), (2, 1, 3)$ , die übrigen haben Signum 1, nämlich  $(1, 2, 3), (2, 3, 1), (3, 1, 2)$ . Für eine  $(3 \times 3)$ -Matrix gilt daher

$$\det A = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{13}a_{22}a_{31} - a_{12}a_{21}a_{33}.$$

Es gilt offenbar  $\det E_n = 1$ . Für eine rechte obere (oder linke untere) Dreiecksmatrix  $R = (r_{ij})$  gilt

$$(7.4) \quad \det R = \prod_{i=1}^n r_{ii}$$

Das ist leicht einzusehen, weil nur Permutationen mit  $\pi(1) = 1$  nichtverschwindende Werte liefern können. Bei  $\pi(2)$  ist es ähnlich:  $\pi(2) = 1$  ist durch  $\pi(1)$  schon vergeben, bleibt also nur  $\pi(2) = 2$ , um etwas Nichtverschwindendes zu erreichen.

Für eine  $(n \times n)$ -Matrix  $A$  bezeichnen wir mit  $A_{ij} \in \mathbb{K}^{(n-1) \times (n-1)}$  die Matrix, die aus  $A$  durch Streichen der  $i$ -ten Zeile und der  $j$ -ten Spalte hervorgeht.

**Satz 7.8 (Entwicklungsatz von Laplace)** *Die Determinante einer Matrix  $A \in \mathbb{K}^{n \times n}$  lässt sich mit den folgenden Formeln nach einer Zeile oder einer Spalte „entwickeln“*

$$\det A = \sum_{j=1}^n (-1)^{i+j} a_{ij} \det A_{ij} \quad (\text{Entwicklung nach der } i\text{-ten Zeile})$$

$$\det A = \sum_{i=1}^n (-1)^{i+j} a_{ij} \det A_{ij} \quad (\text{Entwicklung nach der } j\text{-ten Spalte}).$$

Ein richtiger Satz ist das eigentlich nicht: Man stellt in der Leibnizschen Definition der Determinante nur fest, in welchen Summanden das Element  $a_{ij}$  vorkommt. Der Rechenaufwand zur Berechnung der Determinante nach der Laplace-Formel ist genauso gewaltig wie nach der Definition.

Für  $n = 3$  erhalten wir bei Entwicklung nach der ersten Spalte

$$\begin{aligned} \det A &= a_{11} \det \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} - a_{21} \det \begin{pmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{pmatrix} + a_{31} \det \begin{pmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{pmatrix} \\ &= a_{11}(a_{22}a_{33} - a_{32}a_{23}) - a_{21}(a_{12}a_{33} - a_{32}a_{13}) + a_{31}(a_{12}a_{23} - a_{22}a_{13}). \end{aligned}$$

Sind  $a_1, \dots, a_n \in \mathbb{K}^{n,1}$  die Spaltenvektoren von  $A$ , so schreiben wir  $A = (a_1 | a_2 | \dots | a_n)$

**Satz 7.9** *Die Determinante ist eine alternierende Multilinearform in den Spalten von  $A$ . Das heißt:*

(a) *Ist  $b \in \mathbb{K}^{n,1}$  ein beliebiger Spaltenvektor, so gilt*

$$\det(a_1 | \dots | a_{i-1} | a_i + b | a_{i+1} | \dots) = \det A + \det(a_1 | \dots | a_{i-1} | b | a_{i+1} | \dots).$$

(b) *Für  $\alpha \in \mathbb{K}$  gilt*

$$\det(a_1 | \dots | a_{i-1} | \alpha a_i | a_{i+1} | \dots) = \alpha \det A.$$

(c) *Besitzt  $A$  zwei identische Spalten, so gilt  $\det A = 0$ .*

*Beweis:* (a) In der Definition der Determinante (7.3) kommt in jedem Summanden genau ein Element der  $i$ -ten Spalte vor. Wenn also die  $i$ -te Spalte durch  $a_i + b$  ersetzt wird, ist dieses Element von der Form  $a_{ij} + b_j$  und kann mit dem Distributivgesetz auseinandergezogen werden.

(b) Wird die  $i$ -te Spalte durch  $\alpha a_i$  ersetzt, erscheint in jedem Summanden von (7.3) ein  $\alpha a_{ij}$  an Stelle von  $a_{ij}$ . Dieses  $\alpha$  kann daher aus der Summe ausgeklammert werden.

(c) Wir verwenden Induktion über  $n$ . Für  $n = 2$  ist die Behauptung richtig. Denn wenn  $A \in \mathbb{K}^{2 \times 2}$  zwei identische Spalten besitzt, verschwindet ihre Determinante. Für den Schluss  $n \rightarrow n + 1$  entwickeln wir die Determinante mit Satz 7.8 nach der ersten Zeile. Enthält  $A_{1,l}$  die beiden identischen Spalten, so verschwindet ihre Determinante nach Induktionsvoraussetzung. Sind  $l, l'$  die beiden identischen Spalten, so besitzen  $(-1)^{l+1} \det A_{l,1}$  und  $(-1)^{l'+1} \det A_{l',1}$  entgegengesetztes Vorzeichen.  $\square$

**Bemerkung 7.10** Die Determinante ist ebenso eine alternierende Multilinearform in den Zeilen der Matrix. Die Aussagen (a)-(c) im letzten Satz bleiben richtig, wenn man das Wort Spalte durch Zeile ersetzt. Die Beweise sind genauso einfach.  $\square$

**Bemerkung 7.11** In einer alternierenden Multilinearform wechselt das Vorzeichen, wenn die Komponenten vertauscht werden. Entsteht  $A'$  aus  $A$  durch Vertauschung zweier Zeilen oder Spalten, so gilt  $\det A = -\det A'$ . Zu beweisen brauchen wir dieses Prinzip nur für eine alternierende Multilinearform  $a(v_1, v_2)$  in zwei Komponenten. Es gilt

$$\begin{aligned} 0 &= a(v_1 + v_2, v_1 + v_2) = a(v_1 + v_2, v_1) + a(v_1 + v_2, v_2) = a(v_1, v_1) + a(v_2, v_1) + a(v_1, v_2) + a(v_2, v_2) \\ &= a(v_2, v_1) + a(v_1, v_2) \end{aligned}$$

$\square$

Man beachte, dass bei  $\alpha A$  das  $\alpha$  in jeder Zeile erscheint, daher  $\det \alpha A = \alpha^n \det A$ .

**Satz 7.12 (Multiplikationssatz für Determinanten)** Für  $(n \times n)$ -Matrizen  $A, B$  gilt  $\det AB = \det A \cdot \det B$ .

*Beweis:* Seien  $A = (a_{ik})$ ,  $B = (b_{kj})$ ,  $C = (c_{ij})$  mit  $c_{ij} = \sum_{k=1}^n a_{ik} b_{kj}$ . Mit der Definition der Determinante gilt dann

$$\begin{aligned} \det AB &= \sum_{\pi} \text{sign}(\pi) c_{1\pi(1)} \cdots c_{n\pi(n)} \\ &= \sum_{\pi} \text{sign}(\pi) \left( \sum_{j_1=1}^n a_{1j_1} b_{j_1\pi(1)} \right) \cdots \left( \sum_{j_n=1}^n a_{1j_n} b_{j_n\pi(n)} \right) \\ &= \sum_{\pi} \text{sign}(\pi) \sum_{1 \leq j_1, \dots, j_n \leq n} a_{1j_1} \cdots a_{nj_n} b_{j_1\pi(1)} \cdots b_{j_n\pi(n)} \\ &= \sum_{1 \leq j_1, \dots, j_n \leq n} a_{1j_1} \cdots a_{nj_n} \cdot \left[ \sum_{\pi} \text{sign}(\pi) b_{j_1\pi(1)} \cdots b_{j_n\pi(n)} \right] \end{aligned}$$

Den Ausdruck in den eckigen Klammern können wir interpretieren als die Determinante der Matrix, die in der ersten Zeile die Zeile  $j_1$  von  $B$ , in der zweiten Zeile die Zeile  $j_2$  von  $B$  besitzt, oder allgemein: In der  $i$ -ten Zeile steht die Zeile  $j_i$ . Bezeichnen wir diese Matrix mit  $B(j_1, \dots, j_n)$ , so ist

$$\det AB = \sum_{1 \leq j_1, \dots, j_n \leq n} a_{1j_1} \cdots a_{nj_n} \cdot \det B(j_1, \dots, j_n).$$

Nach Bemerkung 7.10 verschwindet  $B(j_1, \dots, j_n)$  höchstens dann nicht, wenn die Zeilen paarweise verschieden sind. Daher brauchen wir die Summe nur über alle Permutationen der Zahlen  $1, \dots, n$  zu erstrecken

$$\det AB = \sum_{\pi} a_{1\pi(1)} \cdots a_{n\pi(n)} \cdot \det B(\pi(1), \dots, \pi(n)).$$

Die Matrix  $B(\pi(1), \dots, \pi(n))$  geht aus einer Permutation der Zeilen von  $B$  hervor. Aus Bemerkung 7.11 wissen wir, dass

$$\det B(\pi(1), \dots, \pi(n)) = \text{sign } \pi \det B.$$

Damit

$$\det AB = \sum_{\pi} a_{1\pi(1)} \cdots a_{n\pi(n)} \cdot \text{sign } \pi \det B. = \det A \cdot \det B.$$

$\square$

**Korollar 7.13** Für eine reguläre Matrix gilt

$$\det A^{-1} = \frac{1}{\det A}.$$

*Beweis:* Nach der Multiplikationsformel gilt

$$1 = \det E_n = \det(AA^{-1}) = \det A \cdot \det A^{-1}.$$

□

Die beiden bisher vorgestellten Möglichkeiten zur Berechnung der Determinante, nämlich die Definition und der Entwicklungssatz von Laplace, benötigen  $n!$  Summanden und kommen ab  $n \geq 4$  kaum noch in Frage. Stattdessen lässt sich die Determinante sehr einfach aus dem Gaußschen Eliminationsverfahren bestimmen. Wir hatten dort die Matrix  $A$  durch eine Folge von Umformungen auf eine rechte obere Dreiecksmatrix gebracht. Genauer gibt es Permutationsmatrizen  $P_k$  und Eliminationsmatrizen  $G_k$  für  $k = 1, \dots, n-1$  mit

$$(7.5) \quad R = G_{n-1}P_{n-1} \dots G_1 P_1 A$$

mit einer rechten oberen Dreiecksmatrix  $R$ . Mit der Matrix  $P_k$  wird die  $k$ -te Zeile mit einer anderen Zeile vertauscht oder es ist  $P_k = E_n$ , wenn keine Vertauschung notwendig ist. Die Anzahl der echten Vertauschungen sei  $l$ .  $G_k$  ist ein Produkt von Matrizen  $G_{ik,\alpha}$ , um die  $k$ -te Spalte zu eliminieren. Die  $G_{ik,\alpha}$  sind linke untere Dreiecksmatrizen mit lauter Einser in der Hauptdiagonalen, daher  $\det G_k = 1$ . Aus dem Multiplikationssatz für Determinanten und (7.5) folgt daher

$$(7.6) \quad \det A = (-1)^l \det R = (-1)^l \prod_{i=1}^n r_{ii}.$$

Wir fassen die wichtigsten Eigenschaften der Determinante zusammen:

- (a) Eine Matrix ist genau dann regulär, wenn  $\det A \neq 0$ .
- (b) Es gilt  $\det A = \det A^T$ .
- (c) Ist  $\mathbb{K} = \mathbb{C}$ , so  $\det \bar{A} = \overline{\det A}$ .
- (d) Entsteht  $A'$  aus  $A$ , indem zwei Zeilen oder zwei Spalten von  $A$  vertauscht werden, so gilt  $\det A' = -\det A$ .

*Beweis:* (a) folgt aus (7.6).

(b) folgt aus (7.5), indem man diese Formel transponiert.

(c) Das beweist man direkt aus der Definition. □

**7.6 Adjunkte und Cramersche Regeln** Sei  $A \in \mathbb{K}^{n \times n}$ . Wie im vorigen Abschnitt bezeichnen wir mit  $A_{ij}$  die Matrix in  $\mathbb{K}^{(n-1) \times (n-1)}$ , die aus der Matrix  $A$  hervorgeht, wenn wir die  $i$ -te Zeile und  $j$ -te Spalte streichen. Das Element

$$\hat{a}_{ij} = (-1)^{i+j} \det A_{ji}$$

heißt  $(i, j)$ -ter Komplementärwert der Matrix  $A$  ( $A_{ji}$  ist kein Schreibfehler!). Die Matrix

$$\hat{A} = \begin{pmatrix} \hat{a}_{11} & \dots & \hat{a}_{1n} \\ \vdots & & \vdots \\ \hat{a}_{n1} & \dots & \hat{a}_{nn} \end{pmatrix}$$

Heißt *Komplementärmatrix* oder *Adjunkte* zu  $A$ .

**Lemma 7.15** Es gilt

$$\hat{A}A = A\hat{A} = \begin{pmatrix} \det A & & \\ & \ddots & \\ & & \det A \end{pmatrix} = \det A \cdot E_n.$$

*Beweis:* Aus der Definition der  $\hat{a}_{jk}$  erhalten wir

$$\sum_{j=1}^n a_{ij} \hat{a}_{jk} = \sum_{j=1}^n (-1)^{j+k} a_{ij} \det A_{kj} = \det A',$$

wobei  $A'$  aus  $A$  hervorgeht, indem die  $k$ -te Zeile von  $A$  durch die  $i$ -te Zeile von  $A$  ersetzt wird. Das folgt aus dem Laplaceschen Entwicklungssatz 7.8. Falls  $k = i$ , so wurde an der Matrix nichts verändert und es gilt  $\det A' = \det A$ . Andernfalls besitzt  $A'$  zwei gleiche Zeilen und ihre Determinante verschwindet nach Bemerkung 7.10.  $\square$

**Satz 7.16 (Cramersche Regeln)** Sei  $A \in \mathbb{K}^{n \times n}$  regulär.

(a) Für die Inverse von  $A$  gilt

$$A^{-1} = \frac{\hat{A}}{\det A}.$$

(b) Für die Lösung des linearen Gleichungssystems  $Ax = b$  gilt

$$x_i = \frac{\det(a_1 | \dots | a_{i-1} | b | a_{i+1} | \dots | a_n)}{\det A}.$$

*Beweis:* (a) Das folgt aus dem vorigen Lemma wegen  $\det A \neq 0$ .

(b) Es gilt

$$x = A^{-1}b = \frac{\hat{A}b}{\det A}.$$

Der  $i$ -te Eintrag von  $\hat{A}b$  ist

$$\sum_{j=1}^n \hat{a}_{ij} b_j = \sum_{j=1}^n (-1)^{i+j} b_j \det A_{ji}$$

und die letzte Summe ist gleich der behaupteten Determinante nach dem Laplaceschen Entwicklungssatz.  $\square$

Die Cramerschen Regeln dienen theoretischen Zwecken, weil sie erlauben, die Inverse und die Lösung eines LGS geschlossen hinzuschreiben. Für  $n > 2$  sind sie aber viel zu aufwendig.

Für  $n = 2$  bekommt man für die Inverse eine Formel, die man auswendig können sollte,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{\hat{A}}{\det A} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

## 8 Euklidische und unitäre Vektorräume

In diesem Kapitel werden nur endlich dimensionale Vektorräume über  $\mathbb{K} = \mathbb{R}$  oder  $\mathbb{K} = \mathbb{C}$  betrachtet. Der Querstrich bezeichnet die komplexe Konjugation ( $z = x + iy$ ,  $\bar{z} = x - iy$ ). Wenn der zugrunde liegende Vektorraum reell ist, so hat er keine Bedeutung.

**8.1 Skalarprodukte** Sei  $V$  ein linearer Raum über  $\mathbb{K}$ . Eine Abbildung  $(\cdot, \cdot) : V \times V \rightarrow \mathbb{K}$  heißt *inneres Produkt* oder *Skalarprodukt* in  $V$ , wenn die folgenden Bedingungen

$$(a) (\alpha_1 x_1 + \alpha_2 x_2, x_3) = \alpha_1(x_1, x_3) + \alpha_2(x_2, x_3) \quad (\text{Linearität}),$$

$$(b) (x_1, x_2) = \overline{(x_2, x_1)} \quad (\text{Antisymmetrie}),$$

$$(c) (x, x) > 0 \quad \text{für } x \neq 0 \quad (\text{Definitheit}),$$

erfüllt sind, wobei  $\alpha_i \in \mathbb{K}$  und  $x_i \in V$ .

Wegen (b) ist  $(x, x) \in \mathbb{R}$ . Aus (a) und (b) folgt, dass das innere Produkt eine *Sesquilinearform* ist, d.h. es ist linear in der ersten Komponente und *antilinear* in der zweiten,

$$(x_1, \alpha_2 x_2 + \alpha_3 x_3) = \overline{(\alpha_2 x_2 + \alpha_3 x_3, x_1)} = \overline{\alpha_2}(x_1, x_2) + \overline{\alpha_3}(x_1, x_3).$$

Im Fall reeller Räume ist das innere Produkt eine Bilinearform.

Mit Hilfe des Skalarprodukts definieren wir später Schnittwinkel zweier sich schneidender Geraden sowie Abstände zwischen zwei Punkten des Vektorraums. Mit

$$\|x\| = (x, x)^{1/2}$$

können wir die „Entfernung“ des Punktes  $x$  zum Nullpunkt definieren.

Im  $\mathbb{R}^n$  ist das Standardprodukt

$$\langle x, y \rangle = \sum_{k=1}^n x_k y_k, \quad |x| = \|x\| = \left( \sum_{k=1}^n |x_k|^2 \right)^{1/2}.$$

Nach dem Satz des Pythagoras ist  $|x|$  gerade die Länge des Vektors  $x$ . Man beachte die Notation:  $(\cdot, \cdot)$  ist ein allgemeines Skalarprodukt,  $\langle \cdot, \cdot \rangle$  ist reserviert für das Standardprodukt im  $\mathbb{K}^n$ , das im Fall  $\mathbb{K} = \mathbb{C}$  so ausschaut:

$$\langle x, y \rangle = \sum_{k=1}^n x_k \overline{y_k}.$$

Im Komplexen wird in der zweiten Komponente des Produkts komplex konjugiert, damit  $\langle x, x \rangle$  reell und  $\geq 0$  ist.

Ein Vektorraum mit Skalarprodukt heißt *euklidischer Vektorraum* ( $\mathbb{K} = \mathbb{R}$ ) bzw. *unitärer Vektorraum* ( $\mathbb{K} = \mathbb{C}$ ). Wir sprechen von einem Raum mit Skalarprodukt, wenn wir es offen lassen, ob der Raum reell oder komplex ist.

**Lemma 8.1 (Cauchy-Ungleichung)** *In einem Raum mit Skalarprodukt gilt für alle  $x, y$*

$$|(x, y)| \leq \|x\| \|y\|.$$

*Beweis:* Aus den Axiomen für das innere Produkt erhalten wir

$$\begin{aligned} 0 \leq (\alpha x + y, \alpha x + y) &= |\alpha|^2 \|x\|^2 + (\alpha x, y) + (y, \alpha x) + \|y\|^2 \\ &= |\alpha|^2 \|x\|^2 + 2\operatorname{Re}(\alpha(x, y)) + \|y\|^2. \end{aligned}$$

Wir können  $x \neq 0$  voraussetzen und wählen  $\alpha = -\overline{(x,y)}/\|x\|^2$ , also

$$0 \leq \|\alpha x + y\|^2 = \|y\|^2 - \frac{|\langle x, y \rangle|^2}{\|x\|^2}.$$

Damit ist die Ungleichung bewiesen.  $\square$

**Lemma 8.2**  $\|x\| = (x, x)^{1/2}$  ist eine Norm auf  $V$ , sie besitzt die Eigenschaften

- (a)  $\|x\| > 0$  für  $x \neq 0$  (Definitheit),
- (b)  $\|\alpha x\| = |\alpha| \|x\|$  für alle  $\alpha \in \mathbb{K}$  (positive Homogenität),
- (c)  $\|x + y\| \leq \|x\| + \|y\|$  (Dreiecksungleichung).

*Beweis:* Die beiden ersten Normaxiome folgen direkt aus der Definition der Sesquilinearform, die Dreiecksungleichung beweist man mit Hilfe der Cauchy-Ungleichung

$$\begin{aligned} \|x + y\|^2 &= \|x\|^2 + (x, y) + (y, x) + \|y\|^2 \|x\|^2 + 2\operatorname{Re}(x, y) + \|y\|^2 \\ &\leq \|x\|^2 + 2\|x\| \|y\| + \|y\|^2 = (\|x\| + \|y\|)^2. \end{aligned}$$

$\square$

Aus der Dreiecksungleichung folgt die *umgekehrte Dreiecksungleichung*

$$|\|x\| - \|y\|| \leq \|x - y\|.$$

Dies folgt aus

$$\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\|.$$

Die andere Richtung beweist man, indem man die Rollen von  $x$  und  $y$  vertauscht.

**8.2 Orthogonalität** Sei  $V$  ein Vektorraum mit Skalarprodukt. Zwei Vektoren  $x, y \in V$  heißen *orthogonal*, wenn  $(x, y) = 0$ . Wir schreiben dafür  $x \perp y$ .

**Satz 8.3 (Pythagoras)** (a) In einem euklidischen oder unitären Vektorraum gilt

$$x \perp y \Rightarrow \|x\|^2 + \|y\|^2 = \|x + y\|^2.$$

(b) In einem euklidischen Vektorraum gilt auch die Umkehrung:

$$\|x\|^2 + \|y\|^2 = \|x + y\|^2 \Rightarrow x \perp y.$$

*Beweis:* (a)  $\|x + y\|^2 = (x + y, x + y) = \|x\|^2 + (x, y) + (y, x) + \|y\|^2 = \|x\|^2 + \|y\|^2$ .

(b) Im euklidischen Fall gilt in der letzten Formel  $(x, y) + (y, x) = 2(x, y)$ , so dass wir auf  $x \perp y$  schließen können. Dagegen ist bei unitären Räumen  $(x, y) + (y, x) = (x, y) + \overline{(x, y)} = 2\operatorname{Re}(x, y)$  und wir erhalten in diesem Fall nur, dass  $(x, y)$  rein imaginär ist.  $\square$

Eine Menge von Vektoren  $x_1, \dots, x_k$  heißt *Orthogonalsystem*, wenn die Vektoren nicht verschwinden und paarweise orthogonal sind, also  $(x_i, x_j) = 0$  für  $i \neq j$  erfüllt ist. Ein Orthogonalsystem heißt *Orthonormalsystem*, wenn zusätzlich  $\|x_i\| = 1$  für  $i = 1, \dots, k$  erfüllt ist. Aus einem Orthogonalsystem  $x_1, \dots, x_k$  erhalten wir mit der Normierung  $y_i = x_i/\|x_i\|$  ein Orthonormalsystem  $y_1, \dots, y_k$ .

Die Vektoren in einem Orthogonalsystem sind linear unabhängig, denn in

$$\alpha_1 x_1 + \dots + \alpha_k x_k = 0$$

können wir von rechts mit  $x_j$  multiplizieren und die Linearität des Skalarprodukts ausnutzen,

$$0 = (\alpha_1 x_1 + \dots + \alpha_k x_k, x_j) = \alpha_1(x_1, x_j) + \dots + \alpha_k(x_k, x_j) = \alpha_j(x_j, x_j).$$

Es folgt  $\alpha_j = 0$ .

Nun wollen wir eine l.u. Menge von Vektoren  $u_1, \dots, u_k$  so linear kombinieren, dass eine Orthogonalsystem  $x_1, \dots, x_k$  entsteht mit  $\text{span}\{u_1, \dots, u_i\} = \text{span}\{x_1, \dots, x_i\}$ ,  $1 \leq i \leq k$ . Wir setzen  $x_1 = u_1$ . Anschließend bestimmen wir  $\alpha \in \mathbb{K}$  so, dass

$$\alpha x_1 + u_2 \perp x_1 \Rightarrow \alpha = -(u_2, x_1)/\|x_1\|^2.$$

Mit diesem  $\alpha$  ist dann  $x_2 = \alpha x_1 + u_2 \perp x_1$ . Allgemeiner verwenden wir den folgenden

**Satz 8.4 (Orthogonalisierungsverfahren von Gram-Schmidt)** *Sei  $V$  ein Vektorraum mit Skalarprodukt  $(\cdot, \cdot)$  und sei  $u_1, \dots, u_k$  eine l.u. Menge von Vektoren in  $V$ . Dann erhält man durch*

$$(8.1) \quad x_1 = u_1, \quad x_{i+1} = u_{i+1} - \sum_{j=1}^i \frac{(u_{i+1}, x_j)}{\|x_j\|^2} x_j \text{ für } i = 1, \dots, k-1$$

ein Orthogonalsystem mit

$$(8.2) \quad \text{span}\{u_1, \dots, u_i\} = \text{span}\{x_1, \dots, x_i\} \text{ für } 1 \leq i \leq k.$$

Insbesondere sind die Vektoren  $x_i \neq 0$  und können mit  $y_i = x_i/\|x_i\|$  zu einem Orthonormalsystem  $y_1, \dots, y_k$  normiert werden.

*Beweis:* Wir zeigen die Orthogonalität der Vektoren  $x_1, \dots, x_k$  sowie (8.2) mit Hilfe von (8.1) durch Induktion über  $k$ . Für  $k = 1$  ist  $x_1 = u_1$  und (8.2) erfüllt. Sei also die Behauptung für  $k$  erfüllt, insbesondere dürfen wir (8.2) für dieses  $k$  verwenden sowie die Orthogonalität der Vektoren  $x_1, \dots, x_k$ . Mit dem Ansatz

$$(8.3) \quad x_{k+1} = u_{k+1} + \alpha_1 x_1 + \dots + \alpha_k x_k$$

erhalten wir aus der Multiplikation mit dem Vektor  $x_i$

$$(x_{k+1}, x_i) = (u_{k+1}, x_i) + \alpha_i(x_i, x_i),$$

denn wegen der Induktionsvoraussetzung gilt  $(x_j, x_i) = 0$  für  $j \neq i$ . Es gilt daher  $(x_{k+1}, x_i) = 0$  genau dann, wenn wir  $\alpha_i = -(u_{k+1}, x_i)/(x_i, x_i)$  wählen, das ist gerade (8.1). Wäre  $x_{k+1} = 0$ , so  $u_{k+1} \in \text{span}\{x_1, \dots, x_k\} = \text{span}\{u_1, \dots, u_k\}$  im Widerspruch zur vorausgesetzten linearen Unabhängigkeit der  $u_1, \dots, u_{k+1}$ .  $\text{span}\{x_1, \dots, x_{k+1}\} = \text{span}\{u_1, \dots, u_{k+1}\}$  folgt aus (8.3).  $\square$

Für die Orthogonalisierung mit einem Computerprogramm ist das hier vorgestellte Verfahren die denkbar schlechteste Möglichkeit, weil Rundungsfehler die Orthogonalität stören und einmal gestörte Orthogonalität zu größeren Fehlern in den folgenden Schritten führt (=Aufschaukelung von Rundungsfehlern). Besser ist daher das *modifizierte Gram-Schmidt-Verfahren* oder das *Householder-Verfahren*.

Sei  $V$  ein Vektorraum mit Skalarprodukt  $(\cdot, \cdot)$  und  $U$  ein Unterraum von  $V$ . Die Menge

$$U^\perp = \{x \in V : (x, u) = 0 \text{ für alle } u \in U\}$$

heißt *orthogonales Komplement* von  $U$  in  $V$ . Gilt für einen Vektor  $x \in V$ , dass  $(x, u) = 0$  für alle  $u \in U$ , so sagen wir, dass  $x$  senkrecht auf  $U$  steht und schreiben  $x \perp U$ .

Als kleines Beispiel betrachten wir den  $\mathbb{R}^2$  mit den drei prinzipiellen Unterräumen  $\{0\}, g, \mathbb{R}^2$ , wobei  $g$  eine Gerade durch den Nullpunkt in Richtung  $x \in \mathbb{R}^2$  bezeichnet. Es gilt dann  $\{0\}^\perp = \mathbb{R}^2$ ,

$\mathbb{R}^2^\perp = \{0\}$ . Alle Vektoren, die auf  $x$  senkrecht stehen, bilden das orthogonale Komplement von  $g$ . Mit  $y \perp x$ ,  $y \neq 0$ , gilt dann  $g^\perp = \{\alpha y : \alpha \in \mathbb{R}\}$ .

Allgemeiner gilt in einem beliebigen Vektorraum mit Skalarprodukt  $\{0\}^\perp = V$ ,  $V^\perp = \{0\}$ . Zur Berechnung von  $U^\perp$  für einen nichttrivialen Unterraum  $U$  von  $V$  mit  $\dim V = n$  wählen wir eine Basis  $u_1, \dots, u_r$  von  $U$  und ergänzen sie nach dem Basisergänzungssatz 5.9 mit  $u_{r+1}, \dots, u_n$  zu einer Basis von  $V$ . In dieser Reihenfolge der Vektoren wenden wir den Gram-Schmidt-Algorithmus 8.1 an und normieren die erhaltenen Vektoren, was zu einem Orthonormalsystem  $x_1, \dots, x_n$  von  $V$  führt. Wegen (8.2) gilt  $U = \text{span}\{x_1, \dots, x_r\}$  und die Vektoren in  $U' = \text{span}\{x_{r+1}, \dots, x_n\}$  stehen senkrecht auf  $U$ . Daher ist  $U' \subset U^\perp$ . Jeder Vektor aus  $V$  lässt sich als eine Linearkombination  $v = \sum_{i=1}^n \alpha_i x_i$  schreiben. Sei  $u = \sum_{i=1}^r \alpha_i x_i$ . Ist  $u = 0$ , so ist  $u \in U'$ , andernfalls ist  $u \in U$  mit  $(u, u) > 0$ . Damit ist  $U' = U^\perp$  gezeigt.

Aus dieser Konstruktion lassen sich alle wichtigen Eigenschaften von  $U^\perp$  ablesen:

- Satz 8.5** (a)  $U^\perp$  ist Unterraum von  $V$ ,
- (b)  $U \cap U^\perp = \{0\}$ ,
- (c)  $\dim V = \dim U + \dim U^\perp$ .

**Beispiele 8.6** (i) Sei  $V = \mathbb{R}^3$  mit dem Standard-Skalarprodukt  $\langle \cdot, \cdot \rangle$  versehen. Sei  $U = \{(x, y, z)^T : 2x + 3y + 4z = 0\}$  eine Ebene. Dann ist  $U^\perp = \text{span}\{(2, 3, 4)^T\}$  wegen

$$\langle (2, 3, 4)^T, (x, y, z)^T \rangle = 0 \Leftrightarrow 2x + 3y + 4z = 0.$$

(ii) Allgemeiner nennen wir einen Unterraum  $U$  eines endlich dimensionalen Vektorraums  $V$  Hyperebene, wenn  $\dim U = \dim V - 1 > 0$ . Da hier das Skalarprodukt nicht eingeht, gilt diese Definition auch in allgemeinen Vektorräumen über beliebigen Körpern. Im Falle von euklidischen oder unitären Vektorräumen gestatten diese Hyperebenen eine einfache Darstellung mit Hilfe des Skalarprodukts. Wie in der Konstruktion des orthogonalen Komplements beschrieben erhalten wir  $U^\perp = \text{span}\{x\}$  mit  $x \neq 0$ . Dann gilt

$$U = \{u \in V : \langle x, u \rangle = 0\} \Leftrightarrow U = \{u \in V : x_1 u_1 + \dots + x_n u_n = 0\}.$$

Man nennt dies die *Hessische Normalenform* der Hyperebene  $U$ . Anders ausgedrückt: Die Hyperebene kann charakterisiert werden durch einen beliebigen Vektor  $x \neq 0$ , der senkrecht auf  $U$  steht und Normale von  $U$  genannt wird.

(iii) Sei  $V = \mathbb{C}^3$  versehen mit dem Standard-Skalarprodukt  $\langle \cdot, \cdot \rangle$ . Für  $U = \text{span}\{(1, i, 0)^T, (0, 0, 1)^T\} = \text{span}\{x_1, x_2\}$  wollen wir das orthogonale Komplement bestimmen. Die beiden erzeugenden Vektoren sind bereits orthogonal. Durch Probieren finden wir heraus, dass  $e_2 = (0, 1, 0)^T$  nicht im Bild dieser beiden Vektoren ist. Nach (8.1) erhalten wir

$$\begin{aligned} x_3 &= e_2 - \frac{\langle e_2, x_1 \rangle}{\langle x_1, x_1 \rangle} x_1 - \frac{\langle e_2, x_2 \rangle}{\langle x_2, x_2 \rangle} x_2 \\ &= \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} - \frac{\left\langle \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ i \\ 0 \end{pmatrix} \right\rangle}{\left\langle \begin{pmatrix} 1 \\ i \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ i \\ 0 \end{pmatrix} \right\rangle} \begin{pmatrix} 1 \\ i \\ 0 \end{pmatrix} - \frac{\left\langle \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\rangle}{\left\langle \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\rangle} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} - \frac{-i}{2} \begin{pmatrix} 1 \\ i \\ 0 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} i \\ 1 \\ 0 \end{pmatrix} \end{aligned}$$

Damit ist  $U^\perp = \text{span}\{(i, 1, 0)^T\}$ .

(iv) Ist  $x = (x_1, x_2)^T \in \mathbb{K}^2$ , so gilt für  $x^\perp = (-\bar{x}_2, \bar{x}_1)^T$ , dass  $\langle x, x^\perp \rangle = 0$ .  $\square$

Für einen Unterraum  $U$  des Raums  $V$  hatten wir mit Hilfe einer Orthonormalbasis  $x_1, \dots, x_n$  von  $V$  mit  $U = \text{span}\{x_1, \dots, x_r\}$  den Unterraum  $U^\perp = \text{span}\{x_{r+1}, \dots, x_n\}$  konstruiert. Entwickeln wir ein beliebiges  $v \in V$  nach dieser Basis,  $v = \sum_{i=1}^n \alpha_i x_i$ , so erhalten wir mit

$$(8.4) \quad u = \sum_{i=1}^r \alpha_i x_i, \quad u^\perp = \sum_{i=r+1}^n \alpha_i x_i$$

eine Zerlegung

$$v = u + u^\perp \text{ mit } u \in U, u^\perp \in U^\perp$$

$u$  und  $u^\perp$  sind nach Konstruktion eindeutig bestimmt.

Die *Orthogonalprojektion* von  $V$  auf  $U$  ist die Abbildung

$$p_U : V \rightarrow U \subset V, \quad v = u + u^\perp \mapsto u.$$

**Satz 8.7** Für die Orthogonalprojektion  $p_U$  eines Vektorraums  $V$  auf einen Unterraum  $U$  gilt:

- (a)  $p_U$  ist linear mit  $p_U^2 = p_U \circ p_U = p_U$ .
- (b) Bild  $p_U = U$ , Kern  $p_U = U^\perp$ .
- (c) Es gilt  $\|p_U v\| \leq \|v\|$ .

*Beweis:* Die Eigenschaften folgen aus (8.4).  $\square$

Die Berechnung der Orthogonalprojektion erfolgt ebenfalls über (8.4). Es gilt

$$(u, x_j) = (\sum_{i=1}^n \alpha_i x_i, x_j) = \alpha_j (x_j, x_j) = \alpha_j.$$

Damit können wir durch einfaches multiplizieren mit  $x_j$  das  $\alpha_j$  rekonstruieren. Daher

$$(8.5) \quad p_U(v) = \sum_{i=1}^r (u, x_i) x_i.$$

**Beispiel 8.8** Sei  $V = \mathbb{R}^4$  versehen mit dem Standard-Produkt  $\langle \cdot, \cdot \rangle$ . Sei

$$U = \text{span} \left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \right\}, \quad v = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Gesucht ist die Orthogonalprojektion von  $v$  auf  $U$ . Wir bestimmen eine Orthonormalbasis von  $U$ :

$$v_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \quad |v_1| = \sqrt{3},$$

$$v_2 = \begin{pmatrix} 2 \\ 0 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ 1 \\ 0 \end{pmatrix}, \quad |v_2| = \sqrt{3},$$

$$v_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} - \frac{1}{3} \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} - \frac{1}{3} \begin{pmatrix} 1 \\ -1 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \frac{2}{3} \begin{pmatrix} -1 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \quad |v_3| = \frac{2}{\sqrt{3}}.$$

Damit erhalten wir die Orthonormalbasis von  $U$

$$x_1 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad x_2 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ -1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad x_3 = \frac{1}{\sqrt{3}} \begin{pmatrix} -1 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}.$$

Gemäß (8.5) folgt

$$p_U(v) = \langle v, x_1 \rangle x_1 + \langle v, x_2 \rangle x_2 + \langle v, x_3 \rangle x_3 = \frac{1}{3} (3, 5, 8, 13, 0)^T.$$

□

**8.3 Orthogonale und unitäre Matrizen** In diesem Abschnitt betrachten wir nur die Vektorräume  $\mathbb{R}^n$  und  $\mathbb{C}^n$  versehen mit dem zugehörigen Standard-Produkt.

Eine reelle bzw. komplexe  $(n \times n)$ -Matrix heißt *orthogonal* bzw. *unitär*, wenn

$$A^T A = E_n \text{ bzw. } \bar{A}^T A = E_n.$$

Dies bedeutet, dass  $A$  regulär ist mit  $A^{-1} = A^T$  bzw.  $A^{-1} = \bar{A}^T$ . Damit gilt auch  $A \bar{A}^T = E_n$  (im Reellen hat der Querstrich wie immer keine Bedeutung). Wir bezeichnen mit  $a_i$  die Spaltenvektoren von  $A$ ,  $A = (a_1 | \dots | a_n)$ . Dann bedeutet  $A^T A = E_n$  im Reellen, dass

$$\langle a_i, a_j \rangle = \delta_{ij} := \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{falls } i \neq j \end{cases}.$$

Die Spaltenvektoren der Matrix bilden damit ein Orthonormalsystem. Interpretieren wir  $A A^T = E_n$  auf die gleiche Weise, kommen wir zur analogen Schlussfolgerung, dass auch die Zeilenvektoren ein Orthonormalsystem bilden.

Im Komplexen können wir genauso folgern wegen

$$(\bar{A}^T A)_{ij} = \sum_k \bar{a}_{ki} a_{kj} = \langle a_j, a_i \rangle.$$

Wir formulieren diese Ergebnisse nur für den komplexen Fall, im Reellen gilt der folgende Satz völlig analog.

**Satz 8.9** Sei  $A \in \mathbb{C}^{n \times n}$ . Die folgenden Aussagen sind äquivalent:

- (a)  $A$  ist eine unitäre Matrix.
- (b)  $A$  ist regulär mit  $A^{-1} = \bar{A}^T$ .
- (c) Die Spaltenvektoren (bzw. Zeilenvektoren) bilden eine Orthonormalbasis des  $\mathbb{C}^n$  bezüglich des Standard-Produkts.

**Beispiele 8.10** (i) Im  $\mathbb{R}^2$  sind die Drehmatrizen mit Winkel  $\omega$

$$A = \begin{pmatrix} \cos \omega & -\sin \omega \\ \sin \omega & \cos \omega \end{pmatrix}$$

offenbar orthogonal.

(ii) Im  $\mathbb{R}^n$  ist eine Hyperebene durch einen Vektor  $w \in \mathbb{R}^n \setminus \{0\}$  gegeben:  $U = \{x : \langle w, x \rangle = 0\}$  (s. Beispiel 8.6 (ii)). Hier können wir  $|w| = 1$  voraussetzen. Dann besitzt die Spiegelung an dieser Hyperebene die Darstellungsmatrix

$$S = E_n - 2ww^T.$$

Dabei ist  $A = ww^T$  die  $(n \times n)$ -Matrix mit Einträgen  $a_{ij} = w_i w_j$ . Ist  $x \in \mathbb{R}^n$ , so gilt  $x = z + \alpha w$  mit  $z \in U$ , denn  $U = \text{span}\{w\}^\perp$ . Die Spiegelung an  $U$  muss diesen Vektor abbilden auf  $Sx = z - \alpha w$  und das ist der Fall:

$$\begin{aligned} Sx &= (E_n - 2ww^T)(z + \alpha w) = z + \alpha w - 2(ww^T)(z + \alpha w) \\ &= z + \alpha w - 2w(w^T z) - 2\alpha w(w^T w). \end{aligned}$$

Im Reellen gilt für Spaltenvektoren  $x, y$ , dass  $x^T y = \langle x, y \rangle$ . Damit ist  $w^T z = 0$  wegen  $w \perp z$  und  $w^T w = 1$  wegen  $|w| = 1$ . Insgesamt erhalten wir  $Sx = z - \alpha w$  wie behauptet. Man rechnet leicht nach  $S^2 = E_n$  sowie  $S = S^T$ . Damit ist  $S$  orthogonal.  $\square$

Für beliebige reelle  $(n \times n)$ -Matrizen  $A$  gilt

$$\langle Ax, y \rangle = \langle x, A^T y \rangle \quad \text{für alle } x, y \in \mathbb{R}^n$$

wegen

$$\langle Ax, y \rangle = \sum_{k=1}^n (Ax)_k y_k = \sum_{k=1}^n \sum_{j=1}^n a_{kj} x_j y_k = \sum_{k=1}^n \sum_{j=1}^n a_{jk}^T x_j y_k = \langle x, A^T y \rangle.$$

Damit gilt für eine orthogonale  $(n \times n)$ -Matrix  $A$

$$\langle Ax, Ay \rangle = \langle x, A^T Ay \rangle = \langle x, y \rangle,$$

insbesondere auch für  $y = x$ :  $|Ax| = |x|$ . Damit erhält eine orthogonale Matrix das Skalarprodukt und damit auch die Längen von Vektoren. Die zugehörigen orthogonalen Selbstabbildungen  $f(x) = Ax$  erhalten damit alle Strukturen, die in einem euklidischen Vektorraum vorhanden sind. Die ganze Herleitung gilt auch im unitären Raum  $\mathbb{C}^n$ .

Wir zeigen: Hat  $(n \times n)$ -Matrix  $A$  die Eigenschaft

$$|Ax| = |x| \quad \text{für alle } x \in \mathbb{K}^n,$$

so ist sie bereits orthogonal bzw. unitär. Im Reellen gilt

$$|Ax + Ay|^2 = |x + y|^2 \Rightarrow (Ax, Ay) = (x, y) \text{ wegen } |Ax| = |x|, |Ay| = |y|,$$

woraus  $(x, A^T Ay) = (x, y)$  folgt. Wir können hier für  $x$  die kanonischen Einheitsvektoren einsetzen und erhalten  $A^T Ay = y$  für alle  $y$  und damit  $A^T A = E_n$ . Im Komplexen folgt mit gleicher Überlegung nur  $\text{Re}(Ax, Ay) = \text{Re}(x, y)$ . Wir können hier aber  $x$  durch  $ix$  ersetzen und erhalten dann auch  $\text{Im}(Ax, Ay) = \text{Im}(x, y)$ . Der Rest verläuft genauso wie zuvor.

## 9 Das Eigenwertproblem und die Jordansche Normalform

### 9.1 Das Eigenwertproblem

**Ähnliche Matrizen** Zwei Matrizen  $A, B \in \mathbb{K}^{n \times n}$  heißen *ähnlich*, wenn es eine reguläre Matrix  $T \in \mathbb{K}^{n \times n}$  gibt mit

$$(9.1) \quad A = T^{-1}BT.$$

Ähnlichkeit ist eine Äquivalenzrelation auf dem Raum der  $(n \times n)$ -Matrizen. Wegen  $A = E_n A E_n$  ist  $A$  zu sich selber ähnlich. In der Definition (9.1) wurde implizit die Symmetrie der Ähnlichkeit vorausgesetzt. Wir können aber in (9.1) nach  $B$  auflösen,  $B = TAT^{-1}$ , also ist auch  $B$  zu  $A$  ähnlich. Ist  $A$  zu  $B$  und  $B$  zu  $C$  ähnlich, so

$$A = T^{-1}BT, \quad B = T'^{-1}CT' \Rightarrow A = T^{-1}T'^{-1}CT'T = (T'T)^{-1}C(T'T).$$

**Nullstellen von Polynomen** Sei

$$p(x) = a_n x^n + \dots + a_1 x + a_0, \quad a_i \in \mathbb{C}, \quad a_n \neq 0$$

ein Polynom vom Grad  $n$ . Wir sagen,  $q$  besitzt in  $x_0$  eine Nullstelle der *Vielfachheit*  $k$ , wenn es ein Polynom  $q$  vom Grade  $n - k$  gibt mit

$$(9.2) \quad p(x) = (x - x_0)^k q(x) \quad \text{mit } q(x_0) \neq 0.$$

Der Fundamentalsatz der Algebra besagt, dass die Summe der Vielfachheiten der Nullstellen gerade  $n$  ergibt. Im Reellen ist dieser Satz nicht richtig, wie das Polynom  $p(x) = x^2 + 1$  beweist, das im Reellen keine Nullstellen besitzt.

In (9.2) können wir aufgrund dieses Fundamentalsatzes auch die  $n - k$  Nullstellen von  $q$  ausklammern. Sind  $x_1, \dots, x_n$  die Nullstellen von  $p$ , die hier nicht alle verschieden sein müssen, so können wir schreiben

$$(9.3) \quad p(x) = a_n(x - x_1) \dots (x - x_n).$$

Wir betrachten Eigenwertprobleme nur über dem Körper  $\mathbb{C}$ . Wenn eine Matrix reellwertig ist, ist sie auch eine Matrix über  $\mathbb{C}$ .

Sei  $A \in \mathbb{C}^{n \times n}$ .  $\lambda \in \mathbb{C}$  heißt *Eigenwert* von  $A$ , wenn

$$Ax = \lambda x \quad \text{für ein } x \in \mathbb{C}^n \setminus \{0\}.$$

$x$  ist dann *Eigenvektor* zu  $\lambda$ . Insbesondere ist  $U = \text{span}\{x\}$  ein *invarianter Raum*, d.h.  $AU \subset U$ .

$\lambda$  ist genau dann Eigenwert, wenn die Matrix  $A - \lambda E_n$  singulär ist und damit das *charakteristische Polynom* von  $A$

$$\phi(\mu) = \det(A - \mu E_n)$$

in  $\lambda$  eine Nullstelle besitzt.

Die Größe

$$\sigma(\lambda) = \text{Vielfachheit der Nullstelle } \lambda \text{ in } \phi$$

heißt *algebraische Vielfachheit* von  $\lambda$ . Nach dem im vorigen Abschnitt Gesagten ist die Summe der algebraischen Vielfachheiten der Eigenwerte  $n$ . Der Vektorraum

$$L(\lambda) = \{x \in \mathbb{C}^n : Ax = \lambda x\}$$

heißt *Eigenraum* zu  $\lambda$ . Ferner heißt

$$\rho(\lambda) = \dim L(\lambda)$$

*geometrische Vielfachheit* von  $\lambda$ .  $\rho(\lambda)$  ist die Zahl der linear unabhängigen Eigenvektoren zu  $\lambda$ .

**Satz 9.1** (a) Ist  $p(\mu)$  ein Polynom und gilt  $Ax = \lambda x$  für ein  $x \neq 0$ , so besitzt  $p(A)$  ebenfalls den Eigenvektor  $x$  zum Eigenwert  $p(\lambda)$ .

(b)  $\lambda$  ist genau dann Eigenwert von  $A$ , wenn  $\bar{\lambda}$  Eigenwert von  $\bar{A}$  ist. Insbesondere: Ist die Matrix  $A$  reellwertig, so ist mit einem komplexen Eigenwert  $\lambda$  von  $A$  auch  $\bar{\lambda}$  Eigenwert von  $A$ .

(c)  $A$  und  $A^T$  besitzen die gleichen Eigenwerte.

(d) Die Determinante von  $A$  stimmt mit dem Produkt aller Eigenwerte von  $A$  überein.

(e) Ähnliche Matrizen besitzen das gleiche charakteristische Polynom, also auch die gleichen Eigenwerte. Wenn

$$B = T^{-1}AT$$

und  $A$  besitzt den Eigenwert  $\lambda$  mit Eigenvektor  $x$ , so besitzt  $B$  den Eigenwert  $\lambda$  mit Eigenvektor  $T^{-1}x$ .

*Beweis:* (a) Aus  $Ax = \lambda x$  folgt  $A^kx = \lambda^k x$  und

$$p(A)x = a_m A^m x + \dots + a_0 x = p(\lambda)x.$$

(b) Nach Satz 7.14(b) gilt

$$\det(\bar{A} - \bar{\lambda}E_n) = \overline{\det(A - \lambda E_n)}.$$

(c) Das folgt aus 7.14(c)

(d) In

$$\phi(\mu) = \det(A - \mu E_n) = (-1)^n(\mu - \lambda_1) \dots (\mu - \lambda_n)$$

(vergleiche (9.3)) setze man  $\mu = 0$ .

(e) Mit dem Determinantenmultiplikationssatz folgt

$$\begin{aligned} \det(B - \lambda E_n) &= \det(T^{-1}AT - \lambda E_n) = \det(T^{-1}(A - \lambda E_n)T) \\ &= \det T^{-1} \det(A - \lambda E_n) \det T = \det(A - \lambda E_n). \end{aligned}$$

Ferner gilt

$$BT^{-1}x = T^{-1}Ax = T^{-1}(\lambda x) = \lambda T^{-1}x.$$

□

**Beispiel 9.2** Das Jordan-Kästchen der Länge  $\nu$  zum Eigenwert  $\lambda$  ist definiert durch

$$(9.4) \quad C_\nu(\lambda) = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix} \in \mathbb{C}^{\nu \times \nu}.$$

Wegen

$$\det(C_\nu(\mu) - \lambda E_n) = (\mu - \lambda)^\nu$$

ist  $\lambda$  Eigenwert mit  $\sigma(\lambda) = \nu$ , aber  $x = e_1$  ist einziger Eigenvektor von  $C_\nu$ , also  $\rho(\lambda) = 1$ . □

Damit ist gezeigt, dass algebraische und geometrische Vielfachheit nicht übereinstimmen müssen. Es gilt aber  $\rho(\lambda) \leq \sigma(\lambda)$ .

**9.2 Die Jordansche Normalform** Es sei an die Definition des Jordan-Kästchens  $C_\nu(\lambda)$  in (9.4) erinnert.

**Satz 9.3** Sei  $A \in \mathbb{C}^{n \times n}$ ,  $\lambda_1, \dots, \lambda_k$  seien die Eigenwerte von  $A$  mit geometrischen bzw. algebraischen Vielfachheiten  $\rho(\lambda_i)$  und  $\sigma(\lambda_i)$ . Zu jedem  $\lambda_i$  gibt es Zahlen  $\nu_1^{(i)}, \dots, \nu_{\rho(\lambda_i)}^{(i)}$  mit

$$\sigma(\lambda_i) = \nu_1^{(i)} + \dots + \nu_{\rho(\lambda_i)}^{(i)}$$

und eine reguläre Matrix  $T \in \mathbb{C}^{n \times n}$  mit  $J = T^{-1}AT$ ,

$$J = \begin{pmatrix} C_{\nu_1^{(1)}}(\lambda_1) & & & \\ & \ddots & & 0 \\ & & C_{\nu_{\rho(\lambda_1)}^{(1)}}(\lambda_1) & \\ & & & C_{\nu_1^{(2)}}(\lambda_2) \\ 0 & & & & \ddots \\ & & & & C_{\nu_{\rho(\lambda_k)}^{(k)}}(\lambda_k) \end{pmatrix}$$

$J$  ist bis auf die Reihenfolge der Jordan-Kästchen eindeutig bestimmt.

*Beweis:* Der Beweis ist sehr aufwändig.  $\square$

**Diagonalsierbare Matrizen** Eine Matrix heißt *diagonalsierbar*, wenn für alle Eigenwerte  $\lambda_i$  gilt  $\rho(\lambda_i) = \sigma(\lambda_i)$ . Wenn man dann mehrfache Eigenwerte auch mehrfach zählt, folgt wegen  $\nu_j^{(i)} = 1$ ,

$$J = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Anders ausgedrückt: Im diagonalsierbaren Fall gibt es eine Basis aus Eigenvektoren  $\{x_1, \dots, x_n\}$  und die Matrix  $T$  hat die Gestalt

$$T = (x_1 | \dots | x_n).$$

**Beispiel 9.4** Wir bestimmen die Eigenwerte und Eigenvektoren der Matrix

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 2 \\ 0 & 0 & 2 \end{pmatrix}.$$

Wir berechnen das charakteristische Polynom

$$\begin{aligned} \det(A - \lambda E_3) &= \det \begin{pmatrix} 1 - \lambda & 2 & 1 \\ 1 & 2 - \lambda & 2 \\ 0 & 0 & 2 - \lambda \end{pmatrix} \\ &= (1 - \lambda)(2 - \lambda)^2 - 1 \cdot 2(2 - \lambda) = (2 - \lambda)((1 - \lambda)(2 - \lambda) - 2) \\ &= (2 - \lambda)(2 - 3\lambda + \lambda^2 - 2) = (2 - \lambda)\lambda(\lambda - 3). \end{aligned}$$

Wir haben also die drei einfachen Eigenwerte  $\lambda_1 = 2$ ,  $\lambda_2 = 0$ ,  $\lambda_3 = 3$ .

Die Kernvektoren von  $A - \lambda_i E_3$  bestimmen wir mit dem Gauß-Algorithmus.

$$A - 2E_3 = \begin{pmatrix} -1 & 2 & 1 \\ 1 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix} \quad \rightarrow \quad \begin{pmatrix} -1 & 2 & 1 \\ 0 & 2 & 3 \\ 0 & 0 & 0 \end{pmatrix}.$$

Die ersten beiden Spaltenvektoren spannen das Bild auf. Wir setzen daher  $x_3 = 1$  und erhalten aus  $(A - 2E_3)x = 0$  für die anderen Komponenten  $x_2 = -\frac{3}{2}$ ,  $x_1 = -2$ . Man kann hier noch die Probe machen:

$$Ax = \begin{pmatrix} -4 \\ -3 \\ 2 \end{pmatrix} = 2 \begin{pmatrix} -2 \\ -3/2 \\ 1 \end{pmatrix}.$$

$$A - 0E_3 = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 2 & 2 \\ 0 & 0 & 2 \end{pmatrix} \quad \rightarrow \quad \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 2 \end{pmatrix} \quad \rightarrow \quad \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Hier spannen die Spalten 1 und 3 das Bild auf. Wir setzen daher  $x_2 = 1$  und erhalten  $x_3 = 0$  und  $x_1 = -2$ . Die Probe kann man leicht im Kopf durchführen.

$$A - 3E_3 = \begin{pmatrix} -2 & 2 & 1 \\ 1 & -1 & 2 \\ 0 & 0 & -1 \end{pmatrix} \quad \rightarrow \quad \begin{pmatrix} -2 & 2 & 1 \\ 0 & 0 & 5/2 \\ 0 & 0 & -1 \end{pmatrix} \quad \rightarrow \quad \begin{pmatrix} -2 & 2 & 1 \\ 0 & 0 & 5/2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Wie zuvor setzen wir  $x_2 = 1$  und erhalten  $x_3 = 0$ ,  $x_1 = 1$ .

Insgesamt erhalten wir eine Basis aus Eigenvektoren

$$T = \begin{pmatrix} -2 & -2 & 1 \\ -3/2 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

und es gilt

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 3 \end{pmatrix} = T^{-1}AT.$$

□

## 10 Folgen und Reihen

**10.1 Definition und Beispiele** Eine Abbildung  $a : \mathbb{N} \rightarrow \mathbb{R}$  heißt (*reelle*) *Zahlenfolge*. Statt  $a(n)$  schreiben wir kürzer  $a_n$  und bezeichnen die ganze Folge mit  $(a_n)_{n \in \mathbb{N}}$  oder einfach  $(a_n)$ , was aber nicht darüber hinwegtäuschen soll, dass unsere Zahlenfolgen immer unendlich viele Folgenglieder besitzen. Eher als die konkreten Werte der  $a_n$  interessiert uns das Verhalten der Folge für große  $n$ .

**Beispiele 10.1** (i)  $a_n = n$  oder  $(a_n) = (1, 2, 3, \dots)$  ist die Folge der natürlichen Zahlen, deren Folgenglieder beliebig groß werden.

(ii)  $a_n = 1/n$  oder  $(a_n) = (1, \frac{1}{2}, \frac{1}{3}, \dots)$  ist eine Folge, deren Glieder der Null beliebig nahe kommen.

(iii) Die Folge  $a_n = (-1)^n + \frac{1}{n}$  oder  $(a_n) = (0, \frac{3}{2}, -\frac{2}{3}, \frac{5}{4}, -\frac{4}{5}, \dots)$  wechselt nach dem ersten Folgenglied das Vorzeichen. Man sagt auch: Die Folge alterniert. Für große  $n$  wechselt sie zwischen Werten, die nahe bei  $\pm 1$  liegen.  $\square$

**10.2 Beschränktheit und Konvergenz von Zahlenfolgen** Die Folge heißt *beschränkt*, wenn es eine Zahl  $M$  gibt mit  $|a_n| \leq M$  für alle  $n \in \mathbb{N}$ . An sich ist der Wertebereich  $M_a$  der Folge  $(a_n)$ , nämlich

$$M_a = \{a_1, a_2, a_3, \dots\}$$

deutlich von der Folge zu unterscheiden, weil es bei der Folge auch auf die Reihenfolge der Folgenglieder ankommt. Im Fall der Beschränktheit verhalten sich beide Begriffe gleich: Die Folge  $(a_n)$  ist genau dann beschränkt wenn der Wertebereich  $M_a$  eine beschränkte Menge ist. Da endliche Mengen reeller Zahlen immer beschränkt sind, sind auch Folgen mit nur endlich vielen Werten beschränkt. Von den Beispielfolgen in 10.1 sind die Folgen (ii) und (iii) durch 2 beschränkt. Die Folge der natürlichen Zahlen in (i) ist ein Beispiel für eine unbeschränkte Folge.

Eine Folge  $(a_n)$  ist genau dann *konvergent gegen  $a \in \mathbb{R}$* , wenn es zu jedem  $\varepsilon > 0$  ein  $N \in \mathbb{N}$  gibt, das von  $\varepsilon$  abhängen darf, so daß für alle  $n \geq N$  gilt  $|a_n - a| < \varepsilon$ . In diesem Fall heißt  $a$  *Grenzwert* oder *Limes* von  $(a_n)$  und wir schreiben

$$\lim_{n \rightarrow \infty} a_n = a \quad \text{oder} \quad a_n \rightarrow a \quad \text{für } n \rightarrow \infty.$$

Formal kann man die Definition der Konvergenz so schreiben:

$$a_n \rightarrow a \iff \forall \varepsilon > 0 \exists N \in \mathbb{N} \forall n \geq N |a_n - a| < \varepsilon.$$

Wir können uns den schwierigen Konvergenzbegriff auf vielfältige Weise verdeutlichen. Wir sagen, eine Eigenschaft trifft für *fast alle*  $n \in \mathbb{N}$  zu, wenn sie für alle bis auf endlich viele  $n$  zutrifft. Die Eigenschaft, größer als 100 zu sein, trifft für fast alle natürlichen Zahlen zu, aber die Eigenschaft, geradzahlig zu sein, trifft nicht auf fast alle natürlichen Zahlen zu. Die Menge

$$B_\varepsilon(a) = (a - \varepsilon, a + \varepsilon) = \{x : |x - a| < \varepsilon, \varepsilon > 0,$$

heißt  $\varepsilon$ -Umgebung der reellen Zahl  $a$ .

Die folgenden Aussagen sind zu  $a_n \rightarrow a$  äquivalent.

(i) Zu jedem  $m \in \mathbb{N}$  gibt es ein  $N \in \mathbb{N}$ , so dass  $|a_n - a| < \frac{1}{m}$  für alle  $N \geq N$ .

(ii) Für jedes  $\varepsilon > 0$  liegen fast alle Folgenglieder in  $B_\varepsilon(a)$

Beweis von (i): Aus  $a_n \rightarrow a$  folgt (i). Sei also umgekehrt (i) erfüllt und  $\varepsilon > 0$  vorgegeben. Es gibt dann ein  $m \in \mathbb{N}$  mit  $\frac{1}{m} \leq \varepsilon$ . Für dieses  $m$  bekommen wir aus (i) ein  $N$  und für alle  $n \geq N$  gilt  $|a_n - a| < \frac{1}{m} \leq \varepsilon$ .

Beweis von (ii):  $a_n \in B_\varepsilon(a)$  ist gleichbedeutend mit  $|a_n - a| < \varepsilon$ . Gilt dies für fast allen  $n$ , so gilt es für eine endliche Menge  $M \subset \mathbb{N}$  nicht. Endliche Mengen haben ein maximales Element, nennen wir es hier  $N - 1$ . Damit gilt  $|a_n - a| < \varepsilon$  für alle  $n \geq N$ .

**Beispiel 10.2** Für die Folge

$$a_n = \frac{2n+1}{n+1}$$

erhalten wir

$$|a_n - 2| = \left| \frac{2n+1}{n+1} - 2 \right| = \left| \frac{2n+1 - 2(n+1)}{n+1} \right| = \frac{1}{n+1}$$

Zu jedem  $\varepsilon > 0$  gibt es ein  $N$  mit  $N > \frac{1}{\varepsilon}$ . Für  $n \geq N$  gilt dann  $\frac{1}{n+1} < \varepsilon$ . Damit liegen in jedem  $B_\varepsilon(2)$  alle bis auf endlich viele Folgenglieder und  $\lim_{n \rightarrow \infty} a_n = 2$ .  $\square$

**10.3 Häufungspunkte von Folgen** Ein Punkt  $a \in \mathbb{R}$  heißt *Häufungspunkt* der Folge, wenn für alle  $\varepsilon > 0$  in jedem  $B_\varepsilon(a) = (a - \varepsilon, a + \varepsilon)$  unendlich viele Folgenglieder liegen.

**Beispiel 10.3** Sei  $a_n = (-1)^n + \frac{1}{n}$ . Wegen  $a_{2n} = 1 + \frac{1}{2n}$  gilt  $|a_{2n} - 1| \leq \frac{1}{2n}$ . Zu jedem  $\varepsilon > 0$  gibt es daher unendlich viele Folgenglieder, die in  $B_\varepsilon(1)$  liegen. Damit ist 1 Häufungspunkt der Folge. Genauso erhält man mit  $a_{2n-1} = -1 + \frac{1}{2n-1}$ , daß auch  $-1$  Häufungspunkt der Folge ist. Einen Grenzwert besitzt die Folge nicht, weil weder in  $B_1(1)$  noch in  $B_1(-1)$  alle bis auf endlich viele Folgenglieder liegen.  $\square$

**Satz 10.4** (a) Existiert der Grenzwert einer Folge, so ist er eindeutig bestimmt.

(b) Eine konvergente Folge ist beschränkt und besitzt genau einen Häufungspunkt, nämlich den Grenzwert der Folge.

*Beweis:* (a) Angenommen, für eine Folge  $(a_n)$  gilt  $\lim_{n \rightarrow \infty} a_n = a$  und  $\lim_{n \rightarrow \infty} a_n = b$  mit  $a \neq b$ . Für  $0 < \varepsilon = |a - b|/2$  sind  $B_\varepsilon(a)$  und  $B_\varepsilon(b)$  disjunkt und können demnach nicht beide alle bis auf endlich viele Folgenglieder enthalten.

(b) Ist  $\lim_{n \rightarrow \infty} a_n = a$ , so wählen wir in der Definition der Konvergenz  $\varepsilon = 1$ . Damit genügen alle bis auf endlich viele Folgenglieder der Abschätzung  $|a_n| < |a| + 1$ . Die übrigen Folgenglieder bilden eine endliche Menge. Endliche Mengen sind immer beschränkt.

Besitzt eine Folge mehr als einen Häufungspunkt, so können wir zwei Häufungspunkte mit dem Argument aus (a) durch  $\varepsilon$ -Umgebungen trennen. In jeder  $\varepsilon$ -Umgebung liegen dann unendlich viele Folgenglieder, was die Konvergenz der Folge ausschließt.  $\square$

Die Begriffe Grenzwert, Häufungspunkt und Beschränktheit hängen nicht von endlichen Abschnitten der Folge ab. Lassen wir endlich viele Folgenglieder weg oder fügen endlich viele Folgenglieder hinzu, so ändert das nichts an ihrem Grenzwert, an ihren Häufungspunkten oder an ihrer Beschränktheit.

#### 10.4 Verträglichkeit mit den arithmetischen Operationen

**Satz 10.5** Seien  $(a_n), (b_n)$  Folgen mit  $\lim_{n \rightarrow \infty} a_n = a$  und  $\lim_{n \rightarrow \infty} b_n = b$ . Dann sind auch die Folgen  $(a_n + b_n)$ ,  $(a_n \cdot b_n)$  und, falls  $b_n, b \neq 0$ , auch  $(a_n/b_n)$  konvergent und es gilt

$$a_n + b_n \rightarrow a + b, \quad a_n b_n \rightarrow ab, \quad \frac{a_n}{b_n} \rightarrow \frac{a}{b} \quad \text{für } n \rightarrow \infty.$$

*Beweis:* Sei  $\varepsilon > 0$  vorgegeben. Nach Definition der Konvergenz gibt es  $N_1 \in \mathbb{N}$  mit  $|a_n - a| < \varepsilon$  für alle  $n \geq N_1$  und  $N_2 \in \mathbb{N}$  mit  $|b_n - b| < \varepsilon$  für alle  $n \geq N_2$ . Für  $n \geq \max\{N_1, N_2\}$  sind dann beide Ungleichungen erfüllt. Für diese  $n$  folgt aus der Dreiecksungleichung

$$|a_n + b_n - (a + b)| \leq |a_n - a| + |b_n - b| < 2\varepsilon.$$

Damit liegen in jeder Umgebung  $B_{2\varepsilon}(a+b)$  alle bis auf endlich viele Folgenglieder, also  $\lim_{n \rightarrow \infty} (a_n + b_n) = a + b$ .

Da eine konvergente Folge beschränkt ist, gilt  $|b_n| \leq M$ . Aus der Dreiecksungleichung folgt für  $n \geq \max\{N_1, N_2\}$

$$\begin{aligned}|a_n b_n - ab| &= |a_n b_n - ab_n + ab_n - ab| \leq |a_n b_n - ab_n| + |ab_n - ab| \\ &\leq M|a_n - a| + |a||b_n - b| < (M + |a|)\varepsilon,\end{aligned}$$

was  $\lim_{n \rightarrow \infty} a_n b_n = ab$  impliziert.

Für die Konvergenz des Quotienten genügt es  $\frac{1}{b_n} \rightarrow \frac{1}{b}$  nachzuweisen. Die Aussage folgt dann aus der Konvergenz des Produkts. Zu  $\varepsilon = |b|/2$  gibt es ein  $N_3$  mit

$$|b_n| = |b - b + b_n| \geq |b| - |b - b_n| > |b| - |b|/2 = |b|/2$$

für alle  $n \geq N_3$ . Für  $n \geq \max\{N_1, N_2, N_3\}$  gilt

$$\left| \frac{1}{b_n} - \frac{1}{b} \right| = \left| \frac{b - b_n}{b_n b} \right| \leq \frac{2}{|b|^2} |b_n - b|,$$

$$\text{also } \lim_{n \rightarrow \infty} \frac{1}{b_n} = \frac{1}{b}. \quad \square$$

**Beispiel 10.6** Den Grenzwert der Folge

$$a_n = \frac{2n^3 + 2n^2 + n}{n^3 + 1} = \frac{2 + \frac{2}{n} + \frac{1}{n^2}}{1 + \frac{1}{n^3}}$$

können wir leicht mit diesem Satz bestimmen, weil Zähler und Nenner gegen 2 bzw. 1 konvergieren, also  $a_n \rightarrow 2$ .  $\square$

## 10.5 Grenzwerte wichtiger Folgen

**Satz 10.7 (Bernoulli-Ungleichung)** Für jede reelle Zahl  $a \geq -1$  und für jedes  $n \in \mathbb{N}_0$  gilt

$$(B_n) \quad (1+a)^n \geq 1 + na.$$

*Beweis:* In diesem Fall können wir die Induktion mit  $n_0 = 0$  verankern, denn  $(1+a)^0 = 1$ . Für  $n \geq 0$  gilt unter Verwendung der Induktionsvoraussetzung  $(B_n)$

$$\begin{aligned}(1+a)^{n+1} &= (1+a)^n(1+a) \\ &\geq (1+na)(1+a) = 1 + na + a + na^2 \\ &\geq 1 + (n+1)a.\end{aligned}$$

$\square$

Nun bestimmen wir die Grenzwerte einiger prominenter Folgen. Für die geometrische Folge  $a_n = q^n$  für  $q \in \mathbb{R}$  gilt

$$q^n \rightarrow 0 \text{ falls } |q| < 1, \quad |q|^n \text{ ist unbeschränkt für } |q| > 1.$$

Ist nämlich  $|q| = 1 + x$  mit  $x > 0$ , so folgt aus der Bernoulli-Ungleichung

$$|q|^n \geq 1 + nx.$$

Ist dagegen  $|q| < 1$ , so ist aufgrund der letzten Abschätzung  $|q|^{-n} \geq 1 + nx$ , also  $|q|^n \leq 1/(1+nx) \rightarrow 0$ .

Man kann das letzte Beispiel noch verschärfen. Es gilt für beliebiges, aber fest gewähltes  $m \in \mathbb{N}$

$$(10.1) \quad \lim_{n \rightarrow \infty} n^m q^n = 0 \text{ falls } |q| < 1.$$

Anschaulich bedeutet dies, daß  $q^n$  „schneller“ gegen Null konvergiert als  $n^m$  gegen unendlich geht. Der Beweis ist mit unseren bisherigen Mitteln nur sehr aufwendig zu erbringen und wird noch zurückgestellt (siehe 10.13).

Es gilt

$$(10.2) \quad \lim_{n \rightarrow \infty} \sqrt[n]{a} = 1$$

für jede reelle Zahl  $a > 0$ . Für den Beweis sei zunächst  $a \geq 1$ . Dann ist  $b_n = \sqrt[n]{a} - 1 \geq 0$  und aus der Bernoulli-Ungleichung folgt

$$a = (1 + b_n)^n \geq 1 + nb_n.$$

Damit  $b_n \leq (a - 1)/n \rightarrow 0$  und  $\lim_{n \rightarrow \infty} \sqrt[n]{a} = 1$ . Für  $0 < a < 1$  gilt  $\lim_{n \rightarrow \infty} \sqrt[n]{a^{-1}} = 1$  und nach Satz 10.4  $\lim_{n \rightarrow \infty} \sqrt[n]{a} = 1/1 = 1$ .

Es gilt  $\lim_{n \rightarrow \infty} \sqrt[n]{n} = 1$ . Analog zum vorigen Fall setzen wir  $b_n = \sqrt[n]{n} - 1 \geq 0$ . Mit der binomischen Formel folgt

$$n = (1 + b_n)^n \geq 1 + \binom{n}{2} b_n^2,$$

für  $n \geq 2$  also  $b_n^2 \leq 2/n \rightarrow 0$  und  $b_n \rightarrow 0$ . Aufgrund von Satz 10.4 gilt für fest gewähltes  $m \in \mathbb{N}$

$$(10.3) \quad \lim_{n \rightarrow \infty} \sqrt[m]{n^m} = \lim_{n \rightarrow \infty} \sqrt[n]{n} \dots \lim_{n \rightarrow \infty} \sqrt[n]{n} = 1.$$

**10.6 Konvergenz monotoner Folgen** Wir bezeichnen eine Folge  $(a_n)$  als *monoton wachsend (fallend)*, wenn für alle  $n$  die Bedingung  $a_n \leq a_{n+1}$  bzw.  $a_n \geq a_{n+1}$  erfüllt ist. Eine Folge heißt *strengh monoton wachsend oder fallend*, wenn für alle  $n$  die strikte Ungleichung erfüllt ist. Konvergiert eine monoton wachsende Folge  $(a_n)$  gegen  $a$ , so schreiben wir  $a_n \nearrow a$ , konvergiert sie monoton fallend, so  $a_n \searrow a$ .

**Satz 10.8** *Eine beschränkte, monoton wachsende oder fallende Folge ist konvergent.*

*Beweis:* Sei  $(a_n)$  monoton wachsend und beschränkt. Dann ist die zugehörige Menge  $\{a_n\}_{n \in \mathbb{N}}$  nach oben beschränkt und besitzt ein Supremum  $a$ , für das also  $a_n \leq a$  gilt. Aus der Definition des Supremums folgt, daß es zu jedem  $\varepsilon > 0$  ein  $N \in \mathbb{N}$  gibt mit  $a_N + \varepsilon \geq a$ , denn andernfalls wäre  $a - \varepsilon$  ebenfalls eine obere Schranke. Da die Folge monoton wachsend ist, gilt  $0 \leq a - a_n \leq \varepsilon$  für alle  $n \geq N$  und somit  $\lim_{n \rightarrow \infty} a_n = a$ .  $\square$

**Beispiel 10.9** Dieser Satz wird häufig verwendet, um die Konvergenz rekursiv definierter Folgen nachzuweisen. Als ein Beispiel betrachten wir die Folge

$$a_{n+1} = \sqrt{6 + a_n}, \quad a_0 = 0.$$

Durch Induktion über  $n$  zeigen wir, daß die Folge streng monoton wachsend ist. Der Induktionsanfang  $a_1 > a_0$  ist richtig. Ist  $a_n > a_{n-1}$ , so  $a_{n+1} = \sqrt{6 + a_n} > \sqrt{6 + a_{n-1}} = a_n$ .

Ebenfalls durch Induktion wird bewiesen, daß die Folge durch 3 nach oben beschränkt ist. Für  $a_0$  ist das richtig. Gilt  $a_n < 3$ , so ist  $a_{n+1} = \sqrt{6 + a_n} < \sqrt{6 + 3} = 3$ .

Damit haben wir gezeigt, daß die Folge konvergiert. Der Grenzwert kann mit einer Methode bestimmt werden, die in Kapitel 11 erläutert wird.  $\square$

**10.7 Teilfolgen und der Satz von Bolzano-Weierstraß** Sei  $(a_n)_{n \in \mathbb{N}}$  eine Folge. Für eine streng monoton wachsende Folge  $(n_k)_{k \in \mathbb{N}}$  natürlicher Zahlen heißt  $(a_{n_k})_{k \in \mathbb{N}}$  *Teilfolge* von  $(a_n)_{n \in \mathbb{N}}$ . Eine Teilfolge besteht ebenfalls aus unendlich vielen Elementen und ist daher selber eine Folge.

Kehren wir zur Folge  $a_n = (-1)^n + \frac{1}{n}$  zurück. Mit  $n_k = 2k$  ist  $a_{n_k} = 1 + \frac{1}{2k}$  eine Teilfolge, die gegen 1 konvergiert. Durch Auswahl einer Teilfolge können wir in diesem Beispiel einen Häufungspunkt zum Grenzwert der Teilfolge machen. Daß dies immer möglich ist, zeigt der folgende Satz.

**Satz 10.10** *Sei  $(a_n)_{n \in \mathbb{N}}$  eine Folge mit einem Häufungspunkt  $a$ . Dann existiert eine Teilfolge  $(a_{n_k})_{k \in \mathbb{N}}$  von  $(a_n)_{n \in \mathbb{N}}$  mit  $\lim_{k \rightarrow \infty} a_{n_k} = a$ .*

*Beweis:* Wir bestimmen die Folgenglieder  $a_{n_k}$  induktiv. Seien  $a_{n_1}, \dots, a_{n_k}$  mit  $(n_i)_{i=1, \dots, k}$  streng monoton wachsend bereits konstruiert. Zu  $\varepsilon = 1/(k+1)$  liegen in  $B_{1/(k+1)}(a)$  unendlich viele Folgenglieder. Aus diesen wählen wir ein beliebiges  $a_{n_{k+1}}$  mit  $n_{k+1} > n_k$  aus. Dann gilt  $|a_{n_{k+1}} - a| < 1/(k+1)$ , woraus  $\lim_{k \rightarrow \infty} a_{n_k} = a$  folgt.  $\square$

**Satz 10.11 (Bolzano-Weierstraß)** *Jede beschränkte Folge besitzt einen Häufungspunkt. Insbesondere enthält jede beschränkte Folge eine konvergente Teilfolge. Ferner besitzt eine beschränkte Folge einen größten und einen kleinsten Häufungspunkt.*

*Beweis:* Sei  $a_n \in (c, d)$  für alle  $n$ . Die Menge

$$M = \{x \in \mathbb{R} : a_n > x \text{ für höchstens endlich viele } n\}$$

ist nichtleer, weil  $d \in M$ , und sie ist nach unten beschränkt durch  $c$ . Wir zeigen, daß  $a = \inf M$  ein Häufungspunkt und zwar der größte Häufungspunkt ist. Nach Definition des Infimums ist für beliebiges  $\varepsilon > 0$   $a + \varepsilon \in M$  und  $a - \varepsilon \notin M$ . Es gibt daher höchstens endlich viele Folgenglieder mit  $a_n > a + \varepsilon$  und es gibt unendlich viele Folgenglieder mit  $a_n > a - \varepsilon$ . Daher ist  $a$  Häufungspunkt. Angenommen, es gibt einen weiteren Häufungspunkt  $b > a$ . Dann wählt man einen Punkt  $\xi$  zwischen  $a$  und  $b$ . Da oberhalb von  $\xi$  nur endlich viele Folgenglieder liegen, kann  $b$  kein Häufungspunkt sein.  $\square$

**10.8 Das Cauchy-Kriterium** Eine Folge  $(a_n)$  heißt *Cauchy-Folge*, wenn es zu jedem  $\varepsilon > 0$  ein  $N \in \mathbb{N}$  gibt mit

$$|a_m - a_n| < \varepsilon \quad \text{für alle } m, n \geq N.$$

Wie wir gleich sehen werden, ist eine Folge genau dann konvergent, wenn sie eine Cauchy-Folge ist. Dennoch ist der Begriff der Cauchy-Folge oft nützlich, weil in ihrer Definition der Grenzwert der Folge nicht vorkommt.

**Satz 10.12** *Ein Folge ist genau dann konvergent, wenn sie eine Cauchy-Folge ist.*

*Beweis:* Sei  $a_n \rightarrow a$ . Zu  $\varepsilon > 0$  sei  $N \in \mathbb{N}$  mit  $|a_n - a| < \varepsilon$  für alle  $n \geq N$ . Für  $m, n \geq N$  folgt dann aus der Dreieckungleichung

$$|a_m - a_n| = |a_m - a + a - a_n| \leq |a_m - a| + |a_n - a| < 2\varepsilon.$$

Ist umgekehrt  $(a_n)$  eine Cauchy-Folge, so wählen wir in der Definition der Cauchy-Folge  $\varepsilon = 1$  und erhalten für alle  $n$  größer gleich dem zugehörigen  $N$

$$|a_n| \leq |a_n - a_N| + |a_N| < 1 + |a_N|.$$

Die Cauchy-Folge ist damit beschränkt,

$$|a_n| \leq \max\{|a_1|, \dots, |a_{N-1}|, 1 + |a_N|\}.$$

Nach dem Satz von Bolzano-Weierstraß hat die Folge daher eine konvergente Teilfolge  $a_{n_k} \rightarrow a$  für  $k \rightarrow \infty$ . Wir zeigen, dass die gesamte Folge gegen  $a$  konvergiert. Sei  $\varepsilon > 0$  vorgegeben. Dann gibt es ein  $N \in \mathbb{N}$  mit  $|a_m - a_n| < \varepsilon$  für alle  $m, n \geq N$  und wegen der Konvergenz der Teilfolge ein  $n_k \geq N$  mit  $|a - a_{n_k}| < \varepsilon$ . Aus der Dreiecksungleichung folgt dann für alle  $n \geq N$

$$|a_n - a| \leq |a_n - a_{n_k}| + |a_{n_k} - a| < 2\varepsilon$$

und damit die Konvergenz der gesamten Folge gegen  $a$ .  $\square$

**10.9 Limes superior, Limes inferior und bestimmte Divergenz** Wir bezeichnen den größten Häufungspunkt  $a^*$  einer beschränkten Folge  $(a_n)$  als *Limes superior* und den kleinsten Häufungspunkt  $a_*$  als *Limes inferior* der Folge und schreiben

$$a^* = \limsup_{n \rightarrow \infty} a_n, \quad a_* = \liminf_{n \rightarrow \infty} a_n.$$

Das Verhalten unbeschränkter Folgen soll im folgenden weiter präzisiert werden. Eine Folge  $(a_n)$  divergiert bestimmt gegen unendlich, Schreibweise  $\lim_{n \rightarrow \infty} a_n = \infty$ , wenn es zu jedem  $M \in \mathbb{R}$  ein  $N \in \mathbb{N}$  gibt mit  $a_n \geq M$  für alle  $n \geq N$ . Die bestimmte Divergenz gegen  $-\infty$  ist analog definiert. Beispielsweise gilt  $\lim_{n \rightarrow \infty} n = \infty$ , aber  $b_n = (-1)^n n$  divergiert nicht bestimmt.

Diese Begriffsbildung lässt sich auch auf Häufungspunkte übertragen. Wir sagen, daß eine Folge  $(a_n)$  den *uneigentlichen Häufungspunkt*  $\infty$  hat, wenn eine Teilfolge von  $(a_n)$  bestimmt gegen  $\infty$  divergiert. Dies ist äquivalent zur Bedingung: Zu jedem  $M \in \mathbb{R}$  gibt es ein  $n \in \mathbb{N}$  mit  $a_n \geq M$ . Ferner schreiben wir in diesem Fall auch  $\limsup a_n = \infty$ .

**10.10 Definition und Beispiele von Reihen** Ein altes Problem der Analysis ist es, einer Reihe  $\sum_{n=1}^{\infty} a_n$  mit reellen Zahlen  $a_n$  einen „Wert“ zuzuordnen. Ein typisches Beispiel ist die unendliche Reihe  $1 - 1 + 1 - 1 + \dots$ , die der Theologe Giordano Bruno als Modell für die Erschaffung der Welt aus dem Nichts angesehen hat: Wir erhalten  $(1 - 1) + (1 - 1) + \dots = 0$ , aber auch  $1 - (1 - 1) - (1 - 1) - \dots = 1$ . Bruno ist später als Ketzer verbrannt worden, aber nicht deshalb.

Diese Konfusion hat sich erledigt, weil wir der Reihe die *Folge der Partialsummen* zuordnen

$$s_n = \sum_{k=1}^n a_k.$$

Konvergiert die Folge  $(s_n)_{n \in \mathbb{N}}$  gegen  $s$ , so nennen wir die Reihe *konvergent* mit Grenzwert  $s$  und schreiben  $s = \sum_{n=1}^{\infty} a_n$ . Konvergiert die Folge  $(s_n)$  nicht, so sagen wir, dass die Reihe divergiert. Im obigen Beispiel ist  $(s_n) = (1, 0, 1, 0, \dots)$  und die Reihe ist nicht konvergent.

Divergiert  $(s_n)$  bestimmt gegen unendlich, so schreiben wir  $\sum_{n=1}^{\infty} a_n = \infty$ .

So wie eine Reihe eine Folge erzeugt, kann man einer Folge  $(s_n)$  auch eine Reihe zuordnen, nämlich

$$a_1 = s_1, \quad a_2 = s_2 - s_1, \quad a_3 = s_3 - s_2, \quad \dots$$

Die Folge der Partialsummen der Reihe  $\sum a_n$  ist also gerade  $(s_n)$ . Die Begriffe „Reihe“ und „Folge“ sind damit vollständig äquivalent. Aus Satz 10.4 erhält man insbesondere, dass die Summe zweier konvergenter Reihen wieder konvergent ist. Das Produkt zweier Reihen wird später betrachtet.

Es dürfte klar sein, dass man das Konvergenzverhalten einer Reihe nicht ändert, wenn man endlich viele Reihenglieder weglässt oder hinzufügt. Die Konvergenztheorie von Reihen der Form  $\sum_{n=k}^{\infty} a_n$  mit  $k \in \mathbb{Z}$  hängt daher nicht von  $k$  ab.

Fundamental ist die *geometrische Reihe*  $\sum_{n=0}^{\infty} q^n$ , für deren Partialsummen  $(s_n)_{n \in \mathbb{N}_0}$  nach der geometrischen Summenformel

$$s_n = \sum_{k=0}^n q^k = \frac{1 - q^{n+1}}{1 - q}, \quad q \neq 1$$

gilt. Daher ist

$$\sum_{n=0}^{\infty} q^n = \begin{cases} \frac{1}{1-q} & \text{für } |q| < 1 \\ \infty & \text{für } q \geq 1 \end{cases}$$

Eine prominentes Beispiel für eine bestimmt divergente Reihe ist die *harmonische Reihe*

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{1}{n} &= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) + \dots \\ &\geq 1 + \frac{1}{2} + 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{8} + \dots \end{aligned}$$

Damit gilt für die Folge der Partialsummen  $\lim_{n \rightarrow \infty} s_n = \infty$  und die Reihe divergiert bestimmt.

Notwendig für die Konvergenz einer Reihe ist, dass die  $(a_n)$  eine Nullfolge bilden, denn die Partialsummen können höchstens dann konvergieren, wenn  $|s_n - s_{n+1}| \rightarrow 0$ . Das letzte Beispiel zeigt aber auch, dass diese Bedingung nicht hinreichend ist.

**10.11 Alternierende Reihen** Ist eine Reihe *alternierend*, haben also die Reihenglieder wechselndes Vorzeichen, so gilt:

**Satz 10.13 (Leibniz-Kriterium)** Sind die Reihenglieder von der Form  $a_n = (-1)^n b_n$  und ist  $(b_n)$  eine streng monoton fallende Nullfolge, so ist die Reihe  $\sum_{n=k}^{\infty} a_n$  konvergent und für den Reihenrest gilt

$$r_l = \sum_{n=l+1}^{\infty} a_n = \theta a_{l+1} \quad \text{mit } 0 < \theta < 1.$$

*Beweis:* Wir können annehmen, dass die Reihe bei  $n = 0$  beginnt. Dann ist  $a_0 = b_0 > 0$  und wir erhalten für die ungeraden Partialsummen

$$s_{2n+1} = (b_0 - b_1) + (b_2 - b_3) + \dots + (b_{2n} - b_{2n+1}).$$

$(s_{2n+1})$  ist also monoton wachsend. Entsprechend sind die geraden Partialsummen

$$s_{2n} = b_0 - (b_1 - b_2) - \dots - (b_{2n-1} - b_{2n})$$

monoton fallend. Wegen  $a_{2n+1} < 0$  gilt ferner  $0 < s_{2n+1} < s_{2n} < b_0$ . Nach Satz 10.6 haben die monotonen und beschränkten Folgen  $(s_{2n+1})$  und  $(s_{2n})$  jeweils einen Grenzwert, der wegen  $|s_{2n+1} - s_{2n}| \rightarrow 0$  in beiden Fällen der gleiche sein muß. Damit ist dieser Grenzwert  $s$  auch Grenzwert der Reihe. Es gilt  $0 < s < b_0 = a_0$ , also  $s = \theta a_0$  für  $0 < \theta < 1$ . Die gleiche Überlegung können wir für den Reihenrest machen, der ja wiederum eine alternierende Reihe ist. Damit ist auch die Fehlerabschätzung bewiesen.  $\square$

Der Satz ist nicht richtig, wenn  $(b_n)$  eine nichtnegative Nullfolge ist, die Monotonie ist wesentlich.

**Beispiel 10.14** Leibniz hat bewiesen, dass

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \dots$$

Das können wir im Moment nicht nachvollziehen, aber wir können das Ergebnis mit Hilfe der Fehlerabschätzung für den Reihenrest überprüfen. In der ersten Zeile führen wir die Partialsummen  $s_n$  auf und in der zweiten die vom Satz garantierte Abschätzung für den Reihenrest, also  $|a_{n+1}|$ , der exakte Wert ist  $\frac{\pi}{4} = 0.785\dots$

$s_n$	1.00	0.666	0.866	0.723	0.834	0.734
$ r_n  \leq$	0.333	0.200	0.143	0.111	0.100	0.083

Zur Berechnung einer Milliarden Stellen von  $\pi$  ist dieses Verfahren offenbar nicht geeignet.  $\square$

**10.12 Absolute Konvergenz von Reihen und Cauchy-Kriterium** Nach dem Leibniz-Kriterium ist die alternierende harmonische Reihe

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$$

konvergent, aber sie ist nicht robust gegen eine veränderte Auswertung. Wir können ja versuchen, zunächst die Teilreihe der positiven Gliedern auszurechnen und darauf die Reihe mit den negativen Gliedern aufzuaddieren. Es gilt aber

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{1}{2n-1} &= 1 + \frac{1}{3} + \left(\frac{1}{5} + \frac{1}{7}\right) + \left(\frac{1}{9} + \frac{1}{11} + \frac{1}{13} + \frac{1}{15}\right) + \dots \\ &\geq 1 + \frac{1}{3} + 2 \cdot \frac{1}{8} + 4 \cdot \frac{1}{16} + \dots \end{aligned}$$

und die Teilreihe ist divergent. Wir nennen eine Reihe  $\sum_{n=1}^{\infty} a_n$  absolut konvergent, wenn die Reihe  $\sum_{n=1}^{\infty} |a_n|$  konvergent ist. Zum Beispiel ist die geometrische Reihe  $\sum_{n=1}^{\infty} q^n$  absolut konvergent für alle  $|q| < 1$ .

Die Folge der Partialsummen einer Reihe ist nach Satz 10.8 genau dann konvergent, wenn sie eine Cauchy-Folge ist. Angewendet auf Reihen sieht das folgendermaßen aus: Zu jedem  $\varepsilon > 0$  gibt es ein  $N \in \mathbb{N}$ , so dass für alle  $n \geq m \geq N$  gilt

$$\left| \sum_{k=m}^n a_k \right| < \varepsilon.$$

Man nennt diese Bedingung auch *Cauchy-Kriterium für Reihen*. Mit diesem Kriterium lässt sich leicht zeigen, dass eine absolut konvergente Reihe auch konvergent ist,

$$\left| \sum_{k=m}^n a_k \right| \leq \sum_{k=m}^n |a_k|.$$

Da die Partialsummen von  $\sum |a_n|$  eine Cauchy-Folge bilden, gilt gleiches für die Partialsummen von  $\sum a_n$ . Die Abschätzung

$$\left| \sum_{n=1}^{\infty} a_n \right| \leq \sum_{n=1}^{\infty} |a_n|.$$

folgt aus der analogen Abschätzung für die Partialsummen und Grenzübergang.

Ohne Beweis sei vermerkt, dass man eine absolut konvergente Reihe beliebig umordnen, also die Reihe in beliebiger Reihenfolge auswerten kann, ohne den Grenzwert zu verändern.

### 10.13 Kriterien für die absolute Konvergenz von Reihen

**Satz 10.15** (a) Majorantenkriterium: Ist  $|a_n| \leq c_n$  mit  $\sum_{n=1}^{\infty} c_n$  konvergent, so ist  $\sum_{n=1}^{\infty} a_n$  absolut konvergent.

(b) Quotientenkriterium: Sei  $a_n \neq 0$ . Existiert eine Zahl  $q$  mit  $0 < q < 1$  und

$$\left| \frac{a_{n+1}}{a_n} \right| \leq q < 1 \quad \text{für fast alle } n,$$

so ist die Reihe  $\sum_{n=1}^{\infty} a_n$  absolut konvergent. Ist dagegen

$$\left| \frac{a_{n+1}}{a_n} \right| \geq 1 \quad \text{für fast alle } n,$$

so ist die Reihe divergent.

(c) Wurzelkriterium: Existiert eine Zahl  $q$  mit  $0 < q < 1$  und

$$\sqrt[n]{|a_n|} \leq q < 1 \quad \text{für fast alle } n,$$

so ist die Reihe  $\sum_{n=1}^{\infty} a_n$  absolut konvergent. Ist dagegen

$$\sqrt[n]{|a_n|} \geq 1 \quad \text{für unendlich viele } n,$$

so ist die Reihe divergent.

*Beweis:* (a) Es ist klar, dass die Folge der Partialsummen der Reihe  $\sum_{n=1}^{\infty} |a_n|$  beschränkt bleibt, wenn die Reihe  $\sum_{n=1}^{\infty} c_n$  konvergiert.

(b) Da das Weglassen von endlich vielen Gliedern nichts am Konvergenzverhalten einer Reihe ändert, können wir annehmen, dass die Bedingung für alle  $n \in \mathbb{N}_0$  erfüllt ist. Aus  $|a_{n+1}| \leq q|a_n|$  folgt  $|a_n| \leq q^n|a_0|$ . Damit ist  $c_n = q^n|a_0|$  wegen  $0 < q < 1$  eine konvergente Majorante der Reihe. Ist  $|a_{n+1}| \geq |a_n|$ , so folgt entsprechend  $|a_n| \geq |a_0|$  für alle  $n$ . Damit ist die Folge  $(a_n)$  keine Nullfolge und die Reihe  $\sum_{n=1}^{\infty} a_n$  ist nicht konvergent.

(c) Für fast alle Folgenglieder gilt  $|a_n| \leq q^n$ , die Behauptung folgt wieder aus der Konvergenz der geometrischen Reihe. Aus der zweiten Bedingung folgt  $|a_n| \geq 1$  für unendlich viele  $n$ . Damit ist  $(a_n)$  keine Nullfolge.  $\square$

Implizit wurde hier auch das *Minorantenkriterium* verwendet: Ist  $|a_n| \geq b_n \geq 0$  und die Reihe  $\sum b_n$  divergent, so ist auch die Ausgangsreihe  $\sum a_n$  nicht absolut konvergent.

Für die harmonische Reihe  $\sum_{n=1}^{\infty} \frac{1}{n}$  gilt

$$\left| \frac{a_{n+1}}{a_n} \right| < 1, \quad \sqrt[n]{|a_n|} < 1 \quad \text{für alle } n,$$

aber die harmonische Reihe ist divergent. Sowohl im Quotienten- als auch im Wurzelkriterium ist es also ganz wesentlich, dass das  $q < 1$  unabhängig von  $n$  gewählt werden kann.

**Beispiele 10.16** (i)  $\sum_{n=1}^{\infty} n^m q^n$  ist absolut konvergent für  $m \in \mathbb{N}_0$  und  $|q| < 1$ . Mit (10.3) gilt  $\sqrt[n]{n^m} \rightarrow 1$  für  $n \rightarrow \infty$  und

$$\sqrt[n]{n^m |q|^n} = \sqrt[n]{n^m} |q| \leq (1 + \varepsilon)q \quad \text{für alle } n \geq N(\varepsilon).$$

Wir wählen hier  $\varepsilon$  so klein, dass  $(1 + \varepsilon)|q| < 1$  und haben damit das Wurzelkriterium erfüllt. Da die Glieder einer konvergenten Reihe eine Nullfolge bilden, haben wir auch  $\lim_{n \rightarrow \infty} n^m q^n = 0$  und damit (10.1) gezeigt.

(ii) Ein typisches Beispiel für die Anwendung des Quotientenkriteriums ist die Reihe  $\sum \frac{q^n}{n!}$ . Mit

$$\left| \frac{a_{n+1}}{a_n} \right| = \frac{|q|^{n+1} n!}{(n+1)! |q|^n} = \frac{|q|}{n+1}$$

haben wir das Quotientenkriterium für alle  $q \in \mathbb{R}$  erfüllt.

(iii) Die Reihe  $\sum_{n=1}^{\infty} \frac{1}{n^2}$  ist konvergent, aber sowohl das Wurzel- als auch das Quotientenkriterium versagen. Beide Kriterien beruhen ja auf einer Majorisierung durch die geometrische Reihe, was ein ziemlich grober Klotz ist.  $\square$

## 11 Funktionen und Stetigkeit

**11.1 Beispiele von Funktionen** Für eine Menge  $D \subset \mathbb{R}$  bezeichnen wir die Abbildungen  $f : D \rightarrow \mathbb{R}$  als *Funktionen*.  $D$  heißt *Definitionsbereich*,

$$\mathcal{R}(f) = f(D) = \{y = f(x) \text{ für ein } x \in D\}$$

heißt *Wertebereich* der Funktion  $f$ . Gilt  $f(x) = 0$ , so heißt  $x$  *Nullstelle* von  $f$ . In den meisten Fällen ist  $D$  ein Intervall oder die Vereinigung von Intervallen.

Seien  $a_n, a_{n-1}, \dots, a_1, a_0$  reelle Zahlen. Dann heißt

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$$

*Polynom*. Ist  $a_n \neq 0$ , so heißt  $n$  der *Grad* von  $p$  und wir schreiben  $\text{grad } p = n$ . Ein Polynom ist für jedes  $x \in \mathbb{R}$  definiert.

Sind  $p(x)$  und  $q(x)$  Polynome, so heißt  $r(x) = \frac{q(x)}{p(x)}$  *rationale Funktion*. Eine rationale Funktion ist außerhalb der Nullstellen des Nennerpolynoms  $p(x)$  definiert.

Eine Funktion, die sich aus Wurzelausdrücken und rationalen Funktionen zusammensetzt, heißt *algebraische Funktion*. Ein Beispiel für eine algebraische Funktion ist  $f(x) = \sqrt{1 - x^2}$ . Beim Definitionsbereich algebraischer Funktionen ist zu beachten, dass Wurzeln nur aus nichtnegativen Zahlen gezogen werden. In unserem Beispiel ist daher  $D = [-1, 1]$ .

Seien  $I_1, \dots, I_n$  disjunkte Intervalle und  $D = \cup I_k$ . Eine Funktion  $f : D \rightarrow \mathbb{R}$  mit  $f$  konstant auf jedem  $I_k$  heißt *stückweise konstante Funktion*.

**11.2 Grenzwerte von Funktionen** Wir hatten bereits Häufungspunkte von Zahlenfolgen definiert.  $a$  hieß Häufungspunkt der Folge  $(a_n)$ , wenn in jeder Umgebung von  $a$  unendlich viele Folgenglieder liegen. Diesen Begriff können wir auch auf Mengen reeller Zahlen übertragen. Ist  $A \subset \mathbb{R}$ , so heißt  $a \in \mathbb{R}$  *Häufungspunkt von A*, wenn in jeder  $\varepsilon$ -Umgebung  $B_\varepsilon(a) = (a - \varepsilon, a + \varepsilon)$  unendlich viele Punkte von  $A$  liegen.

Sei  $\xi$  Häufungspunkt des Definitionsbereichs  $D$  der Funktion  $f$ . Wir sagen,  $f$  konvergiert gegen  $a$  für  $x \rightarrow \xi$ ,

$$\lim_{x \rightarrow \xi} f(x) = a \quad \text{oder} \quad f(x) \rightarrow a \text{ für } x \rightarrow \xi,$$

wenn für alle Folgen  $(x_n)$  mit  $x_n \rightarrow \xi$  und  $x_n \neq \xi$  gilt  $f(x_n) \rightarrow a$ . Die Definition wird sinngemäß auch für die Werte  $\xi = \pm\infty$  verwendet: Wir schreiben  $\lim_{x \rightarrow \infty} f(x) = a$ , wenn für jede Folge  $(x_n)$ , die bestimmt gegen  $\infty$  divergiert, gilt  $\lim f(x_n) = a$ .

**Beispiel 11.1** Für  $f(x) = \frac{x}{1+x}$  gilt  $\lim_{x \rightarrow 0} f(x) = 0$  und  $\lim_{x \rightarrow \infty} f(x) = 1$ .  $\square$

**11.3 Stetigkeit von Funktionen** Sei  $f : D \rightarrow \mathbb{R}$  eine Funktion. Die Funktion  $f$  heißt *stetig* in  $\xi \in D$ , wenn für jede Folge  $(x_n)$  mit  $x_n \rightarrow \xi$  gilt  $f(x_n) \rightarrow f(\xi)$ .  $f$  heißt *stetig in D*, wenn  $f$  in jedem Punkt von  $D$  stetig ist.

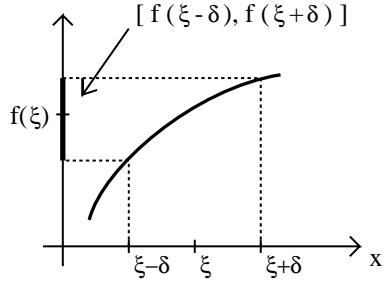
Ist  $\xi \in D$  Häufungspunkt von  $D$ , so ist die Stetigkeit von  $f$  äquivalent dazu, dass der Grenzwert  $\lim_{x \rightarrow \xi} f(x)$  existiert und mit  $f(\xi)$  übereinstimmt. Anschaulich kommen die Werte von  $f(x)$  dem Wert  $f(\xi)$  immer näher, wenn  $x$  dem Punkt  $\xi$  immer näher kommt. Diese Vorstellung lässt sich präzise fassen.

**Satz 11.2** Die Funktion  $f$  ist genau dann stetig in  $\xi \in D$ , wenn es zu jedem  $\varepsilon > 0$  ein  $\delta > 0$  gibt, so dass für alle  $x \in D$  mit  $|x - \xi| < \delta$  folgt

$$|f(x) - f(\xi)| < \varepsilon.$$

*Beweis:* Sei  $f$  stetig in  $\xi$ . Angenommen, die Bedingung des Satzes ist nicht erfüllt. Dann gibt es ein  $\varepsilon > 0$ , so dass für alle  $\delta > 0$  ein  $x_\delta \in D$  existiert mit  $|x_\delta - \xi| < \delta$  und  $|f(x_\delta) - f(\xi)| \geq \varepsilon$ . Speziell können wir hier  $\delta = \frac{1}{n}$  wählen und  $x_\delta$   $x_n$  nennen. Dann gilt  $x_n \rightarrow \xi$ , aber  $f(x_n) \not\rightarrow f(\xi)$ . Widerspruch!

Nun zeigen wir die umgekehrte Richtung. Sei  $(x_n)$  eine Folge mit  $x_n \rightarrow \xi$ . Zu vorgegebenem  $\varepsilon > 0$  gibt es ein  $\delta > 0$ , so dass  $|x - \xi| < \delta$  gerade  $|f(x) - f(\xi)| < \varepsilon$  impliziert. Wegen  $x_n \rightarrow \xi$  gibt es ein  $N \in \mathbb{N}$  mit  $|x_n - \xi| < \delta$  für alle  $n \geq N$ , daher  $|f(x_n) - f(\xi)| < \varepsilon$ . Damit ist  $f(x_n) \rightarrow f(\xi)$  erfüllt.  $\square$



**Beispiele 11.3** (i) Jede konstante Funktion ist stetig, denn wenn  $f(x) = a$  für alle  $x \in D$ , so folgt aus  $x_n \rightarrow \xi$ , dass  $a = f(x_n) = f(\xi)$ .

(ii) Die Funktion  $f(x) = x$  ist stetig, denn wenn  $x_n \rightarrow \xi$ , so trivialerweise  $x_n = f(x_n) \rightarrow f(\xi) = \xi$ .

(iii) Der Absolutbetrag  $f(x) = |x|$  ist stetig. Er stimmt für  $x \neq 0$  mit  $-x$  oder  $x$  überein. Beide Funktionen sind nach (ii) stetig. Für  $\xi = 0$  ist der Nachweis der Stetigkeit auch kein Problem: Ist  $x_n \rightarrow 0$ , so auch  $|x_n| \rightarrow 0$ .

(iv) Die Signum-Funktion ist im Punkt 0 unstetig. Für Folgen  $(x_n)$  mit  $x_n \nearrow 0$  folgt  $f(x_n) \rightarrow -1$  und für Folgen mit  $x_n \searrow 0$  gilt  $f(x_n) \rightarrow 1$ . Dagegen ist  $f(0) = 0$ . Dieses Argument kann auf alle Funktionen mit einer „Sprungstelle“ wie etwa stückweise konstante Funktionen übertragen werden.

(v) Die *Dirichlet-Funktion*  $f : [0, 1] \rightarrow \mathbb{R}$  mit

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

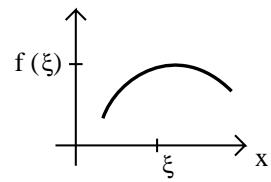
ist in jedem Punkt unstetig. Denn ist  $\xi$  irrational, so gibt es nach Abschnitt 4.3 eine Folge rationaler Zahlen  $(x_n)$  mit  $x_n \rightarrow \xi$  und daher  $1 = f(x_n) \not\rightarrow f(\xi) = 0$ . Ist  $\xi$  dagegen rational, so ist beispielsweise  $x_n = \xi + \sqrt{2}/n$  irrational mit  $x_n \rightarrow \xi$  und es folgt  $0 = f(x_n) \not\rightarrow f(\xi) = 1$ .

(vi) Die Funktion  $f : (0, 1] \rightarrow \mathbb{R}$  mit

$$f(x) = \begin{cases} 0 & \text{für } x \text{ irrational} \\ \frac{1}{k} & \text{für } x = \frac{k}{l} \text{ mit } k, l \in \mathbb{N} \text{ teilerfremd} \end{cases}$$

ist in jedem rationalen Punkt unstetig. Denn wie in (v) gezeigt wurde, gibt es zu jeder rationalen Zahl  $\xi$  eine Folge von Irrationalzahlen  $(x_n)$  mit  $x_n \rightarrow \xi$ , aber  $0 = f(x_n) \not\rightarrow f(\xi) > 0$ . Ist  $\xi$  irrational, so brauchen wir nur eine Folge rationaler Zahlen  $x_n = \frac{k_n}{l_n}$  zu betrachten mit  $x_n \rightarrow \xi$ . Da es nur endlich viele Zahlen der Form  $\frac{k}{l}$  mit  $k, l \leq K$  gibt, muss zwangsläufig  $k_n$  bestimmt gegen unendlich divergieren. Daher ist  $f$  in jedem irrationalen Punkt stetig.  $\square$

Anschaulich bedeutet die Stetigkeit einer Funktion, dass man ihren Graphen in einem Zug, ohne abzusetzen, zeichnen kann. Gilt für eine in  $\xi$  stetige Funktion  $f$ , dass  $f(\xi) > 0$ , so gibt es eine Umgebung  $B_\varepsilon(\xi) = (\xi - \varepsilon, \xi + \varepsilon)$  mit  $f(x) > 0$  für alle  $x \in B_\varepsilon(\xi)$ . Dies ist anschaulich klar, lässt sich aber auch leicht indirekt beweisen. Denn andernfalls gäbe es in jedem  $B_{1/n}(\xi)$  ein  $x_n$  mit  $f(x_n) \leq 0$ . Diese  $x_n$  bilden eine Folge mit  $x_n \rightarrow \xi$  und  $f(x_n) \rightarrow f(\xi) > 0$ , was einen Widerspruch bedeutet.



Wir nennen eine Funktion *von rechts stetig*, wenn die Einschränkung von  $f$  auf die Menge

$$D^+ = \{x \in D : x \geq \xi\}$$

in  $\xi$  stetig ist. In diesem Fall schreiben wir

$$f(\xi+) = f(\xi + 0) = \lim_{x \rightarrow \xi^+} f(x) = \lim_{x \searrow \xi} f(x).$$

Die Stetigkeit von links wird ganz analog definiert und bezeichnet.

**Beispiel 11.4** Die Funktion  $f(x) = [x]$  = größte ganze Zahl  $\leq x$  ist für jedes  $p \in \mathbb{Z}$  unstetig, wegen  $f(p+) = p$  ist sie in jedem Punkt von rechts stetig.  $\square$

#### 11.4 Stetigkeit und arithmetische Operationen

**Satz 11.5** Sind  $f, g : D \rightarrow \mathbb{R}$  stetig und sind  $\alpha, \beta \in \mathbb{R}$ , so sind auch  $\alpha f + \beta g$ ,  $fg$  und, sofern  $g \neq 0$  in  $D$ , auch  $f/g$  stetig. Ist  $f$  auf dem Bildbereich von  $g$  definiert und stetig, so ist auch die Komposition  $f \circ g(x) = f(g(x))$  stetig.

*Beweis:* Ist  $x_n \rightarrow \xi$ , so gilt  $f(x_n) \rightarrow f(\xi)$ ,  $g(x_n) \rightarrow g(\xi)$ . Aus den Regeln für die Konvergenz von Zahlenfolgen in Satz 10.4 folgt dann, dass auch  $\alpha f(x_n) + \beta g(x_n) \rightarrow \alpha f(\xi) + \beta g(\xi)$ ,  $f(x_n)g(x_n) \rightarrow f(\xi)g(\xi)$ ,  $f(x_n)/g(x_n) \rightarrow f(\xi)/g(\xi)$ .

Ist  $f$  auf dem Bildbereich von  $g$  stetig, so folgt aus  $x_n \rightarrow \xi$ , dass  $g(x_n) \rightarrow g(\xi)$ . Für die Folge  $(g(x_n))$  können wir die Stetigkeit von  $f$  im Punkt  $g(\xi)$  verwenden und erhalten  $f(g(x_n)) \rightarrow f(g(\xi))$ .  $\square$

Nach obigem Beispiel sind die Funktionen 1 und  $x$  stetig. Wenden wir auf diese Satz 11.5 an, so erhalten wir, dass alle Polynome und in ihrem Definitionsbereich auch alle rationalen Funktionen stetig sind.

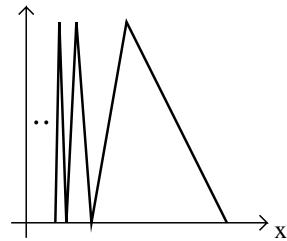
**11.5 Gleichmäßige Stetigkeit** Eine Funktion  $f : D \rightarrow \mathbb{R}$  heißt *gleichmäßig stetig* in  $D$ , wenn es zu jedem  $\varepsilon > 0$  ein  $\delta > 0$  gibt mit

$$|f(x) - f(y)| < \varepsilon \quad \text{für alle } x, y \in D \text{ mit } |x - y| < \delta.$$

Für festes  $x$  liefert diese Definition genau die Stetigkeit von  $f$  in  $x$ . Aus der gleichmäßigen Stetigkeit folgt also die Stetigkeit von  $f$  in  $D$ . Die Bedingung der gleichmäßigen Stetigkeit ist aber stärker als die Stetigkeit von  $f$  in  $D$ , weil das zu findende  $\delta > 0$  bei der gleichmäßigen Stetigkeit nicht von  $x$  und  $y$  abhängen darf. Wir können das auch formal darstellen:

$$\begin{aligned} f \text{ stetig in } D &\Leftrightarrow \forall x \forall \varepsilon > 0 \exists \delta > 0 \forall y (|x - y| < \delta \Rightarrow |f(x) - f(y)| < \varepsilon) \\ f \text{ gleichmäßig stetig in } D &\Leftrightarrow \forall \varepsilon > 0 \exists \delta > 0 \forall x, y (|x - y| < \delta \Rightarrow |f(x) - f(y)| < \varepsilon) \end{aligned}$$

**Beispiel 11.6** Sei  $f : (0, 1] \rightarrow \mathbb{R}$  folgendermaßen definiert. Wir verbinden für jedes  $k \in \mathbb{N}$  die Punkte  $(\frac{1}{2k-1}, 0)$  und  $(\frac{1}{2k}, 1)$  durch eine Strecke und die Punkte  $(\frac{1}{2k}, 1)$  und  $(\frac{1}{2k+1}, 0)$  ebenfalls durch eine Strecke. Wir erhalten den nebenstehenden Graphen.  $f$  ist offenbar stetig, aber wir müssen das  $\delta$  immer kleiner wählen, je näher wir mit dem  $x$  zur 0 kommen.  $f$  ist also nicht gleichmäßig stetig in  $D$ .  $\square$



Es ist keine Zufall, dass im obigen Beispiel das Definitionsspektrum nach links offen war:

**Satz 11.7** Eine auf einem beschränkten und abgeschlossenen Intervall definierte stetige Funktion ist dort gleichmäßig stetig.

*Beweis:* Angenommen, eine stetige Funktion  $f$  ist auf dem beschränkten und abgeschlossenen Intervall  $D$  stetig, aber nicht gleichmäßig stetig. Es gibt dann ein  $\varepsilon_0 > 0$ , für das wir kein zugehöriges  $\delta > 0$  finden können. Zu jedem  $\delta_n = \frac{1}{n}$  gibt es also Punkte  $x_n, y_n \in D$  mit  $|x_n - y_n| < 1/n$  und  $|f(x_n) - f(y_n)| \geq \varepsilon_0$ . Nach dem Satz von Bolzano-Weierstraß gibt es eine in  $D$  konvergente Teilfolge von  $(x_n)$ , die wir der Einfachheit halber wieder mit  $(x_n)$  bezeichnen. Es gilt also  $x_n \rightarrow \xi$  und wegen  $|x_n - y_n| < 1/n$  auch  $y_n \rightarrow \xi$ . Wegen der Stetigkeit der Funktion  $f$  in  $\xi$  folgt  $|f(x_n) - f(y_n)| \rightarrow 0$ , was einen Widerspruch ergibt.  $\square$

## 11.6 Der Zwischenwertsatz

**Satz 11.8** Ist  $f : [a, b] \rightarrow \mathbb{R}$  stetig, so gibt es zu jedem  $y$  im Intervall zwischen  $f(a)$  und  $f(b)$  mindestens ein  $\xi \in [a, b]$  mit  $f(\xi) = y$ .

Wir beweisen den Zwischenwertsatz konstruktiv mit einem Verfahren zur Nullstellenbestimmung. Der allgemeine Fall folgt, indem wir für  $F(x) = f(x) - y$  eine Nullstelle bestimmen.

**Algorithmus 11.9 (Bisektionsverfahren)** Sei  $a < b$  und  $f$  stetig auf  $[a, b]$  mit  $f(a) < 0, f(b) > 0$ . Mit den Startwerten  $a_0 = a$  und  $b_0 = b$  bestimmen wir Folgen  $(a_n), (b_n)$  durch:

Seien  $a_n, b_n$  bereits bestimmt. Setze  $m = (a_n + b_n)/2$  sowie

$$\begin{aligned} f(m) \geq 0 &\Rightarrow a_{n+1} = a_n, b_{n+1} = m \\ f(m) < 0 &\Rightarrow a_{n+1} = m, b_{n+1} = b_n \end{aligned}$$

Man bricht ab, wenn  $|f(m)|$  kleiner als eine vorgegebene Schranke ist.

Für den Beweis des Zwischenwertsatzes brechen wir nicht ab.  $(a_n)$  ist dann eine monoton steigende Folge und  $(b_n)$  ist monoton fallend. Da zudem die Intervalllänge in jedem Schritt halbiert wird, schachteln die beiden Folgen genau eine reelle Zahl  $x$  ein. Wegen der Stetigkeit von  $f$  gilt sowohl  $f(x) = \lim f(a_n) \leq 0$  als auch  $f(x) = \lim f(b_n) \geq 0$ , also  $f(x) = 0$ .

Als Anwendung dieses Satzes beweisen wir: Jedes Polynom ungeraden Grades besitzt mindestens eine Nullstelle. Wir können den führenden Koeffizienten des Polynoms zu 1 normieren und haben

$$p(x) = x^n + q(x), \quad q(x) = a_{n-1}x^{n-1} + \dots + a_1x + a_0.$$

Mit  $r = 1 + |a_{n-1}| + \dots + |a_1| + |a_0|$  folgt dann

$$\begin{aligned} |q(\pm r)| &\leq |a_{n-1}|r^{n-1} + \dots + |a_1|r + |a_0| \\ &\leq (|a_{n-1}| + \dots + |a_1| + |a_0|)r^{n-1} \\ &= (r-1)r^{n-1} < r^n. \end{aligned}$$

Es folgt  $p(r) \geq r^n - |q(r)| > 0$  und, da  $n$  als ungerade vorausgesetzt wurde,  $p(-r) \leq -r^n + |q(-r)| < 0$ . Nach dem Zwischenwertsatz besitzt  $p$  daher eine Nullstelle in  $[-r, r]$ .

## 11.7 Monotone Funktionen und Stetigkeit der Umkehrfunktion

$f : D \rightarrow \mathbb{R}$  heißt *monoton wachsend*, wenn für alle  $x, y \in D$

$$(11.1) \quad x \leq y \Leftrightarrow f(x) \leq f(y).$$

$f$  heißt *streng monoton wachsend*, wenn in (11.1) „ $\leq$ “ ersetzt werden kann durch „ $<$ “. Die Begriffe „monoton fallend“ und „streng monoton fallend“ sind analog definiert.

**Satz 11.10** Ist  $f[a, b] \rightarrow \mathbb{R}$  streng monoton wachsend und stetig mit  $f(a) = \alpha$ ,  $f(b) = \beta$ , so ist auch die Umkehrfunktion  $f^{-1} : [\alpha, \beta] \rightarrow [a, b]$  mit  $f^{-1}(f(x)) = x$  für alle  $x \in [a, b]$  streng monoton wachsend und stetig.

*Beweis:* Da  $f$  streng monoton wachsend ist, ist  $f$  injektiv. Ferner folgt aus dem Zwischenwertsatz 11.6, dass die Umkehrfunktion im angegebenen Bereich existiert. In die Beziehung

$$x_1 < x_2 \Leftrightarrow f(x_1) < f(x_2)$$

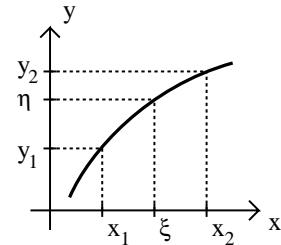
setzen wir  $x_1 = f^{-1}(y_1)$ ,  $x_2 = f^{-1}(y_2)$  ein und erhalten, dass auch  $f^{-1}$  streng monoton wachsend ist.

Nun zeigen wir die Stetigkeit von  $f^{-1}$ . Sei zunächst  $\eta = f(\xi)$  ein Punkt aus dem offenen Intervall  $(\alpha, \beta)$ . Aufgrund der strengen Monotonie von  $f$  ist dann  $\xi \in (a, b)$ . Für genügend kleines  $\varepsilon$  sind auch  $x_1 = \xi - \varepsilon$ ,  $x_2 = \xi + \varepsilon \in (a, b)$ . Für  $y_1 = f(x_1)$ ,  $y_2 = f(x_2)$  gilt  $y_1 < \eta < y_2$ . Daher gibt es ein  $\delta > 0$  mit

$$y_1 < \eta - \delta < \eta < \eta + \delta < y_2,$$

also

$$|y - \eta| < \delta \Rightarrow |f^{-1}(y) - \xi| < \varepsilon.$$



Damit ist  $f^{-1}$  stetig in  $\eta$ . Ist  $\eta$  ein Randpunkt, kann man mit halbseitigen Umgebungen entsprechend verfahren.  $\square$

Die Funktion  $f(x) = x^n : [0, a] \rightarrow [0, a^n]$  ist für jedes  $a > 0$  zwischen den angegebenen Bereichen bijektiv, stetig und streng monoton wachsend. Die Umkehrfunktion  $f^{-1}(y) = \sqrt[n]{y}$  ist damit ebenfalls stetig. Ferner sind alle algebraischen Funktionen als Kompositionen von Wurzel- und rationalen Funktionen mit Satz 11.5 in ihrem Definitionsbereich stetig.

## 11.8 Stetiges Bild eines beschränkten und abgeschlossenen Intervalls

**Satz 11.11 (Weierstraß)** Das stetige Bild eines beschränkten und abgeschlossenen Intervalls ist wieder ein beschränktes und abgeschlossenes Intervall, insbesondere nimmt jede stetige Funktion  $f$  auf einem beschränkten und abgeschlossenen Intervall  $[a, b]$  Maximum und Minimum an, es gibt also  $\xi_1, \xi_2 \in [a, b]$  mit  $f(\xi_1) \leq f(x) \leq f(\xi_2)$  für alle  $x \in [a, b]$ .

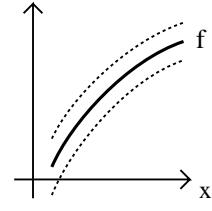
*Beweis:* Sei  $d = \inf \mathcal{R}(f)$ , wobei  $d = -\infty$  gesetzt wird, falls die Bildmenge nach unten unbeschränkt ist. Nach Definition des Infimums gibt es eine Folge  $(x_n)$  mit  $f(x_n) \rightarrow d$ . Im Falle  $d = -\infty$  ist damit gemeint, dass die Folge nach  $-\infty$  bestimmt divergiert. Nach dem Satz von Bolzano-Weierstraß 11.11 gibt es eine Teilfolge  $(x_{n_k})$ , die gegen ein  $\xi_1 \in [a, b]$  konvergiert. Da  $f$  stetig ist, gilt  $d = \lim f(x_{n_k}) = f(\xi_1)$ . Damit ist  $d$  endlich und  $\xi_1$  das gesuchte Minimum.

Da die Existenz des Maximums genauso bewiesen wird, können wir auf das Intervall  $[\xi_1, \xi_2]$  den Zwischenwertsatz anwenden. Damit ist das Bild von  $f$  das ganze Intervall  $[f(\xi_1), f(\xi_2)]$ .  $\square$

Die Beispiele  $f(x) = \frac{1}{x}$  für  $x \in (0, 1]$  und  $f(x) = x$  für  $x \in \mathbb{R}$  zeigen, dass an der Voraussetzung, dass das zugrunde liegende Intervall abgeschlossen und beschränkt sein muss, nicht gerüttelt werden darf.

**11.9 Punktweise und gleichmäßige Konvergenz von Funktionenfolgen** Seien  $f_n : D \rightarrow \mathbb{R}$ . Wir sagen, die Folge  $(f_n)$  konvergiert punktweise gegen  $f : D \rightarrow \mathbb{R}$ , wenn für alle  $x \in D$  gilt  $f_n(x) \rightarrow f(x)$ . Die Folge  $(f_n)$  konvergiert gleichmäßig gegen  $f$ , wenn es zu jedem  $\varepsilon > 0$  eine  $N \in \mathbb{N}$  gibt mit  $|f_n(x) - f(x)| < \varepsilon$  für alle  $n \geq N$  und für alle  $x \in D$ .

Wir können die punktweise Konvergenz auch mit Hilfe von  $\varepsilon$  und  $N$  definieren:  $f_n \rightarrow f$  punktweise ist genau dann erfüllt, wenn es für alle  $x \in D$  und alle  $\varepsilon > 0$  ein  $N \in \mathbb{N}$  gibt, das von  $\varepsilon$  und  $x$  abhängen darf, mit  $|f_n(x) - f(x)| < \varepsilon$ . In der gleichmäßigen Konvergenz darf das  $N$  dagegen *nicht* von  $x$  abhängen. Wir können uns die gleichmäßige Konvergenz daher so vorstellen, dass wir um  $f$  einen  $\varepsilon$ -Schlauch legen, in dem alle bis auf endlich viele  $f_n$  liegen müssen.



**Beispiel 11.12** Sei  $D = [0, 1]$  und  $f_n(x) = x^n$ . Für  $0 \leq x < 1$  gilt  $x^n \rightarrow 0$ . Der punktweise Limes der Folge ist daher die Funktion  $f$  mit  $f(x) = 0$  für  $0 \leq x < 1$  und  $f(1) = 1$ . Diese Konvergenz ist jedoch nicht gleichmäßig, denn wenn um die Grenzfunktion ein  $\varepsilon$ -Schlauch mit  $\varepsilon \leq \frac{1}{2}$  gelegt wird, so liegt kein  $f_n$  komplett in diesem Schlauch.  $\square$

**Satz 11.13** Der gleichmäßige Limes stetiger Funktionen ist stetig.

*Beweis:* Sei  $\xi \in D$  und  $\varepsilon > 0$ . Es gibt ein  $n \in \mathbb{N}$  mit  $|f_n(\xi) - f(\xi)| < \varepsilon/3$  für alle  $x \in D$ . Da dieses  $f_n$  stetig ist, gibt es zu  $\varepsilon/3$  ein  $\delta > 0$  mit  $|f_n(\xi) - f_n(x)| < \varepsilon/3$  für alle  $x \in D$  mit  $|x - \xi| < \delta$ . Für diese  $x$  folgt

$$\begin{aligned} |f(\xi) - f(x)| &\leq |f(\xi) - f_n(\xi)| + |f_n(\xi) - f_n(x)| + |f_n(x) - f(x)| \\ &< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon. \end{aligned}$$

Damit ist  $f$  im Punkt  $\xi$  stetig.  $\square$

**11.10 Anwendung der Stetigkeit auf die Konvergenz von Zahlenfolgen** Wir können stetige Funktionen als konvergenzerhaltende Abbildungen ansehen, denn aus  $x_n \rightarrow \xi$  folgt  $f(x_n) \rightarrow f(\xi)$ . Wir gewinnen dadurch neue Sätze über die Konvergenz von Zahlenfolgen. Ist beispielsweise  $a_n \rightarrow a$  und  $a_n \geq 0$ , so gilt  $\sqrt[k]{a_n} \rightarrow \sqrt[k]{a}$ .

Wir betrachten rekursiv definierte Folgen der Form

$$(11.2) \quad a_{n+1} = f(a_n), \quad a_0 \in \mathbb{R} \text{ vorgegeben,}$$

mit stetigem  $f$ . Wenn  $(a_n)$  konvergiert, so kann man auf beiden Seiten von (11.2) zum Grenzwert übergehen und erhält  $a = f(a)$ , der Grenzwert ist also immer ein Fixpunkt von  $f$ .

**Algorithmus 11.14 (Babylonisches Wurzelziehen)** Sei  $b > 0$ . Für  $a_0 > 0$  untersuchen wir die Folge

$$(11.3) \quad a_{n+1} = \frac{1}{2}a_n + \frac{b}{2a_n}.$$

Ist  $a_n = \sqrt{b}$ , so ist auch  $a_{n+1} = \sqrt{b}$  und die Folge stagniert.

Sei also  $0 < a_n \neq \sqrt{b}$ . Aus  $(a_n - \sqrt{b})^2 > 0$  erhalten wir  $a_n^2 - 2a_n\sqrt{b} + b > 0$ . In der letzten Ungleichung können wir durch  $2a_n$  teilen,

$$a_{n+1} = \frac{1}{2}a_n + \frac{b}{2a_n} > \sqrt{b}.$$

Damit gilt für jeden Startwert  $0 < a_0 \neq \sqrt{b}$ , dass  $a_n > \sqrt{b}$  für alle  $n \geq 1$ . Wir zeigen, dass die Folge  $(a_n)$  ab  $n = 1$  streng monoton fallend ist. Aus  $a_n > \sqrt{b}$  folgt  $(a_n)^2 > b$  und

$$-a_n + \frac{b}{a_n} < 0 \quad \Rightarrow \quad a_{n+1} - a_n = -\frac{a_n}{2} + \frac{b}{2a_n} < 0.$$

Da die Folgenglieder nicht negativ werden können, ist die Folge monoton fallend und nach unten beschränkt. Sie besitzt daher einen eindeutigen Grenzwert  $a > 0$ . Da die rechte Seite als rationale Funktion in  $a_n$  stetig auf  $D = \{x > 0\}$  ist, muss  $a$  ein Fixpunkt der Iterationsvorschrift sein, also  $a = \frac{1}{2}a + \frac{1}{2}b/a$  und damit  $a = \sqrt{b}$ .

Dieses Verfahren wird wegen seiner Schnelligkeit noch heute in Computern für die Wurzelberechnung verwendet. Man sehe: Für  $b = 2$  und  $a_0 = 2$  ist  $a_1 = 1.5$  und  $a_2 = 1.416\dots$ . Der exakte Wert ist  $\sqrt{2} = 1.414\dots$ . Man kann zeigen, dass sich bei vernünftigen Startwerten  $a_0 \sim \sqrt{b}$  die Zahl der gültigen Stellen in jedem Schritt verdoppelt.

### 11.11 Potenzreihen

Die wichtigsten Reihen der Analysis sind die *Potenzreihen*

$$(11.4) \quad p(x) = \sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \dots, \quad a_n \in \mathbb{R}.$$

Da eine Potenzreihe für jedes  $x \in \mathbb{R}$  eine Reihe ist, übertragen sich die Begriffe Konvergenz und absolute Konvergenz. Da die Konvergenz einer Reihe auf die Konvergenz der Partialsummen zurückgeführt wird, übernehmen wir auch den Begriff der gleichmäßigen Konvergenz aus dem letzten Abschnitt.

Wir erinnern daran, dass wir mit  $\limsup a_n$  den größten Häufungspunkt der Folge  $(a_n)$  bezeichnet haben. Ist die Folge nach oben beschränkt, so existiert der Limes Superior nach dem Satz von Bolzano-Weierstraß 11.11. Ist die Folge nach oben unbeschränkt, so schreiben wir  $\limsup a_n = \infty$ .

**Satz 11.15** Sei

$$L = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|}$$

und  $R = \frac{1}{L}$ , wobei  $1/0$  als  $R = \infty$  und  $1/\infty$  als  $R = 0$  interpretiert wird. Dann ist die Reihe (11.4) für  $|x| < R$  absolut konvergent und für  $|x| > R$  divergent. Die Reihe konvergiert gleichmäßig in jedem Intervall  $|x| \leq r$  mit  $r < R$ . Über die Konvergenz für  $|x| = R$  lässt sich keine allgemeine Aussage machen.

Existiert der Grenzwert

$$Q = \lim_{n \rightarrow \infty} \frac{|a_{n+1}|}{|a_n|},$$

so gilt  $L = Q$ .

**Bemerkung 11.16** Nach Satz 11.13 ist  $p(x)$  stetig für  $|x| \leq r$ . Da  $r < R$  beliebig gewählt werden kann, ist  $p(x)$  für alle  $|x| < R$  stetig.  $\square$

*Beweis:* Sei zunächst  $0 < L < \infty$ . Für die Konstante  $L'$  des Wurzelkriteriums gilt dann

$$L' = \limsup \sqrt[n]{|a_n x^n|} = \limsup |x| \sqrt[n]{|a_n|} \begin{cases} < 1, & \text{falls } |x| < 1/L \\ > 1, & \text{falls } |x| > 1/L \end{cases}.$$

In beiden Fällen folgt Konvergenz oder Divergenz aus dem Wurzelkriterium. Ist  $L = 0$ , so ist das Wurzelkriterium für alle  $x$  erfüllt. Für  $L = \infty$  liegt nur Konvergenz im Punkt 0 vor.

Für  $|x| \leq r < R$  lässt sich die Potenzreihe unabhängig von  $x$  durch die geometrische Reihe abschätzen. Damit ist die Konvergenz gleichmäßig.

Den zweiten Teil des Satzes beweist man genauso mit Hilfe des Quotientenkriteriums an Stelle des Wurzelkriteriums.  $\square$

**Beispiele 11.17** (i) Nach (10.3) gilt für jedes  $m \in \mathbb{Z}$   $\sqrt[m]{n^m} \rightarrow 1$ . Die Potenzreihen  $\sum n^m x^n$  haben daher alle den gleichen Konvergenzradius  $R = 1$ . Für  $m = 0$  erhalten wir die geometrische Reihe, die für  $x = \pm 1$  divergent ist. Für  $m = -1$  ist für  $x = 1$  die harmonische Reihe divergent, für  $x = -1$  erhalten wir die konvergente alternierende harmonische Reihe. Über das Konvergenzverhalten am Rande lässt sich also in der Tat keine allgemeine Aussage machen.

(ii) Für die Reihe  $\sum n! x^n$  folgt mit dem Wurzelkriterium  $a_{n+1}/a_n = n + 1 \rightarrow \infty$ . Der Konvergenzradius ist daher  $R = 0$ .  $\square$

**11.12 Das Cauchy-Produkt von Potenzreihen** Das Produkt zweier Potenzreihen ergibt sich dadurch, dass man jedes Glied der einen Reihe mit jedem Glied der anderen Reihe multipliziert und das Ergebnis nach Potenzen ordnet.

**Satz 11.18** Sind die Potenzreihen  $f(x) = \sum_{n=0}^{\infty} a_n x^n$  und  $g(x) = \sum_{n=0}^{\infty} b_n x^n$  für  $|x| < R$  konvergent, so ist auch das Produkt

$$f(x)g(x) = \sum_{n=0}^{\infty} c_n x^n, \quad c_n = a_0 b_n + a_1 b_{n-1} + \dots + a_{n-1} b_1 + a_n b_0,$$

mindestens im Bereich  $|x| < R$  konvergent.

Auf den etwas technischen Beweis soll verzichtet werden.

**11.13 Die Exponentialfunktion** Die Exponentialfunktion wird durch die Potenzreihe

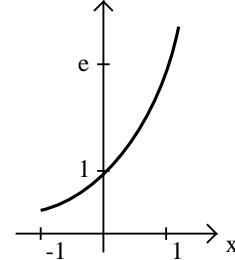
$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

dargestellt. Speziell bezeichnen wir

$$e = \exp(1) = \sum_{n=0}^{\infty} \frac{1}{n!}$$

als Eulersche Zahl. Mit  $a_n = \frac{1}{n!}$  folgt

$$\frac{a_{n+1}}{a_n} = \frac{1}{n+1}$$



Die Exponentialfunktion

und nach dem Quotientenkriterium in Satz 10.13 konvergiert die Reihe auf ganz  $\mathbb{R}$  und stellt dort eine stetige Funktion dar. Eine alternative Darstellung der Exponentialfunktion und der Eulerschen Zahl ist gegeben durch

$$(11.5) \quad \exp(x) = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n.$$

Dies folgt mit Hilfe der binomischen Formel

$$\left(1 + \frac{x}{n}\right)^n = \sum_{k=0}^n \binom{n}{k} \frac{x^k}{n^k}$$

und wegen

$$\binom{n}{k} \frac{1}{n^k} = \frac{1}{k!} \frac{n(n-1)\dots(n-k+1)}{n \cdot n \dots n} = \frac{1}{k!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \dots \left(1 - \frac{k-1}{n}\right)$$

gilt

$$\binom{n}{k} \frac{1}{n^k} \leq \frac{1}{k!}, \quad \binom{n}{k} \frac{1}{n^k} \rightarrow \frac{1}{k!} \text{ für } n \rightarrow \infty.$$

Historisch trat die Eulersche Zahl zuerst im Zusammenhang mit der Definition  $e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n$  auf. Verzinsen wir einen Geldbetrag der Größe 1 in einem Jahr mit einem Zinssatz von 100%, so erhalten wir nach einem Jahr den Betrag 2. Erfolgt die Zinszahlung bei gleichem Zinssatz auch zwischenzeitlich, so verzinst sich das Kapital durch den Zinseszinseffekt besser. Bei monatlicher Zinszahlung bekommen wir  $\left(1 + \frac{1}{12}\right)^{12} = 2,613\dots$ . Machen wir die Zeiträume der Verzinsung kürzer und kürzer, so erhalten wir bei kontinuierlicher Verzinsung  $e = 2,71\dots$  am Jahresende.

**Satz 11.19** Die Exponentialfunktion  $\exp : \mathbb{R} \rightarrow \mathbb{R}_+$  ist bijektiv, streng monoton wachsend und genügt der Funktionalgleichung

$$\exp(x+y) = \exp(x)\exp(y) \quad \text{für alle } x, y \in \mathbb{R}.$$

*Beweis:* Wir bilden das Produkt  $\exp x \exp y$ , indem wir jeden Summanden mit jedem Summanden multiplizieren und das Ergebnis nach Potenzen ordnen,

$$\exp(x)\exp(y) = \sum d_n \quad \text{mit } d_n = \sum_{i=0}^n \frac{x^i}{i!} \frac{y^{n-i}}{(n-i)!} = \frac{1}{n!} \sum_{i=0}^n \binom{n}{i} x^i y^{n-i} = \frac{1}{n!} (x+y)^n.$$

Aus der Definition der Exponentialfunktion folgt sofort, dass sie streng monoton wachsend für nichtnegative  $x$  ist. Ferner gilt  $\exp(x) \geq 1+x$ , daher  $\lim_{x \rightarrow \infty} \exp(x) = \infty$ . Für negative  $x$  erhalten wir die Behauptung aus

$$\exp(x)\exp(-x) = \exp(x-x) = 1,$$

also  $\exp(-x) = \exp(x)^{-1}$ . Damit gilt insbesondere  $\lim_{x \rightarrow -\infty} \exp(x) = 0$ . Wegen des Zwischenwertesatzes ist die Exponentialfunktion bijektiv zwischen den angegebenen Bereichen.  $\square$

Aus der Exponentialreihe erschließen wir ferner für jedes  $n \in \mathbb{N}$

$$(11.6) \quad \lim_{x \rightarrow \infty} \frac{\exp(x)}{x^n} = \infty, \quad \lim_{x \rightarrow \infty} \frac{\exp(-x)}{x^{-n}} = 0,$$

denn es gilt für  $x > 0$

$$\exp(x) > \frac{x^{n+1}}{(n+1)!},$$

daher

$$\frac{\exp(x)}{x^n} > \frac{x}{(n+1)!} \rightarrow \infty, \quad 0 < x^n \exp(-x) < \frac{(n+1)!}{x} \rightarrow 0 \quad \text{für } x \rightarrow \infty.$$

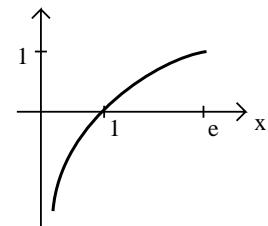
Kurz: Die Exponentialfunktion wächst für  $x \rightarrow \infty$  schneller als jede Potenz und sie fällt für  $x \rightarrow -\infty$  schneller gegen Null als jede negative Potenz.

**11.14 Der Logarithmus** Nach Satz 11.7 besitzt die Exponentialfunktion eine stetige, streng monoton wachsende Umkehrfunktion, die wir als (*natürlichen*) *Logarithmus*  $\ln : \mathbb{R}_+ \rightarrow \mathbb{R}$  bezeichnen. Es ist daher  $\ln : \mathbb{R}_+ \rightarrow \mathbb{R}$  und

$$y = \exp(x) \Leftrightarrow x = \ln y.$$

**Satz 11.20** Der natürliche Logarithmus hat die Eigenschaft

$$(11.7) \quad \ln xy = \ln x + \ln y \quad x, y \in \mathbb{R}_+.$$



Der Logarithmus

*Beweis:* Es gilt

$$\exp(\ln(xy)) = xy = \exp(\ln x)\exp(\ln y) = \exp(\ln x + \ln y).$$

Da die Funktion  $\exp$  bijektiv ist, folgt hieraus die Behauptung.  $\square$

Für jedes  $n$  gilt

$$(11.8) \quad \lim_{x \rightarrow \infty} \frac{\ln x}{\sqrt[n]{x}} = 0, \quad \lim_{x \searrow 0} \sqrt[n]{x} \ln x = 0.$$

Wir beweisen dies, indem wir für  $x > 0$   $x = \exp(ny)$  mit  $y \in \mathbb{R}$  setzen. Es gilt dann

$$\frac{\ln x}{\sqrt[n]{x}} = \frac{ny}{\exp y} \rightarrow 0 \quad \text{für } y \rightarrow \infty, \quad \sqrt[n]{x} \ln x = \exp(y)ny \rightarrow 0 \quad \text{für } y \rightarrow -\infty.$$

Also: Der Logarithmus geht für  $x \rightarrow \infty$  langsamer gegen unendlich als jede Wurzel. Ferner ist die bestimmte Divergenz gegen  $-\infty$  für  $x \searrow 0$  nur schwach ausgeprägt.

Mit Hilfe des Logarithmus wollen wir nun allgemeine Potenzen definieren. Für  $a > 0$  folgt aus der Additionseigenschaft (11.7) des Logarithmus  $\ln a^n = n \ln a$ . Wegen  $\sqrt[n]{a} \dots \sqrt[n]{a} = a$  folgt auch  $\ln \sqrt[n]{a} = \frac{1}{n} \ln a$ . Damit gilt  $\ln a^r = r \ln a$  für alle rationalen  $r$ . Nehmen wir hier auf beiden Seiten die Exponentialfunktion, so gilt  $a^r = \exp(r \ln a)$ . Damit können wir unsere alte, nur für rationale  $r$  gültige Exponentiation auf ganz  $\mathbb{R}$  fortsetzen durch die Definition

$$a^x = \exp(x \ln a), \quad a, x \in \mathbb{R}, \quad a > 0.$$

Entsprechend schreiben wir für  $a = e$  kürzer  $e^x$  statt  $\exp(x)$ . Es gilt dann für  $a, b > 0$  und beliebige  $x, y \in \mathbb{R}$

- (i)  $a^{x+y} = a^x a^y$ ,
- (ii)  $(a^x)^y = a^{xy}$ ,
- (iii)  $a^x b^x = (ab)^x$ .

Die Beweise folgen aus der Definition. (i) erhalten wir mit

$$a^{x+y} = \exp((x+y) \ln a) = \exp(x \ln a) \exp(y \ln a) = a^x a^y,$$

(ii) mit

$$(a^x)^y = \exp(y \ln a^x) = \exp(xy \ln a) = a^{xy},$$

und (iii) mit

$$a^x b^x = \exp(x \ln a) \exp(x \ln b) = \exp(x \ln ab) = (ab)^x.$$

**Beispiele 11.21** Die Logarithmus-Funktion ist ein wichtiges technisches Hilfsmittel, um das Verhalten komplizierter algebraischer Ausdrücke zu untersuchen.

(i) Zur Bestimmung des Grenzwertes der Folge  $\sqrt[n]{n!}$  betrachten wir die Logarithmen der Folgenglieder

$$\ln \sqrt[n]{n!} = \frac{1}{n} (\ln 1 + \ln 2 + \dots + \ln n).$$

Die Logarithmen der Folgenglieder sind also Mittelwerte einer Folge, die bestimmt gegen unendlich divergiert. Damit divergieren auch die Mittelwerte und somit auch  $\sqrt[n]{n!}$  bestimmt gegen unendlich.

(ii) Ein weiteres Beispiel ist das Verhalten der Funktion  $x^{1/x}$  für  $x \rightarrow \infty$ . Der Logarithmus dieser Funktion ist  $\ln x^{1/x} = \ln x/x \rightarrow 0$  für  $x \rightarrow \infty$  nach (11.8), daher  $x^{1/x} \rightarrow 1$  für  $x \rightarrow \infty$ .  $\square$

**11.15 Hyperbelfunktionen** Aus der Exponentialfunktion lassen sich weitere Funktionen ableiten

$$\cosh(x) = \frac{1}{2}(e^x + e^{-x}) \quad (\text{Cosinus hyperbolicus}),$$

$$\sinh(x) = \frac{1}{2}(e^x - e^{-x}) \quad (\text{Sinus hyperbolicus}),$$

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} \quad (\text{Tangens hyperbolicus}),$$

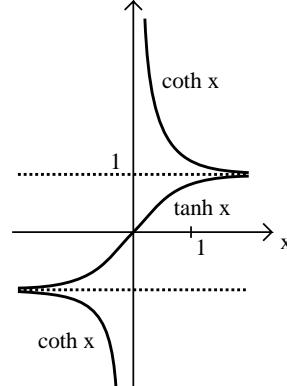
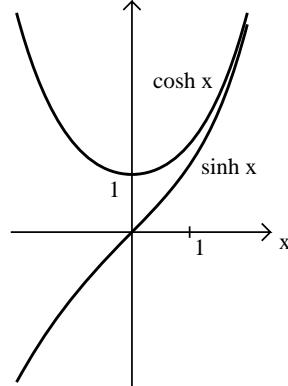
$$\coth(x) = \frac{\cosh(x)}{\sinh(x)} \quad (\text{Cotangens hyperbolicus}).$$

Da  $\sinh$  für  $x = 0$  eine Nullstelle hat, ist  $\coth$  nur für  $x \neq 0$  definiert.

Direkt aus der Additionseigenschaft der Exponentialfunktion beweist man die *Additionstheoreme*

$$\cosh(x + y) = \cosh(x) \cosh(y) + \sinh(x) \sinh(y),$$

$$\sinh(x + y) = \sinh(x) \cosh(y) + \cosh(x) \sinh(y).$$



Die Hyperbelfunktionen  $\sinh x$  und  $\cosh x$  Die Hyperbelfunktionen  $\tanh x$  und  $\coth x$

**11.16 Die Trigonometrischen Funktionen** Die altbekannte Definition von Sinus und Cosinus findet man in der nebenstehenden Zeichnung. Der Punkt  $(\sin x, \cos x)$  liegt auf dem Einheitskreis,  $x$  ist dabei die „Länge“ des Kreisbogens von  $A$  nach  $B$ . Wenn wir einmal davon absehen, dass wir die Länge gekrümmter Kurven bisher nicht definiert haben, kann man aus der Zeichnung alle wichtigen Eigenschaften der beiden Winkelfunktionen ableSEN. Ist  $\pi = 3,1415\dots$  die Länge des Halbkreises, so gilt

$$(11.9) \quad \sin 0 = 0, \sin \frac{\pi}{2} = 1, \sin \pi = 0, \sin(x + 2\pi) = \sin x,$$

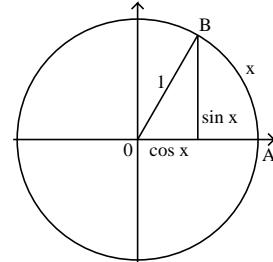
$$(11.10) \quad \cos 0 = 1, \cos \frac{\pi}{2} = 0, \cos \pi = -1, \cos(x + 2\pi) = \cos x,$$

Da die Winkelfunktionen den Einheitskreis parametrisieren, gilt

$$(11.11) \quad \sin^2 x + \cos^2 x = 1.$$

Da diese Definition der Winkelfunktionen auf der geometrischen Anschauung beruht, lassen sich konkrete Werte wie beispielsweise  $\cos 1$  damit nicht berechnen. Wir verwenden daher Potenzreihen und setzen

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



Aus dem Quotientenkriterium in Satz 11.11 folgt, dass die beiden Reihen auf ganz  $\mathbb{R}$  konvergent sind und dort stetige Funktionen darstellen. Wir müssen nun zeigen, dass für die so definierten Funktionen die Eigenschaften (11.9)-(11.11) gelten. Klar ist  $\sin 0 = 0$  und  $\cos 0 = 1$ . Wie das Additionstheorem für die Exponentialfunktion leitet man (11.11) sowie die *Additionstheoreme*

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

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

her.

Wir definieren  $\frac{\pi}{2}$  als erste positive Nullstelle des Cosinus. Die Cosinus-Reihe ist alternierend und es gilt

$$\frac{x^{2n}}{(2n)!} > \frac{x^{2n+2}}{(2n+2)!} \quad \text{für alle } n \in \mathbb{N} \text{ und } 0 < x \leq 3.$$

Die Absolutbeträge der Glieder sind daher ab  $n = 1$  streng monoton fallend und nach dem Leibniz-Kriterium ist

$$C_2(x) = 1 - \frac{x^2}{2} < \cos x < 1 - \frac{x^2}{2} + \frac{x^4}{24} = C_4(x) \quad \text{für } 0 < x \leq 3.$$

Die Unterfunktion  $C_2$  besitzt daher eine Nullstelle für  $\alpha = \sqrt{2}$ , die Oberfunktion  $C_4$  für  $\beta = \sqrt{6 - 2\sqrt{3}}$ . Damit gilt

$$1,4 < \alpha < \frac{\pi}{2} < \beta < 1,6.$$

Mit den Additionstheoremen und (11.11) gilt dann  $\sin \frac{\pi}{2} = 1$  und

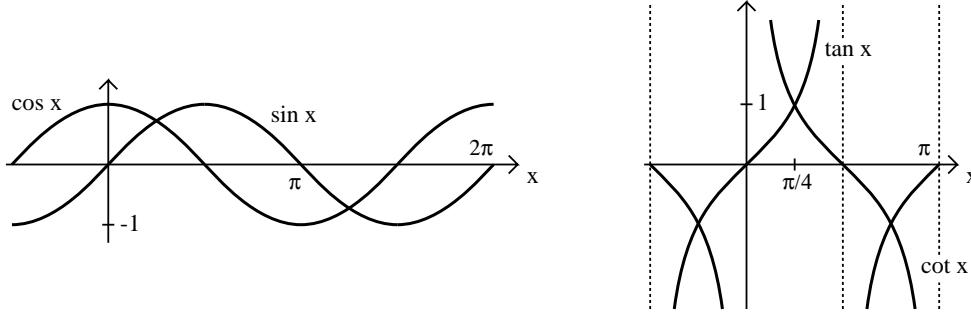
$$(11.12) \quad \sin(x + \frac{\pi}{2}) = \cos x, \quad \cos(x + \frac{\pi}{2}) = -\sin x.$$

Wenden wir diese Beziehungen sukzessive an, haben wir (11.9) und (11.10) vollständig bewiesen. Es fehlt allerdings noch, dass  $\frac{\pi}{2}$  tatsächlich der Länge des Viertelkreises entspricht.

Nun untersuchen wir das Monotonieverhalten von Sinus und Cosinus. Aufgrund der Definition von  $\pi/2$  als erster Nullstelle des Cosinus ist  $\cos x > 0$  in  $[0, \pi/2)$ . Wegen  $\sin \pi/2 = 1$  und  $\sin^2 x + \cos^2 x = 1$  muss wegen des Zwischenwertsatzes auch  $\sin x > 0$  in  $(0, \pi/2)$  gelten. Aus dem Additionstheorem des Cosinus folgt daher für  $0 \leq x < x + y \leq \pi/2$

$$\cos(x + y) = \cos x \cos y - \sin x \sin y \leq \cos x \cos y < \cos x.$$

Der Cosinus ist also im Intervall  $[0, \pi/2]$  streng monoton fallend und entsprechend ist der Sinus in diesem Intervall streng monoton wachsend. Zusammen mit (11.12) haben wir einen vollständigen Überblick über das Monotonieverhalten der beiden trigonometrischen Funktionen.



Wir definieren *Tangens* und *Cotangens* durch

$$\tan x = \frac{\sin x}{\cos x} \quad \text{für } x \neq (2k+1)\frac{\pi}{2}, k \in \mathbb{Z},$$

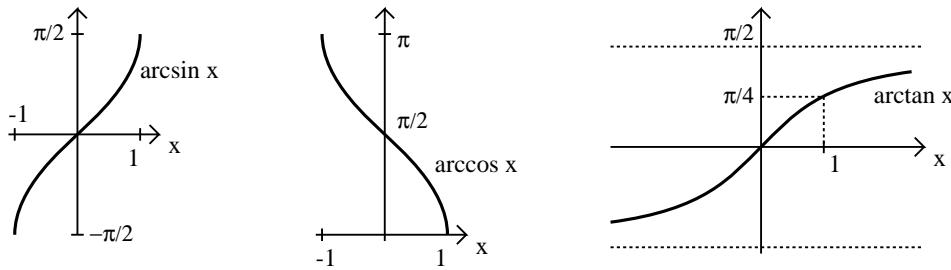
$$\cot x = \frac{\cos x}{\sin x} \quad \text{für } x \neq k\pi, k \in \mathbb{Z}.$$

Beide Funktionen sind  $\pi$ -periodisch.

Die *Arcusfunktionen* sind Umkehrfunktionen der trigonometrischen Funktionen. Da diese alleamt periodisch sind, müssen sie auf ein Intervall eingeschränkt werden, auf dem sie streng monoton

sind. Bei allen vier trigonometrischen Funktionen hat man sich dabei auf ein Intervall geeinigt und spricht vom *Hauptwert* der Umkehrfunktion.

Der Sinus bildet das Intervall  $[-\pi/2, \pi/2]$  bijektiv auf das Intervall  $[-1, 1]$  ab. Für  $y \in [-1, 1]$  bezeichnen wir die Lösung  $x \in [-\pi/2, \pi/2]$  von  $\sin x = y$  als *Arcussinus* von  $y$  und schreiben  $y = \arcsin x$ . Der Arcussinus ist also auf dem Intervall  $[-1, 1]$  definiert, stetig und streng monoton steigend. Selbstverständlich hat die Gleichung  $\sin x = y$  unendlich viele Lösungen, die als *Nebenwerte* des Arcussinus bezeichnet werden und besonders gekennzeichnet werden müssen.



Auf die gleiche Weise definiert man die Hauptwerte der anderen Winkelfunktionen durch mehr oder weniger willkürliche Festlegung des Definitionsbereichs und kommt dann zu

$$y = \arcsin x, \quad |y| \leq \frac{\pi}{2}, \quad (|x| \leq 1),$$

$$y = \arccos x, \quad 0 \leq y \leq \pi, \quad (|x| \leq 1),$$

$$y = \arctan x, \quad |y| \leq \frac{\pi}{2}, \quad (x \in \mathbb{R}),$$

$$y = \operatorname{arccot} x, \quad 0 \leq y \leq \pi, \quad (x \in \mathbb{R}).$$

## 12 Differentiation

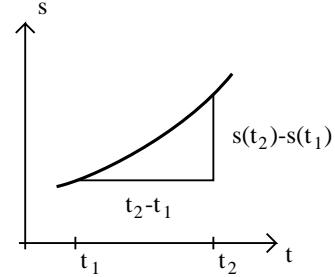
### 12.1 Definition der Differenzierbarkeit

**Beispiel 12.1** Die geradlinige Bewegung eines Massepunktes wird beschrieben durch eine Funktion  $s(t)$ , wobei  $t$  die Zeit und  $s(t)$  den zurückgelegten Weg des Massepunktes bezeichnet. Ist  $s(t)$  linear, so ist

$$\frac{s(t)}{t} = \text{constant}$$

die Geschwindigkeit. Ist  $s(t)$  nichtlinear und sind  $t_1, t_2$  zwei Zeitpunkte, so ist  $s(t_2) - s(t_1)$  der im Zeitraum  $t_2 - t_1$  zurückgelegte Weg und damit

$$\frac{s(t_2) - s(t_1)}{t_2 - t_1}$$

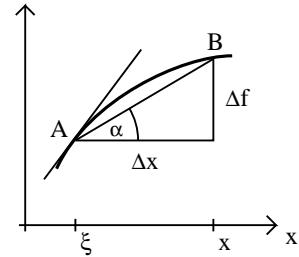


die Durchschnittsgeschwindigkeit im Zeitraum  $(t_1, t_2)$ .  $\square$

Für eine Funktion  $f$  heißt

$$m = \frac{\Delta f}{\Delta x} = \frac{f(x) - f(\xi)}{x - \xi} = \tan \alpha$$

*Differenzenquotient.* Er gibt die *Steigung*  $m$  der Sekante durch die Punkte  $A$  und  $B$  an. Wandert nun  $x$  nach links zum Punkt  $\xi$ , so läuft  $B$  nach  $A$  und die Sekante geht in die Tangente im Punkt  $A$  über.



Sei  $f$  in einer Umgebung von  $\xi \in \mathbb{R}$  definiert.  $f$  heißt in  $\xi$  *differenzierbar*, wenn der Grenzwert

$$f'(\xi) = \frac{df(\xi)}{dx} = \lim_{x \rightarrow \xi} \frac{f(x) - f(\xi)}{x - \xi} = \lim_{h \rightarrow 0} \frac{f(\xi + h) - f(\xi)}{h}$$

existiert.  $f'(\xi)$  heißt *Ableitung* von  $f$  in  $\xi$ .

**Beispiel 12.2** Für  $f(x) = ax + b$  erhalten wir

$$f'(\xi) = \lim_{h \rightarrow 0} \frac{a(\xi + h) + b - (a\xi + b)}{h} = a.$$

$\square$

Physikalisch gibt  $f'(\xi)$  die Momentangeschwindigkeit eines Massepunktes an. Geometrisch ist  $f'(\xi)$  die Steigung der Tangenten im Punkt  $(\xi, f(\xi))$ . Die Tangente  $y(x)$  läuft ebenfalls durch diesen Punkt und besitzt die gleiche Steigung wie  $f$ . Aus  $y(\xi) = f(\xi)$  und  $y'(\xi) = f'(\xi)$  folgt die *Tangentengleichung*

$$y(x) = f(\xi) + f'(\xi)(x - \xi).$$

*Einseitige Ableitungen* lassen sich analog zu einseitigen Grenzwerten definieren

$$f'_+(\xi) = \lim_{h \rightarrow 0+} \frac{f(\xi + h) - f(\xi)}{h} \quad (\text{rechtsseitige Ableitung})$$

$$f'_-(\xi) = \lim_{h \rightarrow 0-} \frac{f(\xi + h) - f(\xi)}{h} \quad (\text{linksseitige Ableitung})$$

Wenn  $f'_+(\xi)$  und  $f'_-(\xi)$  existieren und übereinstimmen, ist  $f$  in  $\xi$  differenzierbar.

**Beispiel 12.3** Für  $f(x) = |x|$  erhalten wir

$$f'_+(0) = \lim_{h \rightarrow 0^+} \frac{|h| - |0|}{h} = 1, \quad f'_-(0) = \lim_{h \rightarrow 0^-} \frac{|h| - |0|}{h} = -1.$$

Da die einseitigen Ableitungen verschieden sind, ist  $|x|$  im Nullpunkt nicht differenzierbar.  $\square$

**Satz 12.4** Sei  $f$  in  $\xi$  differenzierbar. Dann gilt eine lokale Lipschitzbedingung, nämlich

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

für alle  $x$  in einer genügend kleinen Umgebung von  $\xi$ , also  $x \in [\xi - h_0, \xi + h_0]$  für ein  $h_0 > 0$ . Insbesondere ist  $f$  stetig in  $\xi$ . Ist  $f'(\xi) > 0$ , so gilt

$$(12.1) \quad f(\xi - h) < f(\xi) < f(\xi + h) \quad \text{für alle } 0 < h \leq h_0.$$

*Beweis:* Für  $x$  in einer genügend kleinen Umgebung von  $\xi$  folgt aus der Definition der Differenzierbarkeit

$$\left| \frac{f(x) - f(\xi)}{x - \xi} \right| \leq |f'(\xi)| + 1 = K.$$

Ist  $f'(\xi) > 0$ , so gilt für genügend kleine  $|h|$

$$\frac{f(\xi + h) - f(\xi)}{h} > 0.$$

Daraus folgt (12.1).  $\square$

Allgemeiner erfüllt eine Funktion  $f : I \rightarrow \mathbb{R}$  eine globale Lipschitzbedingung, wenn

$$|f(x) - f(y)| \leq K|x - y| \quad \text{für alle } x, y \in I.$$

$f$  heißt dann auch lipschitzstetig. Eine lipschitzstetige Funktion ist gleichmäßig stetig: Für  $\varepsilon > 0$  kann man  $\delta = \frac{\varepsilon}{K}$  verwenden.

## 12.2 Differenzierbarkeit und arithmetische Operationen

**Satz 12.5** Sind die Funktionen  $f, g$  in  $\xi$  differenzierbar, so sind für  $\alpha, \beta \in \mathbb{R}$  auch die Funktionen  $\alpha f + \beta g$  sowie  $fg$  und, falls  $g \neq 0$ ,  $f/g$  in  $\xi$  differenzierbar. Für diese Ableitungen gilt

- (a)  $(\alpha f + \beta g)'(\xi) = \alpha f'(\xi) + \beta g'(\xi)$ , (Linearität),
- (b)  $(fg)'(\xi) = f'(\xi)g(\xi) + f(\xi)g'(\xi)$  (Produktregel),
- (c)  $\left( \frac{f}{g} \right)'(\xi) = \frac{f'(\xi)g(\xi) - f(\xi)g'(\xi)}{g^2(\xi)}$  (Quotientenregel).

*Beweis:* (a) Die Linearität der Ableitung folgt aus der Linearität des Differenzenquotienten.

(b) Es gilt

$$\begin{aligned} \frac{f(\xi + h)g(\xi + h) - f(\xi)g(\xi)}{h} &= \frac{f(\xi + h)g(\xi + h) - f(\xi + h)g(\xi) + f(\xi + h)g(\xi) - f(\xi)g(\xi)}{h} \\ &= f(\xi + h) \frac{g(\xi + h) - g(\xi)}{h} + g(\xi) \frac{f(\xi + h) - f(\xi)}{h}. \end{aligned}$$

Da  $f$  in  $\xi$  stetig ist, folgt die Behauptung durch Grenzübergang.

(c) Wir zeigen die Behauptung nur für  $f = 1$ , der allgemeine Fall folgt dann aus der Produktregel.

$$\frac{1}{h} \left( \frac{1}{g(\xi+h)} - \frac{1}{g(\xi)} \right) = \frac{1}{h} \frac{-g(\xi+h) + g(\xi)}{g(\xi+h)g(\xi)}.$$

Wegen der Stetigkeit von  $g$  in  $\xi$  können wir auch hier den Grenzübergang  $h \rightarrow 0$  durchführen und erhalten die Behauptung.  $\square$

Als Anwendung dieses Satzes zeigen wir  $(x^n)' = nx^{n-1}$  durch vollständige Induktion. Für  $n = 0$  erhalten wir die konstante Funktion, für die  $1' = 0$  aus der Definition der Differenzierbarkeit folgt. Ebenso bestimmt man  $x' = 1$ . Als Induktionsannahme dürfen wir  $(x^n)' = nx^{n-1}$  verwenden. Aus der Produktregel folgt dann

$$(x^{n+1})' = (x \cdot x^n)' = 1 \cdot x^n + xnx^{n-1} = (n+1)x^n.$$

Damit ist die Ableitung eines Polynoms

$$\frac{d}{dx} (a_n x^n + \dots + a_1 x + a_0) = na_n x^{n-1} + \dots + a_1.$$

Zur Bestimmung der Ableitung von  $x^{-n}$  verwenden wir die Quotientenregel

$$\frac{d}{dx} \frac{1}{x^n} = -\frac{nx^{n-1}}{x^{2n}} = -nx^{-n-1}.$$

### 12.3 Kettenregel

**Satz 12.6** Seien  $I, J$  Intervalle mit  $f : I \rightarrow \mathbb{R}$  und  $g : J \rightarrow \mathbb{R}$  sowie  $f(I) \subset J$ . Ist  $f$  an der Stelle  $\xi$  und  $g$  an der Stelle  $f(\xi)$  differenzierbar, so ist auch  $h = g \circ f$ ,  $h(x) = g(f(x))$ , an der Stelle  $\xi$  differenzierbar mit

$$h'(\xi) = g'(f(\xi))f'(\xi) \quad (\text{Kettenregel}).$$

*Beweis:* Sei zunächst  $f'(\xi) \neq 0$ . Sei  $(x_n)$  eine Folge mit  $x_n \rightarrow \xi$  und  $x_n \neq \xi$ . Da  $f$  in  $\xi$  stetig ist, folgt  $y_n = f(x_n) \rightarrow y = f(\xi)$ . Wegen  $f'(\xi) \neq 0$  folgt aus (12.1)  $y_n \neq y$  für genügend große  $n$ . Durch Erweiterung des Differenzenquotienten um  $f(x_n) - f(\xi) = y_n - y$  erhalten wir

$$\frac{h(x_n) - h(\xi)}{x_n - \xi} = \frac{(g(y_n) - g(y))(f(x_n) - f(\xi))}{(y_n - y)(x_n - \xi)} \rightarrow g'(y)|_{y=f(\xi)} f'(\xi).$$

Ist  $f'(\xi) = 0$ , so erhalten wir aus dem ersten Teil von Satz 12.4

$$\begin{aligned} \left| \frac{h(x) - h(\xi)}{x - \xi} \right| &= \left| \frac{g(f(x)) - g(f(\xi))}{x - \xi} \right| \\ &\leq K \left| \frac{f(x) - f(\xi)}{x - \xi} \right| \rightarrow 0 \quad \text{wegen } f'(\xi) = 0. \end{aligned}$$

$\square$

**Beispiele 12.7** (i) Ist  $f$  differenzierbar, so bestimmen wir die Ableitung von  $f^n$ , indem wir  $g(y) = y^n$  setzen. Aus der Kettenregel folgt dann

$$\frac{d}{dx} f^n(x) = \frac{d}{dy} y^n f'(x) = n f^{n-1}(x) f'(x).$$

(ii) Die Ableitung von  $\frac{1}{f}$  kann man ebenfalls aus der Kettenregel erschließen,

$$\frac{d}{dx} \left( \frac{1}{f(x)} \right) = \frac{d}{dy} y^{-1} f'(x) = -\frac{f'(x)}{f(x)^2}.$$

$\square$

## 12.4 Mittelwertsätze

**Satz 12.8 (Rolle)** *f sei im Intervall  $[a, b]$  stetig und in  $(a, b)$  differenzierbar mit  $f(a) = f(b)$ . Dann gibt es einen Punkt  $\xi \in (a, b)$  mit  $f'(\xi) = 0$ .*

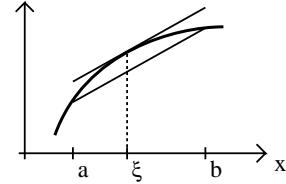
*Beweis:* Ist  $f$  konstant, so ist die Behauptung jedenfalls richtig. Ist  $f$  nicht konstant, so nimmt  $f$  in  $\xi \in (a, b)$  das Maximum oder Minimum an. Ist  $\xi$  das Maximum, so gilt  $f(\xi) \geq f(x)$  und aus dem Differenzenquotienten erschließen wir, dass  $f'_-(\xi) \geq 0$  und  $f'_+(\xi) \leq 0$ , also  $f'(\xi) = 0$ .  $\square$

**Satz 12.9 (Mittelwertsatz der Differentialrechnung)**

*f sei im Intervall  $[a, b]$  stetig und in  $(a, b)$  differenzierbar.*

*Dann gibt es ein  $\xi \in (a, b)$  mit*

$$\frac{f(b) - f(a)}{b - a} = f'(\xi).$$



*Beweis:* Wir betrachten die Hilfsfunktion

$$g(x) = f(x) - \alpha(x - a), \quad \alpha = \frac{f(b) - f(a)}{b - a}.$$

Es gilt  $g(a) = g(b) = f(a)$ . Nach dem Satz von Rolle gibt es ein  $\xi \in (a, b)$  mit  $0 = g'(\xi) = f'(\xi) - \alpha$ , woraus die Behauptung folgt.  $\square$

Erfüllt  $f$  die Voraussetzungen des Mittelwertsatzes und besitzt  $f'$  ein Vorzeichen im Intervall  $(a, b)$ , so kann man auf das Monotonieverhalten der Funktion schließen:

$$(12.2) \quad f'(x) \geq 0 \Leftrightarrow f \text{ ist monoton wachsend},$$

$$(12.3) \quad f'(x) \leq 0 \Leftrightarrow f \text{ ist monoton fallend},$$

$$f'(x) > 0 \Rightarrow f \text{ ist streng monoton wachsend},$$

$$f'(x) < 0 \Rightarrow f \text{ ist streng monoton fallend}.$$

Dabei erhalten wir die Implikation von links nach rechts, indem wir den Mittelwertsatz auf ein beliebiges Teilintervall anwenden. Die Implikation von rechts nach links folgt aus der Definition des Differenzenquotienten. Eine streng monoton wachsende Funktion muß nicht zwangsläufig  $f'(x) > 0$  erfüllen, wie das Beispiel  $f(x) = x^3$  zeigt.

Kombinieren wir (12.2) und (12.3), so erhalten wir eine weitere wichtige Folgerung aus dem Mittelwertsatz

$$(12.4) \quad f'(x) = 0 \text{ in } (a, b) \Rightarrow f \text{ ist konstant in } (a, b).$$

## 12.5 Ableitung von Potenzreihen

**Satz 12.10** *Die Potenzreihe  $p(x) = \sum_{n=0}^{\infty} a_n x^n$  ist im Inneren ihres Konvergenzbereichs  $|x| < R$  differenzierbar und darf dort gliedweise differenziert werden,*

$$p'(x) = \sum_{n=1}^{\infty} n a_n x^{n-1}.$$

Für die Exponentialfunktion, deren Reihe ja auf ganz  $\mathbb{R}$  konvergiert, erhalten wir mit diesem Satz

$$\frac{d}{dx} e^x = \sum_{n=1}^{\infty} \frac{n}{n!} x^{n-1} = e^x.$$

Entsprechend gilt für Sinus und Cosinus, deren Potenzreihen auf Seite 96 angegeben sind

$$\sin' x = \sum_{n=0}^{\infty} \frac{(-1)^n (2n+1)x^{2n}}{(2n+1)!} = \cos x,$$

$$\cos' x = \sum_{n=0}^{\infty} \frac{(-1)^n 2nx^{2n-1}}{(2n)!} = -\sin x.$$

## 12.6 Ableitung der Umkehrfunktion

**Satz 12.11** *f sei im Intervall I stetig und streng monoton. Ist die Umkehrfunktion  $\phi = f^{(-1)}$  in  $a = f(\xi)$  differenzierbar mit  $\phi'(a) \neq 0$ , so ist f in  $\xi$  differenzierbar mit*

$$f'(\xi) = \frac{1}{\phi'(a)} = \frac{1}{\phi'(f(\xi))}.$$

*Beweis:* Sei  $x_n \rightarrow \xi$  mit  $x_n \neq \xi$ . Da f streng monoton ist, gilt  $y_n = f(x_n) \neq f(\xi)$ . Mit  $\lim y_n = a$  folgt

$$\frac{f(x_n) - f(\xi)}{x_n - \xi} = \frac{y_n - a}{\phi(y_n) - \phi(a)} \rightarrow \frac{1}{\phi'(a)}.$$

□

**Beispiele 12.12** (i)  $f(x) = \sqrt[n]{x}$  ist die Umkehrfunktion von  $\phi(y) = y^n$ . Daher

$$(\sqrt[n]{x})' = \frac{1}{ny^{n-1}} = \frac{1}{n(\sqrt[n]{x})^{n-1}} = \frac{1}{n}x^{-1+1/n}.$$

(ii) Der Logarithmus ist die Umkehrfunktion der Exponentialfunktion. Daher für  $x > 0$  und  $y = \ln x$

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

(iii) Mit  $x^\alpha = \exp(\alpha \ln x)$  folgt damit für  $x > 0$

$$(x^\alpha)' = \frac{d}{dx} \exp(\alpha \ln x) = \exp(\alpha \ln x) \cdot \frac{d}{dx} \alpha \ln x = \alpha x^\alpha \cdot \frac{1}{x} = \alpha x^{\alpha-1}.$$

□

Die Schreibweise

$$f' = \frac{df}{dx} = \lim_{\Delta x \rightarrow 0} \frac{\Delta f}{\Delta x}$$

erinnert uns daran, dass die Ableitung aus dem Grenzübergang des Differenzenquotienten hervorgegangen ist und es verwundert nicht, dass man mit den Symbolen  $df$  und  $dx$  so rechnen kann wie mit reellen Zahlen:

$$\text{Kettenregel} \quad z(y(x)) : \quad \frac{dz}{dx} = \frac{dz}{dy} \frac{dy}{dx} \quad (\text{Erweiterung des Bruchs})$$

$$\text{Umkehrfunktion} \quad x(y) : \quad \frac{dx}{dy} = \frac{1}{\frac{dy}{dx}} = \frac{1}{(y'(x))}_{|x=x(y)} \quad (\text{Division des Bruchs})$$

Gleichzeitig macht diese Heuristik uns auf die Ähnlichkeit zwischen Kettenregel und der Ableitung der inversen Funktion aufmerksam. Rein formal können wir schreiben

$$x = \phi(f(x)) \Rightarrow 1 = x' = \phi'(f(x))f'(x),$$

woraus die Ableitung der Inversen folgt. Alle diese Überlegungen sind natürlich nicht mathematisch streng zu verstehen, aber als Gedächtnisstütze nützlich.

**12.7 Höhere Ableitungen** Besitzt  $f$  eine Ableitung  $f'(x)$ , so ist auch  $f'$  eine Funktion in  $x$ , auf die die Definition der Ableitung angewendet werden kann. Für diese *höheren Ableitungen* schreiben wir

$$\frac{d}{dx}f = \frac{df}{dx} = f' = f^{(1)}, \quad \frac{d^n}{dx^n}f = f^{(n)} = \frac{d}{dx}f^{(n-1)}.$$

**Beispiel 12.13** Für das Polynom

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_2 x^2 + a_1 x + a_0$$

erhalten wir

$$p'(x) = n a_n x^{n-1} + (n-1) a_{n-1} x^{n-2} + \dots + 2 a_2 x + a_1,$$

$$p''(x) = n(n-1) a_n x^{n-2} + (n-1)(n-2) a_{n-1} x^{n-3} + \dots + 2 a_2$$

und schließlich  $p^{(n)}(x) = n! a_n$ ,  $p^{(n+1)}(x) = 0$ .  $\square$

**12.8 Der Satz von Taylor** Wir nennen eine Funktion *m-mal stetig differenzierbar*, wenn die Ableitungen bis zur Ordnung  $m$  existieren und stetig sind.

**Satz 12.14** Sei  $I$  ein Intervall und  $f$  sei für ein  $n \in \mathbb{N}_0$  auf  $I$   $n+1$ -mal stetig differenzierbar. Für  $a, x \in I$  gilt die Taylorentwicklung

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \dots + \frac{1}{n!}f^{(n)}(a)(x-a)^n + R_n(x; a)$$

mit dem Restglied nach Lagrange

$$R_n(x; a) = \frac{(x-a)^{n+1}}{(n+1)!} f^{(n+1)}(\xi) \quad \text{für ein } \xi \in (a, x) \text{ oder } \xi \in (x, a).$$

*Beweis:* Wir können  $x > a$  annehmen, der Fall  $x < a$  verläuft völlig analog. Setze

$$g(t) = f(x) - f(t) - f'(t)(x-t) - \frac{f''(t)}{2!}(x-t)^2 - \dots - \frac{f^{(n)}(t)}{n!}(x-t)^n - m \frac{(x-t)^{n+1}}{(n+1)!}$$

mit  $t \in (a, x)$ . Wir können  $m$  so wählen, dass  $g(a) = 0$ . Da auch  $g(x) = 0$  gilt, gibt es nach dem Satz von Rolle ein  $\xi \in (a, x)$  mit  $g'(\xi) = 0$ . Daher

$$\begin{aligned} 0 = g'(\xi) &= 0 - f'(\xi) - f''(\xi)(x-\xi) + f'(\xi) - \frac{f'''(\xi)}{2!}(x-\xi)^2 + f''(\xi)(x-\xi) - \dots \\ &\quad - \frac{f^{(n+1)}(\xi)}{n!}(x-\xi)^n + \frac{f^{(n)}(\xi)}{(n-1)!}(x-\xi)^{n-1} + m \frac{(x-\xi)^n}{n!} \\ &= - \frac{f^{(n+1)}(\xi)}{n!}(x-\xi)^n + m \frac{(x-\xi)^n}{n!}. \end{aligned}$$

Damit ist  $m = f^{(n+1)}(\xi)$ . Wir setzen in  $g(t)$  den Wert  $t = a$  ein und erhalten die behauptete Formel.  $\square$

Als Spezialfall bekommen wir für  $n = 1$

$$f(x) = f(a) + (x-a)f'(\xi)$$

den Mittelwertsatz.

Wir können

$$f(x) = T_n(x; a) + R_n(x; a)$$

schreiben mit dem *Taylorpolynom*  $T_n(x; a)$  vom Grade  $\leq n$  und dem Restglied  $R_n(x; a)$ . Ist das Intervall  $I$  beschränkt und abgeschlossen, so ist die stetige Funktion  $f^{(n+1)}$  beschränkt und das Restglied lässt sich in der Form

$$|R_n(x; a)| \leq c|x - a|^{n+1}$$

abschätzen. Der Satz von Taylor besagt demnach, dass man eine  $n + 1$ -mal stetig differenzierbare Funktion durch ein Polynom vom Grad  $\leq n$  bis auf einen Fehler approximieren kann.

Häufig schreibt man in der Taylorentwicklung  $h$  statt  $x - a$ :

$$f(a + h) = f(a) + f'(a)h + \frac{1}{2}f''(a)h^2 + \dots + \frac{1}{n!}f^{(n)}(a)h^n + R_n(x; a)$$

mit dem Restglied

$$R_n(h; a) = \frac{h^{n+1}}{(n+1)!}f^{(n+1)}(a + \xi) \quad \text{für ein } \xi \in (0, h).$$

**Beispiele 12.15** (i) Für kleines  $|h|$  verwenden Ingenieure

$$\sqrt{1+h} \sim 1 + \frac{1}{2}h.$$

Um dies einzusehen, entwickeln wir die Wurzelfunktion nach Taylor. Mit

$$(\sqrt{x})' = \frac{1}{2}x^{-1/2}, \quad (\sqrt{x})'' = -\frac{1}{4}x^{-3/2}.$$

liefert die Taylor-Formel für  $n = 1$  und Entwicklungspunkt  $a = 1$

$$\sqrt{1+h} = 1 + \frac{1}{2}h - \frac{1}{8}(1+\xi)^{-3/2}h^2.$$

mit  $0 < \xi < h$  für  $h > 0$  und  $h < \xi < 0$  für  $h < 0$ . Für  $h > 0$  können wir das Restglied abschätzen durch

$$\left| -\frac{1}{8}(1+\xi)^{-3/2}h^2 \right| \leq \frac{h^2}{8}.$$

(ii) Es gilt

$$(\ln x)' = \frac{1}{x}, \quad (\ln x)'' = -\frac{1}{x^2},$$

also

$$\ln(1+h) = \ln 1 + \frac{1}{1}h - \frac{1}{2}\frac{h^2}{(1+\xi)^2}.$$

Da das Restglied ein Vorzeichen besitzt, erhalten wir mit  $\ln 1 = 0$  die Ungleichung

$$\ln(1+h) \leq h \quad \text{für } -1 < h < \infty.$$

(iii) Wir wollen die Funktion

$$f(x) = (2x + x^2)e^{-x}$$

durch das Taylorpolynom 2. Grades approximieren. Mit

$$f'(x) = -(2x + x^2)e^{-x} + (2 + 2x)e^{-x} = (2 - x^2)e^{-x}$$

$$f''(x) = -(2 - x^2)e^{-x} - 2xe^{-x} = (-2 - 2x + x^2)e^{-x}$$

$$f'''(x) = -(-2 - 2x + x^2)e^{-x} + (-2 + 2x)e^{-x} = (4x - x^2)e^{-x}$$

erhalten wir für das Taylorpolynom 2. Ordnung mit Entwicklungspunkt  $a = 0$

$$T_2(x; 0) = 2x - x^2.$$

Auf dem Intervall  $[0, 1]$  können wir das Restglied folgendermaßen mit einem  $\xi \in (0, x)$  abschätzen:

$$|R_2(x; 0)| = \frac{x^3}{6} |f'''(\xi)| \leq \frac{x^3}{6} |4\xi - \xi^2| e^{-\xi}$$

Es gilt  $e^{-\xi} \leq 1$  auf dem Intervall  $[0, x]$ . Die Funktion  $g(\xi) = 4\xi - \xi^2$  ist auf dem Intervall  $[0, 1]$  nichtnegativ und streng monoton steigend. Mit  $0 \leq g(\xi) \leq g(x)$  folgt daher

$$|R_2(x; 0)| \leq \frac{x^3}{6} (4x - x^2) \quad \text{für } 0 \leq x \leq 1.$$

□

**12.9 Die Landauschen Symbole** Seien  $f, g$  in einer Umgebung des Punktes  $\xi$  definiert. Wir sagen  $f$  ist gleich groß  $O$  von  $g$  und schreiben

$$f(x) = O(g(x)), \quad x \rightarrow \xi,$$

wenn es ein  $M \in \mathbb{R}$  gibt mit

$$\left| \frac{f(x)}{g(x)} \right| \leq M \quad \text{in einer Umgebung von } \xi.$$

Wir sagen  $f$  ist gleich klein  $o$  von  $g$  und schreiben

$$f(x) = o(g(x)), \quad x \rightarrow \xi,$$

wenn

$$\lim_{x \rightarrow \xi} \frac{f(x)}{g(x)} = 0.$$

Wenn beispielsweise  $\lim_{x \rightarrow \xi} g(x) = 0$ , so bedeutet

$f = O(g)$   $f$  geht so schnell oder schneller gegen Null als  $g$ ,

$f = o(g)$   $f$  geht schneller gegen Null als  $g$ .

Man nennt  $O$  und  $o$  die *Landauschen Symbole*. Die Bezeichnungen  $O$  und  $o$  werden sinngemäß auch für  $\pm\infty$  angewendet. Beispielsweise bedeutet  $f(x) = O(x^n)$ ,  $x \rightarrow \infty$ , dass es ein  $M \in \mathbb{R}$  gibt mit  $|f(x)| \leq Mx^n$  für genügend große  $x$ .

Die Landauschen Symbole gestatten suggestive Schreibweisen, weil sie den Approximationss- oder Wachstumsaspekt hervorheben. Die Ableitung

$$(12.5) \quad \lim_{x \rightarrow \xi} \frac{f(x) - f(\xi)}{x - \xi} = f'(\xi)$$

lässt sich äquivalent schreiben

$$(12.6) \quad f(x) - f(\xi) = f'(\xi)(x - \xi) + h(x)$$

mit einer Funktion  $h(x) = o(|x - \xi|)$ , denn wenn wir in (12.6) durch  $x - \xi$  teilen, konvergiert der Ausdruck  $h(x)/(x - \xi)$  immer noch gegen Null und wir erhalten (12.5). Statt (12.6) schreibt man noch kürzer

$$f(x) - f(\xi) = f'(\xi)(x - \xi) + o(|x - \xi|).$$

Wenn es in der Taylorentwicklung nicht auf die explizite Gestalt des Restglieds ankommt, verwendet man auch die Schreibweise

$$f(x) = T_n(x; a) + O(|x - a|^{n+1}),$$

das Taylorpolynom approximiert  $f$  bis auf einen Fehler der Ordnung  $O(|x - a|^{n+1})$ .

Unbestimmte Ausdrücke der Form  $f(x)/g(x)$  mit  $f(x), g(x) \rightarrow 0$  für  $x \rightarrow \xi$  untersucht man am besten mit Hilfe der Taylorentwicklung um den Punkt  $\xi$  unter Verwendung der Landauschen Symbole.

Für  $\sin x/x$  ist  $\sin x = x + O(x^3)$  und daher

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = \lim_{x \rightarrow 0} \frac{x + O(x^3)}{x} = 1.$$

**Beispiel 12.16** Wir bestimmen für  $a \in \mathbb{R}$

$$L(a) = \lim_{x \rightarrow 0} \frac{e^{-x^2} - 1 + x \sin x}{\sqrt{1 - x^2 + ax^2} - 1}.$$

Dieser Ausdruck ist von der Form „ $\frac{0}{0}$ “. Für die Exponentialfunktion im Zähler verwenden wir

$$e^{-t} = 1 - t + \frac{1}{2}t^2 + O(t^3)$$

und erhalten damit für den Zähler insgesamt

$$e^{-x^2} - 1 + x \sin x = 1 - x^2 + \frac{1}{2}x^4 + O(x^6) - 1 + x(x - \frac{1}{6}x^3 + O(x^5)) = \frac{1}{3}x^4 + O(x^6).$$

Für den Wurzausdruck liefert Taylorentwicklung

$$\sqrt{1+t} = 1 + \frac{1}{2}t - \frac{1}{8}t^2 + O(t^3),$$

für den Nenner gilt daher

$$1 - \frac{1}{2}x^2 - \frac{1}{8}x^4 + O(x^6) + ax^2 - 1 = (a - \frac{1}{2})x^2 - \frac{1}{8}x^4 + O(x^6).$$

Damit ist  $L(a) = 0$  für  $a \neq \frac{1}{2}$  und  $L(a) = \frac{8}{3}$  für  $a = \frac{1}{2}$ .  $\square$

**12.10 Relative Extrema** Eine Funktion  $f$  sei in einer Umgebung des Punktes  $\xi$  definiert.  $\xi$  heißt *relatives Minimum* von  $f$ , wenn es eine Umgebung  $U$  von  $\xi$  gibt mit  $f(\xi) \leq f(x)$  für alle  $x \in U$ . In einem *relativen Maximum* gilt analog  $f(\xi) \geq f(x)$ . Ein Minimum oder Maximum heißt *strikt*, wenn statt  $\leq$  oder  $\geq$  die strikte Ungleichung gilt. Zu beachten ist bei dieser Definition, dass  $f$  mindestens in einem Intervall  $(\xi - \varepsilon, \xi + \varepsilon)$  definiert sein muss. Liegt der Extremwert am Rande des Definitionsbereichs von  $f$ , so sprechen wir von einem *einseitigen* Minimum oder Maximum.

**Satz 12.17 (Notwendige Bedingungen für einen Extremwert)** Die Funktion  $f : (a, b) \rightarrow \mathbb{R}$  besitze in  $\xi \in (a, b)$  ein relatives Minimum oder Maximum.

- (a) Ist  $f$  einmal stetig differenzierbar, so gilt  $f'(\xi) = 0$ .
- (b) Ist  $f$  zweimal stetig differenzierbar, so gilt zusätzlich

$$f''(\xi) \geq 0 \text{ falls } \xi \text{ Minimum, } f''(\xi) \leq 0 \text{ falls } \xi \text{ Maximum.}$$

*Beweis:* (a) Sei  $\xi$  ein relatives Minimum. Für  $h > 0$  gilt dann

$$\frac{1}{h}(f(\xi + h) - f(\xi)) \geq 0, \quad -\frac{1}{h}(f(\xi - h) - f(\xi)) \leq 0.$$

Durch Grenzübergang folgt  $f'(\xi) = 0$ .

(b) Für ein relatives Minimum  $\xi$  gilt nach (a)  $f'(\xi) = 0$ . Die Taylor-Entwicklung für  $n = 1$  lautet daher

$$f(x) = f(\xi) + \frac{1}{2}f''(a)(x - \xi)^2, \quad a \in (x, \xi),$$

daher

$$0 \leq f(x) - f(\xi) = \frac{1}{2}f''(a)(x - \xi)^2, \quad a \in (x, \xi).$$

Für eine Folge  $(x_n)$  mit  $x_n \rightarrow \xi$  folgt für die zugehörigen  $a_n$ , dass  $a_n \rightarrow \xi$ . Wegen der Stetigkeit von  $f''$  erhalten wir  $f''(\xi) \geq 0$ .  $\square$

**Satz 12.18 (Hinreichende Bedingung für einen Extremwert)** *Für die zweimal stetig differenzierbare Funktion  $f$  seien in  $\xi \in (a, b)$  die Bedingungen  $f'(\xi) = 0$  sowie  $f''(\xi) > 0$  ( $f''(\xi) < 0$ ) erfüllt. Dann besitzt  $f$  in  $\xi$  ein striktes relatives Minimum (Maximum).*

*Beweis:* Ähnlich wie im Beweis von Satz 12.10(b) bekommen wir aus der Taylorentwicklung um den Punkt  $\xi$  wegen  $f'(\xi) = 0$ ,

$$f(x) - f(\xi) = \frac{1}{2}f''(a)(x - \xi)^2, \quad a \in (x, \xi).$$

Ist nun  $f''(\xi) > 0$ , so ist wegen der Stetigkeit von  $f''$  auch  $f''(a) > 0$  für alle  $a$  in einer genügend kleinen Umgebung von  $\xi$ . Daraus folgt die Behauptung.  $\square$

Bei einseitigen Extremwerten gibt es nur notwendige und hinreichende Bedingungen erster Ordnung, also nur für die erste Ableitung von  $f$ . Besitzt ein einmal stetig differenzierbares  $f$  ein Minimum an der Stelle  $a$ , so folgt für  $h > 0$   $f(a + h) - f(a) \geq 0$  und damit  $f'(a) \geq 0$ . Gilt  $f'(a) > 0$ , so folgt aus dem Differenzenquotienten, dass  $f$  in  $a$  ein striktes relatives Minimum besitzt. Zu beachten ist dabei, dass die Vorzeichen für den rechten Randpunkt sich umkehren.

Bei der Bestimmung der globalen Extremwerte einer differenzierbaren Funktion sind alle Nullstellen der ersten Ableitungen und die Randpunkte zu untersuchen, bei unbeschränktem Definitionsbereich zusätzlich das Verhalten im Unendlichen.

**Beispiel 12.19** Wir untersuchen die Funktion

$$f(x) = \frac{x^2}{x+1} \quad \text{in } I = (-1, \infty).$$

Es gilt für  $x > -1$

$$f'(x) = \frac{2x}{x+1} - \frac{x^2}{(x+1)^2} = 0 \Leftrightarrow 2x(x+1) - x^2 = 0$$

mit einziger Lösung  $x = 0$  in  $I$ . Da  $f' < 0$  in  $(-1, 0)$  und  $f' > 0$  in  $(0, \infty)$ , ist  $f$  ist streng monoton fallend in  $(-1, 0)$  und streng monoton wachsend in  $(0, \infty)$ .  $x = 0$  ist daher das globale Minimum. Klar ist  $\lim_{x \rightarrow -1^+} f(x) = \infty$ ,  $\lim_{x \rightarrow \infty} f(x) = \infty$ .  $\square$

## 13 Weitere Themen der Analysis

**13.1 Komplexe Wurzeln** Mit Hilfe von Polarkoordinaten können wir die komplexen Wurzeln, also die Lösungen der Gleichung  $z^n = \alpha$  für  $\alpha \in \mathbb{C}$ , leicht bestimmen. Ist  $r = |\alpha| \neq 0$  und  $\phi = \arg \alpha$ , so lässt sich  $\alpha$  in der Form

$$\alpha = r(\cos \phi + i \sin \phi)$$

schreiben. Da bei der komplexen Multiplikation die Beträge multipliziert und die Argumente addiert werden, haben wir genau  $n$  Lösungen, die alle den Betrag  $\sqrt[n]{r}$  und die Argumente  $(\phi + 2k\pi)/n$  für  $k = 0, 1, \dots, n-1$  besitzen.

Die Lösungen von  $z^n = 1$  werden *komplexe Einheitswurzeln* genannt,

$$z_k = \cos \frac{2k\pi}{n} + i \sin \frac{2k\pi}{n}.$$

Sie liegen auf dem komplexen Einheitskreis und bilden dort ein reguläres  $n$ -Eck.

Da man im Komplexen kein klares Verfahren hat, um die Wurzel eindeutig zu machen, ist man im Gegensatz zum Reellen übereingekommen, alle Lösungen von  $z^n = \alpha$  als komplexe Wurzeln  $\sqrt[n]{\alpha}$  zu bezeichnen.

**Beispiel 13.1** Wir bestimmen alle Lösungen der Gleichung  $z^6 - iz^3 = 1$ . Mit  $w = z^3$  folgt  $w^2 - iw = 1$  und

$$(w - \frac{i}{2})^2 = 1 - \frac{1}{4} = \frac{3}{4} \Rightarrow w_{\pm} = \frac{i}{2} \pm \frac{1}{2}\sqrt{3}.$$

Es gilt  $w_{\pm} = \cos \phi_{\pm} + i \sin \phi_{\pm}$  mit  $\phi_+ = \pi/6$  und  $\phi_- = 5\pi/6$ . Damit bekommen wir die 6 Lösungen

$$\cos\left(\frac{\pi}{18} + \frac{2k\pi}{3}\right) + i \sin\left(\frac{\pi}{18} + \frac{2k\pi}{3}\right), \cos\left(\frac{5\pi}{18} + \frac{2k\pi}{3}\right) + i \sin\left(\frac{5\pi}{18} + \frac{2k\pi}{3}\right), k = 1, 2, 3.$$

□

### 13.2 Polynome und Partialbruchzerlegung

Für komplexe Zahlen  $a_n, a_{n-1}, \dots, a_0$  heißt

$$(13.1) \quad p(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_0$$

(komplexes) *Polynom*. Ist  $a_n \neq 0$ , so heißt  $\text{grad } p = n$  der *Grad* von  $p$ .

Zunächst untersuchen wir die Division mit Rest, die auch als *Euklidischer Algorithmus* bezeichnet wird.

**Satz 13.2 (Euklidischer Algorithmus)** Sei  $p$  ein Polynom vom Grad  $m$  und  $q$  ein Polynom vom Grad  $n$  mit  $m \geq n$ . Dann gibt es eindeutig bestimmte Polynome  $s$  und  $r$  mit  $\text{grad } s = m - n$  und  $\text{grad } r < n$  mit

$$p = qs + r.$$

**Beispiel 13.3** Sei

$$p(z) = iz^5 + z^3 - z^2 + 1, \quad q(z) = z^2 - 1.$$

Wir bringen zuerst den höchsten Koeffizienten von  $p$  zum Verschwinden,

$$p(z) - iz^3 q(z) = (1 - i)z^3 - z^2 + 1,$$

fahren auf diese Weise fort,

$$p(z) - iz^3 q(z) - (1 - i)zq(z) = -z^2 + (1 - i)z + 1,$$

und erhalten

$$s(z) = iz^3 + (1 + i)z - 1, \quad r(z) = (1 + i)z$$

□

**Lemma 13.4** (a) Sei  $p$  ein Polynom vom Grad  $n$ . Für jedes  $\xi \in \mathbb{C}$  gibt es eindeutig bestimmte komplexe Zahlen  $b_n, \dots, b_0$ ,  $b_n \neq 0$ , mit

$$(13.2) \quad p(z) = b_n(z - \xi)^n + b_{n-1}(z - \xi)^{n-1} + \dots + b_1(z - \xi) + b_0.$$

In diesem Fall bezeichnen wir  $\xi$  als Entwicklungspunkt des Polynoms  $p$ .

(b) Besitzt das Polynom  $p$  vom Grade  $n$  eine Nullstelle  $\xi \in \mathbb{C}$ , so gibt es ein eindeutiges Polynom  $q$  vom Grad  $n - 1$  mit

$$p(z) = (z - \xi)q(z).$$

*Beweis:* (a) Wir multiplizieren die Darstellung (13.2) mit der binomischen Formel aus und vergleichen die Koeffizienten mit (13.1), was

$$b_k = \sum_{i=k}^n a_i \binom{i}{k} \xi^{i-k}, \quad \text{insbesondere } b_0 = p(\xi), \quad b_n = a_n,$$

ergibt.

(b) Ist  $\xi$  eine Nullstelle, so folgt  $b_0 = 0$  in der Darstellung (13.2). Wir können  $z - \xi$  ausklammern, es verbleibt das gesuchte Polynom  $q$ .  $\square$

Das Lemma bleibt für reelle Polynome sinngemäß richtig. Insbesondere ist das Polynom  $q$  in (b) reell, wenn die Nullstelle  $\xi$  reell ist. Für reelle Polynome gilt  $\overline{p(z)} = p(\bar{z})$ . Ist daher  $\xi$  Nullstelle des reellen Polynoms  $p$ , so ist auch  $\bar{\xi}$  Nullstelle. Nichtreelle Nullstellen reeller Polynome treten also immer paarweise auf. Aus Lemma 13.4 erhalten wir daher

$$p(z) = (z - \xi)(z - \bar{\xi})q(z) = r(z)q(z),$$

wobei  $r(z) = z^2 - 2\operatorname{Re}\xi z + |\xi|^2$  ein reelles quadratisches Polynom ist. Mit  $p$  und  $r$  ist damit auch  $q$  reell.

Im Reellen hat die Gleichung  $x^2 = -1$  keine Lösung. Historisch gesehen wurden die komplexen Zahlen deshalb eingeführt, weil man glaubte, dass im Körper der komplexen Zahlen jedes Polynom eine Nullstelle besitzt. Dieser Glaube erwies sich erst relativ spät als begründet, als Gauß den folgenden Satz bewies.

**Satz 13.5 (Fundamentalsatz der Algebra)** Jedes nichtkonstante Polynom besitzt mindestens eine Nullstelle.

Einen einfachen Beweis tragen wir in Abschnitt 13.5 nach.

Wenden wir den Fundamentalsatz und Lemma 13.4 sukzessive an, so hat jedes Polynom vom Grad  $n$  genau  $n$  Nullstellen und genügt der Darstellung

$$p(z) = a_n z^n + \dots + a_0 = a_n(z - \xi_1) \dots (z - \xi_n).$$

Dabei können die Nullstellen  $\xi_i$  auch mehrfach auftreten.

Die wichtigste Anwendung des Hauptsatzes der Algebra ist eine Darstellung rationaler Funktionen, die *Partialbruchzerlegung* genannt wird. Ist  $r(z) = q(z)/p(z)$  eine rationale Funktion, so können wir wegen des Euklidischen Algorithmus  $m = \operatorname{grad} q < n = \operatorname{grad} p$  annehmen. Durch Kürzen des Bruches können wir ferner den höchsten Koeffizienten von  $p$  zu 1 normieren. Nach dem Fundamentalsatz hat  $p$  die Darstellung

$$p(z) = (z - \xi_1)^{l_1} (z - \xi_2)^{l_2} \dots (z - \xi_k)^{l_k}$$

mit den Nullstellen  $\xi_1, \dots, \xi_k$  und  $\sum l_k = n$ .

**Satz 13.6** Jede rationale Funktion  $r(z) = q(z)/p(z)$  mit  $m = \text{grad } q < n = \text{grad } p$  lässt sich eindeutig als Summe von Partialbrüchen schreiben,

$$\frac{q(z)}{p(z)} = \sum_{i=1}^k \left( \frac{a_{i1}}{z - \xi_i} + \frac{a_{i2}}{(z - \xi_i)^2} + \dots + \frac{a_{il_i}}{(z - \xi_i)^{l_i}} \right).$$

*Beweis:* Wir verwenden vollständige Induktion über den Nennergrad  $n$ . Für  $n = 1$  ist  $q$  konstant und daher nichts zu beweisen. Für den Induktionsschritt dürfen wir annehmen, dass es die behauptete Partialbruchzerlegung gibt für Polynome mit  $n = \text{grad } p > \text{grad } q$ . Sei also jetzt  $r(z) = q(z)/p(z)$  mit  $\text{grad } p = n + 1 > \text{grad } q$ . Ist  $\xi$  eine  $l$ -fache Nullstelle von  $p$ , so

$$p(z) = (z - \xi)^l s(z) \quad \text{mit } s(\xi) \neq 0.$$

Es gilt

$$\frac{q(z)}{s(z)} - \frac{q(\xi)}{s(\xi)} = \frac{q(z)s(\xi) - s(z)q(\xi)}{s(z)s(\xi)} = \frac{(z - \xi)t(z)}{s(z)} \quad \text{mit } \text{grad } t \leq n - 1,$$

weil  $\xi$  Nullstelle von  $q(z)s(\xi) - s(z)q(\xi)$  ist. Wir haben also

$$(13.3) \quad \frac{q(z)}{(z - \xi)^l s(z)} - \frac{q(\xi)}{(z - \xi)^l s(\xi)} = \frac{t(z)}{(z - \xi)^{l-1} s(z)}.$$

Wegen  $\text{grad}((z - \xi)^{l-1} s(z)) = n > \text{grad } t(z)$  können wir auf der rechten Seite die Induktionsvoraussetzung anwenden und haben die Existenz der Partialbruchzerlegung bewiesen.

Zum Nachweis der Eindeutigkeit nehmen wir an, dass es zwei Partialbruchzerlegungen mit Koeffizienten  $a_{ij}$  und  $b_{ij}$  gibt. Wir bilden die Differenz dieser Zerlegungen und erhalten eine Zerlegung der Nullfunktion mit Koeffizienten  $a_{ij} - b_{ij}$ . Diese multiplizieren wir mit  $(z - \xi_j)^{l_j}$ . Der Grenzwert  $z \rightarrow \xi_j$  liefert dann  $a_{il_j} = b_{il_j}$ . Durch Multiplikation mit  $(z - \xi_j)^{l_j-1}$  lässt sich dieses Argument für die nächstniedrigere Potenz wiederholen.  $\square$

Bei der praktischen Durchführung der Partialbruchzerlegung setzt man wie im Satz angegeben an. Indem man die rechte Seite auf den Hauptnenner bringt, lassen sich die Koeffizienten der Partialbruchzerlegung durch Koeffizientenvergleich bestimmen. Alternativ können wir einzelne Werte für  $z$  in den Ansatz einsetzen, was zu einem linearen Gleichungssystem für die gesuchten Koeffizienten führt. Dieses in jedem Fall mühsame Verfahren kann man sich etwas erleichtern, indem man beachtet, dass der höchste Koeffizient der Zerlegung in (13.3) durch  $q(\xi)/s(\xi)$  gegeben ist.

**Beispiel 13.7** Für  $r(z) = \frac{z+1}{z(z-1)^2}$  setzen wir an

$$r(z) = \frac{a}{z} + \frac{b_2}{(z-1)^2} + \frac{b_1}{z-1}.$$

Die höchsten Koeffizienten erhalten wir aus (13.3) oder direkt durch folgende Überlegung. Um beispielsweise  $a$  zu bestimmen, multiplizieren wir obige Gleichung mit  $z$  und führen den Grenzübergang  $z \rightarrow 0$  durch. Damit hängt die Berechnung von  $a$  nicht von den anderen Unbekannten ab und wir erhalten

$$a = \lim_{z \rightarrow 0} zr(z) = \lim_{z \rightarrow 0} \frac{z+1}{(z-1)^2} = 1.$$

Auf die gleiche Weise gilt

$$b_2 = \lim_{z \rightarrow 1} (z-1)^2 r(z) = \lim_{z \rightarrow 1} \frac{z+1}{z} = 2.$$

Da dieses Verfahren für den letzten Koeffizienten versagt, bestimmen wir ihn durch Einsetzen eines beliebigen  $z$ . Für  $z = 2$  ist

$$b_1 = r(2) - \frac{a}{2} - b_2 = \frac{3}{2} - \frac{1}{2} - 2 = -1$$

und damit

$$\frac{z+1}{z(z-1)} = \frac{1}{z} + \frac{2}{(z-1)^2} - \frac{1}{z-1}.$$

□

Sind  $p, q$  reelle Polynome mit  $n = \text{grad } p > m = \text{grad } q$ , so lässt sich die Partialbruchzerlegung auch im Reellen durchführen, indem man komplexe konjugierte Nullstellen von  $p$  zu einem reellen quadratischen Polynom zusammenfasst. Sind  $\xi_1, \dots, \xi_k$  die reellen Nullstellen von  $p$ , so gilt

$$\frac{q(z)}{p(z)} = \sum_{i=1}^k \sum_{j=1}^{l_i} \frac{a_{ij}}{(z-\xi_i)^j} + \sum_{i=1}^h \sum_{j=1}^{m_i} \frac{b_{ij}z + c_{ij}}{(z^2 + \alpha_i z + \beta_i)^j}.$$

In dieser Darstellung sind alle Größen reell. Die Polynome  $z^2 + \alpha_i z + \beta_i$  bestimmt man aus  $(z-\xi)(z-\bar{\xi})$ . Am einfachsten bestimmt man die reelle Partialbruchzerlegung, indem man erst die komplexe berechnet und dann die komplexe konjugierten Terme zusammenfasst.

**Beispiel 13.8** Für  $r(z) = \frac{z^3}{(z^2 + 1)^2}$  setzen wir an

$$(13.4) \quad r(z) = \frac{a_2}{(z-i)^2} + \frac{a_1}{z-i} + \frac{b_2}{(z+i)^2} + \frac{b_1}{z+i}.$$

Die höchsten Koeffizienten bestimmen wir mit

$$a_2 = \lim_{z \rightarrow i} (z-i)^2 r(z) = \frac{i}{4}, \quad b_2 = \lim_{z \rightarrow -i} (z+i)^2 r(z) = -\frac{i}{4}.$$

Nun setzen wir  $z = 0$  und  $z = 2i$  in (13.4) ein und erhalten das lineare Gleichungssystem

$$-a_1 + b_1 = 0, \quad 3a_1 + b_1 = 2,$$

mit Lösung  $a_1 = b_1 = 1/2$ . Die komplexe Partialbruchzerlegung lautet dann

$$r(z) = \frac{i}{4(z-i)^2} - \frac{i}{4(z+i)^2} + \frac{1}{2(z-i)} + \frac{1}{2(z+i)}.$$

Um auf die reelle Zerlegung zu kommen, addieren wir die komplexe konjugierten Summanden gleicher Ordnung

$$r(x) = -\frac{x}{(x^2+1)^2} + \frac{x}{x^2+1}.$$

□

### 13.3 Konvergenz komplexer Zahlenfolgen

Der Kreis um  $a \in \mathbb{C}$  mit Radius  $\varepsilon$

$$B_\varepsilon(a) = \{z \in \mathbb{C} : |z-a| < \varepsilon\} \subset \mathbb{C}$$

heißt  $\varepsilon$ -Umgebung von  $a$ . Eine Folge  $(z_n)$ ,  $z_n \in \mathbb{C}$ , konvergiert gegen  $\xi \in \mathbb{C}$ , wenn in jeder  $\varepsilon$ -Umgebung von  $\xi$  fast alle Folgenglieder liegen.

**Satz 13.9** Mit  $z_n = x_n + iy_n$  und  $\xi = a + ib$  gilt  $z_n \rightarrow \xi$  genau dann, wenn  $x_n \rightarrow a$  und  $y_n \rightarrow b$  in  $\mathbb{R}$ .

*Beweis:* Für  $z = x + iy$  gilt wegen  $|z| = \sqrt{x^2 + y^2}$

$$(13.5) \quad |x|, |y| \leq |z| \leq |x| + |y|.$$

$z_n \rightarrow \xi$  ist äquivalent zu

$$|z_n - \xi| < \varepsilon \quad \text{für alle } n \geq N.$$

Mit (13.5) folgt daraus auch  $|x_n - a|, |y_n - b| < \varepsilon$  und damit  $x_n \rightarrow a$  und  $y_n \rightarrow b$ .

Gilt umgekehrt  $x_n \rightarrow a$  und  $y_n \rightarrow b$ , so folgt wieder aus (13.5) für genügend große  $n$

$$|z_n - \xi| < 2\varepsilon,$$

was  $z_n \rightarrow \xi$  impliziert.  $\square$

Für Reihen komplexer Zahlen wird Konvergenz wie im Reellen mit der Konvergenz der Partialsummen definiert. Entsprechend heißt  $\sum z_n$  absolut konvergent, wenn  $\sum |z_n|$  konvergiert. Nach dem letzten Satz ist dies äquivalent dazu, dass die beiden reellen Reihen  $\sum \operatorname{Re} z_n$  und  $\sum \operatorname{Im} z_n$  absolut konvergent sind. Daher bleiben Majoranten-, Wurzel- und Quotientenkriterium für die absolute Konvergenz komplexer Reihen gültig.

**13.4 Stetigkeit** Sei  $D \subset \mathbb{C}$ ,  $f : D \rightarrow \mathbb{C}$ .  $f$  heißt stetig in  $\xi \in D$ , wenn für alle Folgen  $(z_n)$  mit  $z_n \in D$  und  $z_n \rightarrow \xi$  gilt  $f(z_n) \rightarrow \xi$ .  $f$  heißt stetig in  $D$ , wenn  $f$  in jedem Punkt von  $D$  stetig ist.

Wie im Reellen beweist man, dass auch das  $\varepsilon, \delta$ -Kriterium äquivalent zur Stetigkeit ist:  $f$  ist genau dann stetig in  $\xi \in D$ , wenn es zu jedem  $\varepsilon > 0$  ein  $\delta > 0$  gibt mit  $|f(z) - f(\xi)| < \varepsilon$  für alle  $z$  mit  $|z - \xi| < \delta$ .

Die aus dem Reellen bekannten Sätze über die gleichmäßige Konvergenz von Funktionenfolgen und -reihen bleiben mit gleichem Beweis richtig. Aus diesem Themenkreis benötigen wir nur den folgenden

**Satz 13.10** Die Funktionen  $f_n : D \rightarrow \mathbb{C}$  seien stetig für alle  $n \in \mathbb{N}$ . Wenn  $|f_n(z)| \leq a_n$  für alle  $z \in D$  und die Reihe  $\sum a_n$  konvergent ist, so konvergiert die Reihe  $\sum f_n(z)$  gleichmäßig absolut gegen eine stetige Funktion  $f : D \rightarrow \mathbb{C}$ .

**13.5 Potenzreihen** Die Konvergenz einer komplexen Potenzreihe

$$p(z) = \sum_{n=0}^{\infty} a_n z^n, \quad a_n \in \mathbb{C}$$

lässt sich leicht aus den Sätzen 13.10 und 11.15 ableiten. Wir benötigen eine konvergente Majorante, die sich aus dem Wurzel- oder Quotientenkriterium des Satzes 11.15 ableiten lässt. Satz 11.15 bleibt also für komplexe Potenzreihen gültig, insbesondere haben wir Konvergenz gegen eine stetige Grenzfunktion innerhalb eines Kreises vom Radius  $\frac{1}{L}$  und Divergenz außerhalb dieses Kreises.

Jede reelle Potenzreihe  $f(x) = \sum a_n x^n$  lässt sich auf die komplexe Zahlenebene mit gleichem Konvergenzradius fortsetzen, indem man einfach  $x \in \mathbb{C}$  einsetzt. Auf diese Weise bekommen wir die komplexe Exponentialfunktion sowie den komplexen Sinus und Cosinus

$$\exp z = \sum_{n=0}^{\infty} \frac{z^n}{n!}, \quad \sin z = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{(2n+1)!}, \quad \cos z = \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n}}{(2n)!}.$$

Indem wir  $iz$  in die Exponentialfunktion einsetzen, erhalten wir durch Koeffizientenvergleich die *Eulersche Gleichung*

$$(13.6) \quad e^{iz} = \cos z + i \sin z$$

und damit zwei weitere wichtige Gleichungen

$$(13.7) \quad \cos z = \frac{1}{2}(e^{iz} + e^{-iz}), \quad \sin z = \frac{1}{2i}(e^{iz} - e^{-iz}).$$

Die Funktionalgleichung für die Exponentialfunktion

$$(13.8) \quad e^{z+w} = e^z e^w \quad \text{für } z, w \in \mathbb{C},$$

folgt mit gleichem Beweis wie im Reellen. Hieraus erhalten wir  $e^z e^{-z} = 1$ , insbesondere  $e^z \neq 0$ .

Eine anschauliche Vorstellung vom Verhalten der komplexen Exponentialfunktion bekommen wir, indem wir in (13.6) ein reelles  $y$  einsetzen

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

Für den Absolutbetrag von  $e^z$  ist daher nur der Realteil verantwortlich, der Imaginärteil bestimmt die Richtung von  $e^z$ . Die Exponentialfunktion ist damit  $2\pi$ -periodisch in  $y$ -Richtung. Aus der letzten Gleichung erhalten wir eine elegante Version der Polardarstellung komplexer Zahlen

$$z = r e^{i\phi}, \quad \text{mit } r = |z| \text{ und } \phi = \arg z.$$

Komplexe Multiplikation und Division lassen sich hiermit schön veranschaulichen,

$$zw = r s e^{i(\phi+\psi)}, \quad \frac{z}{w} = \frac{r}{s} e^{i(\phi-\psi)}.$$

Für reelle  $\phi$  notieren wir noch einige Folgerungen,

$$|e^{i\phi}| = 1, \quad e^{i\phi} = e^{i(\phi+2k\pi)} \text{ für } k \in \mathbb{Z}, \quad \overline{e^{i\phi}} = e^{-i\phi} = (e^{i\phi})^{-1}.$$

Die meisten Rechenregeln für die reellen trigonometrischen Funktionen lassen sich unter Verwendung der komplexen Beziehungen (13.7),(13.8) jetzt sehr viel einfacher herleiten. Die Additionstheoreme für Cosinus und Sinus erhält man aus

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

indem man hier Real- und Imaginärteile betrachtet.

Als letztes tragen wir den Beweis des Fundamentalsatzes der Algebra nach. Mit  $K_r(\xi)$  bezeichnen wir die Kreislinie um  $\xi$  mit Radius  $r$ . Für eine beliebige stetige Funktion  $f : \mathbb{C} \rightarrow \mathbb{C}$ , die keine Nullstelle in  $K_r(\xi)$  besitzt, definieren wir die *Drehungszahl*  $d(K_r(\xi), f)$ , indem wir gedanklich mit  $f(z)$  den Kreis im Gegenuhrzeigersinn entlanglaufen und dabei beobachten, wie oft sich  $f(z)$  um den Nullpunkt dreht. Beispielsweise umrundet  $f(z) = z$  den Nullpunkt einmal, wenn wir den Kreis  $K_1(0)$  entlanglaufen, also  $d(K_1(0), z) = 1$ . Aus der Darstellung

$$z^n = |z|^n e^{in\phi}, \quad \phi = \arg z,$$

erhalten wir  $d(K_r(0), z^n) = n$  für alle  $r > 0$ .

Die Drehungszahl hängt stetig von  $f$  ab: Kleine Störungen von  $f$  verändern die Drehungszahl nicht. Das folgende Lemma ist daher anschaulich klar.

**Lemma 13.11** Ist  $|f(z)| \geq a > 0$  und  $|g(z)| \leq a/2$  auf  $K_r(\xi)$ , so gilt  $d(K_r(\xi), f) = d(K_r(\xi), f+g)$ .

Ist

$$p(z) = z^n + q(z) \quad \text{mit } q(z) = a_{n-1} z^{n-1} + \dots + a_0$$

ein komplexes Polynom, so folgt mit

$$M = |a_{n-1}| + |a_{n-2}| + \dots + |a_0|$$

für  $|z| = R \geq 1$  die Abschätzung  $|q(z)| \leq MR^{n-1}$ . Wegen  $|z^n| = R^n$  gilt nach dem Lemma für genügend großes  $R$  folglich  $d(K_R, p) = d(K_R, z^n) = n$ .

Ist  $\xi \in \mathbb{C}$  ein Punkt mit  $a = |p(\xi)| > 0$ , so gibt es wegen der Stetigkeit von  $p$  ein  $\delta > 0$  mit  $|p(z) - p(\xi)| < a/2$  für alle  $z$  mit  $|z - \xi| < \delta$ . Daraus folgt aus dem Lemma  $d(K_{\delta/2}(\xi), p) = d(K_{\delta/2}(\xi), p(\xi)) = 0$ . Wir betrachten eine stetige Deformation, die den Kreis  $K_R(0)$  in  $K_{\delta/2}(\xi)$  überführt, beispielsweise kann man den Kreis zuerst auf den Radius  $\delta/2$  schrumpfen lassen und ihn anschließend verschieben. Die Drehungszahl hängt stetig von einer solchen Deformation ab, ist aber immer ganzzahlig. Da sie im Verlauf der Deformation von  $n$  auf 0 springt, kann sie nicht immer definiert sein. Dies ist aber nur dann der Fall, wenn  $p$  eine Nullstelle besitzt.

Man kann dieses Argument noch etwas verfeinern und erhält dann: Die Drehungszahl liefert die Zahl der Nullstellen im umschlossenen Bereich.

### 13.6 Taylorreihen

Wir nennen

$$f(x) = \sum_{n=0}^{\infty} a_n (x - a)^n$$

*Potenzreihe mit Entwicklungspunkt  $a$ .* Da es sich hier um nichts weiter als die bereits bekannte Potenzreihe handelt, die lediglich um  $a$  verschoben ist, gelten die Sätze 11.15 und 12.10 sinngemäß. Mit

$$L = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|}$$

ist die Reihe konvergent in

$$D = \{x \in \mathbb{R} : |x - a| < R = \frac{1}{L}\}.$$

$f$  kann in  $D$  unendlich oft gliedweise differenziert werden, insbesondere gilt

$$(13.9) \quad f^{(n)}(a) = n! a_n,$$

also

$$(13.10) \quad f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^n.$$

**Satz 13.12** Sei  $f(x) = \sum_{n=0}^{\infty} a_n (x - a)^n$  in einer Umgebung von  $a$  konvergent. Dann stimmt das Taylorpolynom  $T_n(x; a)$  mit dem  $n$ -ten Abschnitt der Reihe überein,

$$T_n(x; a) = \sum_{k=0}^n a_k (x - a)^k.$$

Ist umgekehrt  $f$  unendlich oft differenzierbar mit  $R_n(x; a) \rightarrow 0$  gleichmäßig in Umgebung von  $a$ , so lässt sich in dieser Umgebung  $f$  als Reihe (13.10) darstellen.

Der Beweis folgt unmittelbar aus (13.9) und (13.10).

Der Satz von Taylor gibt uns aufgrund des Restgliedes eine Fehlerabschätzung, wenn nur ein Reihenabschnitt ausgewertet werden soll. Als ein Beispiel wollen wir die Zahl  $e$  mit Hilfe von

$$e = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \frac{1}{120} + \frac{e^\xi}{6!} = 2.716\dots + \frac{e^\xi}{6!}, \quad \xi \in (0, 1),$$

angenähert bestimmen. Wegen  $e < 3$  gilt

$$\frac{e^\xi}{6!} \leq \frac{e^1}{6!} \leq \frac{3}{6!} = 0.00595\dots,$$

also  $|e - 2.716\dots| \leq 0.006$ , der genaue Wert ist  $e = 2.718\dots$

Den Abschluß bildet die Potenzreihe des Logarithmus.

**Satz 13.13** Für  $|x| < 1$  besitzt der Logarithmus die Reihendarstellung

$$\ln(1+x) = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} x^n = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$$

*Beweis:* Mit  $\ln'(1+x) = 1/(1+x)$  können wir die höheren Ableitungen leicht bestimmen

$$\ln^{(n)}(1+x) = \frac{(-1)^{n-1}(n-1)!}{(1+x)^n}.$$

Die angegebene Reihe errechnet sich damit aus (13.10) mit  $a = 0$ . Nach dem Wurzel- oder Quotientenkriterium ist die Reihe in der Tat für  $|x| < 1$  konvergent.  $\square$

### 13.7 Differenzengleichungen Am Beispiel der Fibonacci-Folge

$$F_n = F_{n-1} + F_{n-2}, \quad n \geq 3, \quad F_1 = F_2 = 1,$$

leiten wir ein allgemeines Verfahren für die explizite Bestimmung rekursiv definierter Folgen her. Dazu verwenden wir die Konvention, dass die Folge für alle  $n \in \mathbb{Z}$  definiert ist mit  $F_n = 0$  für  $n \leq 0$ . Ist  $a(n)$  eine Aussage, die für alle  $n \in \mathbb{Z}$  wahr oder falsch ist, so

$$[a(n)] = \begin{cases} 1 & \text{falls } a(n) \text{ wahr} \\ 0 & \text{falls } a(n) \text{ falsch} \end{cases}.$$

Die Fibonacci-Folge wird daher in der Form

$$F_n = F_{n-1} + F_{n-2} + [n = 1], \quad n \in \mathbb{Z},$$

geschrieben. Mit  $F_n = 0$  für  $n \leq 0$  folgt dann  $F_1 = 1$  und  $F_2 = 1$ , wir erhalten also die Fibonacci-Folge zurück.

Im ersten Schritt des Verfahrens ordnen wir der Folge eine Potenzreihe

$$(13.11) \quad F(z) = \sum F_n z^n$$

zu, wobei die Summe sich hier wie im Folgenden über alle  $n \in \mathbb{Z}$  erstreckt. Aus der Definitionsgleichung erhalten wir

$$F(z) = \sum F_{n-1} z^n + \sum F_{n-2} z^n + \sum [n = 1] z^n = zF(z) + z^2 F(z) + z,$$

daher

$$(13.12) \quad F(z) = \frac{z}{1 - z - z^2}.$$

Man beachte, dass all diese Operationen rein formal, also ohne Konvergenzbetrachtungen durchgeführt werden. In dieser Hinsicht stellt das beschriebene Verfahren gar keine Anwendung der Analysis dar. Hier wie in den meisten Fällen lassen sich die Umformungen allerdings auch analytisch deuten: Wegen  $F^n \leq 2^n$  hat die Reihe (13.11) einen positiven Konvergenzradius und ist die Taylorreihe der rationalen Funktion in (13.12).

$F(z)$  in (13.12) heißt *erzeugende Funktion* der Potenzreihe (13.11).

Im zweiten Schritt wird für (13.12) eine Partialbruchzerlegung in etwas modifizierter Form durchgeführt. Wir verwenden dazu das folgende Lemma.

**Lemma 13.14** Das dem Polynom

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_1 z + 1, \quad a_n \neq 0,$$

zugeordnete reflektierte Polynom

$$p^R(z) = a_n + a_{n-1} z + \dots + a_1 z^{n-1} + z^n$$

besitze die  $n$  Nullstellen  $\alpha_1, \dots, \alpha_n$ . Dann gilt

$$p(z) = (1 - \alpha_1 z)(1 - \alpha_2 z) \dots (1 - \alpha_n z).$$

*Beweis:* Es gilt  $p(z) = z^n p^R(\frac{1}{z})$ . Durch Ausmultiplizieren erhalten wir

$$p(z) = z^n \left( \frac{1}{z} - \alpha_1 \right) \dots \left( \frac{1}{z} - \alpha_n \right) = (1 - \alpha_1 z) \dots (1 - \alpha_n z).$$

□

Wir bestimmen die Partialbruchzerlegung von (13.12) in der Form

$$(13.13) \quad \frac{z}{1 - z - z^2} = \frac{a}{1 - \phi z} + \frac{b}{1 - \hat{\phi} z},$$

wobei nach dem Lemma  $\phi$  und  $\hat{\phi}$  die Nullstellen des reflektierten Polynoms  $p^R(z) = z^2 - z - 1$  zu  $p(z) = 1 - z - z^2$  sind. Für diese erhalten wir

$$\phi = \frac{1 + \sqrt{5}}{2}, \quad \hat{\phi} = \frac{1 - \sqrt{5}}{2}.$$

Die Koeffizienten in (13.13) werden völlig analog zur üblichen Partialbruchzerlegung bestimmt. Wir multiplizieren mit  $1 - \phi z$  und werten an der Stelle  $z = 1/\phi$  aus

$$a = \frac{z}{1 - \hat{\phi} z} \Big|_{z=1/\phi} = \frac{1}{\phi - \hat{\phi}} = \frac{1}{\sqrt{5}}.$$

Für  $b$  erhalten wir analog  $b = -1/\sqrt{5}$ , insgesamt

$$(13.14) \quad \sum F_n z^n = \frac{z}{1 - z - z^2} = \frac{1}{\sqrt{5}(1 - \phi z)} - \frac{1}{\sqrt{5}(1 - \hat{\phi} z)}.$$

Aus der geometrischen Reihe folgt für  $\alpha \neq 0$

$$\frac{1}{1 - \alpha z} = \sum_{n=0}^{\infty} \alpha^n z^n$$

und daher mit Koeffizientenvergleich in (13.14)

$$F_n = \frac{\phi^n - \hat{\phi}^n}{\sqrt{5}}.$$

Dies ist die *Binetsche Darstellung* der Fibonacci-Zahlen, die aber bereits früher von L. Euler angegeben wurde.

Es dürfte klar sein, das mit dem hier beschriebenen Verfahren auch allgemeine Rekursionen angegangen werden können. Ist  $(f_n)$  eine beliebige Folge, so nennen wir

$$f(z) = \sum f_n z^n$$

die *zugeordnete Potenzreihe*. Eine elementare Funktion mit Reihendarstellung  $f$  heißt *erzeugende Funktion*. Im Falle einfacher Rekursionsgleichungen wie der Fibonacci-Folge wird die erzeugende Funktion rational sein und wir können Partialbruchzerlegung verwenden. Treten im Nennerpolynom mehrfache Nullstellen auf, so entstehen Terme der Form

$$(13.15) \quad \frac{1}{(1-\alpha z)^m} = \sum_{n=0}^{\infty} \binom{m+n-1}{n} \alpha^n z^n,$$

was man mit Induktion über  $m$  beweist.

**Beispiel 13.15** Wir bestimmen die Lösung der Rekursion

$$f_n = f_{n-1} + 2f_{n-2} + (-1)^n, \quad n \geq 2, \quad f_0 = f_1 = 1.$$

In diesem Fall ist bereits  $f_0 = 1$ , was nichts ausmacht, wenn wir die Rekursion so umschreiben, dass sie für alle  $n \in \mathbb{Z}$  richtig ist. Wir setzen

$$f_n = f_{n-1} + 2f_{n-2} + (-1)^n [n \geq 0] + [n = 1].$$

und aus  $f_{-2} = f_{-1} = 0$  folgt  $f_0 = 1, f_1 = 1$ . Damit wird die Folge für alle  $n \in \mathbb{Z}$  korrekt dargestellt. Für die zugeordnete Reihe gilt

$$\begin{aligned} f(z) &= \sum f_n z^n = \sum f_{n-1} z^n + 2 \sum f_{n-2} z^n + \sum (-1)^n z^n [n \geq 0] + \sum z^n [n = 1] \\ &= zf(z) + 2z^2 f(z) + \frac{1}{1+z} + z, \end{aligned}$$

daher

$$f(z) = \frac{1+z(1+z)}{(1+z)(1-z-2z^2)} = \frac{1+z+z^2}{(1-2z)(1+z)^2}.$$

In

$$f(z) = \frac{a}{1-2z} + \frac{b}{1+z} + \frac{c}{(1+z)^2}$$

erhalten wir

$$a = \frac{1 + \frac{1}{2} + \frac{1}{4}}{\frac{9}{4}} = \frac{7}{9}, \quad c = \frac{1 - 1 + 1}{1+2} = \frac{1}{3}.$$

Durch Einsetzen von beispielsweise  $z = 0$  folgt  $b = -\frac{1}{9}$  und mit (13.15)

$$f_n = \frac{7}{9} 2^n - \frac{1}{9} (-1)^n + \frac{1}{3} (n+1)(-1)^n = \frac{7}{9} 2^n + \left(\frac{1}{3} n + \frac{2}{9}\right) (-1)^n.$$

□

Wir betrachten nun einen Spezialfall, nämlich die *homogene lineare Differenzengleichung der Ordnung  $k$*

$$(13.16) \quad f_n = a_{k-1} f_{n-1} + \dots + a_1 f_{n-k+1} + a_0 f_{n-k}.$$

In diesem Fall benötigen wir  $k$  Anfangswerte, beispielsweise die Kenntnis von  $f_0, \dots, f_{k-1}$ , um die Rekursion starten zu können. Wir setzen  $a_0, \dots, a_{k-1} \in \mathbb{C}$  voraus und betrachten die Lösungsmenge ebenfalls in  $\mathbb{C}$ . Die Menge der Lösungen von (13.16) bilden dann einen linearen Vektorraum über  $\mathbb{C}$ , denn wenn wir zwei Lösungen addieren oder eine Lösung mit einer komplexen Zahl multiplizieren, erhalten wir ebenfalls eine Lösung. Da wir  $f_0, \dots, f_{k-1}$  frei wählen können und bei jeder Wahl genau eine Lösung bekommen, ist die Dimension dieses Vektorraums genau  $k$ . Um eine Basis des Lösungsraums zu konstruieren, untersuchen wir den Ansatz  $f_n = \alpha^n$  mit einer komplexen Zahl  $\alpha$ .

Setzen wir dies ein und teilen durch  $\alpha^{n-k}$ , so erfüllt  $f_n$  genau dann die Rekursionsgleichung, wenn  $\alpha$  Nullstelle des *charakteristischen Polynoms*

$$q(\alpha) = \alpha^k - a_{k-1}\alpha^{k-1} - \dots - a_1\alpha - a_0$$

ist.  $q$  stimmt mit dem zuvor als  $p^R$  bezeichneten Polynom überein. Sind die Nullstellen  $\alpha_1, \dots, \alpha_k$  von  $q$  alle verschieden, so ist die allgemeine Lösung der Rekursion (13.16)

$$(13.17) \quad f_n = c_1\alpha_1^n + \dots + c_k\alpha_k^n, \quad c_j \in \mathbb{C}.$$

Die Konstanten  $c_1, \dots, c_k$  werden aus den Anfangsbedingungen für die  $f_n$  bestimmt.

Ist  $\alpha_j$  eine  $r$ -fache Nullstelle des charakteristischen Polynoms mit  $r > 1$ , so lässt sich leicht nachrechnen, dass  $f_n = c_{j,1}\alpha_j^n, c_{j,2}n\alpha_j^n, \dots, c_{j,r}n^{r-1}\alpha_j^n$  ebenfalls Lösungen der Rekursion sind. Da ein Polynom vom Grade  $k$  genau  $k$  Nullstellen besitzt, haben wir also immer  $k$  partikuläre Lösungen, deren Konstanten aus den  $k$  Anfangsbedingungen bestimmt werden können.

**Beispiele 13.16** (i) Für die Fibonacci-Folge ist das charakteristische Polynom  $p(\alpha) = \alpha^2 - \alpha - 1$  mit den bekannten Nullstellen  $\alpha_1 = \phi, \alpha_2 = \hat{\phi}$ . In der allgemeinen Lösung

$$f_n = c_1\phi^n + c_2\hat{\phi}^n$$

liefern die Anfangsbedingungen  $f_0 = 0, f_1 = 1$  das lineare Gleichungssystem

$$c_1 + c_2 = 0, \quad c_1\phi + c_2\hat{\phi} = 1,$$

mit Lösung  $c_1 = -c_2 = 1/\sqrt{5}$ .

(ii) Das oben behandelte Beispiel

$$f_n = f_{n-1} + 2f_{n-2} + (-1)^n, \quad n \geq 2, \quad f_0 = f_1 = 1.$$

ist nicht von der Form (13.16). Wir können die Rekursion erneut auf  $f_{n-1}$  anwenden und erhalten

$$f_n = (f_{n-2} + 2f_{n-3} + (-1)^{n-1}) + 2f_{n-2} + (-1)^n = 3f_{n-2} + 2f_{n-3}.$$

Das charakteristische Polynom

$$p(\alpha) = \alpha^3 - 3\alpha - 2 = (\alpha - 2)(\alpha + 1)^2$$

hat für  $\alpha_2 = -1$  eine doppelte Nullstelle. Das führt auf die allgemeine Lösung

$$f_n = c_12^n + c_2(-1)^n + c_3n(-1)^n.$$

Aus der Anfangsbedingung  $f_0 = f_1 = 1, f_2 = 4$ , erhalten wir wie oben  $c_1 = 7/9, c_2 = 2/9, c_3 = 1/3$ .

□

## 14 Schnelle Fourier-Transformation

**14.1 Die Lagrangesche Interpolationsaufgabe** Mit  $\mathbb{P}_n$  bezeichnen wir den Raum der komplexen Polynome  $p : \mathbb{C} \rightarrow \mathbb{C}$  vom Grad  $\leq n$ . Gegeben seien  $n+1$  verschiedene Stützstellen  $x_j \in \mathbb{C}$ ,  $j = 0, \dots, n$ , und  $n+1$  nicht notwendig verschiedene Werte  $y_0, \dots, y_n \in \mathbb{C}$ . In der *Lagrangeschen Interpolationsaufgabe* ist ein Polynom  $p \in \mathbb{P}_n$  gesucht mit

$$(14.1) \quad p(x_j) = y_j, \quad j = 0, 1, \dots, n.$$

Die  $y_j$  können wir uns als Werte  $y_j = f(x_j)$  einer vorgegebenen Funktion  $f$  vorstellen. Wir sagen dann, dass  $p$  die Funktion  $f$  *interpoliert*.

Die Dimension von  $\mathbb{P}_n$  ist  $n+1$ . Wir haben daher in der Lagrangeschen Interpolationsaufgabe  $n+1$  Bedingungen gestellt, aber auch  $n+1$  Freiheiten zur Verfügung. Zur Lösung des Interpolationsproblems definieren wir die *Lagrange-Basis*  $\{l_j\}_{j=0, \dots, n}$ ,  $l_j \in \mathbb{P}_n$ , durch

$$(14.2) \quad l_j(x) = \frac{(x - x_0) \dots (x - x_{j-1})(x - x_{j+1}) \dots (x - x_n)}{(x_j - x_0) \dots (x_j - x_{j-1})(x_j - x_{j+1}) \dots (x_j - x_n)}.$$

Es gilt dann  $l_i(x_j) = \delta_{ij}$  und die Interpolationsaufgabe (14.1) wird gelöst durch

$$(14.3) \quad p(x) = \sum_{j=0}^n y_j l_j(x) \in \mathbb{P}_n.$$

**Satz 14.1** Die Interpolationsaufgabe (14.1) wird eindeutig gelöst durch das Lagrangesche Interpolationspolynom (14.3).

*Beweis:* Gäbe es zwei Lösungen  $p_1, p_2 \in \mathbb{P}_n$  von  $p(x_j) = y_j$ , so gilt für  $q = p_1 - p_2 \in \mathbb{P}_n$ , dass  $q(x_j) = 0$ . Damit hat  $q$   $n+1$  Nullstellen und muss das Nullpolynom sein. Daher ist  $p_1 = p_2$ .  $\square$

Ist  $y_j = f(x_j)$  für ein Polynom  $f \in \mathbb{P}_n$ , so gilt für das Interpolationspolynom  $p = f$ , weil die Interpolationsaufgabe eindeutig lösbar ist. Polynome vom Grad  $\leq n$  werden also in der Interpolation reproduziert.

**14.2 Schnelle Polynommultiplikation** Sind  $p, q \in \mathbb{P}_{n-1}$ , so gilt mit

$$p(x) = a_{n-1}x^{n-1} + \dots + a_1x + a_0, \quad q(x) = b_{n-1}x^{n-1} + \dots + b_1x + b_0,$$

für das Produkt

$$r(x) = p(x)q(x) = \sum_{l=0}^{2n-2} c_l x^l, \quad c_l = \sum_{j+k=l} a_j b_k, \quad l = 0, \dots, 2n-2,$$

wobei die Konvention  $a_j, b_j = 0$  für  $j > n-1$  verwendet wurde. Da jeder Koeffizient des einen Polynoms mit jedem Koeffizienten des anderen Polynoms multipliziert wird, benötigen wir für die direkte Polynommultiplikation genau  $n^2$  Multiplikationen und  $O(n^2)$  Additionen.

Eine Alternative zur direkten Multiplikation ist die im vorigen Abschnitt besprochene Polynominterpolation nach Lagrange. Demnach ist ein Polynom vom Grade  $n-1$  durch die Werte  $p(x_j)$  an  $n$  verschiedenen Stützstellen  $x_1, \dots, x_n$  eindeutig bestimmt. Da das Produkt  $r(x) = p(x)q(x)$  im Raum  $\mathbb{P}_{2n-2}$  liegt, genügt es, die Polynome  $p$  und  $q$  an  $2n-1$  verschiedenen Stellen auszuwerten, die  $2n-1$  Produkte  $p(x_j)q(x_j)$  zu berechnen und anschließend das Interpolationspolynom zu diesen Produkten zu bestimmen. Da das Interpolationspolynom eindeutig bestimmt ist, wird durch diese Vorgehensweise das Produktpolynom reproduziert. Dieses Verfahren ist zunächst weniger effektiv als die direkte Methode, denn selbst wenn das Interpolationspolynom nach der raffiniertesten Methode bestimmt wird, werden dazu  $O(n^2)$  Operationen benötigt.

Die Idee der schnellen Polynommultiplikation besteht in einer vorteilhaften Wahl der Stützstellen  $x_j$ , die eine simultane Auswertung von  $p(x_j)$  gestattet.

Um ein Polynom an  $n$  Stellen auszuwerten, verwenden wir die komplexen Einheitswurzeln

$$\omega_n = \exp\left(\frac{2\pi i}{n}\right), \quad \omega_n^k = \exp\left(\frac{2k\pi i}{n}\right), \quad k = 0, \dots, n-1.$$

Für diese gilt

$$(14.4) \quad \omega_n^k = \omega_n^{k+ln} \quad \forall l \in \mathbb{Z}$$

wegen  $\exp(z) = \exp(z + 2\pi i)$ .

Sei nun  $n$  gerade und  $p$  ein Polynom vom Grade  $\leq n-1$ . Wir schreiben

$$\begin{aligned} p(x) &= a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \dots + a_2x^2 + a_1x + a_0 \\ &= (a_{n-2}x^{n-2} + \dots + a_2x^2 + a_0) + x(a_{n-1}x^{n-2} + \dots + a_3x^2 + a_1) \\ &= p_g(x^2) + xp_u(x^2). \end{aligned}$$

Um  $p(\omega_n^k)$ ,  $k = 0, \dots, n-1$  zu bestimmen, müssen die Polynome  $p_g$  und  $p_u$ , die nur noch den Grad  $n/2 - 1$  besitzen, an den Stellen  $\omega_n^{2k} = \omega_{n/2}^k$  ausgewertet werden, mit (14.4) sind dies  $\omega_{n/2}^0, \dots, \omega_{n/2}^{n/2-1}$ . Ferner lassen sich die Auswertungen von  $p_g$  und  $p_u$  zweimal verwenden, nämlich für  $k = 0, \dots, n/2 - 1$  und für  $k + n/2$  wegen  $\omega_n^{2k} = \omega_n^{2k+2n/2}$ .

Da das beschriebene Verfahren offenbar rekursiv durchgeführt werden kann, wenn  $n$  eine Zweierpotenz ist, nehmen wir nun  $n = 2^l$  an. Das Programm

**procedure**  $FFT(n, p, \omega, a)$

bestimmt die Auswertung eines Polynoms vom Grade  $n-1$  in den  $n$  Punkten  $\omega_n^0, \dots, \omega_n^{n-1}$ . Die Zweierpotenz  $n$  gibt die Länge des Problems an, in der Inputvariablen  $p$  ist das auszuwertende Polynom als  $n$ -Vektor  $p = (a_0, \dots, a_{n-1})$  gespeichert. Ferner haben wir die Inputvariable  $\omega = \omega_n$ . Das Ergebnis ist der  $n$ -Vektor  $a = (p(\omega^0), \dots, p(\omega^{n-1}))$ . Die schnelle Polynomauswertung kann damit folgendermaßen implementiert werden:

```

recursive subroutine  $FFT(n, p, \omega, a)$ 
  if  $n = 1$  then
     $a(0) = p(0)$ 
  else
     $n_2 = n/2$ 
     $p_g = (a_0, a_2, \dots, a_{n-2})$ 
     $p_u = (a_1, a_3, \dots, a_{n-1})$ 
    call  $FFT(n_2, p_g, \omega^2, g)$ 
    call  $FFT(n_2, p_u, \omega^2, u)$ 
    do  $k = 0, n_2 - 1$ 
       $a(k) = g(k) + \omega^k u(k)$ 
       $a(k + n_2) = g(k) - \omega^k u(k)$ 
    enddo
  endif
end

```

Zur Einsparung von Rechenoperationen wurde von der Beziehung  $\omega_n^{k+n/2} = \omega_n^k \omega_n^{n/2} = -\omega_n^k$  Gebrauch gemacht.

Nun bestimmen wir die Anzahl der Multiplikationen  $M(n)$  für diesen Algorithmus. Da die Zahlen  $\omega^k$  einmal berechnet und anschließend abgespeichert werden können, fallen in der Do-Schleife nur die  $n/2$  Multiplikationen  $\omega^k u(k)$  an. Da  $FFT$  zweimal mit Inputlänge  $n/2$  aufgerufen wird, genügt  $M(n)$  der Rekursion

$$M(n) = 2M\left(\frac{n}{2}\right) + \frac{n}{2}$$

für  $n = 2^l$  daher

$$M(2^l) = 2M(2^{l-1}) + 2^{l-1}.$$

Iterieren wir diese Beziehung  $l$  mal, so erhalten wir

$$M(2^l) = l2^{l-1} + M(1)2^l.$$

Da im Then-Teil des obigen Programms keine Operationen anfallen, ist  $M(1) = 0$  und daher

$$M(n) = \frac{n}{2} \log n,$$

wobei mit  $\log$  der Logarithmus zur Basis 2 bezeichnet wird. Da für die Anzahl der Addditionen eine analoge rekursive Beziehung gilt, erhalten wir für die Gesamtzahl an Operationen ebenfalls  $O(n \log n)$ .

Nun wenden wir uns dem Interpolationsproblem zu: Zu Daten  $b_0, \dots, b_{n-1}$  ist ein Polynom  $p \in \mathbb{P}_{n-1}$  gesucht mit  $p(\omega_n^k) = b_k$  für  $k = 0, \dots, n-1$ . Wie bereits zu Anfang dieses Abschnitts gezeigt wurde, existiert ein solches Polynom und ist eindeutig bestimmt. Um dieses Problem anzugehen, deuten wir zunächst die Polynomauswertung als Multiplikation einer Matrix mit einem Vektor. Zu  $\alpha = (\alpha_0, \dots, \alpha_{n-1})^T \in \mathbb{C}^n$  mit paarweise verschiedenen  $\alpha_k$  definieren wir die zugehörige *Vandermondsche Matrix* als

$$V(\alpha) = \begin{pmatrix} 1 & \alpha_0 & \alpha_0^2 & \cdots & \alpha_0^{n-1} \\ 1 & \alpha_1 & \alpha_1^2 & \cdots & \alpha_1^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_{n-1} & \alpha_{n-1}^2 & \cdots & \alpha_{n-1}^{n-1} \end{pmatrix}.$$

Für ein Polynom

$$p(x) = \sum_{j=0}^{n-1} a_j x^j$$

gilt dann

$$(14.5) \quad b := (p(\alpha_0), \dots, p(\alpha_{n-1}))^T = V(\alpha)a, \quad a = (a_0, \dots, a_{n-1})^T.$$

Die Auswertung von  $p$  an den Stellen  $\alpha_0, \dots, \alpha_{n-1}$  ist also nichts anderes als die Bestimmung von  $V(\alpha)a$ . Die Rekonstruktion der Koeffizienten  $a_0, \dots, a_{n-1}$  aus den Daten  $b_k = p(\alpha_k)$  (=Interpolation) ist das dazu inverse Problem und wird durch

$$a = V(\alpha)^{-1}b$$

gelöst. Bei der speziellen Wahl  $\alpha = (\omega_n^0, \dots, \omega_n^{n-1})^T$  kann die inverse Matrix zu  $V(\alpha)$  leicht angegeben werden. Für eine komplexe Zahl  $\beta$  schreiben wir  $[\beta] = (\beta^0, \dots, \beta^{n-1})^T$ .

**Satz 14.2** Für die  $n$ -te Einheitswurzel  $\omega_n = \exp(2\pi i/n)$  gilt

$$V([\omega_n])^{-1} = \frac{1}{n} V([\omega_n^{-1}]).$$

*Beweis:* Für die Matrix  $W = V([\omega_n])V([\omega_n^{-1}])$  gilt

$$w_{jk} = \sum_{l=0}^{n-1} \omega_n^{jl} \omega_n^{-kl} = \sum_{l=0}^{n-1} (\omega_n^{j-k})^l.$$

Für  $j = k$  erhalten wir  $w_{jj} = n$ . Für  $j \neq k$  ist  $0 < |j - k| < n$  und damit  $\omega_n^{j-k} \neq 1$ . Wir können daher die geometrische Summenformel anwenden, also

$$(14.6) \quad \sum_{l=0}^{n-1} (\omega_n^{j-k})^l = \frac{\omega_n^{n(j-k)} - 1}{\omega_n^{j-k} - 1} = 0$$

wegen  $\omega_n^{n(j-k)} = (\omega_n^n)^{j-k} = 1$ .  $\square$

Der Algorithmus zur Bestimmung des Interpolationspolynoms  $p \in \mathbb{P}_{n-1}$  aus den Daten  $b := (p(1), p(\omega_n^1), \dots, p(\omega_n^{n-1}))$ ,

**subroutine**  $FFI(n, b, \omega, p)$ ,

hat als Input die Länge  $n = 2^l$ , den  $n$ -Vektor  $b$  und  $\omega = \omega_n$ . Das gesuchte Polynom wird auf dem  $n$ -Vektor  $p = (a_0, \dots, a_{n-1})$  ausgegeben.

```

subroutine  $FFI(n, b, \omega, p)$ 
  call  $FFT(n, b, \omega^{-1}, p)$ 
   $p = n^{-1}p$ 
end
```

Der Algorithmus zur schnellen Multiplikation von Polynomen  $p, q \in \mathbb{P}_k$  ergibt sich nun fast von selbst. Wir bestimmen die kleinste Zahl  $l$  mit  $n = 2^l > 2k$  und rufen mit  $\omega = \omega_n$

$FFT(n, p, \omega, a), \quad FFT(n, q, \omega, b)$

auf. Anschließend berechnen wir  $c(k) = a(k)b(k)$  für  $k = 0, \dots, n-1$ , was  $n$  Multiplikationen entspricht. Mit dem Aufruf von

$FFI(n, c, \omega, r)$

stehen auf dem Vektor  $r$  die Koeffizienten des gesuchten Produkts  $p(x)q(x)$ . Der gesamte Algorithmus benötigt immer noch  $O(n \log n)$  und wegen  $n \leq 4k$  auch  $O(k \log k)$  Rechenoperationen.

In der vorgestellten Form ist der Algorithmus aber noch problematisch. Zunächst bringt der Algorithmus nur dann einen beträchtlichen Gewinn gegenüber dem Standardverfahren, wenn  $\text{grad } p$  und  $\text{grad } q$  von gleicher Größenordnung sind. Beispielsweise hat das Standardverfahren im Extremfall  $\text{grad } p = 1$  und  $\text{grad } q = n$  die Komplexität  $O(n)$ , während die schnelle Polynommultiplikation immer noch von der Ordnung  $O(n \ln n)$  ist. Dies wird man sicherlich als weniger gravierend ansehen. Bedeutsamer ist dagegen die implizit verwendete Definition des Wortes „Operation“. Auch wenn die Koeffizienten der beteiligten Polynome ganzzahlig sind, werden im Algorithmus nicht rationale komplexe Zahlen benötigt, die auf einem Rechner gar nicht exakt dargestellt werden können. Statt dessen müssen die Operationen mit einer Gleitkommaarithmetik näherungsweise bestimmt werden und es dürfte klar sein, dass man für großes  $n$  auch eine umfangreiche Gleitkommaarithmetik benötigt.

**14.3 Schnelle Multiplikation natürlicher Zahlen** Liegen natürliche Zahlen  $a, b$  in einem Stellenwertsystem zur Basis  $g$  vor und besitzen diese Zahlen eine Länge  $\leq K$ , so lassen sie sich mit  $O(K^2)$  Operationen multiplizieren, wobei eine Operation aus der Multiplikation zweier Ziffern oder der Addition zweier ganzer Zahlen besteht. Alternativ können wir  $a$  und  $b$  die Polynome

$$p_a(x) = \sum_{i=0}^{K-1} a_i x^i, \quad p_b(x) = \sum_{i=0}^{K-1} b_i x^i$$

zuordnen, wobei  $a_i, b_i \in \{0, 1, \dots, g - 1\}$  die zugehörigen Ziffern sind. Mit der schnellen Polynommultiplikation kann  $p_c(x) = p_a(x)p_b(x)$  bestimmt werden.  $p_c(g)$  ist dann das gesuchte Produkt der beiden Zahlen, aus dem noch der Übertrag entfernt werden muss. Mit dem im vorigen Abschnitt dargestellten Multiplikationsverfahren erhalten wir aufgrund von Rundungsfehlern ein Polynom  $\tilde{p}_c$ , dessen Koeffizienten Gleitkommazahlen sind. Die verwendete Gleitkommaarithmetik muss so ausgelegt sein, dass aus  $\tilde{p}_c$  durch Rundung das korrekte Polynom  $p_c$  entsteht.

Da die Rundungsfehleranalyse der diskreten Fouriertransformation aufgrund ihrer rekursiven Struktur recht aufwendig ist, soll hier nur das Endergebnis referiert werden. Wir betrachten den Spezialfall  $K = 2^k$  und  $g = 2^l$ , mit dem Zahlen der Bitlänge  $n \leq \frac{1}{2}Kl$  miteinander multipliziert werden können. Verwenden wir für die schnelle Fouriertransformation eine Gleitkommaarithmetik der Genauigkeit  $2^{-m}$ , so erhält man nach Rundung von  $\tilde{p}_c$  das exakte Polynom  $p_c$ , falls

$$m \geq 3k + 2l + \log k + 7/2$$

erfüllt ist. Für  $k \geq 7$  erhalten wir hieraus die bequemere Abschätzung

$$m \geq 4k + 2l.$$

Um Zahlen der Bitlänge  $n = 2^{13} = 8192$  miteinander zu multiplizieren, können wir hier  $l = 8$ ,  $k = 11$  setzen und erhalten  $m = 54$ , was von einer doppelt genauen Gleitkommaarithmetik geleistet wird. Für die meisten praktisch relevanten Fälle lassen sich damit zwei  $n$ -stellige Zahlen in  $O(n \log n)$  Gleitkommaoperationen miteinander multiplizieren. Für die Multiplikation noch größerer Zahlen kann auch die Gleitkommamultiplikation mit Hilfe der schnellen Fouriertransformation beschleunigt werden. Für den Gesamtalgorithmus haben wir dann eine Komplexität von  $O(n \log n \log(\log n))$ , die allerdings auch nur bis zu einer sehr großen Zahl  $n$  richtig ist.

In der *modalen Fouriertransformation* von Schönhage und Strassen (1971) werden kommutative unitäre Ringe mit  $n$ -ter Einheitswurzel verwendet, die letztlich aus ganzen Zahlen bestehen. Damit kann man zwei  $n$ -stellige Zahlen in  $O(n \log n \log(\log n))$  miteinander multiplizieren, in diesem Fall ohne Einschränkung an  $n$ .

## 15 Diskrete Wahrscheinlichkeitsräume

**15.1 Grundbegriffe**  $\Omega$  sei eine endliche oder abzählbar unendliche Menge. Eine Abbildung  $P : \Omega \rightarrow [0, 1]$  heißt *Wahrscheinlichkeitsmaß*, wenn  $\sum_{\omega \in \Omega} P(\omega) = 1$ . In diesem Fall wird  $(\Omega, P)$  als *Wahrscheinlichkeitsraum* bezeichnet, der aus *elementaren Ereignissen*  $\omega \in \Omega$  besteht.

Anschaulich besteht ein Wahrscheinlichkeitsraum aus den möglichen Ausgängen eines wiederholbaren Zufallsexperiments, wie dem Werfen eines Würfels ( $\Omega = \{1, \dots, 6\}$ ), dem Ziehen einer Karte aus einem Kartenspiel ( $\Omega = \{1, \dots, 52\}$ ), oder dem Werfen einer Münze ( $\Omega = \{K, Z\}$ ).  $P(\omega)$  gibt dann die relative Häufigkeit an, mit der  $\omega$  eintritt.  $P(\omega) = 1$  bedeutet demnach, dass  $\omega$  immer eintritt, wegen  $\sum P(\omega) = 1$  treten dann alle anderen elementaren Ereignisse niemals ein. In den oben genannten Beispielen kann man davon ausgehen, dass alle elementaren Ereignisse *gleichverteilt* sind und demnach  $P(\omega) = 1/|\Omega|$  gilt, wobei  $|\Omega|$  die Anzahl der Elemente von  $\Omega$  bezeichnet.

$A \subset \Omega$  heißt *Ereignis*, die *Wahrscheinlichkeit* von  $A$  ist dann definiert durch

$$P(A) = \sum_{\omega \in A} P(\omega).$$

Aus dieser Definition folgt für  $A, B, A_i \subset \Omega$

$$P(A^c) = 1 - P(A), \quad \text{insbesondere } P(\emptyset) = 0, \quad P(\Omega) = 1$$

$$A \subset B \Rightarrow P(A) \leq P(B)$$

$$P(A \setminus B) = P(A) - P(A \cap B)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i) \quad \text{mit Gleichheit bei disjunkten } A_i$$

Im Fall einer Gleichverteilung bedeutet die Definition

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{„Zahl der günstigen Fälle“}}{\text{„Zahl der möglichen Fälle“}}$$

Wir behandeln nun die vier grundsätzlichen Abzählprobleme bei Gleichverteilung. In einer Urne liegen  $N$  Kugeln, die wir uns von 1 bis  $N$  nummeriert denken. Es werden  $n$  Kugeln sukzessive gezogen, was *Stichprobe* genannt wird. Wie viele solcher Stichproben es gibt, hängt von der Art der Ziehung ab und davon, ob die Reihenfolge der gezogenen Kugeln berücksichtigt wird.

**I Stichproben in Reihenfolge mit Rücklegen** Nach jedem Ziehen wird die Kugel wieder zurückgelegt, unterschiedliche Reihenfolgen werden mitgezählt. Dies ist äquivalent dazu, aus  $n$  Urnen jeweils eine Kugel zu ziehen. Die Zahl der Möglichkeiten ist demnach  $N^n$ .

**II Stichproben in Reihenfolge ohne Rücklegen** Für die erste Kugel gibt es  $N$  Möglichkeiten, für die nächste  $N - 1$ . Für  $n \leq N$  gibt es daher  $N(N - 1) \dots (N - n + 1)$  Möglichkeiten, für  $n > N$  keine.

**III Stichproben ohne Reihenfolge ohne Rücklegen** In diesem Fall stimmt die Zahl der Möglichkeiten mit der Anzahl der  $n$ -elementigen Teilmengen von  $\{1, \dots, N\}$  überein, das sind  $\binom{N}{n}$ .

**IV Stichproben ohne Reihenfolge mit Rücklegen** Wir normieren die Stichprobe, indem wir die gezogenen Kugeln nach Größe ordnen. Demnach ist die Zahl der Elemente der Menge

$$M = \{(\omega_1, \omega_2, \dots, \omega_n) : \omega_1 \leq \omega_2 \leq \dots \leq \omega_n, \omega_i \in \{1, \dots, N\}\}$$

zu bestimmen. Wir addieren die  $i$ -te Komponente eines Vektors in  $M$  mit  $i - 1$  und erhalten so einen Vektor der Menge

$$M' = \{(\omega'_1, \omega'_2, \dots, \omega'_n) : \omega'_1 < \omega'_2 < \dots < \omega'_n, \omega_i \in \{1, \dots, N + n - 1\}\}$$

Da aber auch umgekehrt aus jedem Vektor in  $M'$ , wenn man von seiner  $i$ -ten Komponente  $i - 1$  abzieht, ein Element von  $M$  entsteht, sind beide Mengen gleich groß. Damit gibt es  $\binom{N+n-1}{n}$  Stichproben ohne Reihenfolge mit Rücklegen.

**Beispiele 15.1** (i) Es werden vier nicht unterscheidbare Würfel gleichzeitig geworfen. Wie groß ist die Wahrscheinlichkeit  $p$ , dass die vier erscheinenden Augenzahlen verschieden sind?

Der zugrunde liegende Wahrscheinlichkeitsraum ist die Zahl aller möglichen Würfe, also  $6^4$ . Um die Zahl der günstigen Fälle zu bestimmen, macht man sich klar, dass es für den ersten Würfel 6 Möglichkeiten gibt, für den zweiten 5 usw. Es liegt also Typ II vor und wir erhalten

$$p = \frac{6 \cdot 5 \cdot 4 \cdot 3}{6^4} = \frac{5}{18}.$$

(ii) Wie groß ist die Wahrscheinlichkeit  $p$ , beim Zahlenlotto „6 aus 49“ genau vier Richtige zu haben?

Da die Reihenfolge der gezogenen Zahlen keine Rolle spielt, ist der Wahrscheinlichkeitsraum vom Typ III, seine Kardinalität daher  $\binom{49}{6}$ . Man hat genau vier richtige Zahlen, wenn vier Zahlen in der Menge der sechs gezogenen Zahlen liegen und zwei Zahlen außerhalb. Da es in beiden Fällen nicht auf die Reihenfolge ankommt, liegt wieder Typ III vor und wir erhalten

$$p = \frac{\binom{6}{4} \binom{43}{2}}{\binom{49}{6}} = 0,00096\dots$$

□

Das letzte Beispiel ist ein Spezialfall der *hypergeometrischen Verteilung*: Aus einer Urne mit  $S$  schwarzen und  $W$  weißen Kugeln werden  $n \leq N = S + W$  Kugeln ohne Rücklegen gezogen. Die Wahrscheinlichkeit dafür, dass diese Stichprobe aus genau  $s$  schwarzen und  $w = n - s$  weißen Kugeln besteht, beträgt dann

$$h(s; n, N, S) = \binom{S}{s} \binom{W}{w} / \binom{S+W}{n}.$$

Um dies einzusehen, argumentieren wir genauso wie im letzten Beispiel. Die Zahl der möglichen Stichproben (Typ III) ist  $\binom{S+W}{n}$ . Da wir  $s$  schwarze Kugeln erhalten wollen, haben wir dazu  $\binom{S}{s}$  Möglichkeiten. Da die zu ziehenden weißen Kugeln von den gezogenen schwarzen Kugeln unabhängig sind, muss diese Zahl mit der Zahl der Möglichkeiten,  $n - s$  weiße Kugeln zu ziehen, multipliziert werden.

Ist  $R$  eine Aussage über die Elemente eines Wahrscheinlichkeitsraumes, so ist  $P(R)$  die Summe der  $P(\omega)$ , für die  $R(\omega)$  wahr ist.

**15.2 Bedingte Wahrscheinlichkeit und Unabhängigkeit** Mit  $P(A|B)$  bezeichnen wir die Wahrscheinlichkeit des Ereignisses  $A$ , wenn wir bereits wissen, dass  $B$  eingetreten ist. Als Beispiel betrachten wir das Werfen eines Würfels.  $A$  sei das Ereignis, dass eine 6 geworfen wurde,  $B$  das Ereignis, dass eine gerade Zahl geworfen wurde. Wenn also  $B$  bereits eingetroffen ist, bleiben nur noch die Möglichkeiten 2, 4, 6, die Wahrscheinlichkeit für  $A$  ist dann  $1/3$ . Bei Gleichverteilung erhalten wir daher für die bedingte Wahrscheinlichkeit

$$P(A|B) = \frac{|A \cap B|}{|B|}.$$

Mit  $P(A \cap B) = |A \cap B|/|\Omega|$  und  $P(B) = |B|/|\Omega|$  ist dies äquivalent zu

$$(15.1) \quad P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Sei nun  $(\Omega, P)$  ein allgemeiner Wahrscheinlichkeitsraum. Ist  $P(B) > 0$ , so definieren wir die *bedingte Wahrscheinlichkeit*  $P(A|B)$  von  $A$  bei gegebenem  $B$  durch (15.1). Diese Definition kann leicht mit der anschaulichen Bedeutung von  $P$  als relativer Häufigkeit eines Zufallsexperiments erklärt werden. Wir brauchen nur die Fälle des Experiments zu betrachten, in denen das Ereignis  $B$  eintrifft. Dann liegt entweder der günstige Fall  $A \cap B$  oder der ungünstige Fall  $B \setminus A$  vor.

**Lemma 15.2** Sind  $A_1, \dots, A_k$  Ereignisse mit  $P(A_1), \dots, P(A_k) > 0$ , so gilt

$$P(A_1 \cap \dots \cap A_k) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2) \cdot \dots \cdot P(A_k | A_1 \cap A_2 \cap \dots \cap A_{k-1})$$

*Beweis:* Wir verwenden Induktion über  $k$ .  $k = 2$  ist gerade die Definition der bedingten Wahrscheinlichkeit. Für den Induktionsschritt schreiben wir

$$P(A_1 \cap \dots \cap A_{k-1} \cap A_k) = P(A_1 \cap \dots \cap A_{k-1}) \cdot P(A_k | A_1 \cap \dots \cap A_{k-1}).$$

□

**Beispiel 15.3** *Skat* wird mit einem deutschen Kartenspiel zu 32 Karten gespielt. Die drei Spieler erhalten jeweils 10 Karten, die restlichen 2 Karten werden in den *Skat* gelegt. Wie groß ist die Wahrscheinlichkeit, dass jeder Spieler genau ein As besitzt?

Wir nehmen an, dass Spieler 1 die ersten 10 Karten, Spieler 2 die zweiten 10 und Spieler 3 die dritten 10 Karten bekommt. Sei  $A_i$  das Ereignis, dass Spieler  $i$  genau ein As erhält. Nach dem letzten Lemma gilt dann

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2 | A_1) \cdot P(A_3 | A_1 \cap A_2).$$

Es ist  $P(A_1) = \binom{4}{1} \binom{28}{9} / \binom{32}{10}$  und  $P(A_2 | A_1) = \binom{3}{1} \binom{19}{9} / \binom{22}{10}$ , denn nachdem Spieler 1 die 10 Karten mit genau einem As bekommen hat, bleiben noch 22 Karten mit drei Assen übrig. Analog gilt  $P(A_3 | A_1 \cap A_2) = \binom{2}{1} \binom{10}{9} / \binom{12}{10}$ . □

**Satz 15.4** (a) Sei  $P(B) > 0$ . Durch  $P_B(A) = P(A|B)$  ist ein Wahrscheinlichkeitsmaß auf  $\Omega$  definiert.

(b) (Formel von der totalen Wahrscheinlichkeit) Sei  $\{B_1, \dots, B_k\}$  eine disjunkte Zerlegung von  $\Omega$ . Dann gilt für jedes Ereignis  $A$

$$P(A) = \sum_i P(B_i)P(A|B_i),$$

wobei im Falle  $P(B_i) = 0$  das Produkt  $P(B_i)P(A|B_i)$  zu Null gesetzt wird.

(c) (Formel von Bayes) Ist  $P(A) > 0$  und gelten die Voraussetzungen von (ii), so

$$P(B_j | A) = \frac{P(B_j) P(A | B_j)}{\sum_i P(B_i) P(A | B_i)}.$$

*Beweis:* (a) ist klar.

(b) folgt aus

$$P(A) = P(\cup_i (A \cap B_i)) = \sum_i P(A \cap B_i) = \sum_i P(B_i) P(A | B_i).$$

(c) folgt aus (b) wegen

$$P(B_j | A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(B_j) P(A | B_j)}{\sum_i P(B_i) P(A | B_i)}.$$

□

**Beispiel 15.5** Eine Krankheit kommt bei 0,5% der Bevölkerung vor. Es gibt einen Test für diese Krankheit, der 99% der Kranken identifiziert, aber auch bei 2% der Gesunden positiv ist. Wie groß ist die Wahrscheinlichkeit, dass ein positiv Getesteter tatsächlich krank ist?

Sei  $\{1, \dots, N\}$  die Menge der Bevölkerung,  $B_k$  sei die Menge der Kranken,  $B_g$  die der Gesunden. Damit gilt  $|B_k| \approx 0,005N$ ,  $|B_g| \approx 0,995N$ . Ist ferner  $A$  die Menge der positiv getesteten Personen, so haben wir  $|A \cap B_k| \approx 0,99|B_k|$  und  $|A \cap B_g| \approx 0,02|B_g|$ . Bei zufälliger Auswahl einer Person ergeben sich die Wahrscheinlichkeiten

$$P(B_k) = 0,005, \quad P(B_g) = 0,995, \quad P(A \cap B_k) = 0,99 \cdot 0,005, \quad P(A \cap B_g) = 0,02 \cdot 0,995,$$

nach der Formel von Bayes daher

$$P(B_k | A) = \frac{P(A \cap B_k)}{P(A)} = \frac{0,99 \cdot 0,005}{0,99 \cdot 0,005 + 0,02 \cdot 0,995} = \frac{495}{2485} \approx 0,2.$$

Von den positiv getesteten Personen sind demnach nur 20% krank. □

Seien  $A, B$  zwei Ereignisse mit Wahrscheinlichkeiten  $P(A), P(B) > 0$ . Wir können  $P(A)$  als eine Wette auf das Ereignis  $A$  ansehen.  $A$  und  $B$  definieren wir als unabhängig, wenn die Kenntnis von  $B$  an der Wahrscheinlichkeit für das Eintreffen von  $A$  nichts ändert, wenn also  $P(A) = P(A | B)$ . Anders ausgedrückt: Die Kenntnis von  $B$  hilft uns nicht weiter.

Der Begriff der Unabhängigkeit ist problematisch, weil er von der Mehrheit der Menschen nicht akzeptiert wird. Es ist aber eine empirische Tatsache, dass der zweite Wurf einer Münze oder einer Roulettekugel nicht vom ersten Wurf in irgendeiner Weise beeinflusst wird. Gegen diese Tatsache stehen Millionen Menschen, die glauben, dass aus den ausgespielten Zahlen beispielsweise im Lotto oder im Roulette auf die Zahlen der Zukunft geschlossen werden kann. Die gängigsten Einwände gegen die Unabhängigkeit von Ereignissen kann unter der Theorie subsummiert werden, dass jedes Ereignis ein raum-zeitliches Ereignisfeld aufbaut, das auch die anderen Ereignisse beeinflusst. Dem Einwand, dass eine solche Einflussnahme empirisch nicht nachweisbar ist, kann mit Kritik am Versuchsaufbau oder dem Argument begegnet werden, dass die Einflussnahme äußerst schwach ist. Eine argumentative Auseinandersetzung mit diesem Standpunkt ist nicht möglich, weil die Wahrscheinlichkeit ein mathematisches Konstrukt ist, das keine Aussagen über die Tiefenstruktur der Welt erlaubt, zumal die Unabhängigkeit von Ereignissen bei sehr kleinen Teilchen durchaus zweifelhaft ist.

Wir werfen zwei Würfel hintereinander,  $A$  sei das Ereignis, dass im ersten Wurf eine 1 oder 2 erscheint,  $B$  das Ereignis, dass im zweiten Wurf eine 6 kommt. Dann gilt

$$P(A) = \frac{1}{3}, \quad P(B) = \frac{1}{6}, \quad P(A \cap B) = \frac{1}{18}.$$

Damit ist  $P(A | B) = \frac{1}{3} = P(A)$ , was auch anschaulich klar ist, denn das Auftreten von  $B$  ändert nichts an der Wahrscheinlichkeit von  $A$ . Um eine korrekte mathematische Definition zu bekommen, die auch im Falle  $P(B) = 0$  greift, bezeichnen wir zwei Ereignisse  $A$  und  $B$  als *unabhängig*, wenn

$$P(A \cap B) = P(A)P(B)$$

erfüllt ist, denn in diesem Fall gilt  $P(A) = P(A | B)$  und  $P(B) = P(B | A)$ .

**Beispiel 15.6** Wir werfen zwei Würfel hintereinander.  $A$  sei das Ereignis, dass eine gerade Augensumme gewürfelt wurde,  $B$  sei das Ereignis, das die zweite Augenzahl geradzahlig ist. Dann gilt

$$P(A) = \frac{1}{2}, \quad P(B) = \frac{1}{2}, \quad P(A \cap B) = \frac{1}{4},$$

die beiden Ereignisse sind also unabhängig, obwohl  $B$  durchaus  $A$  beeinflusst. Der zugrunde liegende Mechanismus wird deutlicher, wenn wir die Wahrscheinlichkeit für das Auftreten einer geraden Zahl in beiden Würfen zu  $2/5$  abändern. Dann gilt

$$P(A) = \left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2, \quad P(A \cap B) = \left(\frac{2}{5}\right)^2 \neq P(A)P(B).$$

$A$  und  $B$  sind nun nicht mehr unabhängig.  $\square$

Sei  $I$  eine beliebige Indexmenge und  $\{A_i : i \in I\}$  eine Familie von Ereignissen. Die Familie heißt *unabhängig*, wenn für alle endlichen Indexmengen  $\emptyset \neq J \subset I$  gilt

$$(15.2) \quad P\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i).$$

**Beispiel 15.7** Der Sinn dieser Definition wird folgendermaßen deutlich. Wir werfen zweimal hintereinander eine Münze und betrachten die Ereignisse  $A = \text{"beim ersten Wurf K"}$ ,  $B = \text{"beim zweiten Wurf K"}$ ,  $C = \text{"genau ein K"}$ . Dann sind natürlich  $A$  und  $B$  unabhängig, aber, obwohl  $A$  und  $B$  das Ereignis  $C$  bestimmen, sind  $A$  und  $C$  ebenfalls unabhängig wegen  $P(A) = 1/2$ ,  $P(C) = 1/2$ ,  $P(A \cap C) = 1/4$ . Gleichermaßen gilt für  $B$  und  $C$ . Andererseits ist  $P(A \cap B \cap C) = 0$ , aber  $P(A)P(B)P(C) = 1/8$ . Aus der paarweisen Unabhängigkeit folgt daher nicht notwendig die oben definierte Unabhängigkeit.  $\square$

**Satz 15.8** (a) Ist  $\{A_i : i \in I\}$  eine unabhängige Familie von Ereignissen und ist  $P(A_k) = 0$  oder  $P(A_k) = 1$  mit  $k \notin I$ , so ist auch  $\{A_i : i \in I \cup \{k\}\}$  unabhängig.

(b) Ist  $\{A_i : i \in I\}$  unabhängig und ist  $B_i \in \{A_i, A_i^c, \emptyset, \Omega\}$  für jedes  $i \in I$ , so ist auch  $\{B_i : i \in I\}$  unabhängig.

(c) Ist  $I = \{1, \dots, n\}$ , so ist die endliche Familie von Ereignissen  $\{A_i : i \in I\}$  genau dann unabhängig, wenn für jede Wahl von  $B_i \in \{A_i, A_i^c\}$  gilt

$$(15.3) \quad P(B_1 \cap \dots \cap B_n) = P(B_1) \dots P(B_n).$$

*Beweis:* (a) Gehört  $k$  zur Indexmenge  $J$  in (15.2), so ändert sich im Falle  $P(B) = 1$  auf beiden Seiten der Definition nichts wegen  $P(A \cap B) = P(A)$  für alle Mengen  $A$ . Im Falle  $P(B) = 0$  steht hingegen auf beiden Seiten Null.

(b) Wegen (a) ist nur der Fall  $B_i \in \{A_i, A_i^c\}$  näher zu untersuchen. Durch Induktion über  $m$  beweisen wir die Behauptung: Ist  $J \subset I$  endlich und  $|\{j \in J : B_j = A_j^c\}| \leq m$ , so gilt

$$(15.4) \quad P\left(\bigcap_{j \in J} B_j\right) = \prod_{j \in J} P(B_j).$$

Ist  $m = 0$ , so sind alle  $B_i$  gleich  $A_i$  und (15.4) folgt aus (15.2). Sei (15.4) für  $m$  bewiesen und  $J$  eine Indexmenge mit  $|\{j \in J : B_j = A_j^c\}| = m + 1$ . Wir können annehmen, dass  $J = \{1, \dots, n\}$  und  $B_1 = A_1^c$ . Für  $j = 2, \dots, n$  kann dann die Induktionsvoraussetzung verwendet werden. Es gilt daher

$$\begin{aligned} P\left(\bigcap_{j=1}^n B_j\right) &= P\left(\bigcap_{j=2}^n B_j\right) - P\left(A_1 \cap \bigcap_{j=2}^n B_j\right) \\ &= \prod_{j=2}^n P(B_j) - P(A_1) \cdot \prod_{j=2}^n P(B_j) = \prod_{j=1}^n P(B_j). \end{aligned}$$

(c) Die Notwendigkeit der Bedingung ist gerade (b). Auf (15.3) addieren wir die gleiche Bedingung für die Mengen  $B_1^c, B_2, \dots, B_n$  und erhalten

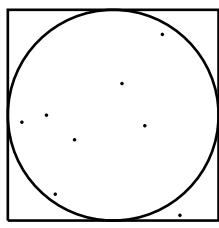
$$\begin{aligned} P\left(\bigcap_{i=2}^n B_i\right) &= P\left(B_1^c \cap \bigcap_{i=2}^n B_i\right) + P\left(\bigcap_{i=1}^n B_i\right) \\ &= P(B_1^c) \prod_{i=2}^n P(B_i) + \prod_{i=1}^n P(B_i) = \prod_{i=2}^n P(B_i). \end{aligned}$$

Auf diese Weise können wir die Bedingung auch für kleinere Indexmengen als  $\{1, \dots, n\}$  nachweisen.

□

### 15.3 Die Monte-Carlo-Methode

**Berechnung des Kugelvolumens** Sei



$$K_n = \{x \in \mathbb{R}^n : x_1^2 + x_2^2 + \dots + x_{n-1}^2 + x_n^2 < 1\}.$$

die Einheitskugel des  $n$ -dimensionalen Raums. In  $n = 2$  Dimensionen sperren wir den Einheitskreis in das Quadrat  $Q_1 = (-1, 1) \times (-1, 1)$  ein. Wir generieren Zufallsvektoren  $(x_1, x_2) \in Q_1$ , die auf  $Q_1$  gleichverteilt sind, und zählen, wie viele davon in den Einheitskreis fallen. Das Verhältnis der Treffer zur Gesamtzahl der Versuche konvergiert dann gegen den gesuchten Flächeninhalt/4.

Bei  $N$  Versuchen erhalten wir für die Kugeln in den Raumdimensionen  $n = 2, 3, 4, 5$ :

$n$	$N = 10^1$	$N = 10^3$	$N = 10^5$	$N = 10^7$
2	3,59999	3,25200	3,14276	3,14222
3	4,80000	4,23199	4,18344	4,18810
4	3,20000	4,97599	4,90384	4,93469
5	9,60000	6,07999	5,28032	5,26119

$\text{Vol}(K_2) = 3,14159\dots$  liefert für die Fehler im Falle  $n = 2$ :

$n$	$K = 10$	$K = 10^3$	$K = 10^5$	$K = 10^7$
2	0.45841	0.11041	0.0117	0.0063

An diesem einfachen Beispiel sehen wir die Vor- und Nachteile der Monte-Carlo-Simulation:

1. Man braucht vom gestellten Problem keine Ahnung zu haben. In unserem Fall lässt sich das Kugelvolumen durch höhere Mathematik exakt bestimmen.
2. Daraus folgt: Auch sehr komplexe Probleme können mit einfachen Monte-Carlo-Programmen behandelt werden.
3. Das Verfahren ist sehr langsam, in der Regel geht der Fehler wie  $\sqrt{N^{-1}}$ .
4. Das Verfahren ist nicht sicher. Es gibt immer Ausreißer, die stark von den angegebenen  $\sqrt{N^{-1}}$  abweichen.

Weiter benötigt man sehr viele Zufallszahlen. Woher nimmt man die?

**Zufallszahlen** Für die Bestimmung des Kugelvolumens reichen gleichverteilte Zufallszahlen über dem Intervall  $[0, 1)$  aus. Durch eine Streckung oder Stauchung dieses Intervalls kann man Zufallszahlen auf jedem anderen Intervall bekommen. Gleichverteilt bedeutet, dass die Zufallszahl  $y$  mit Wahrscheinlichkeit  $b - a$  im Intervall  $(a, b) \subset [0, 1)$  liegt.

Da es sehr schwierig ist, „echte“ Zufallszahlen zu generieren, verwendet man durch eine Formel erzeugte „Pseudo-Zufallszahlen“. Im *Linearen Kongruenz-Generator* gibt man sich eine natürliche Zahl  $m$  vor sowie Zahlen  $a, b \in \{0, 1, \dots, m - 1\}$ . Für eine weitere vorgegebene Zahl  $x_0 \in \{0, 1, \dots, m - 1\}$  als Startwert bestimmt man

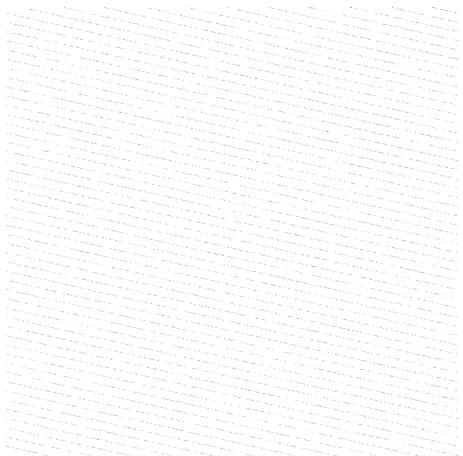
$$x_{i+1} = (ax_i + b) \mod m.$$

Dabei ist  $(\dots) \mod m$  so zu verstehen, dass  $x_{i+1} \in \{0, 1, \dots, m - 1\}$  der Rest ist, der beim Teilen der rechten Seite durch  $m$  entsteht. Die  $x_i$  liegen alle in der Menge  $\{0, 1, \dots, m - 1\}$  und sind periodisch mit einer Periode  $\leq m$ .

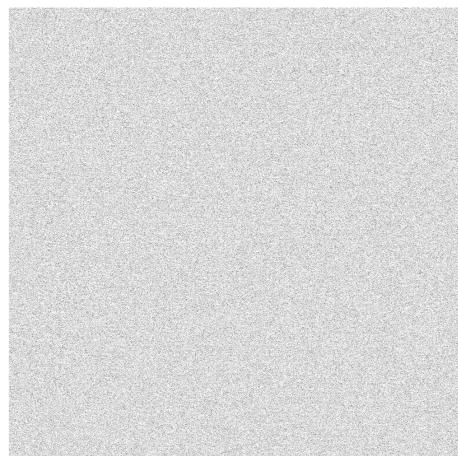
Als einfaches Beispiel nehmen wir  $a = 4$ ,  $b = 1$ ,  $m = 11$  und erhalten

$$x_0 = 1, x_1 = 5, x_2 = 10, x_3 = 8, x_4 = 0, x_5 = 1.$$

Offenbar müssen die Zahlen  $m, a, b$  gut gewählt werden, wenn man Periodenlänge  $m$  haben will.



Raster 1000 x 1000 Linearer Kongruenzgenerator  
a=12453, b=8889, m=247897



Raster 1000 x 1000 IFORT-Compiler  
Weiss: kein Treffer Schwarz: Treffer>4

In der Praxis werden die Zahlen  $m, a, b$  so bestimmt, dass die Periode maximal, also  $m$  ist, und die erzeugten  $x_i$  verschiedene Tests auf Zufälligkeit erfolgreich bestehen. Man kann die Folge in

Binärdarstellung hintereinander schreiben. Es soll dann eine zufällige 0, 1-Folge entstehen, in denen alle Bitmuster in etwa gleich häufig vorkommen.

Man teilt die Folgenglieder  $x_i$  des linearen Kongruenzgenerators durch  $m$  und erhält (hoffentlich) gleichverteilte Zufallszahlen auf dem Intervall  $[0, 1)$ . Da die Beschreibung der Folge kurz ist, wird es immer Anwendungen geben, bei denen die „Zufälligkeit“ dieser Zahlen nicht ausreicht.

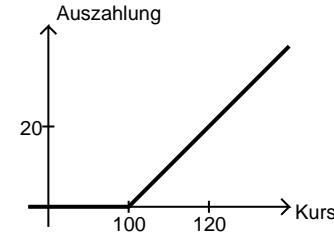
Auch wenn  $m, a, b$  im linearen Kongruenz-Generator gut gewählt werden, sind *Zufallsvektoren* problematisch. In diesem Fall bildet man – wie bei der Berechnung von Flächeninhalten – aus den Zufallszahlen die Folge

$$(x_0, x_1), (x_2, x_3), (x_4, x_5), \dots$$

Das Bild links zeigt die von einem linearen Zufallsgenerator erzeugten Vektoren. Obwohl die zugehörigen Zufallszahlen tatsächlich wie zufällig erscheinen, liegen im Bild viele Vektoren auf einer Geraden und sind daher für die Flächenberechnung ungeeignet. Das Bild rechts zeigt die Zufallsvektoren des fortran90-Compilers ifort. Beide Bilder wurden erzeugt, indem ein  $1000 \times 1000$ -Raster in das Einheitsquadrat gelegt und dann gezählt wird, wie oft ein Zufallsvektor in eines der Teilquadrate fällt. Je öfter dies geschieht, desto dunkler wird das Teilquadrat eingefärbt.

**Fairer Preis von Finanz-Derivaten** Als Beispiel für ein Finanz-Derivat erläutern wir die Option auf eine Aktie. Eine Aktie kostet zum 1.1. eines Jahres 100 Euro. Man kann eine Kaufoption auf diese Aktie erwerben, die einem das Recht gibt, diese Aktie zum 31.12. des gleichen Jahres zum Preis von 100 Euro zu kaufen. Wir nehmen an, dieses Recht kostet 10 Euro. Zum 31.12. kauft man die Aktie zu 100 Euro, sofern der Kurs über 100 Euro liegt, und verkauft sie gleich wieder. Es ergeben sich damit folgende Möglichkeiten zum 31.12.:

Kurs	Gewinn
120	10
100	-10
80	-10



Auszahlungsfunktion einer Kaufoption (ohne Kaufpreis)

Das Bild rechts zeigt die Auszahlungsfunktion dieser Kaufoption. Das Tröstliche daran ist, dass man nicht mehr verlieren kann als seinen Einsatz im Gegensatz zum Aktionär, der mit dieser Aktie im ungünstigsten Fall 100 Euro in den Sand setzt. Daher sieht die Mathematik Optionen durchaus positiv, sofern nur ein kleiner Teil des Vermögens darin angelegt wird.

Auch wenn es mathematisch keinen Unterschied macht, unterscheidet man zwischen Optionen und Optionsscheinen. Erstere werden an Terminbörsen gehandelt und der Preis wird durch Angebot und Nachfrage marktwirtschaftlich ermittelt. Optionsscheine werden dagegen von Handelshäusern begeben, die an der Börse Kauf- und Verkaufskurse stellen. In diesem Fall wird zur Preisfeststellung ein mathematisches Modell benötigt, da ein Markt für die Optionsscheine de facto nicht existiert.

Was ist der faire Wert einer Option? Wir nehmen an, dass der Kurs der Aktie sich im Jahresverlauf zufällig entwickelt mit einer durchschnittlichen Rendite, die der Markttrendite für Anleihen entspricht. Der durchschnittliche Gewinn, der mit einer Option bei diesen zufälligen Kursverläufen erzielt wird, ist gleichzeitig der faire Preis der Option. Gerade für erfahrene Börsianer ist dieser Ansatz erstaunlich, weil die Aktie sich nur besser als der planlose Zufall entwickeln muss, um mit der Option Gewinn zu erzielen.

Wir müssen uns als erstes ein stochastisches Modell des Kursverlaufs verschaffen. Sei  $I$  die Laufzeit der Option in Börsentagen und  $s$  die maximale relative Kursänderung an einem Börsentag, die aus der Vergangenheit des Kurses bestimmt wird.  $s$  beschreibt, wie flatterhaft oder volatil die Aktie in der Vergangenheit gewesen ist. Die Volatilität wird zum einen von der allgemeinen

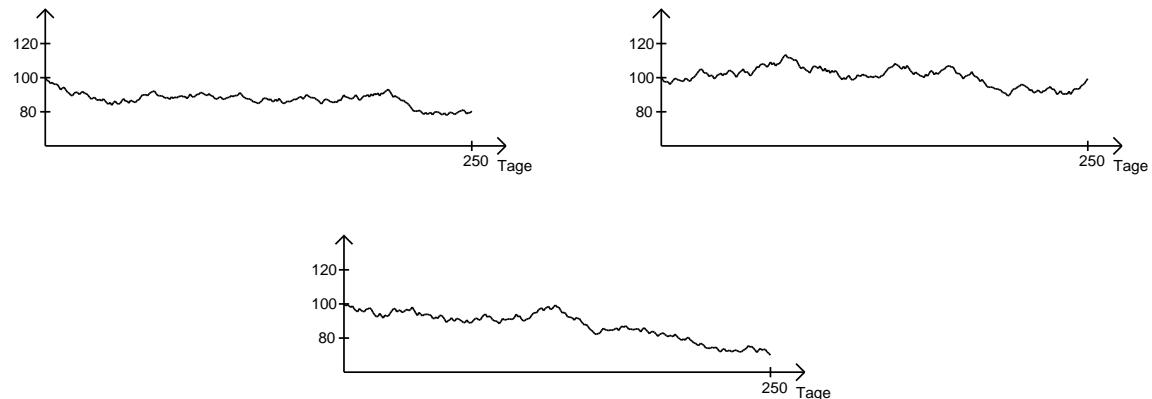
Börsenverfassung bestimmt, zum anderen auch von der Aktiengesellschaft selbst. So überstehen finanziestarke Gesellschaften mit einem wenig anfälligen Geschäftsmodell Krisen besser als Firmen mit einem stark zyklischen Geschäft. Ein Beispiel für Letzteres ist die Stahlindustrie, die bei niedrigem Stahlpreis mit einem Bein im Konkurs steht und bei hohen Preisen finanziell aus dem Vollen schöpft. Die im folgenden Modell unterstellte Hypothese, dass die aus der Vergangenheit bestimmte Volatilität auch für die Laufzeit gültig bleibt, ist also nicht aus der Luft gegriffen. Es ist überdies klar, dass die Volatilität ein wichtiger preisbestimmender Faktor für eine Option darstellt. Denn wenn der Kurs im obigen Beispiel bei 100 Euro verharrt, was der Volatilität  $s = 0$  entspricht, wird man seinen Einsatz verlieren.

$a$  sei der Tageszinssatz des Marktes (z.B.  $a^{250} = 1.01$  bei etwa 250 Börsentagen) und  $x_0$  der Kurs der Aktie am Tag 0. Mit gleichverteilten Zufallszahlen  $y_i$  im Intervall  $(-s, s)$  simulieren wir einen zufälligen Aktienkurs durch

$$x_{i+1} = ax_i + y_i x_i, \quad 0 \leq i \leq I - 1.$$

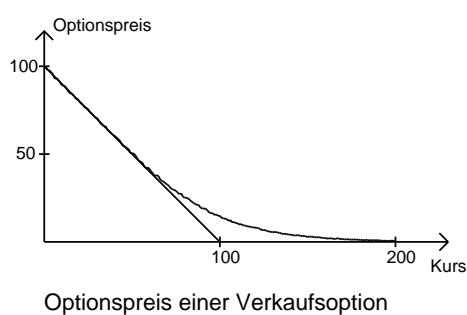
Wenn wir hier  $y_i = 0$  setzen, so erhalten wir mit  $x_i$  den Tageswert einer Anleihe, die sich in unserem Beispiel mit 1% verzinst. Wenn wir also fortwährend in diesem Modell Aktien kaufen, so erhalten wir im Schnitt am Jahresende 1% Gewinn. Aus  $x_I$  kann der Gewinn durch die Option bestimmt werden.

Und nun die Frage an die Leserinnen und Leser. Sind die folgenden Charts „echt“ oder vom Zufallsgenerator erzeugt?



Echte Kurse oder aus dem Zufallsgenerator ?

Wir machen einige 1000 Durchläufe mit den beschriebenen zufälligen Kursverläufen, bestimmen jedesmal  $x_I$  und den daraus resultierenden Gewinn und erhalten mit dem Durchschnitt dieser Gewinne den Erwartungswert und damit den fairen Preis für die Option.



Im nebenstehenden Bild sehen wir den durch die Simulation bestimmten fairen Preis einer Verkaufsoption bei unterschiedlichen aktuellen Kursen. Die stückweise lineare Funktion ist die Auszahlungsfunktion, die in diesem Zusammenhang als *inneren Wert* der Verkaufsoption bezeichnet wird. Was über dem inneren Wert liegt, wird *Aufpreis* genannt. Der Aufpreis ist offenbar am größten, wenn der aktuelle Kurs der Aktie genau auf dem *Basispreis* von 100 Euro liegt.

Mit der hier vorgestellten Methode lassen sich alle Arten von Derivaten und derivathaltigen Geschäften (auch Genußscheine, Wandelanleihen, Futures, Zertifikate usw.) auf alle Arten von Basisobjekten (Aktienindizes, Währungen, landwirtschaftliche Produkte usw.) untersuchen. Dennoch

gilt bei Optionen: Den Optionspreis legt der Markt fest, nicht die Theorie.

Der Verkäufer der Option trägt das Risiko. Mit einer Erweiterung der Theorie lässt sich ein Gegengeschäft des Verkäufers angeben, das ihn im Rahmen der Theorie von jedem Risiko freistellt.

In der Praxis verwendet man normalverteilte an Stelle von gleichverteilten Zufallszahlen und einen kontinuierlichen an Stelle eines diskreten stochastischen Prozesses. In einfachen Fällen kann der Optionspreis durch eine Formel angegeben werden, was man als *Black-Scholes-Theorie* bezeichnet.